KATHOLIEKE UNIVERSITEIT LEUVEN

Fakulteit der Toegepaste Wetenschappen

Departement Elektrotechniek
Kard. Mercierlaan 94, 3030 Heverlee

Mathematical Concepts and Techniques for Modelling of Static and Dynamic Systems

Jury:
Prof.Dr.Ir. J. Delrue, vice-dekaan, voorzitter
Prof.Dr.Ir. J. Vandewalle, promotor
Prof.Dr. A. Bultheel
Dr.Ir. A. Barbé
Prof.Dr.Ir. J. Peperstraete
Prof.Dr.Ir. H. Van Brussel
Prof.Dr.Ir. M. Gevers

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de Toegepaste Wetenschappen
door
Bart De Moor

UDC 519.71

Juni 1988

Preface.

Dit doktoraatswerk is tot stand gekomen met de steun en medewerking van vele mensen.

Eerst en vooral zou ik mijn promotor, *Joos Vandewalle*, willen bedanken. Herhaaldelijk ben ik, gedurende de meer dan 5 jaar dat ik met hem heb samengewerkt, door collega's van binnen en buiten de universiteit gewezen op het benijdenswaardige van deze samenwerking. Dit is meer dan terecht. Niet alleen besteedt Joos veel tijd aan zijn doktoraatsstudenten, zij kunnen ook steeds rekenen op zijn aktieve inbreng in hun onderzoek. Ik heb veel geleerd van Joos' bedachtzame suggesties en de krachtlijnen die hij formuleerde in verband met het gepresenteerde onderzoek. Zijn brede interdisciplinaire wetenschappelijke belangstelling en zijn internationale ervaring hebben op mij een blijvende indruk gemaakt. Ik zal met plezier terugdenken aan onze talrijke diskussies, die niet alleen over wetenschappelijke onderwerpen handelden. Ik hoop dan ook dat dit doktoraatswerk slechts het begin is van een verdere samenwerking.

Ik zou ook de leden van het leescomité, *Andre Barbé* en *Adhemar Bultheel*, willen bedanken voor het nazicht van dit niet onomvangrijk manuscript, alsook de juryleden, de professoren *Jan Delrue*, *Jan Peperstraete*, *Rik Van Brussel* en *Michel Gevers*.

De professoren van de afdeling, *André Oosterlinck*, *René Govaerts*, *Jan Peperstraete* en *Willy Sansen* hebben ervoor gezorgd dat ik in de ideale omstandigheden wetenschappelijk onderzoek kon bedrijven. Ik wil hen dan ook van harte bedanken.
Hierbij wil ik ook het I.W.O.N.L. bedanken, dat gedurende vijf jaren, met een onderbreking voor mijn burgerdienst, gezorgd heeft voor een vlotte behandeling van mijn dossier, alsook het N.F.W.O. omdat het mij herhaaldelijk in staat heeft gesteld om door internationale kontakten mijn wetenschappelijke kennis te vervolledigen.
Voor mijn burgerdienst kon ik terecht bij professor van Steenberghe, die steeds een levendige belangstelling voor mijn aktiviteiten aan de dag heeft gelegd. Hem en zijn medewerkers ben ik van harte dankbaar voor onze samenwerking in de bio-medische signaalverwerking, waarvan ik hoop dat zij in de nabije toekomst kan worden verdergezet.

Mijn kollega's en ex-kollega's, *Yvo Desmedt, Dirk Callaerts, Frank Debondt, Guido Tijskens, Walter Mommaerts, Jan Vanderschoot, Sabine Van Huffel, Chen Jiande, Dirk Van Compernolle, ...* hebben binnen *SISTA* steeds voor een stimulerende en plezante atmosfeer gezorgd. In het bijzonder zou ik hier *Jan Staar en Michel Verhaegen* willen bedanken. Jan's doktoraat was mijn eerste kontakt met het domein van systeemidentifikatie. Ik denk dat Jan vooral de resultaten in hoofdstuk 4 zal appreciëren: ze zijn gebaseerd op zijn ideeën. Ook Michel zou ik willen bedanken voor onze talrijke interessante diskussies.

In particular, I want to thank *Shaohua Tan* for the many stimulating discussions we had together, not only as colleagues, but also as friends. Shaohua taught me that the best way to avoid narrow-mindedness is open-mindedness, not only concerning scientific research, but on the whole spectrum of small and large social, cultural and political issues. He has contributed a lot to this work and the fact that it treats so many subjects, originates, at least partially, in his influence. I hope he will discover some of his own reflections in this thesis.

De meeste resultaten van dit werk zouden niet zijn wat ze zijn zonder de aktieve inbreng van *Jan Swevers, Marc Moonen, Piet Van Mieghem en Lieven Vandenberghe.* Marc en Lieven waren niet alleen aktieve en vooral kritische leden van een informeel leescomité, maar ongetwijfeld zullen zij in deze thesis, vele elementen terugvinden die we samen bediskussiëerd, aangepakt en opgelost hebben. Meer in het bijzonder wil ik Marc bedanken voor zijn aktieve inbreng in de resultaten van de hoofdstukken 5, 7 en 8 en Lieven voor zijn bijdragen tot de hoofdstukken 2, 3, 5 en 8.

Mijn speciale dank gaat ook naar *Daniël Berckmans.* In onze nachtelijke uitstapjes naar het laboratorium voor Agrarische Bouwkunde, heeft hij mij bijgebracht dat de *theorie van systeemidentifikatie* een bezigheid is voor overdag, maar dat *de praktijk ervan* zich vooral 's nachts situeert. Ik wens Daniël en *Geert Taes* dan ook veel geluk wanneer zij zullen proberen sommige van de resultaten van dit werk in de *dagelijkse* praktijk om te zetten.

Ik wil ook speciaal de technische staf van de afdeling en het departement bedanken, *Lou Schol, Elvira Wauters, Ingrid Tokka, Luc Boone* en de vele andere medewerkers in het departement die altijd in de bres zijn gesprongen om praktische en technische moeilijkheden uit de weg te ruimen.

Tijdens de 5 jaar die ik in ESAT heb doorgebracht, werd ook een personal computer-netwerk voor didaktische doeleinden opgestart, waardoor ik in de gelegenheid werd gesteld om aan enkele van mijn didaktische verzuchtingen te voldoen. Ik wil hierbij de vele *medewerkers* danken die gedurende al die tijd hebben ingestaan voor de ontelbare soft- en hardware beslommeringen die met een dergelijk initiatief gepaard gaan. Verder wil ik ook de talrijke *eindwerkstudenten* bedanken. Allen hebben zij op hun manier een steentje bijgedragen tot deze thesis.

Ik zou ook *Ton Backx* van Philips Eindhoven, *Pieter De Groen* van het departement wiskunde van de V.U.B. *Paul Van Dooren* van MBLE-Brussel en *Dirk Horstens* van het departement wiskunde van de K.U.L., willen bedanken voor de interessante diskussies.
Moreover, I would like to thank dr. *Cornelius Los* of the University of Florida in Gainesville, prof. *Roberto Guidorzi* of Bologna University, prof. George Verghese of M.I.T., prof. *Thomas Kailath* and prof. *Gene Golub* of Stanford University for the interesting discussions.

Ik zou wijlen dekaan professor *Snoeys* willen bedanken voor onze talrijke gesprekken op en na de vele vergaderingen die we samen meemaakten en voor de belangstelling die hij steeds in ons onderzoek heeft getoond.
Verder wil ik ook de leden van de bestuursvergaderingen van de V.I.Lv. en L.O.V.A.N. bedanken, niet in het minst *Marcel Van Bael* en *Rik Demey*, voor de vele malen dat zij mij organisatorisch hebben bijgestaan.

Tenslotte zou ik de vele *vrienden* en (toekomstige) *familieleden*, mijn broer *Maarten*, zus *Veerle* en mijn *ouders* willen bedanken. Zij hebben enerzijds ervoor gezorgd dat ik ten gepaste tijde de wetenschappelijke beslommeringen opzij kon schuiven en dat ik anderzijds, mij zorgeloos aan deze beslommeringen kon overleveren.

Insiders weten dat ik tot vandaag eigenlijk getrouwd was met dit doktoraat. Hieraan wordt heel binnenkort paal en perk gesteld. Ik draag dit werk dan ook op aan *Hilde*.

Bart De Moor
Leuven, mei 1988

Contents

List of Figures

The figure on p.1 is from:
P.Bockstaele. *Die waerachtige const der Geometrien 1513 Het oudste gedrukte Nederlandse Meetkundeboek.* uit 'Liber Amicorum' voor G.Bosteels, M.Lamberechts, uitg. door de Vrienden van K.A. Berchem, pp.29, 1979.

The figures on p.34, p.114, p.182, p.254 are from:
V.V.Nalimov. *Faces of Science.* ISI Press, Philadelphia, 1981.

Notations, Abbreviations, Conventions

Throughout the book, square brackets are used for references, which can be found at the end of each chapter.

Number Sets.

- Z^n Set of integer n-tuples
- Q^n Set of rational n-tuples
- R^n Set of real n-tuples
- C^n Set of complex n-tuples
- $\#\mathcal{A}$ Cardinality of the set $\mathcal{A}$
- $\lfloor \alpha \rfloor$ integer truncation of the real number α
- UB The unit ball $UB = \{q \in \mathcal{R}^m \mid q^t \cdot q = 1\}$
- $\bigcap$ Symbol denoting the intersection of 2 sets.
- $\bigcup$ Symbol denoting the union of 2 sets.

Matrices and Vectors.

All matrices and vectors are assumed to be real. Column vectors $a, b, \ldots$ are denoted by small letters. Row vectors are denoted explicitly as the transpose of a column vector. Capitals $A, B, \ldots$ represent matrices, as well as the greek capitals $\Gamma, \Delta, \Lambda, \Sigma$. The letters i, j, k, l, m, n, p, q are integers (used for e.g. indexing). Real numbers are denoted by greek symbols $\alpha, \beta, \ldots$ except when it concerns components of a vector a, denoted by a_i. No distinction is made between the notation of a linear transformation and its matrix representation nor between a vector and the column vector with its coordinates in some basis. Equalities and inequalities between matrices and vectors always hold elementwise. The symbol 0 denotes the zero element for scalars, vectors and matrices, depending on the context.

- $A_{m \times n}$ Matrix with m rows and n columns
- A^t Transpose of a matrix

- A^* complex conjugate transpose of the matrix A
- A^{-1} inverse if square non-singular
- A^{-t} transpose of the inverse = inverse of the transpose.
- A^+ Pseudo-inverse
- $r(A) = rank(A)$ Algebraic rank of the matrix A.
- $cor(A) = corank(A)$ corank of the matrix A: $cor(A) = min(m,n) - r(A)$.
- a^i i-th column of the matrix A
- $a_{n\times 1}$ $n \times 1$ real vector
- a_i i-th component of vector a.
- a^i_j Element (i,j) of the matrix A.
- $span_{row}(A)$ vector space generated by rows of A.
- $span^{\perp}_{row}(A)$ orthogonal complement of A's row space.
- $span_{col}(A)$ vector space generated by columns of A.
- $span^{\perp}_{col}(A)$ orthogonal complement of A's column space.
- $ker(A)$ kernel of A defined as $\{x \mid Ax = 0\}$
- $a^t b$ Euclidean inner product: $\sum_{i=1}^{n} a_i b_i$
- a^+, a^- sign decomposition of a vector
- $\mid x \mid$ absolute value (elementwise)
- $I_{m\times n}$ rectangular identity matrix (ones on the main diagonal, zeros everywhere else)
- I_n square identity matrix
- e^i i-th column of the identity matrix
- $\lambda(A)$ set of eigenvalues of the matrix A
- $\sigma(A)$ set of singular values of the matrix A
- $\underline{A}$ $(\overline{A})$ Matrix obtained from A by omission of last (first) (block-) row.
- $\mid A$ $(A \mid)$ Matrix obtained from A by omission of first (last) (block-)column.
- $\|a\|_2$ 2-norm of a vector: $\sqrt{a^t a}$
- $trace(A) = \sum_i a^i_i$ Trace of a matrix
- $\|A\|^2_F$ Frobeniusnorm of a matrix: $trace(AA^t)$

- $E_q[A]$ Oriented energy of the column vector sequence of the matrix A in the direction of the vector q: $q^t A A^t q$.
- $R_q[A, B]$ Oriented signal-to-signal ratio of the column vector sequences of the matrices A and B in the direction of the vector q.
- $MmR[A, B, r]$ The maximal minimal signal-to-signal ratio over all r-dimensional subspaces of the column vector sequences contained in the matrices A and B.
- $mMR[A, B, r]$ The minimal maximal signal-to-signal ratio over all r-dimensional subspaces of the column vector sequences contained in the matrices A and B.
- $A \otimes B$ Kronecker product of two matrices
- $A \odot B$ Khatri-Rao product of two matrices
- $vec(A)$ column vector obtained from storing the column of A in a long column vector.
- $vecd(A)$ column vector with diagonal elements of A.
- $diag(a_i)$ diagonal matrix with elements a_i.

'Pure noise' quantities are denoted by a 'tilde' : $\tilde{a}, \tilde{\Sigma}$ while 'exact' items are marked with a 'hat' : $\hat{a}, \hat{\Sigma}$. In some cases MATLAB-notation is used to denote submatrices: Let A be an $m \times n$ matrix:

- $A(:, p:q)$ Submatrix consisting of columns p to q.
- $A(p:q, :)$ Submatrix consisting of rows p to q
- $A(p:q, r:s)$ Submatrix consisting of row p to q, and column r to s.

A matrix U is called *orthogonal* if $U^t U$ is diagonal. It is called *orthonormal* if $U^t U$ equals the identity matrix. The *smallest* (*largest*) eigenvector is the eigenvector that is associated with the eigenvalue of largest (smallest) modulus. Similarly, the *largest* (*smallest*) left or right singular vector is the left or right singular vector that corresponds to the largest (smallest) singular value.

Abbreviations

AONI: Allowed Orthant Null Invariant

GLCP: Generalized Linear Complementarity Problem

GSVD: Generalized Singular Value Decomposition

LCP: Linear Complementarity Problem

LLS: Linear Least Squares

maxcor: Maximal Corank

minr: Minimal rank

NI: Null Invariant

NND: Nonnegative Definite

OI: Orthant Invariant

ONI: Orthant Null Invariant

PD: Positive Definite

SVD: Singular Value Decomposition

TLLS: Total Linear Least Squares

WL: Wilson-Ledermann bound

Die waerachtige const der Geometrie leerende hoemē alderhāde breydde lingdē dicte ēn hooch dē Der veldē Beemden Bosschē Berghē Metselrie Paueyselē Torren Huysen Kercken ende alderhan de dinghen meten sal. Hoemen ooc maken sal die wijnroede Om daer mede te roeden alderhande Tō nen Vaten Cuyppen backen ende dier ghelijcke

Die waerachtige const der Geometrien leerende hoemen alderhande breydden, lingden, dicten, ende hoochden
Der velden Beemden Bosschen Berghen Metselrien Paveyselen Torren Huysen Kercken ende alderhande dinghen meten sal.
Hoemen oock maken sal die wijnroede, Om daer mede te roeden alderhande Tonnen Vaten Cuyppen backen ende dier ghelijcke.

Eldest Flemish book about Geometry, by Thomas Van Der Noot, Brussels, April 20, 1513.

Chapter 1

Introduction and outline of the thesis

This doctoral dissertation focuses on the development of *concepts and algorithms for the mathematical analysis and modelling of static and dynamic systems.*

Loosely speaking, static systems are systems without memory capacity. More rigorously, the main difference between static and dynamic systems lies in the notion of *state.* In general, the state of a system can be considered as the information that is needed, to determine uniquely the behavior. Observe however, that this characterization of the state, depends upon what is considered to be an input. As a matter of fact, it will be shown that the distinction between states and inputs is not always trivial. As a preliminary example, think of state variable feedback in control theory. In mathematical models, the dynamics of the behavior will be almost always revealed in the *structural properties* of the equations. For instance, in the case of linear systems, it will be demonstrated that the dynamics reveal themselves in the so-called *shift structure* of certain spaces.

In this dissertation, we shall pay attention to *linear and non-linear static systems* and to *linear dynamic systems.*

- The results for *static linear systems* include the deductive analysis of linear modelling approaches under the influence of additive noise, such as linear and total linear least squares, the computation of approximate intersections between subspaces and the development of a fundamental uncertainty principle of mathematical modelling.

- The results for *static non-linear systems* concentrate around the generalized linear complementarity problem. It is shown how the resulting algorithms and insights allow to model and analyse a wide variety of non-linear phenomena.

- Concerning *linear dynamic systems*, attention will be paid to the problem of identification and realization. Algorithms for the identification of state space models from measurements will be developed.

Throughout the work, *matrix algebra*, its *geometrical* interpretations and *numerical* implementations are the essential tools that will be employed to derive the concepts and the algorithms.

1.1 Geometrical and Numerical Linear Algebra.

1.1.1 Geometrical Linear Algebra.

The label *geometric* is applied for several reasons: The basic ideas are thought of as geometric properties of distinguished subspaces. The geometry was first brought in out of revulsion against the orgy of matrix manipulation which linear control theory and linear algebra consisted of not so long ago. But secondly, the geometric setting often suggests methods of attack which have proved to be intuitive and economical. The essential link between matrix algebra and geometry goes through the notion of a vector space. In the case of real vector spaces, a matrix of real numbers adequately defines a linear map once basis vectors have been fixed for the domain and codomain spaces, hence allowing for a geometrical interpretation of most matrix operations. As an example, the result of a matrix-vector multiplication of the form Ax is another vector which consists of linear combinations of the column vectors of the matrix A. Note however that not everything can be explained by geometry. As an example, geometrical intuition does not provide a straightforward explanation for the fact that the column rank of a matrix equals its row rank.

1.1.2 Numerical Linear Algebra.

It makes me nervous to fly in airplanes,
since I know that they are designed in floating point.
Householder

The accumulation and round-off error in long computerized calculations and in recursive algorithms can really destroy an efficient and theoretically sound computational procedure. A tell-tale example is the Kalman filter divergence. The causes of floating point arithmetic errors are threefold: First, there are *intrinsic* errors due to finite wordlength representation of a given number. Second, binary operations on two numbers may require a longer wordlength. Hardware implementation greatly affects this error (presence of guardbits, double register arithmetic, rounding, truncation,...). These kinds of error are *extrinsic errors*. Finally all these errors propagate through the recursion and accumulate. They are called the *inherent errors* since they inherit their properties from the operation sequence and the given recursion.

The analysis of numerical error propagating mechanisms has become an independent mathematical discipline, including the study of backward and forward stability, sensitiveness and conditioning etc.... Good software development demands a mathematical understanding of the problem to be solved, a flair for algorithmic expression and an appreciation for finite precision arithmetic. More and more, scientific research ultimately results in software. The impact of numerical linear algebra is felt in several ways. The reliance on orthogonality of matrices, the appreciation of the problem's sensitivity and the careful consideration of round-off, have spilled over many areas. A prime example is the increased use of the (generalized) singular value decomposition as an analytical, conceptual, algebraic, geometrical and numerical tool in formulating and deriving theoretical concepts. The (generalized) singular value decomposition, although a catchword in control theory today, is just one small part of the numerical literature which is becoming more and more essential to control engineering. This fact is reflected in the increasing attention for and development of software packages for systems

and control.

In this thesis, we shall not undertake an explicit numerical analysis of the algorithms we use. However, the abundant use of (generalized) singular value decomposition guarantees almost surely the *numerical* reliability of our algorithms. Whenever possible, we shall develop concepts and insights in terms of the singular value decomposition, henceforth permitting an almost trivial numerical reliable implementation while at, the same time, guaranteeing an elegant geometrical interpretation. All the programming, the numerical testing and the practical implementation of our algorithms has been performed in *PRO-MATLAB.* [1]

1.1.3 Linearity.

The one organizational principle for which system theory has provided a thorough understanding is the concept of *linearity.* In this thesis, the assumption of linearity is neither a question of fact nor of evaluation, but a self imposed limitation on the types of operations or devices used. The mathematical reason for this assumption is clear: linear problems are almost always much simpler than their nonlinear generalizations. In certain applications moreover, the linearity assumption may be justified on more rigorous grounds: A linear predictor is the absolute optimal method if the time series is Gaussian. Linearity may be dictated by the simplicity of mechanization. Linear filters are easy to synthesize and there is an extensive relevant theory, with no corresponding wealth of transparant results in nonlinear system theory. Linear theory may merely be used because of the lack of any better approach. An incomplete solution is better than none at all.

One may however not underestimate the meaning of linearity. As an example, the algebraic eigenvalue problem, which is *an sich* a non-linear problem, is considered to be part of linear algebra. As a matter of fact, most computational problems in this thesis reduce to the calculation of the (generalized) singular value decomposition, which is another non-linear factorization of a matrix.

A technique that returns in almost all chapters is an algebraic trick that we have baptised *homogenization.* The concept can best be illustrated by the problem of solving a set of linear equations with a non-zero right hand side:

$$Ax = b$$

Homogenization in this case consists of introducing an additional parameter α and then rewriting the problem in a homogeneous form as :

$$(A \;\; b)\begin{pmatrix} x \\ \alpha \end{pmatrix} = 0$$

The homogeneous problem of increased dimension is then solved. In fact, one solves a kind of generalized problem : The case where $\alpha = 0$ correspond to solutions of $Ax = 0$, which is equivalent to the problem of solving a set of homogeneuous linear equations. The corresponding solutions for x are called the homogeneous solutions or in some applications, solutions at

[1] PRO-MATLAB, VAX/VMX version 2.1.-VMS, December 1, 1986, Mathworks Inc., MA, USA

infinity. The other solutions can be normalized such that $\alpha = 1$. They correspond to finite solutions.

Homogenization will be applied to find all nonnegative solution of sets of linear equations (chapter 2), the solution of the Generalized Linear Complementarity Problem (chapter 3), the Total Linear Least Squares principle (chapter 5) and the derivation and identification of generalized linear state space models (chapter 8). In fact, the advantages of the homogenization technique can be summarized as follows :

1. One finds at once all solutions (if any) of the corresponding homogeneous problem. In a lot of cases, the general solution set is the sum of all solutions of the homogeneous problem plus one particular solution. This is an explicit consequence of the *linearity* of the problem under consideration. As an example, consider again the general solution of a set of linear equations $Ax = b$ where A is a $m \times n$ matrix. Assume that x_1 and x_2 are two solutions. Then, obviously the vector $x_2 - x_1$ is a solution to the homogeneous problem $Ax = 0$. Suppose that the set of solutions of the homogeneous problem is parametrized as $X_h y$ where the columns of X_h generate a basis for the null space of the matrix A. The vector y takes all possible combinations of these basic solution vectors. Then obviously there must exist a vector y_2 such that $x_2 = x_1 + X_h\, y_2$. Hence, we have proved that all solutions to the non-homogeneous problem can be considered as *the sum of a particular solution of the non-homogeneous with all solutions of the homogeneous problem.* This proof readily carries over to the description of the solution set of linear differential and difference equations, to the nonnegative solution of sets of linear equations (chapter 2), to the description of the solution set of the Generalized Linear Complementarity Problem (chapter 3). The solution to the homogeneous problem is often more easily found in terms of matrix computations than in the case of the non-homogeneous problem formulation.

2. The solutions at infinity often have a physical interpretation and the behavior can be much richer by including the properties at infinity. The additional parameters that characterize the possible behavior at infinity also have a stabilizing influence with respect to the numerical sensitivity of the solution. As an example, it will be shown how ill conditioned problems can be solved without any complication via *homogenization* while the classical problem formulation (without additional parameters) leads to numerical unreliability (e.g.least squares etc...).

1.2 Survey of the Thesis.

1.2.1 Relation between the chapters.

This doctoral dissertation contains 9 chapters. You are now reading the first one. The last one consists of the main conclusions of this work and some suggestions and perspectives for future research. Chapter 2 to 8 are interrelated as depicted in figure 1.1. We shall now briefly summarize the contents and main results of each chapter.

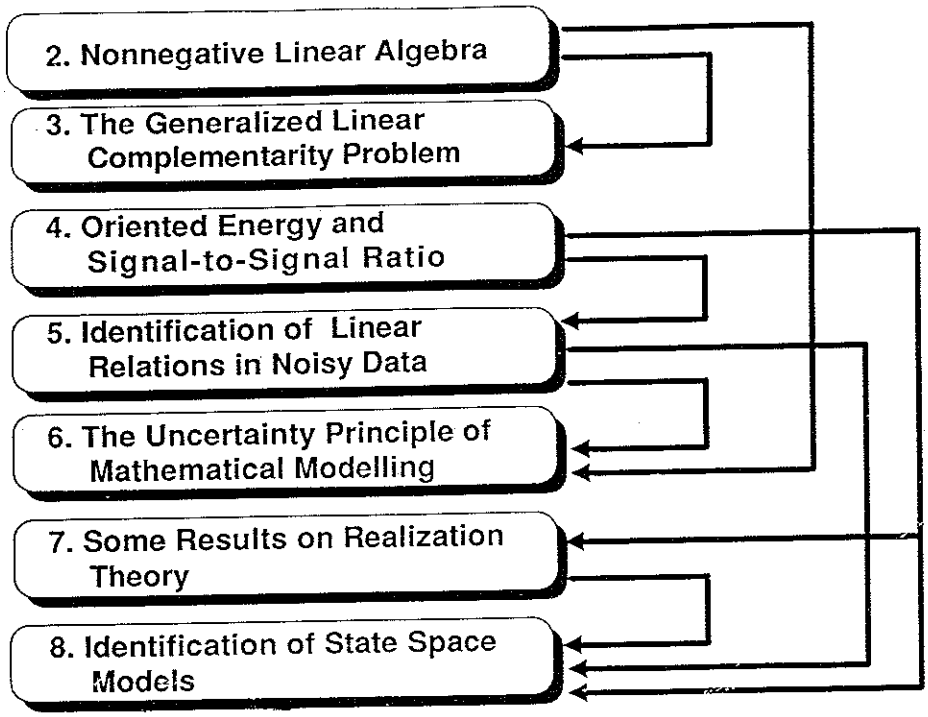


Figure 1.1: Relation between the chapters 2 to 8.

1.2.2 Chapter 2: Nonnegative Linear Algebra.

In this chapter, an algorithm is developed to find nonnegative solutions of sets of linear (in-)equalities. The algorithm essentially consists of a double inductive argument:

- The first induction is based upon updating in each stage the polyhedral cone resulting from the previous step.
- The second induction shows how it is allowed to investigate in each stage the *redundancy* of the vertices that describe the polyhedral solution cone and eliminate vertices that are not *adjacent*.

While these results are known in literature, there appears to be no formal proof nor a global algorithm. Both are provided in this chapter, including necessary and sufficient conditions and tests for extremity and adjacency of the vectors of the polyhedral solution cone.

1.2.3 Chapter 3: The Generalized Linear Complementarity Problem.

The Linear Complementarity Problem has been studied now for more than 20 years in literature. First we show how a slight generalization of the classical problem allows for several extensions:

- It allows to obtain *mathematical models* of piecewise linear systems that are *more general* than the classical ones. The main reason is that certain matrices are allowed to be *singular*.

- Because of the allowed singularity of certain matrices, there is no need for an inversion-type preprocessing step in order to convert the problem to a classical linear complementarity problem. This results in an increased *numerical reliability.*

- An important feature of our approach is that the *complementarity conditions* may be much *more general* than those in the classical linear complementarity problem. As a matter of fact, there is no orthogonality involved anymore and the complementarity conditions are expressed as sums of products.

We derive an algorithm that is based on a threefold induction argument:

- First the problem is converted to the problem of solving a set of linear equations non-negatively. Hereto, the algorithm derived in chapter 2 is employed.

- Next, it is shown how in each stage of the algorithm, it is allowed to eliminate the intermediate solutions that are not complementary. This results in considerable savings concerning the memory requirements.

- Contrary to existing LCP-solvers, our algorithm explicitly finds *all solutions* to the generalized linear complementarity problem, finite or infinite in number. It is shown how in general, the solution set consists of the union of discrete vectors, polytopes and polyhedral cones.

- It is demonstrated how a straightforward investigation of the extremal rays via the so-called *cross-complementarity* conditions allows to gain considerable insight in the geometrical nature of the solution set.

- The robustness of the solution is briefly investigated.

Finally, attention is paid to the many applications of the generalized linear complementarity problem including piecewise linear descriptions via several parametrizations, computing intersections between $3 - D$ geometrical objects, mathematical programming and quadratic optimization and modelling piecewise linear resistive networks. It is moreover shown how the computation of the invariant states of a *neural network* with and without partial a priori information, reduces to the solution of a generalized linear complementarity problem.

1.2.4 Chapter 4: Oriented energy and signal-to-signal ratios.

In a lot of signal processing and identification approaches, the mathematical model consists of certain subspaces, that reveal information concerning the spatial distribution of observed vector sequences. In this chapter, it will be derived how the singular value decomposition contains this information about one vector sequence, while the generalized singular value decomposition allows to compare the relative spatial distribution of 2 vector sequences. It is demonstrated how the maximal minimal signal-to-signal ratio between two vector sequences and the corresponding subspaces, can be computed. It is shown how this framework allows to classify and unify a lot of mathematical modelling techniques and applications.

1.2.5 Chapter 5: Identification of linear relations in noisy data.

First, the intimate relation between orthogonality, linearity and noise is discussed. Two points of view concerning mathematical modelling are developed: the deductive and the inspirational approach.

The chapter contains 2 central results. The *orthogonality* theorem states under which conditions, 'exact' and 'noise' vectors are orthogonal to each other. The *lever theorem* investigates the singular value decomposition of the sum of 2 matrices under certain orthogonality conditions. This theorem is shown to be extremely important in a lot of applications and is the basic idea behind a lot of identification approaches.

- The theorem essentially states that the so-called *short* space of a matrix can be *consistently* estimated if exact data are perturbed by additive noise. The *long* space however is irreversibly lost.
- It allows to prove *consistency* of some well known statistical estimators such as (total) linear least squares.
- Together with the concept of oriented energy, it allows to explain the rationale behind the used inproduct (inverse of the noise covariance matrix) in Gauss-Markov minimum variance approaches.
- It allows to compute the *bias* of canonical angles between *long* spaces.

In a last section, the notion of *canonical angles between subspaces* is analysed in depth. The biasedness is explained and computed. Several algorithms, based upon the generalized singular value decompsosition are provided. They all compute approximate intersection between subspaces, if the data are noisy. While most of them are ultimately based upon the canonical angles, it is shown how a recently derived criterion takes into account the oriented energy of the signal, in contrast with the canonical correlation approach.

1.2.6 Chapter 6: The Uncertainty Principle of Mathematical Modelling.

Based upon some recent ideas of Kalman, we develop in this chapter some algebraic and numerical tools that allow to compute the minimal rank of a symmetric positive definite matrix, when only its diagonal elements can be modified, subject to nonnegative definiteness conditions. The resulting algorithms basically reduce to the computation of the nonnegative solutions of a set of linear equations. This approach allows for considerable geometrical insight into the general solution. The conceptual consequences are farreaching. It is shown how certain polyhedral solution sets allow to quantify the deviation of the observed data from linearity.

1.2.7 Chapter 7: Some results in Realization Theory.

Realization theory is concerned with the axiomatic definition of 'system' and with relating several internal system descriptions (e.g. state descriptions) with external (e.g. input-output descriptions). In this chapter, it is shown how some existing realization algorithms that are based upon the singular value decomposition of block Hankel matrices containing the Markov parameters of a system, follow naturally from the so-called *shift structure* of certain subspaces.

It is shown how there is a one-to-one relation between these subspaces and the eigenvalues that summarize the dynamical behavior. Furthermore, the link between oriented energy and quantitive characterizations of controllability and observability is exploited via the singular values of block Hankel matrices. Using the non-conventional matrix calculus of Kronecker and Khatri-Rao products, we find a parametrization of all block Hankel matrices with a certain set of minimal system poles. It is shown that the kernel of a certain matrix contains all necessary information. The results can be used in design problems, where one wants to optimize controllability and observability of the system.

This theory is not restricted to block Hankel matrices, which is demonstrated by the introduction of *structure exploiting factor analysis*. It is demonstrated how it is possible to find subspaces with shift structure of maximal dimension in noisy data. Attention is paid to matrices with shiftstructure that are at the same time orthonormal.

1.2.8 Chapter 8: Identification of State Space Models.

In a first part of this chapter, some conceptual results on state space models for linear lumped, discrete time, time-invariant finite dimensional systems are derived. The prominent role of a block Hankel matrix, consisting of input-output observations of the systems, is emphasized, The conceptual results include:

- The rank of the input-output block Hankel matrix is analysed in detail. It allows to gain considerable information about the system under study.
- It is shown how the state sequence can be derived from the intersection of row spaces of past and future block Hankel matrices.
- Some interesting phenomena are analysed in depth, such as *rank cancellation*, systems with *delays*, *persistancy of excitation*, the mathematical characterization of *causality*, etc...

Based upon the properties derived in the first part, two kinds of algorithms are derived in a deductive framework. First a heuristic linear and total linear least squares approach are studied. Both exploit the shift structure of certain subspaces. The behavior with noisy data is analysed. Second, several algorithms based upon the computation of an approaximate state sequence are derived, employing the algorithms of chapter 5. It is shown how everything may be converted to the so-called *short* space of the block Hankel matrices. This allows to conjecture consistency of the derived models and results in considerably computational savings, allowing for an on-line implementation of the algorithms.

In a last part, the theory is illustrated with several industrial application examples, including the identification of a power plant, an ecological system, a chemical destillation column and a glass tube production installation.

Chapter 2

Nonnegative Linear Algebra

$$\sum_{i=0}^{d-1}(-1)^i F_i(M) = 1 + (-1)^{d-1}$$
The Euler - Poincaré polytope formula

2.1 Introduction

Nonnegative linear algebra is linear algebra in which one imposes certain equality and inequality constraints on the variables while performing linear algebraic operations such as solving sets of equations. The imposed conditions may be quite straightforward as for instance requiring that the variables that solve a problem must be nonnegative. However, as will become obvious in this and the next chapter, the constraints may be of a much more general nature: examples are inequalities , (partial) orthogonality and complementarity conditions.
The rich variety of geometrical concepts and objects has been studied since ancient times. The thirteenth book of Euclid is devoted to the five regular polytopes known as the *platonic solids*. In his work *On polyhedra*, Archimedes described all semi-regular polyhedra. Important mathematical contributions are due to famous mathematicians as Minkowski, Voronoi, Poincaré, Caratheodory in the beginning of this century and Weyl, Kantorovich, Dantzig, Klee and Khachian in the second half of the century. However, already Euler delivered some marvellous results in combinatorial topology such as his celebrated relation from 1752 : $v-e+f = 2$. It is interesting to note that this formula was already known to Descartes about a hundred years earlier. However, his manuscript was lost and a partial copy of it was only found in 1860 among Leibniz's papers. Poincaré generalized Euler's formula to arbitrary d-polytopes, using a topological proof [14, p.45]. Although today, the field has achieved a certain degree of maturity, it is still the subject of intensive ongoing research. For instance, mathematical programming algorithms for the optimization of linear and quadratic cost functions subject to constraints have resulted in considerable economical benefits. Consider as an example the linear programming problem in which a linear cost function is to be optimized, subject to certain linear (in-)equality constraints on the variables. There is an exciting historical excursion with all kinds of *petites histoires* one can imagine. It starts with Dantzig's *simplex algorithm*[9], which, since its discovery in 1948, is one of the most frequently used algorithms in the world. It goes over the 1979 Soviet *ellipsoid algorithm* of Khachian [9], that has pro-

vided a theoretical justification of the existence of a polynomial time approach and leads to Karmarkar's mysterious approach, which is claimed to be very fast but which, for commercial reasons, is kept secret.
Moreover, the analysis of the geometrical objects themselves, induced by the constraints of economical and transportation problems, is still the subject of intensive research (e.g. multi-index transportation polytopes with(out) side conditions [14]). The study of the figures formed by the vertices and edges of any three-dimensional polytope led to another discipline : *graph theory* which, together with *combinatorial topology*, established itself as an independent branch of mathematics, commonly referred to as *discrete mathematics*. Other applications can be found in the *geometry of numbers*, initiated by Minkowski [14, p.133] and *mathemetical cristallography*. Another quite recent development is that discipline which was baptized as *computational geometry*. It has its roots in nonnegative linear algebra and finds its applications in robotics, VLSI design, architecture and numerous other applications [10].
As a matter of fact, nonnegative linear algebra is part of our daily life (not that a lot of people are aware of it). This page you are reading, the volume generated by this thesis, the room where you are sitting in, the diamond you just offered to your wife and the little pointed sack that contains the Belgian (not French !) frites you are eating, they all are geometrical objects that can be described mathematically employing the tools of nonnegative linear algebra . Loosely speaking, one of the main features of the geometry of nonnegative linear algebra is its departure from the 'smoothness' properties that are shared by the classical linear algebraic objects such as half-spaces, spheres etc .... Objects induced by nonnegavity conditions may be pointed and can posses corners, they may contain broken lines, as in piecewise linear descriptions, they may be convex and multifacetted, etc ...

In this chapter, the problem of *finding nonnegative solutions to a set of linear equations* will be discussed and an algorithm will be proposed and analysed. Strange enough, in basic books on the subject, little attention is paid to this problem, though it is our conviction that the problem and its solution are fundamental both from the *didactical* and *scientific* point of view. Of course, in general, the problem receives attention indirectly since the constraints of linear programming are of the same form. However, implementations as for instance the *simplex algorithm* only take into account part of the solution set, because this can be very complicated. The sources of inspiration of the present work are the following:

- In our research, we have first encountered the problem via the questions that arose from the Uncertainty Principle of Mathematical Modeling, which will be discussed in chapter 6. The same problem provided the abundant amount of inspiration, which lead to the numerous results in chapter 3 (The Generalized Linear Complementarity Problem).

- First a geometrical algorithm was derived and the corresponding properties were rigorously proved [1](see also section 2.3.1 and 2.3.2).

- However, more than 50 years ago, Motzkin, in his Ph.D. thesis of 1936 [5], already proposed a satisfactory approach, which was published as a paper in 1953 [6]. The algorithm is called the *double description method* and it is derived in the context of two-person zero-sum games with a finite number of pure strategies. Our approach bears a lot of similarity with these results. However, the 1953 Motzkin paper contains no proofs which will be given in this chapter.

- Redundancy elimination tests were derived and rigorously proved in [2]. The techniques are similar to those employed in Linear Programming literature [9] [14].
- A lot of useful results can be found in the work of Greenberg [4]. However, some of his results are erroneous and raised quite some polemics in literature [12] [13].

This chapter is organised as follows: Some basic geometrical objects induced by (nonnegativity) constraints (equalities and inequalities) are described in section 2.2. An algorithm to find nonnegative solutions to sets of linear equations is developed in section 2.3., employing a geometrical approach. The obvious generalization to sets of linear inequalities is presented in section 2.4. The results in chapter 3 (The Generalized Linear Complementarity Problem) and in chapter 6 (The Uncertainty Principle of Mathematical Modeling) are fundamentally based upon the algorithm presented in this chapter.

2.2 The geometrical objects of nonnegative linear algebra.

Consider the general linear problem which is central in a lot of applications:

$$\text{Solve for } x \text{ in } Ax \leq b \tag{2.1}$$

The geometrical description of the solution set of this problem will be defined and analysed. For a detailed exposition, the reader is referred to excellent books as [9] [11] [14]. Henceforth, it is assumed that the reader is familiar with the definitions and properties of nonnegative linear algebra, such as affine, nonnegative and convex combinations, polyhedrons, polyhedral cones, polytopes and simplices (see appendix A). However, for completeness and to sharpen the geometrical intuition, the main features of nonnegative linear algebra will be summarized in this section.
First, consider the homogeneous problem:

$$Ax \leq 0 \tag{2.2}$$

The following properties, though trivial, are important.

Lemma 1 *Let x_1 and x_2 be 2 solutions to $Ax \leq 0$.*

1. *Any nonnegative multiple of x_1 and x_2 is also a solution.*
2. *Any nonnegative combination of x_1 and x_2 is also a solution.*

Proof : Trivial □

These properties define a geometrical object, which is called a *polyhedral cone.* More specifically, we have:

Definition 1 Polyhedral Cone
A polyhedral cone is the set of solutions of a finite system of homogeneous linear inequalities.

There is an extremely important characterization of a polyhedral cone, which is due to Hermann Weyl [14] :

Theorem 1 The Weyl theorem
A polyhedral cone is finitely generated.

Proof : See e.g. [14, p.16], [11, p.171] □

Hence, the property of being polyhedral is a finiteness condition on the external representation of a convex set. This deep result implies that any vector of a polyhedral cone can be written as a nonnegative (also called conical) combination of a *finite* number of vectors.

Corollary 1 *Given a finite set of vectors. All nonnegative combinations of these vectors define a polydehral cone, that is said to be* generated *by these vectors.*

Let A be a matrix. The expression '*the polyhedral cone* A' will be used henceforth to denote the polyhedral cone generated by the column vectors of this matrix. Similarly, the notation $x \in A$ where A is a matrix, implies that the vector x belongs to the polyhedral cone generated by the columns of A. However, observe that some redundancy may be present among these column vectors. Possibly, some of these may be written as a nonnegative combination of others. Hence, the question of minimality of the number of vectors to represent a polyhedral cone is an important one. The minimal set of vectors, that is needed to generate the complete cone with nonnegative linear combinations, are the so called extremal rays of the polyhedral cone:

Definition 2 Extremal rays of a polyhedral cone
A vector v is an extreme ray of a polyhedral cone A if there exists a hyperplane $\mathcal{H} = \{x \in \mathcal{R}^n \mid h^t x = 0\}$ *such that* $\mathcal{H} \cap A = \{x \mid x = v\lambda\ ,\ \lambda \geq 0\}$

The first orthant is a particular example of a polyhedral cone, generated by the natural basis vectors defined as $e_i^k = \delta_{ki}$, which are its extremal rays. The question of redundancy will further be discussed in detail in section 2.3.

Let's now have a look at the non-homogeneous problem (2.1). If x_1 and x_2 are solutions, then any convex combination of these vectors is also a solution. Hence, we have:

Lemma 2 *The set of solution vectors of any (finite or infinite) system of linear inequalities is either convex or empty (if the system is inconsistent).*

Proof : Trivial □

Consider a particular solution x_1 satisfying $Ax_1 \leq b$ and the polyhedral solution cone of the homogeneous problem $Ax \leq 0$ generated by the columns of the matrix X_h. Then obviously, also $A(x_1 + X_h y) \leq b$ is satisfied for any nonnegative vector y of appropriate dimension. Hence, the solution to a set of linear inequalities is unbounded (i.e. not contained within a sphere), whenever the homogeneous problem has a non-trivial, non-zero solution. Moreover, any convex combination of particular solutions to the non-homogeneous problem is a particular solution as well. The geometrical object defined by this collection of solutions is called a *polyhedron.*

Definition 3 Polyhedron
A polyhedron is the solution set of a finite system of linear inequalities.

The solutions to the homogeneous problem can be interpreted as solutions *at infinity.* Any vector which can be written as the sum of a solution of the non-homogeneous problem and a 'direction', which is contained within the polyhedral solution cone of the homogeneous problem, is a solution to the non-homogeneous problem. This is stated in the following result:

Theorem 2 The Goldman Resolution Theorem
The solution polyhedron to $Ax \leq b$ is the set of convex combinations of its extreme points plus the nonnegative linear combinations of generators of the extreme rays of the recession cone, defined by $Ax \leq 0$.

However, the case where the homogeneous problem has no non-trivial solution is of particular interest.

Definition 4 Polytope
A polytope is the convex hull of a finite set of points.

Important subsets of a polytope are its *faces* :

Definition 5 Faces of a polytope
Let $\mathcal{H}$ be a supporting hyperplane of the polytope $\mathcal{M}$. The set $\mathcal{F} = \mathcal{M} \cap \mathcal{H}$ is a face of the polytope $\mathcal{M}$, generated by $\mathcal{H}$. If $dim(\mathcal{F}) = i$, then $\mathcal{F}$ is an i-face of the polytope $\mathcal{M}$.

A 0-face is called a *vertex* while a 1-face is called an *edge*. The empty set and $\mathcal{M}$ itself are called *improper* faces. All other faces are called *proper*. The dimension of a polytope (face) equals the dimension of its affine hull. If $dim(\mathcal{M}) = d$, the $(d-1)$-faces are called *facets*. They are the proper faces of maximal dimension. The Euler-Poincaré formula mentioned at the beginning of this chapter, relates the number of faces of all dimensions. Generically, it is the only *linear* relation between the number of vertices. The vertices of a polytope are its only extreme points, as a matter of fact, a polytope is the convex hull of its vertices, which are the points of the polytope that cannot be written as a convex combination of any two other points. An important result is the following:

Theorem 3 The Weyl (1897) - Minkowski (1935) theorem
The set M is a polytope if and only if it is a bounded polyhedron.

Proof : [14, p.18]. □

The Weyl-Minkowski theorem shows that every polytope in a fixed coordinate system can be specified by means of a finite system of linear inequalities. This permits on the one hand, to utilise the well developed apparatus of the theory of linear inequalities to study polytopes and on the other hand, to give an algebraic interpretation of the geometrical properties of polytopes.

This concludes this elementary introduction in which we have described the most important geometrical objects. For a more detailed discussion and much more properties (and proofs!), the reader is referred to appendix A and the literature. We are now ready to start the development of the main topic of this chapter.

2.3 Nonnegative solutions to sets of linear equalities

In this section, we shall restrict our attention to the analysis and solution of the following homogeneous problem:

$$\text{Solve for } x \geq 0 \text{ in } Ax = 0 \tag{2.3}$$

where A is a $m \times n$ real matrix and b is an m-vector. Observe that via *homogenization* any non-homogeneous problem of solving nonnegatively a set of linear equalities can be converted into the homogeneous (in fact more general) form. The nonnegativity constraints however, represent a restriction of the solution of (2.3) with respect to that of (2.2). As a matter of fact, the solution set is still a polyhedral cone, but it is *pointed.*

Definition 6 Pointed Polyhedral Cone
A pointed polyhedral cone is a cone that does not contain a straight line or equivalently , a cone with a unique vertex.

Theorem 4 *The solution set to* $Ax = 0$, $x \geq 0$ *is a pointed polyhedral cone.*

Proof : Trivial □

Note that the problem (2.3) always has at least one solution, $x = 0$, which is called henceforth the *trivial* solution. The problem of the nonnegative solution of a set of linear equations arises in a lot of applications:

- First of all, nonnegativity constraints are natural for a lot of physical and economical quantities, such as weights, prices, energies, etc.... As will be demonstrated in chapter 3, linear least squares estimation with nonnegativity constraints on the variables can be solved via an algorithm, that consists basically of solving nonnegatively a set of linear equations with so called complementarity constraints.

- The results and algorithms to be described in chapter 3 (The Generalized Linear Complementarity Problem) and chapter 6 (The Uncertainty Principle of Mathematical Modeling) are basically applications of solving nonnegatively a set of linear equalities.

- Linear programming consists of the optimization of a linear cost function, subject to a system of linear inequalities. Existing algorithms to solve this problems, such as the celebrated *simplex method* [9], do not 'visit' all solutions to the linear inequalities, but follow only a path of increasing optimality of the cost function, along part of the set of extremal vertices of the solution set. However, in some applications, the explicit determination of all extremal solutions may be of interest, such as minimization problems in various directions as in the case of an economy with variable prices [6]. Moreover, in other economical applications, one is also interested in *all* vertices of the solution polyhedron, as is the case in e.g. transportation polytopes [14]. Intensive research is still going on in order to characterize and count the number of faces of such polytopes.

- Another application where one may be interested in the complete solution set of a system of inequalities is the domain of system identification and adaptive control . Examples can be found in [7, p.234](and the references cited in this book), where it is shown how to update an estimate of unknown variables, in case that the noise is described by bounds via linear inequalities. Based upon the same noise model via linear inequalities, adaptive algorithms are derived in [8], which remarkably enough bear a close resemblance with the Kalman filter formula's. A similar approach for system identification based upon linear inequalities is proposed in [13], where a similar algorithm is discussed as the one that will be studied in this chapter. In [3], it is shown how adaptive control with constrained

inputs leads to a quadratic programming problem. In chapter 5 it is demonstrated how such problems are equivalent with the so called Generalized Linear Complementarity Problem and hence with the results of this chapter.

The algorithm to solve (2.3) will be derived in an inductive way via a geometrical approach in 3 steps.

1. First, the solution for A consisting of one row vector only will be analysed in detail (subsection 2.3.1). The basic observation stems from geometrical considerations about the intersection of a halfspace with a polyhedral cone.

2. Next, this will be generalized for the case of A having multiple rows (subsection 2.3.2). This reduces geometrically to the intersection of several polyhedral cones.

3. However, as will be shown, these first intuitive developments lead to redundant solutions, i.e. solution vectors x that can be written as a convex (or nonnegative) linear combination of other solutions (never rely on your intuition only !). Necessary and sufficient conditons for redundancy elimation will be discussed in subsection 2.3.3.

In section 2.3.4, some very useful observations concerning the implementation will be discussed.

2.3.1 The intersection of a halfspace and a polyhedral cone

Consider in the Euclidean vectorspace $\mathcal{R}^n$ the nonnegative orthant (henceforth called the *first* orthant), which is a polyhedral cone with extremal rays represented by the columns of the identity matrix I_n. The $(n-1)$-faces of this polyhedral cone will be denoted as *orthant planes*. The i-th orthant plane is the one orthogonal to the i-th column e^i of I_n. Consider the hyperplane $\mathcal{H}$ through the origin, defined by an $1 \times n$ row vector a^t:

$$\mathcal{H} = \{x \in \mathcal{R}^n \mid a^t x = 0\}$$

This hyperplane also defines two closed half-spaces $\mathcal{H}^+ = \{x \in \mathcal{R}^n \mid a^t x \geq 0\}$ (called the *positive half-space*) and $\mathcal{H}^- = \{x \in \mathcal{R}^n \mid a^t x \leq 0\}$ (the *negative half-space*). The problem that will now be discussed is the computation of the description of the geometrical object that results from the intersection of the hyperplane $\mathcal{H}$ with the first orthant. From definition 1, it is already known that the solution set will be a *polyhedral cone*. Take two columns e^i and e^j of I_n and determine in which half-space they are lying. This is obviously revealed by the signs of the components a_i and a_j. Three cases may occur:

1. At least one of the 2 components is zero. Obviously, if $a_i = 0$, $a^t e^i = 0$ and hence e^i belongs the the polyhedral solution cone. If $a_j = 0$, $a^t e^j = 0$ and e^j belongs to the polyhedral solution cone.

2. Both a_i and a_j are non-zero, but have a different sign. Assume that $a_i > 0$ and $a_j < 0$. This indicates that e^i lies in the positive half-space $\mathcal{H}^+$ while e^j belongs to the negative half-space $\mathcal{H}^-$. Hence, none of them belongs to the solution set. It is straightforward to show that the vector x, generated by the nonnegative combination $x = e^i \mid a_j \mid + e^j \mid a_i \mid$ is orthogonal to a^t and nonnegative, hence belongs to the polyhedral solution cone.

3. Both a_i and a_j are non-zero, but have the same sign. Obviously, e^i nor e^j are a solution, neither is any nonnegative combination of them.

The preceding recipe is now formalized. As a matter of fact, an even stronger result is proved:

Theorem 5 Nonnegative vectors orthogonal to a vector
Given an $n \times 1$ vector $a \in \mathcal{R}^n$. The polyhedral cone of nonnegative vectors orthogonal to a is generated by the column vectors of a matrix S that is constructed as follows :

1. *For each zero component $a_k = 0$, add to the solution columns the k-th column e^k of the identity matrix I_n.*
2. *For each pair of non-zero components a_k, a_l such that $a_k a_l < 0$, add the vector $e^k \mid a_l \mid + e^l \mid a_k \mid$.*

Before proving the theorem, let's first count the number of columns of the matrix S. Hereto, introduce the index sets $\mathcal{I}_a^+ = \{i \mid a_i > 0\}$, $\mathcal{I}_a^- = \{i \mid a_i < 0\}$, $\mathcal{I}_a^0 = \{i \mid a_i = 0\}$ and their cardinalities $\sigma_a^+ = \#(\mathcal{I}_a^+)$, $\sigma_a^- = \#(\mathcal{I}_a^-)$, $\sigma_a^0 = \#(\mathcal{I}_a^0)$.

Lemma 3 *The matrix S, constructed in theorem 5, counts $p = \sigma_a^+ \sigma_a^- + \sigma_a^0$ columns.*

Proof : Trivial. □

We are now ready to prove theorem 5.
Proof : Let p be the number of columns of the matrix S (lemma 3). It is an easy exercise to prove that any vector, constructed from the recipe described in the theorem, is nonnegative and orthogonal to a, hence belongs to the polyhedral solution cone. It is now proved that every nonnegative vector x, orthogonal to a, can be written as a nonnegative combination of the columns of S:

$$\forall\, x \;:\; a^t x = 0\,,\; x \geq 0 \;\;, \exists\, v \in \mathcal{R}^p\,,\; v \geq 0 \;\text{ such that }\; (\; S \;\; -x\;) \begin{pmatrix} v \\ -1 \end{pmatrix} = 0$$

or equivalently:

$$\forall\, x \;:\; a^t x = 0\,,\; x \geq 0 \;\;, \exists\, w \in \mathcal{R}^{p+1}\,,\; w \geq 0 \;\text{ such that }\; (\; S \;\; -x\;)\, w = 0 \text{ with } w_{p+1} \neq 0$$

The proof is constructive. First observe that every column of S has either one or two nonzero elements. For every vector $x \geq 0$ with $a^t x = 0$, construct w as follows:

1. The last component w_{p+1} is defined as:

$$w_{p+1} = \sum_{k \in \mathcal{I}_a^-} a_k x_k$$

   Observe that this choice guarantees that $w_{p+1} \neq 0$.

2. If the i-th column s^i of S has 2 nonzero elements (assume the k-th and the l-th one)

$$s^k = [\; 0 \;\ldots\; \mid a_l \mid \;\ldots\; \mid a_k \mid \;\ldots\; 0\;]^t$$

   then take $w_i = x_k x_l$

3. If the j-th column s^j of S has 1 nonzero element (assume the k-th one which equals 1), then take $w_j = x_j w_{p+1}$.

Three different kinds of rows may occur in the matrix S:

1. Either the row contains zeros and components $\mid a_i \mid$ where $i \in \mathcal{I}_a^+$.
2. Either the row contains zeros and components $\mid a_j \mid$ where $j \in \mathcal{I}_a^-$.
3. Either the row contains zeros and only one nonzero component equal to 1.

In all these cases, the construction of the vector w is precisely such that $(S - x)\, w = 0$ with $w_{p+1} \neq 0$, $w \geq 0$. □

The recipe of theorem 5 and its proof are clarified by the following example:

Example 1:
Given the vector $a^t = [\, a_1 \;\; -a_2 \;\; 0 \;\; a_4 \;\; 0 \;\; -a_6 \;\; -a_7 \;\; a_8 \,]$ where all $a_i > 0$. Then the matrix S can be constructed from theorem 5. According to lemma 3 , it will contain 11 vectors:

$$S = \begin{pmatrix} a_2 & a_6 & a_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ a_1 & 0 & 0 & a_4 & a_8 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_2 & 0 & 0 & a_6 & a_7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & a_1 & 0 & 0 & 0 & 0 & a_4 & 0 & 0 & a_8 & 0 \\ 0 & 0 & a_1 & 0 & 0 & 0 & 0 & a_4 & 0 & 0 & a_8 \\ 0 & 0 & 0 & 0 & a_2 & 0 & 0 & 0 & 0 & a_6 & a_7 \end{pmatrix}$$

Every nonnegative vector x orthogonal to a can be written as a nonnegative combination of the columns of S. The nonnegative vector w described in the proof of theorem 5, is easily seen to be :

$$w^t = (x_1 x_2 \;\; x_1 x_6 \;\; x_1 x_7 \;\; x_2 x_4 \;\; x_2 x_8 \;\; x_3 w_{12} \;\; x_4 x_6 \;\; x_4 x_7 \;\; x_5 w_{12} \;\; x_6 x_8 \;\; x_7 x_8 \;\; w_{12})$$

with $w_{12} = a_2 x_2 + a_6 x_6 + a_7 x_7$.

2.3.2 Nonnegative vectors orthogonal to several vectors

Now consider instead of one vector a, several n-vectors a^i , $i = 1, \ldots, m$ which are row vectors of an $m \times n$ matrix A. We shall now describe an algorithm to find all nonnegative vectors x orthogonal to the row space of the matrix A: $A\, x = 0$, $x \geq 0$. From definition 1, we already know that the solution set is a polyhedral cone.

Theorem 6 Nonnegativity and orthogonality
Given an $m \times n$ matrix A. The polyhedral cone of all nonnegative vectors x orthogonal to the row space of A, is generated by the columns of the matrix $S = S_1 S_2 \ldots S_l$, $l \leq m$, to be determined via the following procedure:

1. *S_1 is determined from a^1 via theorem 5.*

2. *The matrix S_i ($i \geq 2$) is determined according to theorem 5 for the vector* $[(a^i)^t S_1 \dots S_{i-1}]$

Proof : Since we are looking for all nonnegative vectors x such that $Ax = 0$, we have obviously that $(a^1)^t x = 0$ for any solution x. Hence, from theorem 5, it follows that:

$$\forall\, x \geq 0\ ,\ \exists\, w^1 \geq 0 \ \text{ such that } \ x = S_1 w^1$$

where S_1 is constructed as in theorem 5. Hence, $Ax = 0$, $x \geq 0$ implies $AS_1 w^1 = 0$ and obviously, the first row of the matrix AS_1 is zero. Consider the second row of AS_1, which is $(a^2)^t S_1$. Again from the application of theorem 5 to the matrix AS_1, it follows that :

$$\forall\, w^1 \geq 0\ ,\ \exists\, w^2 \geq 0 \ \text{ such that } \ w^1 = S_2 w^2$$

where S_2 is constructed from the elements of the vector $(a^2)^t$ as in theorem 5. Hence, for all vectors $x \geq 0$ satisfying $Ax = 0$, there exists a vector $w^2 \geq 0$ such that $x = S_1 S_2 w^2$. The first two rows of $AS_1 S_2$ are zero. Proceeding this process, one finds at the k-th stage, corresponding to the k-th row $(a^k)^t$ of A, that there must exist a vector $w^k \geq 0$ such that:

$$x = S_1 S_2 \dots S_k w^k$$

the first k rows of $AS_1 S_2 \dots S_k$ being zero. The procedure stops if for $l \leq m$, $AS_1 \dots S_l = 0$. The solution set is then obviously the polyhedral cone generated from all possible nonnegative combinations of the matrix $S = S_1 \dots S_l$. There is no solution if, for a certain $l \leq m$, there is at least one row of $AS_1 \dots S_l$ with either only strictly negative or strictly positive elements. This completes the proof. □

Let's now clarify theorem and proof with an example, that will also lead us to an observation, which requires some further fundamental analysis:

Example 2:
Consider the matrix

$$A = \begin{pmatrix} -3 & -1 & 3 & 0 & 1 \\ 4 & 2 & 0 & 1 & -3 \\ 1 & 1 & 1 & 0 & -1 \end{pmatrix}$$

Applying the algorithm of theorem 6, results in the following self-explaining sequence of matrices:

$$S_1 = \begin{pmatrix} 3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 \\ 3 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 3 & 0 & 1 & 0 \end{pmatrix} \qquad AS_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 12 & -5 & 6 & -1 & 1 \\ 6 & -2 & 4 & 0 & 0 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 5 & 1 & 0 & 0 & 0 & 0 \\ 12 & 0 & 6 & 1 & 0 & 0 \\ 0 & 0 & 5 & 0 & 1 & 0 \\ 0 & 12 & 0 & 0 & 6 & 1 \\ 0 & 0 & 0 & 5 & 0 & 1 \end{pmatrix} \qquad AS_1S_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 6 & 6 & 8 & -2 & 4 & 0 \end{pmatrix}$$

$$S_3 = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 6 & 6 & 8 & 4 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad S = S_1S_2S_3 = \begin{pmatrix} 60 & 12 & 20 & 4 & 0 \\ 0 & 24 & 30 & 18 & 1 \\ 30 & 6 & 10 & 2 & 0 \\ 30 & 30 & 40 & 20 & 1 \\ 90 & 42 & 60 & 24 & 1 \end{pmatrix}$$

All nonnegative vectors x orthogonal to the row space of the matrix A can now be written as a nonnegative combination of the columns of the matrix S. However, when analysing this matrix in detail, one observes a quite embarassing phenomenon: Let s^i be the i-th column of S, then:

$$s^2 = 1/5s^1 + 24s^5 \qquad s^3 = 1/3s^1 + 30s^5 \qquad s^4 = 1/15s^1 + 18s^5$$

Hence, only the vectors s^1 and s^5 represent extremal rays. This means that the algorithm described in theorem 6 indeed generates a sufficient amount of vectors to generate the polyhedral cone, as was proved, but that *redundant vectors may be generated !* How to avoid this redundancy, is the subject of the next section.

2.3.3 Redundancy Reduction

In this section, necessary and sufficient conditions for extremity are derived and proved, such that redundant vectors in the polyhedral solution cone description can be eliminated. Hence, as the algorithm stops, only extremal rays will remain (or only the *trivial* solution). It turns out that the theorems described below are more or less of an algorithmic nature and hence can be directly inserted in an algorithmic inplementation of theorem 6 [12] [13].

Theorem 7 Necessity and Sufficiency for Extremity
A necessary and sufficient condition for a vector x from the polyhedral solution cone generated by the solutions of $Ax = 0$, $x \geq 0$, to be an extremal ray is, that no other solutions possess zeros at the same position as x: Call $\mathcal{I}_x^0 = \{k \mid x_k = 0\}$, the set of indices k where $x_k = 0$. Then x is an extreme ray of the polyhedral cone if and only if there does not exist a solution y with $\mathcal{I}_x^0 \subseteq \mathcal{I}_y^0$.

Proof :
1. Necessity
Assume x is an extremal ray with zero index set $\mathcal{I}_x^0$. Write $Ax = 0$ as :

$$Ax = \sum_{j \notin \mathcal{I}_x^0} a^j x_j = 0$$

where a^j is the j-th column of A. It will now be shown that the set of vectors $\{a^j \mid j \notin \mathcal{I}_x^0\}$ contains one and only one dependent vector. This means that the kernel of the matrix $[a^j]_{j \notin \mathcal{I}_x^0}$ is one dimensional. In other words, the rank of the matrix $[a^j]_{j \notin \mathcal{I}_x^0}$ formed by the columns a^j, $j \notin \mathcal{I}_x^0$ equals:

$$rank[a^j]_{j \notin \mathcal{I}_x^0} = n - \#(\mathcal{I}_x^0) - 1$$

Suppose that this would *not* be true. Then there should exist a vector $d = [\ d_1\ \ \ldots\ \ d_n]^t$, linearly independent of the vector x such that:

$$\sum_{j \notin \mathcal{I}_x^0} a^j d_j = 0$$

and $d_j = 0$ for $j \in \mathcal{I}_x^0$. Observe that d is not necessarily nonnegative. This would imply the existence of 2 vectors $x + \theta\, d$ and $x - \theta\, d$ that satisfy:

$$A\,(\ x \pm \theta\, d\ ) = 0$$

where θ is a real positive number, small enough to have $x_j \pm \theta\, d_j > 0$ for $j \notin \mathcal{I}_x^0$. A suitable θ is easy to find: If $d_j > 0$, then $(x_j + \theta\, d_j) > 0$ and $(x_j - \theta\, d_j) > 0 \Longleftrightarrow \theta < x_j/d_j$. If $d_j < 0$, then $(x_j - \theta\, d_j) > 0$ and $(x_j + \theta\, d_j) > 0 \Longleftrightarrow \theta < -x_j/d_j$. If $d_j = 0$, obviously $x_j > 0$. Hence, we have :

$$0 < \theta < \min\{x_j/\mid d_j \mid \mid\ j \notin \mathcal{I}_x^0\}$$

Hence there would exist two nonnegative solutions: $x^1 = x + \theta\, d$ and $x^2 = x - \theta\, d$. Because the solution set is a polyhedral cone, also $(x^1 + x^2)/2 = x$ would be a solution, contrary to the assumption that x is an extremal ray.

2. Sufficiency :
Assume the vector x is a solution with zero index set $\mathcal{I}_x^0 = \{i \mid x_i = 0\}$ and no other solution has zeros at the same position. Now construct the hyperplane $c^t v = 0$, consisting of vectors v, where :

$$\begin{aligned} c_j &= 1 \quad \text{if } j \in \mathcal{I}_x^0 \\ &= 0 \quad \text{if } j \notin \mathcal{I}_x^0 \end{aligned}$$

is the normal vector of this hyperplane. We now claim that this is the hyperplane from the definition of extremal ray (definition 2), hence that x is an extreme ray of the polyhedral solution cone. Assume that there exists another solution y, linearly independent of x and lying in the same hyperplane. Then obviously, $c^t y = 0$ but this implies that $y_j = 0$ for $j \in \mathcal{I}_x^0$ such that $\mathcal{I}_x^0 \subseteq \mathcal{I}_y^0$, contradicting the assumption that no other solutions than x have zeros at the same position as x. Hence, there does not exist another solution , linearly independent of x and lying in the same hyperplane. Hence, from definition 2, x is an extreme ray. □

An immediate consequence of the theorem is the following corollary:

Corollary 2 *If $n > m$, and the rows of the $m \times n$ matrix A are linearly independent, then a necessary condition for extremity of a solution x to (2.3) is that the number of zeros in x determined via theorem 6, is greater than or equal to $n - m - 1$: $\#(\mathcal{I}_x^0) \geq n - m - 1$.*

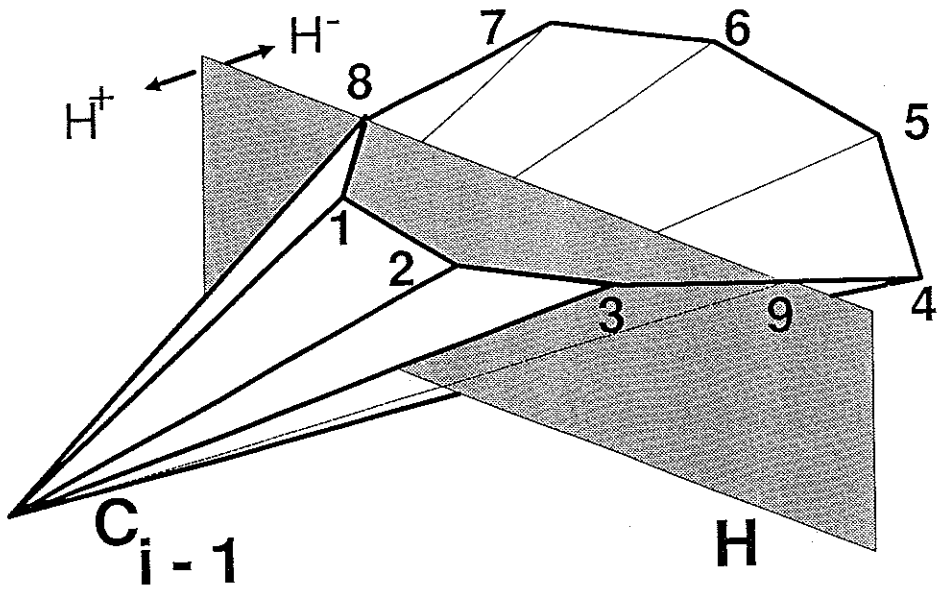


Figure 2.1: Polyhedral Cone associated with first $i-1$ equalities

Proof : From the proof of theorem 5, it follows that

$$rank[a^j]_{j \notin \mathcal{I}_x^0} = n - \#(\mathcal{I}_x^0) - 1$$

As A has m rows, it follows that:

$$n - \#(\mathcal{I}_x^0) - 1 = rank[a^j]_{j \notin \mathcal{I}_x^0} \leq m$$

or equivalently:

$$\#(\mathcal{I}_x^0) \geq n - m - 1$$

□

Another immediate consequence of theorem 7 is the observation that the algorithm, described in theorem 5 only produces *extremal rays* of the polyhedral cone of nonnegative vectors orthogonal to a given vector. Hence no redundancy occurs for this most elementary case. This means that redundancy is caused in the other steps than the first one of the algorithm of theorem 6. This mechanism will now be analysed in detail, via a geometrical visualization. Let $C_{i-1} = S_1 \ldots S_{i-1}$ be the matrix containing (only) extremal rays of the polyhedral cone associated with the first $i-1$ equalities (figure 2.1). Let C_i be the polyhedral cone obtained from C_{i-1} by taking into account a new equation $(a^i)^t x = 0$, $x \geq 0$. This equation defines a hyperplane $\mathcal{H}$ and two half-spaces $\mathcal{H}^+$ and $\mathcal{H}^-$. C_i is obtained from C_{i-1} by the intersection of C_{i-1} with the hyperplane $\mathcal{H}$. The extreme rays of C_i are some of the extreme rays of C_{i-1} (ray 8 in figure 2.1) and rays resulting from some convex combinations of extreme rays of C_{i-1} belonging to $\mathcal{H}^+$ and extreme rays of C_{i-1} belonging to $\mathcal{H}^-$ (ray 9 in figure 2.1). The coefficients of the convex combinations are such that the new ray belongs to the hyperplane $\mathcal{H}$. However, note that extreme rays of C_i are obtained only by combining *adjacent* rays of C_{i-1}. Loosely speaking, these are extreme rays of the polyhedral cone that are each other's neighbour on the 'outside' of the cone. For instance, the rays that are adjacent to ray 5 in figure 2.1, are ray 4 and ray 6. Obviously, if the non-adjacent rays 5 and 2 are nonnegatively

combined, the intersection will not produce an *extremal* ray of the new polyhedral cone C_i. Only, the nonnegative combination of ray 3 and 4 will result in an extremal ray for the new polyhedral cone C_i (ray 9 in figure 2.1.). This notion of adjacency however deserves the following rigorous definition:

Definition 7 Adjacency of extreme rays of a polyhedral cone
Two extreme rays v and w of a polyhedral cone C are adjacent iff there exists a supporting hyperplane $\mathcal{D} = \{x \mid d^t x = 0\ ,\ d^t z \geq 0\ ,\ \forall z \in C\ \}$ such that
$\mathcal{D} \cap C = \{x = v\lambda_1 + w\lambda_2\ ,\ \lambda_1, \lambda_2 \geq 0\}$.

Definition 8 Faces of a polyhedral cone
The set of all convex combinations of two adjacent rays of a polyhedral cone is called a face of that cone.

In the construction of the polyhedral cone C_i from the cone C_{i-1} we have already demonstrated via the little example of figure 2.1 that only *adjacent* rays of C_{i-1} may be combined that belong to different half-spaces associated with the i-th equality. This idea is now formalized:

Theorem 8 Adjacency
A necessary and sufficient condition for two extremal rays x and y of the polyhedral solution cone of $Ax = 0\ ,\ x \geq 0$, to be adjacent, is that there exist no other extremal solutions with zeros at the same positions as the common zeros of x and y. Call $\mathcal{I}^0_{xy} = \{\ k \mid x_k = 0$ and $y_k = 0\}$, the set of indices of the common zeros of x and y. Then x and y are adjacent if and only if there exist no other extreme solutions z with $I^0_{xy} \subseteq I^0_z$.

Proof : The proof goes along the same lines as the proof of theorem 7.
1. Necessity
Rewrite $Ax = 0$ and $Ay = 0$ as:

$$\sum_{j \notin \mathcal{I}^0_{xy}} a^j x_j = 0 \qquad \sum_{j \notin \mathcal{I}^0_{xy}} a^j y_j = 0$$

It will be shown that there does *not* exist another solution vector, linearly independent from x and y (hence not lying in the plane generated by x and y), with zeros at the same positions as the common zeros of x and y. The existence of such a vector would imply that the rays x and y are not adjacent. Hence, it will be demonstrated that:

$$rank[a^j]_{j \notin \mathcal{I}^0_{xy}} = n - \#(\mathcal{I}^0_{xy}) - 2$$

Consider a vector z in the relative interior of the face generated by x and y : $z = x\lambda_1 + y\lambda_2\ ,\ \lambda_1, \lambda_2 > 0$. Obviously, $\sum_{j \notin \mathcal{I}^0_{xy}} a^j z_j = 0$ and $z_j > 0\ ,\ \forall j \notin \mathcal{I}^0_{xy}$. Suppose now that $\text{rank}[z^j]_{j \notin \mathcal{I}^0_{xy}} < n - \#(\mathcal{I}^0_{xy}) - 2$. Then, there exists a vector $d = [d_1\ \ldots\ d_n]^t$, independent of x and y, such that:

$$\sum_{j \notin \mathcal{I}^0_{xy}} a^j d_j = 0$$

and $d_j = 0\ ,\ \forall j \in \mathcal{I}^0_{xy}$. Choosing a sufficiently small θ, one can construct 2 vectors $x^1 = z + \theta\, d$ and $x^2 = z - \theta\, d$, satisfying $\sum_{j \notin \mathcal{I}^0_{xy}} a^j x^1_j = 0$ and $\sum_{j \notin \mathcal{I}^0_{xy}} a^j x^2_j = 0$ and $z_j \pm \theta d_j \geq 0$ (See the proof of theorem 7 for an explicit construction of θ). These two vectors are solutions to the

problem and lie on either side of the plane spanned by x and y. This is in contradiction with the extremity and adjacency of x and y. Hence, $rank[a^j]_{j \notin \mathcal{I}^0_{xy}} = n - \#(\mathcal{I}^0_{xy}) - 2$. Hence all solutions z with $\mathcal{I}^0_{xy} \subseteq \mathcal{I}^0_z$ lie in the face generated by x and y and therefore cannot be extremal.

2. Sufficiency :
Assume the vectors x and y are extreme solutions with common zero index set $\mathcal{I}^0_{xy} = \{i \mid x_i = 0 \text{ and } y_i = 0\}$ and no other extreme solution has zeros at the same position. Now construct the hyperplane $c^t v = 0$, consisting of vectors v, where :

$$\begin{aligned} c_j &= 1 \quad \text{if } j \in \mathcal{I}^0_{xy} \\ &= 0 \quad \text{if } j \notin \mathcal{I}^0_{xy} \end{aligned}$$

is the normal vector of this hyperplane. We now claim that this is the supporting hyperplane from the definition of adjacent rays (definition 7), hence that x and y are adjacent extreme rays of the polyhedral solution cone. Assume that there exist another solution z, linearly independent of x and y and satisfying $c^t z = 0$. This implies that $z_j = 0$ for $j \in \mathcal{I}^0_{xy}$ such that $\mathcal{I}^0_{xy} \subseteq \mathcal{I}^0_z$, contradicting the assumption that no other extreme solutions than x and y have zeros at the same position as the common zeros of x and y. Hence, there does not exist another solution, linearly independent of x and y and lying in the same hyperplane. Hence, from definition 2, x and y are extreme and adjacent. □

As an immediate consequence, it follows that:

Corollary 3 *If the rows of the $m \times n$ matrix A are linearly independent and $n > m + 1$, a necessary condition for extreme solutions x and y to be adjacent, is that the number of common zeros in x and y is greater than or equal to $n - m - 2$: $\#(\mathcal{I}^0_{xy}) \geq n - m - 2$.*

Proof : Is an immediate consequence of the proof of theorem 7. □

Example 3 :
Reconsider example 2 but now employ in each stage the redundancy elemination test provided by theorem 8. The results are contained in the following self-explaining sequence of operations.

$$S_1 = \begin{pmatrix} 3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 \\ 3 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 3 & 0 & 1 & 0 \end{pmatrix} \qquad AS_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 12 & -5 & 6 & -1 & 1 \\ 6 & -2 & 4 & 0 & 0 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 12 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 6 & 1 \\ 0 & 5 & 0 & 1 \end{pmatrix} \qquad S_1S_2 = \begin{pmatrix} 27 & 1 & 0 & 0 \\ 0 & 0 & 9 & 1 \\ 15 & 0 & 1 & 0 \\ 0 & 5 & 0 & 1 \\ 36 & 3 & 6 & 1 \end{pmatrix}$$

$$AS_1S_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 6 & -2 & 4 & 0 \end{pmatrix} \qquad S_3 = \begin{pmatrix} 2 & 0 & 0 \\ 6 & 4 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$S = S_1S_2S_3 = \begin{pmatrix} 60 & 0 \\ 0 & 1 \\ 30 & 0 \\ 30 & 1 \\ 90 & 1 \end{pmatrix}$$

Compare to the solution of example 2: The difference lies in the matrix S_2, where now first the columns were eliminated, that would make convex combinations of columns of S_1 that are not adjacent. As a result, the redundancy has disappeared completely in the final solution. Only the extreme rays of the polyhedral solution cone are retained !

2.3.4 Implementation

The final version of the algorithm to solve (2.3) consists of a straightforward combination of the results of sections 2.3.1, 2.3.2 and 2.3.3. by applying a double *inductive* argument: The main observation is in fact that the nature of the algorithm is essentially *recursive* : for each equality, the same double procedure is applicable :

1. The induction corresponds to finding in each stage the polyhedral subcone of the cone of the previous stage, which is cut off by an additional hyperplane. Some basic rays are the same as before this step, namely those that lie in the hyperplane. Others are obtained by joining 2 rays on each side of the hyperplane and computing the piercing point of the line segment with the hyperplane. This step corresponds to a straightforward implementation of theprem 6.

2. Obviously, it is desirable to obtain a minimal description of the new cone, hence all new rays should be extremal. This is the case if only *adjacent* rays are combined, necessary and sufficient conditions for which have been proved. These conditions apply for each additional equation. This step corresponds to a straightforward implementation of the adjacency tests of theorem 8.

We shall now pay attention to some implementational details and remarks on the computational complexity of the problem.

- First of all, observe that the redundacy test can be implemented using *binary arithmetic* only. All elements in the matrices $C_i = S_1 S_2 \dots S_i$, $i = 1, \dots, m$ only contain nonnegative elements. Represent any positive element by a 1 and keep the zeros. Then one can simply apply the necessary and sufficient conditions of theorem 7 and 8 to these *binary* vectors and matrices.

- Observe that on the average, the magnitude of the numbers in the matrix $C_i = S_1 \dots S_i$ increases as a function of i. The reason is obvious : only sums of products of nonnegative numbers are made. In order to avoid overflow, one should normalise the solutions by appropriate scaling. This is perfectly possible since the extreme rays of any polyhedral cone are fixed up to an arbitrary positive scalar. The following result may provide some guidance in the analysis of the possible occurence of numerical overflow:

  Lemma 4 *Consider the problem of solving x from $Ax = b$ and $x \geq 0$ where A is a $m \times n$ matrix and b a $n \times 1$ vector. Then :*

  $$\mid x_j \mid \leq m! \alpha^{m-1} \beta$$

  where $\alpha = \max_{i,j} \mid a_{ij} \mid$ and $\beta = \max_{j=1,\dots,m} \mid b_j \mid$

  Proof : see [9, lemma 2.1] □

- Adding an additional equation poses no difficulty whatsoever, since one of the most important features of the algorithm is its recursiveness in the equations. The very nature of the algorithm precisely consists in adding one equation at a time and updating the polyhedral solution cone.

An important issue concerns the number of rays that may occur. This may become very large. As a matter of fact, the worst case situation is exponential as a function of the dimensions although, *generically*, no solution is to be expected once $m \geq n$. The worst case situation for the number of vertices is given by [9, p.163]:

$$\begin{pmatrix} m+n \\ m \end{pmatrix} = \frac{(m+n)!}{m!\, n!}$$

Without the redundancy tests of section 2.3.2., the algorithm would be of a *double exponential* complexity of the form $O(\gamma^{\gamma^m})$ for some constant $\gamma > 1$, at least for m *sufficiently* smaller than n. The precise interpretation of *sufficient* is a difficult mathematical (though very intriguing and interesting) problem. It is determined on the one hand by the existence of solutions for small m compared to n and of the generical absence of solutions for $m \geq n$. The double exponentiality arises from analysis of the worst case situation: We may have $n/2$ positive coefficients and $n/2$ negative ones in the first row of A. This implies that the matrix S_1 has $n^2/4$ columns. Now it might be that half of the elements in the second row of AS_1 have positive signs, the other half having negative signs which results in a matrix S_2 that has $n^4/64$ columns. Proceeding in this way, the number $p(m)$ of rays for m equations in the worst case is given by:

$$p(m) = 4(n/4)^{2^m} \tag{2.4}$$

hence the previously used constant γ is the solution to the nonlinear equation $m\ln\gamma + \ln(\ln\gamma) = m\ln 2 + \ln(\ln(n/4))$. However, it is *experimentally* verifiable that the redundancy elimination tests described in section 2.3.3. reduce the complexity to $O(\gamma^m)$ for some constant γ. As a final remark about the complexity, observe that the order in which the equations of the problem (2.3) are processed, may influence the maximal number of extreme rays that occurs during the execution and hence the memory and computational requirements. As an example, consider:

Example 4 :
Solve for $x \geq 0$ in $Ax = 0$ where A is given by:

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 & -1 \end{pmatrix} \tag{2.5}$$

Starting with the first row, immediately leads to the only solution $x = [0\ 0\ 0\ 1\ 0]^t$. However, if first the second row is processed, more operations and memory are required:

$$S_1 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix} \qquad AS_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & 0 \end{pmatrix}$$

leading to $S_2 = [\ 0\ 0\ 0\ 0\ 1\ ]^t$ hence $x = S_1 S_2 = [\ 0\ 0\ 0\ 1\ 0]^t$.
Further research is however needed in order to investigate if there exist optimal strategies (minimizing the number of floating point operations).

2.4 Linear Inequalities

In this section, the problem of finding the nonnegative solutions to a set of linear inequalities will be considered. Two approaches will be discussed. The first one reduces the problem to the nonnegative solution of a set of linear equalities by introducing so-called slack variables. The other one provides a direct algorithm from a slight modification to the algorithm described in section 2.4. The problem is the following: Given A,

$$\text{Solve for } Ax \leq 0 \text{ with } x \geq 0 \tag{2.6}$$

Observe that the non-homogeneous problem $Ax \leq b$, $x \geq 0$ can always be *homogenized* into the form (2.6). From definition 6, we already know that the solution is a pointed polyhedral cone.

2.4.1 Introducing slack variables

Let's first state the following property that will be needed furtheron.

Lemma 5 *Assume that the columns of the matrix $n \times p$ matrix S are the extremal rays of a polyhedral cone. Partition the matrix as :*

$$S = \begin{pmatrix} S^1 \\ S^2 \end{pmatrix} \tag{2.7}$$

where S^1 is a $n_1 \times p$ matrix and S^2 is $(n - n_1) \times p$. Then the columns of S^1 generate a polyhedral cone.

Proof : Trivial from the definition of a polyhedral cone. □

Note that the columns of the matrix S^1 are not necessarily extremal rays of the new polyhedral cone. Indeed, there may be some redundancy between the columns in this sense that some of them are nonnegative combinations of the others.
It is straightforward to convert problem (2.6) to problem (2.3) via the introduction of an m-vector of nonnegative auxiliary variables, $s \geq 0$ that are called *slack* variables in the linear programming literature:

$$Ax + s = 0 \quad x \geq 0 \quad s \geq 0 \tag{2.8}$$

This can be rewritten as :

$$(\ A\ \ I_m)y = 0 \quad y \geq 0 \tag{2.9}$$

where $y^t = [\ x^t \quad s^t]$. This problem is of the form (2.3). One can then solve it for the vector y employing the algorithm of section 2.3. Lemma 5 then allows to simply neglect the slack variables (though in some applications they may contain useful information). However, observe that it is necessary to investigate the remaining rays for redundancy. It is not obvious that this algorithm is the most efficient.

2.4.2 A geometrical approach

It will be shown how 2 slight modifications to the algorithm of section 2.4, allow to compute the extremal rays to the polyhedral solution cone of problem (2.6).

1. Reconsider the geometrical interpretation of figure 2.1. Obviously, each constraint defines again a hyperplane $\mathcal{H}$ and a positive and negative halfspace $\mathcal{H}^+$ and $\mathcal{H}^-$. Instead of retaining only vertices that are lying *in* the hyperplane, all the extremal rays that are lying in the negative half-space $\mathcal{H}^-$ should now be retained as well. If the i-th inequality is being processed, this means that those extreme rays of the polyhedral solution cone $C_{i-1} = S_1 \ldots S_{i-1}$ should be retained that correspond to *non-positive* elements in the vector $(a^i)^t S_1 \ldots S_{i-1}$. The other rays of the polyhedral solution cone C_i are to be formed by convex combinations corresponding to extremal rays of C_{i-1} on either side of the hyperplane as in theorem 8.

2. A second modification concerns the redundancy test. Observe that problem (2.6) is equivalent to : Solve for x in

$$\begin{pmatrix} -I_n \\ A \end{pmatrix} x \leq 0 \tag{2.10}$$

The redundancy theorem 7 and the adjacency theorem 8 should be generalized and applied to the matrix $[\ -I_n \quad A^t]^t$. We shall state these generalizations here without proof, because it is an obvious modification of the proofs of theorem 7 and 8 and because we shall not need explicitly the results in the sequel of this work.

Theorem 9 Necessity and Sufficiency for Extremity
A necessary and sufficient condition for a vector x from the polyhedral solution cone

generated by $Ax \leq 0$, $x \geq 0$ *to be an extreme ray is that no other solution* y *is such that:*

$$\mathcal{I}^0(\begin{pmatrix} -x \\ Ax \end{pmatrix}) \subseteq \mathcal{I}^0(\begin{pmatrix} -y \\ Ay \end{pmatrix}) \tag{2.11}$$

where $\mathcal{I}^0(.)$ *denotes the index set of zero components of a vector.*

Theorem 10 Adjacency
A necessary and sufficient condition for two extreme rays x *and* y *of the polyhedral solution cone of* $Ax \leq 0$, $x \geq 0$ *to be adjacent is that there exist no other solutions* z *that are such that:*

$$(\mathcal{I}^0(\begin{pmatrix} -x \\ Ax \end{pmatrix}) \bigcap \mathcal{I}^0(\begin{pmatrix} -y \\ Ay \end{pmatrix})) \subseteq \mathcal{I}^0(\begin{pmatrix} -z \\ Az \end{pmatrix}) \tag{2.12}$$

Note that theorem 7 and theorem 8 now reduce to special case of theorem 9 and 10. Indeed, for any solution x to $Ax = 0$, $x \geq 0$ the lower part of the partitioned vectors in theorem 9 and 10, is always zero. Hence one should only check the solutions themselves instead of the concatened vectors of theorem 9 and 10 . This constitutes the essential difference between the cases of equalities and inequalities. If S is the polyhedral solution cone to either problem, in the case of equalities, all elements of the matrix product AS are zero, while in the case of inequalities, the matrix produxt AS may contain non-zero elements.

Example 5 :
Consider the same matrix A as in example 3, but with inequalities of the form $Ax \leq 0$, $x \geq 0$. Denote by $(a^i)^t$ the i-th row of the matrix A. Then, the adapted algorithm described above will generate the following sequence of matrices :

$$A = \begin{pmatrix} -3 & -1 & 3 & 0 & 1 \\ 4 & 2 & 0 & 1 & -3 \\ 1 & 1 & 1 & 0 & -1 \end{pmatrix} \qquad S_1 = \begin{pmatrix} 1 & 3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 3 & 1 & 0 \\ 0 & 3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 3 & 0 & 0 & 1 & 0 \end{pmatrix}$$

The matrix S_2 is determined, yet without adjacency elimination tests:

$$(a^2)^t S_1 = [\, 4 \;\; 12 \;\; -5 \;\; 2 \;\; 6 \;\; -1 \;\; 1 \,] \qquad S_2 = \begin{pmatrix} 5 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 12 & 0 & 1 & 2 & 6 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 5 & 0 & 0 & 1 & 0 & 0 \\ 0 & 4 & 0 & 12 & 0 & 0 & 0 & 0 & 2 & 6 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 & 0 & 0 & 1 & 0 \end{pmatrix}$$

The adjacency test is to be applied to the matrix

$$\begin{pmatrix} (a^1)^t S_1 \\ S_1 \end{pmatrix} = \begin{pmatrix} -3 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 3 & 1 & 0 \\ 0 & 3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 3 & 0 & 0 & 1 & 0 \end{pmatrix}$$

As one can verify, only the columns 1, 3, 5, 8, 9, 10, 11, 12 of the matrix S_2 lead to extremal rays of the new solution cone. Omit the other rays and call the new matrix again S_2. The matrix S_3 can now be computed from the vector

$$(a^3)^t S_1 S_2 = [\,-3\ \ 6\ \ -2\ \ -2\ \ 1\ \ 4\ \ 0\ \ 0\,]$$

$$S_3 = \begin{pmatrix} 1 & 6 & 1 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6 & 0 & 1 & 1 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 1 & 1 & 4 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The adjacency is to be checked on the matrix :

$$\begin{pmatrix} (a^1)^t S_1 S_2 \\ (a^2)^t S_1 S_2 \\ S_1 S_2 \end{pmatrix} = \begin{pmatrix} -15 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -5 & 0 & 0 & 0 & 0 & -1 \\ 9 & 27 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 9 & 1 & 1 \\ 0 & 15 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 & 0 & 1 & 0 \\ 12 & 36 & 3 & 3 & 2 & 6 & 1 & 1 \end{pmatrix}$$

Obviously, the columns 4, 8 ,9 , 11, 12 of S_3 make convex combinations of extremal rays that are not adjacent. Hence, they can be omitted. The final polyhedral solution cone is then generated by the following set of extremal rays :

$$S_1 S_2 S_3 = \begin{pmatrix} 9 & 135 & 9 & 60 & 60 & 1 & 1 & 0 & 0 \\ 0 & 0 & 9 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 45 & 0 & 30 & 30 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 30 & 0 & 5 & 1 & 0 \\ 12 & 180 & 18 & 90 & 90 & 3 & 3 & 1 & 1 \end{pmatrix}$$

2.5 Conclusions

In this chapter, it has been shown how all nonnegative solutions of sets of linear equalities and inequalities can be computed. The extremal rays of the polyhedral solution cones are computed. Several necessary and sufficient conditions for extremity and adjacency were derived. All of these results will be applied in chapter 3 (the Generalized Linear Complementarity Problem) and in chapter 6 (The Uncertainty Principle of Mathematical Modelling).

Bibliography

[1] De Moor B., Vandewalle J. *All nonnegative solutions to sets of linear equations and the linear complementarity problem.* Proc. of the International Symposium on Circuits and Systems, p.1072-1079, Philadelphia, USA, May 1987.

[2] De Moor B., Vandenberghe L., Vandewalle J. *The Generalized Linear Complementarity Problem and an Algorithm to find all solutions.* Submitted to Mathematical Programming, March 1988.

[3] Dion J.M., Dugard L., Nguyen M.T. *Multivariable adaptive control with input-output constraints.* Proc. of the 26-th CDC, Los Angeles, December 1987, p.1233.

[4] Greenberg H. *An algorithm for determining redundant inequalities and all solutions to convex polyhedra.* Numer. Math., 24, pp.19-26, 1975.

[5] Motzkin T. *Beitrage zur Theorie der linearen Ungleichungen.* Dissertation, Basel, 1933.

[6] Motzkin T.S., Raiffa H., Thompson G.L., Thrall R.M. *The double description method.* Annals of Math. Stud., 28, pp.51-73, 1953.

[7] Norton J.P. *An introduction to identification.* Academic Press, London, 1986.

[8] Norton J.P. *Identification and Application of Bounded Parameter Models.* Automatica, July 1987.

[9] Papadimitriou C.H., Steiglitz K. *Combinatorial Optimization, Algorithms and Complexity.* Prentice Hall Inc., Englewood Cliffs, New Jersey, 1982.

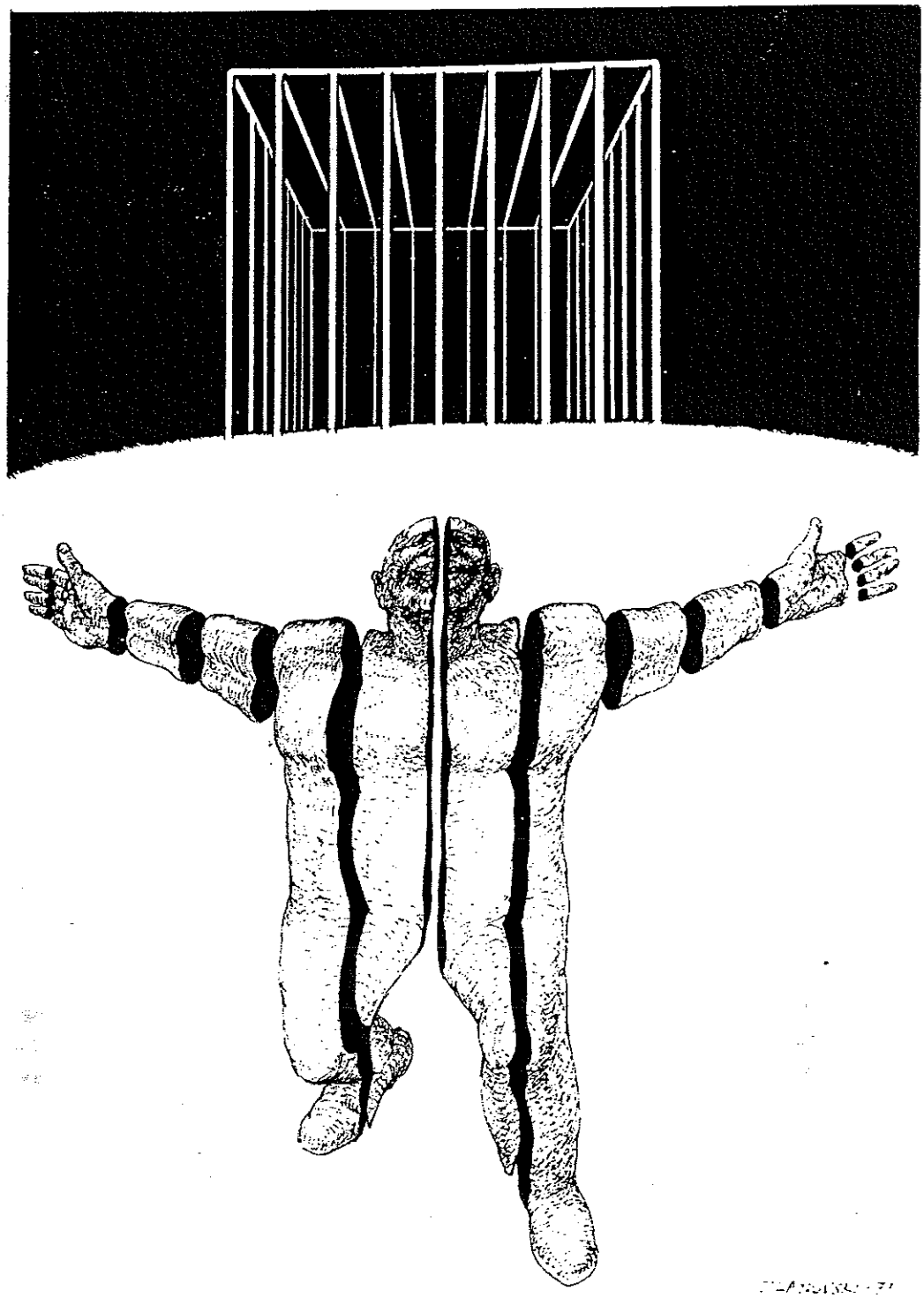
[10] Preparata F.P., Shamos M.I. *Computational Geometry, an Introduction.* Springer Verlag Texts and Monographs in Computer Science, New York, 1985.

[11] Rockafellar R.T. *Convex Analysis.* Princeton University Press, 1970.

[12] Sherman B.F. *A Counterexample to Greenberg's algorithm for solving linear inequalities.* Numer. Math., 27, pp.491-492, 1977.

[13] Walter E., Lahanier H.P. *Exact polyhedral description of the feasible set for bounded error models.* Proc. of the 26-th CDC, Vol.3, p. 1921, Los Angeles, December 1987.

[14] Yemelichev V.A., Kovalev M.M., Kravtsov M.K. *Polytopes, Graphs and Optimization.* Cambridge University Press, Cambridge, England, 1984.



The Complementarity Principle

Chapter 3

The Generalized Linear Complementarity Problem

In the process of world description, to reproduce the integrity of an object,
it is necessary to apply mutually exclusive 'additional' classes of ideas,
each generating its own logically consistent line of judgments,
but still proving logically incompatible with the others.
Bohr's complementarity principle.

In this chapter we provide a farreaching generalization of the 'conventional' Linear Complementarity Problem, which is important from the conceptual point of view. Moreover, an algorithm is developed that allows to determine *all* solutions to the problem, even if there are infinitely many. Third, several clarifying examples and applications are analysed that demonstrate the practical usefulness of the theoretical results.
As will be demonstrated, the Generalized Linear Complementarity Problem (henceforth abbreviated as GLCP) is the central issue in the mathematical treatment of piecewise linear maps, optimization theory and important applications such as electrical circuits and neural nets. *Complementarity* is not quite the same as the *logical law of the excluded middle*: If we consider a statement about the identity of the events $\mathcal{A}$ and $\mathcal{B}$, then the law of the excluded middle reads as follows : $\mathcal{A}$ is either $\mathcal{B}$, or not $\mathcal{B}$. In a complementarity environment, both events can occur but not simultaneously ! This explains the important role that *complementarity* plays in quantum mechanics ! Similarly, in the GLCP, the solutions are required to satisfy certain complementarity constraints that imply that certain solutions or possibilities can occur but not simultaneously. The proposed algorithm will reflect the same principle. In each of its sweeps, certain choices are made : Candidate solutions are retained if they satisfy the conditions while they are rejected if they violate them. This explains at least heuristically, the close connection between the GLCP and important applications such as *optimization theory* and *neural nets* where making decisions is one of the central issues.

This chapter is organised as follows: In section 3.1, the GLCP is introduced and it is shown how the 'conventional' Linear Complementarity Problem reduces to a special case of it . A brief literature survey is presented in section 3.2 (it is brief because there are not too many hard facts in literature that can be used for the GLCP , both because of the more general character of the latter and its recent introduction [12]). In section 3.3 , we describe an *algorithm* to solve the problem. It is based on the insights acquired in chapter 2. Section

3.4 illustrates the usefulness of the GLCP in solving (geometrical) problems with *piecewise linear functions* and the objects defined in chapter 2 such as polyhedral cones, polyhedrons, polytopes etc.... In section 3.5, we show why the generalized linear complementarity problem is of central importance in the *theory of mathematical programming* (linear and quadratic optimization) while section 3.6 explores the application of the GLCP in the *analysis and modelling of non-linear electrical circuits.* The analysis and design of *neural networks* via the GLCP is discussed in section 3.7.

3.1 Linear Complementarity Problems

Let $\mathcal{R}^p_+$ denote the first (i.e. the nonnegative) orthant in the p-dimensional Euclidean space, equipped with the conventional *Euclidean* inner product and the natural basis, which consists of the column vectors of the identity matrix I_p. Let :

$$f(.) : \mathcal{R}^p_+ \rightarrow \mathcal{R}^p : f(v) = Nv - z = w \tag{3.1}$$

denote an affine mapping where N is a given $p \times p$ square matrix and z is a given p-vector.

The *Linear Complementarity Problem* (LCP) is the following:

Given a $p \times p$ matrix N and a $p \times 1$ vector z. Find, or conclude that there are no, vectors v and w, such that $w = Nv - z$ subject to $v \geq 0$, $w \geq 0$ (nonnegativity) and $v^t w = 0$ (orthogonality).

Obviously, the term *complementarity* originates in the combined nonnegativity and orthogonality condition, which causes a complementary zero pattern in the vectors v and w. The problem is really at the heart of numerous other ones: linear and quadratic programming, solving systems of piecewise linear equations, finding the Nash equilibrium of matrix games, free boundary problems of fluid mechanics, structural mechanics, fixed point problems, economic equilibrium theory, modelling of electrical circuits etc ...[1] [10] [19] [20] [24] [25] [29] [30] [34] [36] [41] [43] (and the many references cited in these papers and books of this non-exhaustive list). Briefly, the LCP is considered to be the *Fundamental Problem of Mathematical Programming* [11]. Notwithstanding the fact that the LCP can be considered as a special case of the nonlinear complementarity problem [1, p.297] [24], we shall restrict ourselves in this work to *linear* problems.

However, we shall analyse in this work a more general complementarity problem, which is much more general both in its problem formulation and solution set. Instead of considering $f(x)$ as an *explicit* function of x (an affine mapping as in (3.1)), we shall allow for *implicit* functions of x. Stated in this way, the problem formulation will apparently consist of a number of constraints instead of an explicit functional relationship. This subtle reformulation however is one of the key observations that will turn out to be extremely important: Many more problems and modelling strategies are solvable via this generalized formulation than is the case with the 'conventional' LCP formulation.
To be more specific, the **Generalized Linear Complementarity Problem (GLCP)** is the following:

Given an $m \times n$ matrix Z. Find the $n \times 1$ vector u such that :

$$Zu = 0 \tag{3.2}$$

subject to the *nonnegativity* constraints :

$$u \geq 0 \tag{3.3}$$

and the *complementarity* constraints:

$$\sum \prod u_i = 0 \tag{3.4}$$

which consist of summations of certain products of the nonnegative components u_i of u, depending on the application at hand.

While the literature on the conventional LCP is vast, the GLCP formulation as well as the algorithm to solve it and numerous applications, were recently introduced by us in a series of papers [12] [13] [14] [15] [16] [42]. One of the main achievements of this chapter will be the demonstration of the fact that *in a lot of modelling environments, the GLCP is the most natural representation to obtain piecewise linear models in a structured way.* This will lead to the main results of this chapter, which are:

1. A very simple yet spectacular algorithm which computes *all* solutions to the GLCP, even if there are infinitely many !
2. A survey of properties of the GLCP including results of piecewise linear modelling and numerous examples from applications such as optimization theory, analysis and design of piecewise linear resistive networks, the computation of stationary points of neural nets etc ...

Observe that the epithet *generalized* can be justified from the following observations:

- When the number of columns of Z is odd, a form of the GLCP which is particularly useful is obtained by partitioning Z as:

  $$Z = [\ Z_1 \ \ Z_2 \ \ z^n\ ]$$

  where Z_1 and Z_2 have the same number of columns and z^n is the last column of Z. With a corresponding partitioning of $u^t = [(u^1)^t\ (u^2)^t\ \alpha]$, a frequently occuring GLCP then becomes:
  Find $u^1, u^2,\ \alpha\ \geq 0$ such that:

  $$Z_1 u^1 + Z_2 u^2 = -z^n \alpha \tag{3.5}$$

  subject to the complementarity conditions $(u^1)^t u^2 = 0$.
  For this particular form of the GLCP, observe that:

  1. The conventional LCP is a special case of this GLCP, with $N = Z_1$ being square, $Z_2 = I$ and $\alpha = 1$. However, in the new formulation of the problem, Z_1 and Z_2 are allowed to be singular. More specifically, it may occur that *no* reordering of the columns of Z_1 and Z_2 exists such that the GLCP can be converted into the conventional LCP. Hence, there exist square GLCP's of the form (3.5) that are *not* reconvertable into a conventional LCP.

2. The scalar α may be zero while in the conventional formulation this scalar is always required or assumed to be 1 (or $\neq 0$). The solutions corresponding to $\alpha = 0$ however are important as will be shown. They may correspond to 'solutions at infinity' in geometrical problems.

3. As will be demonstrated, a lot of modelling problems can be translated in a straightforward and highly structured manner into the GLCP, while this is much more difficult for the LCP.

4. The way the problem is generalized reminds of the generalization of regular state space systems into singular state space systems [39]. Indeed, the assumption that one of the matrices in the model should be equal to the identity matrix excludes a priori the potential occurence of important phenomena, that may occur only in the singular case (singular system $\Longleftrightarrow$ GLCP) but are impossible in the non-singular case (regular system $\Longleftrightarrow$ LCP). Moreover, the restriction to the regular case prohibits the use of limiting arguments (singular perturbation analysis ...). While the LCP suffers from these severe a priori limitations, the GLCP does not.

- The adjective *generalized* also refers to the complementarity conditions that are allowed: as will be seen, all kinds of complementarity conditions will be treatable without much complication (section 3.4.1), as long as they are expressible as a summation of products of nonnegative variables. Hence, *complementarity* should be understood in a much more general sense than its most simple occurence as the orthogonality of two nonnegative vectors. However, we insist on the fact that the theoretical development in this chapter applies only to complementarity conditions as expressed by (3.4), i.e. nothing less and nothing more than a *a sum of products of nonnegative variables.* For instance, a constant term, independent of any variable is *not* allowed in the complementarity conditions. Such a term should be dealt with in an additional equation (see example 3).

- A trivial generalization of the GLCP, is obtained when the equality in (3.2) is replaced with inequalities: Given an $m \times n$ matrix Z. Find the $n \times 1$ vector u such that $Zu \leq 0$ subject to *nonnegativity* constraints : $u \geq 0$ and *complementarity* constraints of the form: $\sum \prod u_i = 0$. While at first sight this generalization seems to be an academic *fait divers*, it is very useful in certain design applications (e.g. neural nets).

A purely geometrically inspired algorithm to solve the GLCP was proposed in [12] [13] [14] and will be discussed in this chapter. Its main characteristics, to be contrasted with those of conventional LCP solvers, are:

- It is non-iterative, contrary to most existing LCP-solvers, which are so-called *path followers*, very much like the *simplex algorithm* for linear programming. They trace a piecewise linear path along almost complementary solutions and find only one solution at a time. Therefore a lot of effort has been invested in characterizing the number of solutions of the conventional LCP from computable properties of the matrix N in (3.1) (section 3.2.2). There is no direct motivation to do the same thing for the GLCP, since our GLCP algorithm finds *all* solutions at once. Of course, this does not imply that the theoretical search for such properties for the GLCP would not be interesting in its own right.

- No matrix inversions are required, which is in most LCP solvers the first step in order to convert the GLCP into an LCP. This step can be very ill conditioned or even impossible if the corresponding matrices are singular.

- *All* solutions are computed even if there are infinitely many . In the case of infinitely many solutions, the solution set consists of the union of polyhedral cones, polytopes and discrete vertices. Extreme rays and vertices are explicitely computed by the algorithm. In the case of finitely many solutions, it is straightforward to show that the solution set consists of discrete points, all of which are found. Moreover, it is straightforward to find the nonnegative combinations of solutions that are allowed, i.e. the pairs of so called *cross-complementary* solutions.

- Also the solutions to the 'homogeneous' problem with $\alpha = 0$ are found without the need to solve for a new LCP .

- Updates required by adding additional equations or constraints are most easy to deal with. This follows from the proof of the validity of our algorithm, which is by induction (section 3.3).

3.2 A brief literature survey on the 'conventional' LCP

3.2.1 Algorithms for the conventional LCP

It would lead us too far to treat in detail the several algorithms that have appeared in literature to solve the conventional LCP. Moreover, this would add almost nothing to the insight into the GLCP, which will be solved from a point of view that differs considerably from the common LCP solvers. For a detailed survey of available algorithms for the conventional LCP, the interested reader is referred to the literature of the bibliography at the end of this chapter and the references cited therein. We shall only provide a short enumeration, paying attention mainly to the commom features of these approaches.
Roughly, algorithms for the LCP can be divided in three classes:

1. The first class consists of the so called homotopy algorithms, where the LCP is embedded in a one parameter family of problems [20] [41]. These algorithms generate a set of solutions as function of a parameter along a certain path, in each step extending the already obtained path by a segment, while keeping the solution (almost) complementary. These are the so called complementary pivoting algorithms such as the Katzenelson algorithm, the Cottle algorithm and Lemke's algorithm (for a survey and references, see e.g. [41, chapter 3]).

2. A second class of LCP solvers is founded on a variational approach, which minimizes some constrained cost function, formulated in such a way that its solution coincides with the LCP solution [43]. Hence, one can apply constraint minimization algorithms to solve the LCP, a relation that will briefly be touched upon in section 3.5. However, this approach works properly only if the corresponding object function is convex [41]. A simplex inspired approach is described in [43].

3. The third class are the iterative methods, where the LCP is transformed into a set of nonlinear equations, which are solved by using appropriate fixed point algorithms. As an

example, the modulus algorithm in [41] may be mentioned. It is based on contraction mapping and is appropriate for sparse matrix processing. Its convergence behavior however depends on the conditioning of a certain matrix.

It is however instructive to contrast the features of the above enumerated LCP algorithms with those of the GLCP algorithm that will be described furtheron and were summarized already in the introduction.

1. For almost all algorithms, convergence is assured only for a specific class of matrices (e.g. the P-matrices (see section 2.2.2) for the complementary pivoting algorithms). In some cases, more restrictive requirements are imposed (symmetry, positive definiteness, sparsity, non-singularity, well-conditionedness, etc ...).

2. Most of the algorithms only compute one solution at a time, requiring unavoidable initializations and appropriate starting vectors. Moreover, if the matrix N of (3.1) does not belong to a matrix class for which a priori the number of solutions can be explicitly computed, one is never certain that all solutions have been found.

3.2.2 Counting the number of solutions

A lot of literature is available about the relation between the number of solutions of the (conventional) LCP and a priori verifiable properties of the matrix N of (3.1). The reason is obvious : as most algorithms only find one solution at a time, one can never be sure that the complete solution set is exhausted unless some test or explicit solution count mechanism is available. The problem however, is extremely difficult and very few *general* results are available. Let's mention the most important ones:

- The LCP has a *unique* solution for *every* vector z if and only if all principal minors of N are positive (a matrix with this property is called a P-matrix). This result is due to Murty [33] [34], who also pointed out that when all principal minors of N are non-zero, the solution set consists of a finite number of solutions (see also [25]). The set of P-matrices is large and includes positive definite matrices, totally positive matrices, nonsingular N-matrices etc ...(see [1] for a survey)).

- The classes of matrices that are *strictly monotone* (S-matrices) or *regular* (R-matrices) are defined in [1, p.276]. The following inclusions apply : $P \subset S \subset R$. If N is a regular matrix, then the LCP has *at least* one solution for every vector z [1]. Other classes, such as Q-, Z-, P_0- , (Pre-) Leontieff- matrices, with corresponding properties and inclusion relations, are discussed in [1] [41].

- The following *genericity* result is interesting. It is proved in [36]. *Generically*, the solution set of the conventional LCP is *discrete* (the result is proved for the non-linear complementarity problem for continuously differential maps, of which the affine mapping (3.1) is a special case).

Note that in our algorithm for the explicit solution of the GLCP, in some sense, a priori solution counting becomes redundant since the algorithm will compute in a non-iterative way *all* the solutions of the GLCP, *even if there are infinitely many !* This however does not imply that the theoretical study of the relation between matrix properties and the number of solutions may not provide useful insights into the LCP.

3.3 A new algorithm and its advantages

In this section, the solution set of the GLCP will be discussed both from the theoretical as from the computational point of view. In section 3.3.1, we shall outline the algorithm while algorithmic implementational details are discussed in section 3.3.2. For convenience, let's first restate here the problem:

The Generalized Linear Complementarity Problem
Given an $m \times n$ matrix Z. Find, or conclude that there are no, vectors u such that :

$$Zu = 0 \qquad u \geq 0 \qquad \sum \prod u_i = 0 \tag{3.6}$$

where the sum of products depends on the application.

3.3.1 An algorithm for the solution of the GLCP

The following algorithm allows to compute the general solution set of the GLCP. It is basically the same as the one presented previously for finding all nonnegative solutions to sets of linear (in-)equalities. The only modification is that in each sweep the complementarity conditions are verified for each extremal ray of the polyhedral solution cone. That this is allowed, will be proved with another inductive argument.

The GLCP algorithm

1. The initial solution set equals the *first* orthant, represented by the matrix $S_0 = I_n$.
2. Assume that the matrix S_k represents the solution set for the GLCP consisting of the first k rows of Z, with the required complementarity conditions. Then S_{k+1} is obtained by :
   (a) The application of step 2 of the algorithm of theorem 6 in chapter 2 to solve a homogeneous set of linear equations.
   (b) Eliminating the extremal rays that do not satisfy the complementarity conditions.
3. Having obtained the final solution matrix S_m, determine the sets of column vectors of S_m that are *cross - complementary* i.e. those sets for which any convex or nonnegative combination also generates a vector that satisfies the complementarity conditions.
4. The solution is then generated by the columns of the matrix S_m as described in theorem 1.

Theorem 1 The solution set of the GLCP
The general set of solutions of the GLCP consists of the union of (bounded and unbounded) polyhedra.

Hence, the solution set may contain discrete vectors, polytopes and polyhedral cones and the sums and unions of these vectors. A similar statement for the 'conventional' LCP, can be found in [25].
Before proving the validity of the proposed algorithm, and at the same time, of theorem 1, let's first develop some geometrical insight :

1. Apart from the complementarity conditions, the GLCP reduces to the nonnegative solution of a set of linear equalities. This justifies the use of the algorithm described in section 2.3 and explains already the appearance of polyhedral cones and polytopes in the solution set.

2. The matrices S_k that are computed by the proposed algorithm represent the solutions to the corresponding GLCP problem, which consists of the first k rows of Z. The elimination of the extremal rays that do not satisfy the complementarity conditions is easily justified. Assume that v^k and w^k are 2 extremal rays obtained in sweep k and that the components of v^k , which are all nonnegative, do not satisfy the required complementarity conditions. Then, no nonnegative combination of v^k and w^k will satisfy the required complementarity condition, as is most easily proved. The update of the solution set from S_k to S_{k+1} essentially consists of making nonnegative combinations of the columns of S_k. Hence, none of the combinations that contains v^k will satisfy the complementarity conditions. This justifies the immediate elimination of any extremal ray not satisfying the complementarity conditions.

3. The reason for checking the *cross complementarity* between solution vectors in the final stage, is justified by the following observations: Assume that v and w are two solutions of the GLCP . Then both vectors are nonnegative and satisfy the complementarity conditions:

$$\sum \prod v_i = 0 \qquad \sum \prod w_i = 0$$

Consider now a nonnegative combination $u = \gamma v + \delta w$ with γ , $\delta \geq 0$. Then u is also a solution of the GLCP if its components satisfy the complementarity conditions:

$$\sum \prod u_i = 0$$

The terms in each of the complementarity conditions consist of two kinds of products: Products of components of v or w only and *mixed* products of components of v and w. The former products all satisfy the complementarity conditions because v and w do satisfy them. The latter products however impose additional constraints on the mutual zero pattern of v and w. Only if these additional constraints are satisfied, any nonnegative combination of v and w is allowed, resulting in a polyhedral cone which is part of the solution set. A similar reasoning may be derived by replacing 'nonnegative' combinations by *convex* combinations, resulting in a convex polytope as part of the solution set. The reason to take *convex* combinations could be that a component of the solution vectors is required to have a certain value (e.g. $\alpha = 1$ in (3.5)).

We are now ready to state the proof of the theorem.

Proof : Apart from the complementarity conditions, the GLCP solution set is easily seen to be a polyhedral cone, since it consists of the nonnegative solutions of a set of linear equations. However, the complementarity conditions may forbid nonnegative combinations of certain extremal rays. This may result in the fact that some of the extremal rays of the polyhedral cone become isolated and hence this part of the solution consists of discrete vectors only. Another kind of complementarity condition may be such that certain values are assigned to certain components of any solution vector u. Obviously, this then reduces the corresponding part of the solution set to a polytope. □

Observe that our way of discussing the algorithm and proving the solution theorem, is implicitly based on a threefold inductive approach:

1. The first induction is that on which the nonnegative solution of sets of linear equations (theorem 6, chapter 2) is based.
2. The second induction originates in the extremity and redundancy elimination tests (theorems 7 and 8 , chapter 2).
3. The third one is the induction obtained from the observation that any submatrix of rows of Z, together with the nonnegativity and complementarity conditions, is itself a GLCP of a smaller dimension. Hence, in every sweep one may eliminate the vertices that do not satisfy the complementarity constraints.

The algorithm and the theorem are illustrated in the following examples, that will clarify how the complementarity conditions may influence the geometrical topology of the solution set.

Example 1 :
Solve the GLCP:

$$Zv = \begin{pmatrix} 2 & 1 & -1 & 0 & -1 & -1 \\ 3 & 1 & -2 & -1 & -1 & -1 \\ -2 & -1 & 1 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{pmatrix} = 0$$

subject to the complementarity conditions:

$$\begin{aligned} v_i &\geq 0 \quad i = 1, \ldots, 6 \\ v_1 v_2 v_3 &= 0 \\ v_3 v_4 &= 0 \\ v_4 v_5 &= 0 \end{aligned}$$

The first step of the algorithm delivers the matrix S_1:

$$S_1 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Postmultiplication of the second row of Z with S_1 delivers the vector $[\ -1\ \ 1\ \ 1\ \ -1\ \ 0\ \ 0\ \ -1]$ which then results in the matrix:

$$S_2 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

The redundancy test shows that the columns 3 and 5 of this matrix will take a nonnegative combination of non-adjacent vertices of S_1, hence these combinations are not performed, which then results in the matrix:

$$S_1S_2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 & 0 & 1 \end{pmatrix}$$

Obviously, the third column of S_1S_2 violates the 3rd complementarity condition hence is omitted. Postmultiplying the third row of Z with S_1S_2 results in a zero vector indicating that the third row of Z is redundant. Hence, the solution set is generated by the column vectors of the matrix $S = S_1S_2$:

$$S = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 1 \end{pmatrix}$$

Now, checking the cross-complementarity, one finds that the solution set consists of the *union of three polyhedral cones*, generated by the pairs of extremal rays : 1-2 , 3-5, 4-5. Observe that cross-complementarity is not a transitive relation. The pairs 3-5 and 5-4 are cross-complementary but not the pair 3-4.

Example 2 :
Consider the same GLCP as in example 1, but add the complementarity condition $v_5 = 0$. The solution set now consists of *a discrete solution* (the second column of S) and *a polyhedral cone*, generated by the columns 3-5 of S.

Example 3 :
Consider the same GLCP as in example 1 but add the condition : $v_1 = 1$. It should be emphasized that this represents an additional equation and is *not* to be considered as an additional complementarity condition, since it is not of the form : *a sum of products of variables.* Should this be considered as an additional complementarity condition then one could observe that the vertices 4, 5, 6, 7 of S_1 in example 1, do not satisfy this condition, hence according to the algorithmic recipe, should be omitted. As will become clear from the correct solution, this would result in a loss of certain solutions. The reader may wish to verify that the solution set is generated by the columns 1, 2, 3, 5 of the solution matrix S of example 1. The solution set consists of the *polytope* 1-2, and the *unbounded polyhedron* which is the sum of the third column with any nonnegative multiple of column 5.

3.3.2 Robustness of the solution set

Consider the GLCP as formulated in (3.6). Let $\mathcal{U}$ denote a neighbourhood of a solution u of the GLCP in the n-dimensional Euclidean space of and $\mathcal{Z}$ a neighbourhoud of the matrix Z in the $(m \times n)$- dimensional space of real numbers. Then :

Definition 1 *A solution vector u of the GLCP is called robust if for each neighbourhood $\mathcal{U}$ of u there exists a neighbourhood $\mathcal{V}$ of the matrix Z such that the intersection of $\mathcal{U}$ with the set of all solutions of all GLCP problems with matrix in $\mathcal{V}$ is not empty.*

A (G)LCP is called *robust* if all of its solutions are robust. Robustness is a property of a solution which states whether this solution is stable against slight perturbations of the data. This property is analysed in detail in [25] for the 'classical' LCP. One of the main observations is that the class of robust LCP's is *not open for general n !* Since any LCP can be considered as a special case of a GLCP, the class of robust GLCP's is *not open* as well. In order to provide a proof of this statement, it suffices to consider the following example, which is borrowed from [25].

Example 4 :
Consider the GLCP (which is in fact a LCP if $\alpha = 1$).

$$\begin{pmatrix} 0 & 0 & -1 \\ 0 & -1/\kappa & \epsilon \\ 1 & 1/\kappa & 0 \end{pmatrix} v + \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} w = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} \alpha$$

where $\alpha, \epsilon, \kappa \geq 0$ and v and w are required to be nonnegative and orthogonal (complementary). Depending on the parameters ϵ and κ the solutions are the following:

1. $\epsilon = 0$, $\kappa = +\infty$

v_1	1	1	0
v_2	0	0	1
v_3	1	0	0
w_1	0	0	0
w_2	1	0	0
w_3	0	1	0
α	1	0	0

Solution 2 and 3 are solutions at infinity and they are cross-complementary. It is proved in [25] that the first solution is robust . Hence, if it is required that $\alpha = 1$, which is the case for the classical LCP formulation, then the corresponding GLCP is robust and its solution set consists of one discrete solution.

2. $\epsilon = 0$, $\kappa < +\infty$.

v_1	1	1	0	0
v_2	0	0	κ	κ
v_3	1	0	1	0
w_1	0	0	0	1
w_2	1	0	0	0
w_3	0	1	0	0
α	1	0	1	1

The solutions 1 and 2 are discrete solutions. The solutions 3 and 4 are cross-complementary. If $\alpha = 1$ is required, as would be the case for a classical LCP, they are the vertices of a polytope that belongs to the solution set.

3. $\epsilon \neq 0$, $\kappa < +\infty$.

v_1	1	1	0
v_2	0	0	κ
v_3	1	0	0
w_1	0	0	1
w_2	$1+\epsilon$	0	0
w_3	0	1	0
α	1	0	1

The 3 solutions are discrete solutions.

The fact that the solution $v = [\ 1\ 0\ 1\ ]^t$ appears in the three solution sets, is obviously a necessary condition for the robustness of this solution (the proof of sufficiency requires quite some machinery and is given in [25]). From a comparison between case 2 and 3, one can easily prove that the solution $v = [\ 0\ \kappa\ 1\ ]^t$ is *not* robust. Now, the data of case 2 are a slight perturbation of the data of case 1. As a matter of fact, case 2 reduces to case 1 for $\kappa \longrightarrow \infty$. Since the GLCP of case 1 with the requirement $\alpha = 1$ is known to be robust and that of case 2 is obviously not, we have demonstrated that the class of robust LCP's of size $n = 3$ is not open.

The following example shows how it is not always possible to convert a square GLCP into a classical LCP via matrix inversion. The example will also illustrate the fact that the solution set of a (generalized) linear complementarity problem does *not* change *continuously* as a function of the elements of the matrices.

Example 5 :

$$\begin{pmatrix} -(1+\epsilon) & -1 \\ -1 & -(1+\epsilon) \end{pmatrix} v + \begin{pmatrix} (1+\epsilon) & 1 \\ 1 & (1+\epsilon) \end{pmatrix} w = \begin{pmatrix} -\beta \\ -\beta \end{pmatrix} \alpha$$

with α , β , $\epsilon \geq 0$. Both v and w are required to be nonnegative and have to satisfy the complementarity condition $v^t w = 0$.
The singular values of the two matrices are $2 + \epsilon$ and ϵ, hence for small ϵ they are almost singular. This may cause numerical instability in computing the inverse. Observe that no reordering of the columns exists such that the situation ameliorates. If notwithstanding this bad conditioning and for $\epsilon > 0$, the equation is premultiplied with M^{-1} (where M is the second matrix), then one finds the classical LCP:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} v - \frac{\alpha\beta}{\epsilon+2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = w$$

which obviously is characterized by the identity matrix, which is positive definite, hence is a P-matrix (section 3.2.2). The solution to this LCP is *unique* : $v^t = [\beta\ \beta]$, $w^t = [0\ 0]$, $\alpha = 2+\epsilon$. However, for $\epsilon = 0$, no conversion to an LCP is possible. With our algorithm of section 3.2,

we find four extreme solutions:

v_1	1	β	0	0
v_2	0	0	1	β
w_1	0	0	1	0
w_2	1	0	0	0
α	0	1	0	1

Solutions 1 and 3 are solutions at infinity. The pairs 1-2, 2-4 and 4-3 are cross-complementary. Hence, the solutions set consists of 2 polyhedral cones (1-2, 4-3) and 1 polytope. Observe that the solution set changes abruptly for $\epsilon = 0$.

3.3.3 Implementational aspects

The following remarks provide some nice additional features, some of which are subject to further research:

- A fast implementation of verifying (cross-)complementarity is obviously possible using binary arithmetic only. If the coefficients arising in the complementarity conditions are replaced with 1 if they are non-zero, the multiplications can be replaced by logical **AND** while the additions are to be replaced by logical **OR**. The cross-complementarity of two vertices can be evaluated by employing a simple trick.

  1. Replace the vertices by their binary equivalent (a 1 for a non-zero element, zero otherwise).
  2. Construct a new vector which is the result of an elementwise **OR** of the two binary vectors.
  3. Apply the complementarity test to this new vector.

  As an example, consider the vertices v^1 = [2 0 0 1 0] , v^2 = [0 4 1 0 0] and v^3 = [5 0 0 1 3] that satisfy the complementarity condition $\kappa_1\kappa_4(\kappa_2 + \kappa_3) = 0$. Then their binary equivalents are b^1 = [1 0 0 1 0], b^2 = [0 1 1 0 0] and b^3 = [1 0 0 1 1]. The logical **OR** applied to b^1 and b^2 gives b^{12} = [1 1 1 1 0], which does not satisfy the 'binary' complementarity condition. Hence v^1 and v^2 are not cross-complementary. However, v^1 and v^3 are, since b^{13} = [1 0 0 1 1] satisfies the complementarity condition.

- As the algorithm is such that only nonnegative combinations of nonnegative numbers are computed and accumulated, it is necessary to scale the solution vectors in each stage in order to avoid numerical overflow. This is perfectly possible because the solution vectors represent the extremal rays or vertices up to a nonnegative scalar only. The property of complementarity isn't affected either.

- Further effort could be invested in trying to preprocess certain *easy* equations. In the GLCP approach for piecewise linear descriptions, the complementarity conditions in for instance the λ-parametrization (section 3.4.2) induce a lot of highly structured equations that are part of the GLCP. On the other hand, the GLCP algorithm handles the separate equations one by one. Hence, one could start with those equations that are fixed, i.e. independent of the networks topology. Instead of starting from an initial

matrix, equal to the identity matrix, it pays certainly to investigate whether these structured conditions can not be solved a priori,or better, if the solution cannot be written down immediately.

- Another subject for possible further research is related to all techniques that take into account the sparsity of the matrices (sparse matrix techniques) with possible advantages with respect to speed and memory requirements.

- An important implementational issue is the computational complexity of the proposed algorithm. At present, no theoretical analysis has been performed. There are however some experimental observations available:

  1. As the number of equations m increases with respect to the number of variables n, one observes an increasing zero fill in in the solution matrices. Hence it becomes more and more difficult for the complementarity conditions to be satisfied. Therefore one observes in the initial sweeps typically an exponential growth of the number of solutions as long as $m \ll n$. However, as m goes to n, typically, there will occur a maximum in the number of extremal solutions. After this, the number of allowed solutions decreases rapidly.
  2. The order in which the equations are processed may largely influence the number of intermediate solutions and hence the computational and memory requirements. Several examples will be provided in the sequel.
  3. It would also be instructive to investigate *disaggregate - aggregate* techniques, which by an appropriate partitioning of the matrices of the GLCP, reduce the problem to the solution of smaller problems and then linking together these solutions in order to obtain the solutions to the larger GLCP, taking into account certain relations imposed by the partitioning. Inspiration may be found in comparable partitioning strategies in the literature on linear programming, such as e.g. the Dantzig-Wolfe decomposition for sparse problems [35, p.97].

3.4 Piecewise Linear Descriptions

In this section, it will be demonstrated how to employ the GLCP in modelling systems that are characterized by piecewise linear relations and how to solve complicated geometrical problems. Three different ways to characterize a piecewise linear relation between variables will be discussed, each with their respective (dis-)advantages : the sign decomposition approach, the λ-parametrization and the divide-and-conquer approach. Typically. these parametrizations can be applied in modelling environments in the following way:

- Obtain the piecewise linear model of all constituting components (sub-systems). These models consist of certain equations with complementarity conditions.

- Join these sub-systems together via equations that are imposed by the *topology* of the overall system. This will result in a GLCP.

- The points of interest are then the solutions to this GLCP. Checking the cross complementarity conditions will result in the geometrical description of the solution.

Van Bokhoven provides in [41] a canonical realisation of a piecewise linear relation between 2 variables α and β , which results in an *explicit* relation between α and β expressed by a conventional LCP. However, our results are more general since we shall allow any number of variables in any number of dimensions and the piecewise linear parametrizations result in an *implicit* relation between the variables, which is much more suited when several variables are to be linked together by additional equations.

3.4.1 The sign decomposition

The sign decomposition consists of the decomposition of a vector in 2 nonnegative vectors. It is used in the theory of linear programming [35, p.28]. In this application however, *nonnegativity* is the most important requirement. It will be emphasized however, that the vectors from the sign decomposition are at the same time *complementary*, which is a key observation that results in numerous applications.

Definition 2 Sign Decomposition of a vector
Let x be a vector. Then the sign decomposition is the pair of nonnegative vectors x^+, x^- defined by :

$$x^+ = 1/2\,(\,x + |\,x\,|\,)$$
$$x^- = 1/2\,(\,|\,x\,| - x\,)$$

Lemma 1 *Let x have the sign decomposition x^+, x^-. Then :*

1. $x^+ \geq 0$, $x^- \geq 0$ *(nonnegativity)*
2. $x = x^+ - x^-$
3. $|\,x\,| = x^+ + x^-$
4. $(x^+)^t x^- = 0$ *(complementarity)*

Proof : Trivial from the definition of sign decomposition. □

Another function related to the signs of the components is the signum vector function:

Definition 3 Signum Function of a vector
Let x be a vector. The signum function of x is the pair of nonnegative vectors s^+, s^- defined as:

$$s_i^+ = \begin{cases} 1 & \text{if } x_i \geq 0 \\ 0 & \text{if } x_i < 0 \end{cases} \qquad s_i^- = \begin{cases} 1 & \text{if } x_i < 0 \\ 0 & \text{if } x_i \geq 0 \end{cases}$$

Observe that the convention is adopted that 0 is a positive number. The reason will be clarified in lemma 3. The following complementarity properties are essential:

Lemma 2 *Let x be a vector with sign function x^+, x^- and signum decomposition s^+, s^-. Then:*

$$(x^+)^t x^- = 0 \quad (s^+)^t s^- = 0 \quad (x^+)^t s^- = 0 \quad (x^-)^t s^+ = 0$$

Proof : Trivial. □

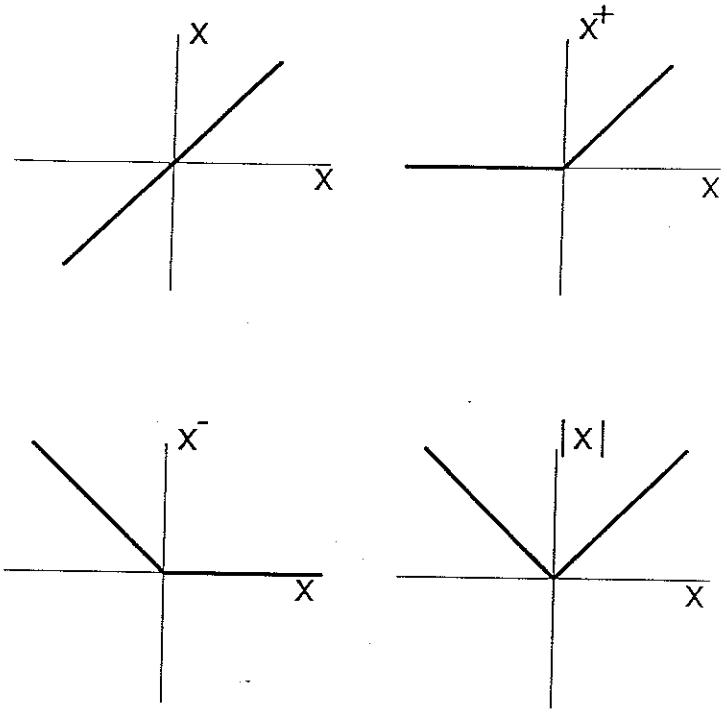


Figure 3.1: The sign decomposition

Lemma 3 *Let x be a vector with signum function s^+, s^-. Let $e = (\ 1\ 1\ \ldots\ 1\ )^t$. Then:*

$$s^+ + s^- = e$$

Proof : Trivial □

It will turn out that these properties (especially the nonnegativity and complementarity of the sign decomposition) are essential in a lot of applications (section 3.5). A nice feature about the sign decomposition is that it allows to treat a vector and its absolute value in the same framework. This is illustrated in the following examples:

Example 6 :
Find the zeros of the following equation:

$$|\ x\ | + sign(x) - 1 = 0 \quad x \in \mathcal{R}$$

where sign(x) is the *signum* function. The problem is visualised in figure 3.2. From lemma 2 and 3, this is equivalent with:

$$x^+ + x^- + s^+ - s^- = 1 \qquad s^+ + s^- = 1$$

with complementarity conditions:

$$x^+x^- = 0 \quad s^+s^- = 0 \quad x^+s^- = 0 \quad x^-s^+ = 0$$

Solving the corresponding GLCP, results in the solutions:

x^+	0	0
s^+	1	0
x^-	0	2
s^-	0	1
α	1	1

Hence, the two zeros are $x_1 = 0$ and $x_2 = -2$.

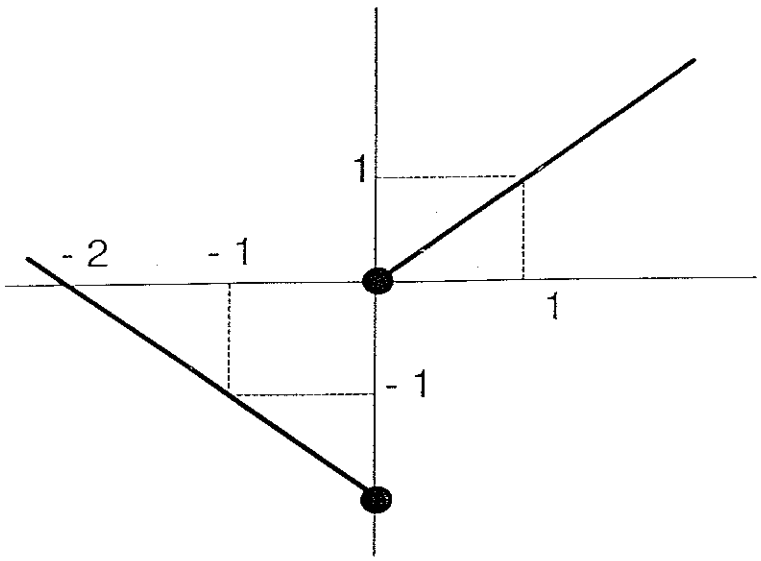


Figure 3.2: Zeros of a piecewise linear function

3.4.2 The λ parametrization

In this section, we derive a parametrization which allows to describe any piecewise linear relation between 2 variables as a GLCP. The main result is stated in the following theorem, which has a constructive proof. Observe that the sign decomposition plays an important role in this parametrization.

Theorem 2 The λ-parametrization
Let the pairs $(\mu_j\ ,\ \nu_j)\ , j = 0,\ldots,k+1$ represent the $k+2$ knots of a connected piecewise linear relation between 2 variables α and β. The first and the last knot are at infinity, representing directions. The piecewise linear equation between α and β can be parametrized as :

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \nu_1 \end{pmatrix} + \begin{pmatrix} \mu_0 \\ \nu_0 \end{pmatrix} \lambda_1^- + \begin{pmatrix} \mu_2 - \mu_1 \\ \nu_2 - \nu_1 \end{pmatrix} \lambda_1^+ \tag{3.7}$$

$$+ \sum_{j=3}^{k} \begin{pmatrix} \mu_j - 2\mu_{j-1} + \mu_{j-2} \\ \nu_j - 2\nu_{j-1} + \nu_{j-2} \end{pmatrix} \lambda_{j-1}^+ \tag{3.8}$$

$$+ \begin{pmatrix} \mu_{k+1} - \mu_k + \mu_{k-1} \\ \nu_{k+1} - \nu_k + \nu_{k-1} \end{pmatrix} \lambda_k^+ \tag{3.9}$$

with $\lambda_j = \lambda_1 - j + 1$ and the k additional constraints :

$$\lambda_j^+ - \lambda_j^- = \lambda_1^+ - \lambda_1^- - j + 1 \qquad j = 2,\ldots,k \tag{3.10}$$

and the complementarity conditions

$$\lambda_j^+ \lambda_j^- = 0, \qquad j = 1,\ldots,k \tag{3.11}$$

Proof: The proof is constructive and can be most easily understood by inspection of figure 3.3. Define a parameter λ, $-\infty < \lambda < +\infty$, which will take on the values $\lambda = (i-1)$ for

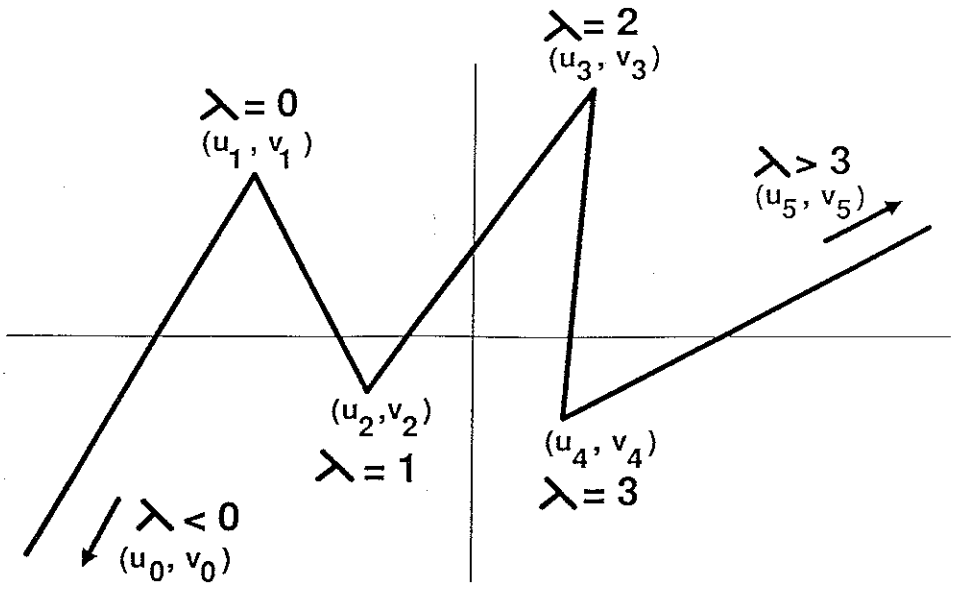


Figure 3.3: Piecewise linear relation between two variables

$(\mu_i \ , \ \nu_i)$, $i = 1, \ldots, k$. The equation of the line through $(\mu_1 \ \ \nu_1)$ with direction $(\mu_0 \ \ \nu_0)$ is given by:

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \nu_1 \end{pmatrix} + \begin{pmatrix} \mu_0 \\ \nu_0 \end{pmatrix} \lambda$$

For $\lambda > 0$, one should make an obvious correction to this equation. Hereto, consider the sign decomposition of $\lambda = \lambda^+ - \lambda^-$. As λ goes from $-\infty$ to $+\infty$, for $\lambda = -\lambda^-$, the equation of the first piece is satisfied. When $\lambda = \lambda^+$, a correction is made in the direction of the second piece. This result is achieved by the following equation:

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \nu_1 \end{pmatrix} + \begin{pmatrix} \mu_0 \\ \nu_0 \end{pmatrix} \lambda^- + \begin{pmatrix} \mu_2 - \mu_1 \\ \nu_2 - \nu_1 \end{pmatrix} \lambda^+$$

This equation gives the first line (3.7) of the theorem. However, when $\lambda > 1$, two correction terms are to be added:

$$- \begin{pmatrix} \mu_2 - \mu_1 \\ \nu_2 - \nu_1 \end{pmatrix} (\lambda - 1)^+$$

$$+ \begin{pmatrix} \mu_3 - \mu_2 \\ \nu_3 - \nu_2 \end{pmatrix} (\lambda - 1)^+$$

The first term subtracts a vector in the 'wrong' direction while the second term makes a correction in the 'right' direction. When proceeding in a similar way, one arrives at the summation term (3.8). The last two terms to be added are the subtraction of a term in the wrong direction and then going to infinity in the 'right' direction. This gives the third line (3.9) of the theorem. The equalities (3.10) are obvious from the definition of the λ_j and the sign decompositon while the complementarity conditions (3.11) follow from lemma 1. □

The following notation will be adopted for the equality constraints (3.10) and complementarity conditions (3.11) :

$$L^+ z^+ + L^- z^- = e \qquad (z^+)^t z^- = 0 \tag{3.12}$$

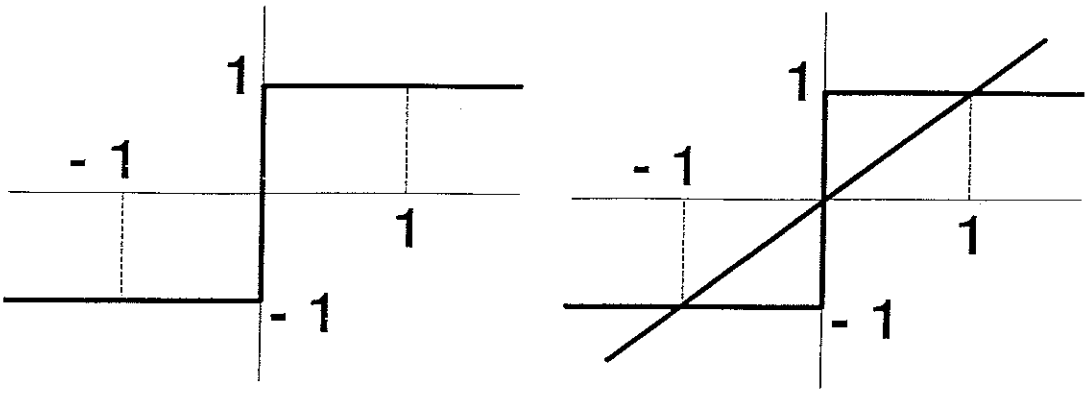


Figure 3.4: Continuous step function and intersection with a line

Here L^+ and L^- are $(k-1)\times k$ matrices and z^+, z^- are $k\times 1$ vectors containing the λ_j^+, resp. λ_j^- parameters for $j = 1,\ldots,k$. The $(k-1)\times 1$ vector e is defined as $e_j = j,\ \ j = 1,\ldots,(k-1)$.

Observe that theorem 2 can easily be generalized to the case where the $k+2$ knots of the piecewise linear relation are vectors in a higher dimensional vectorspace. As a matter of fact, simply replace the coordinate pairs $(\mu_j\ ,\ \nu_j)\ ,\ j = 0,\ldots,k+1$ by vectors of coordinates. The generated object is a piecewise linear manifold.

Example 7 : (continuous) step function
The piecewise linear description of the step function depicted in figure 3.4 is straightforward, employing the parametrization of theorem 2.

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix}0\\ -1\end{pmatrix} + \begin{pmatrix}-1\\ 0\end{pmatrix}\lambda_1^- + \begin{pmatrix}0\\ 2\end{pmatrix}\lambda_1^+ + \begin{pmatrix}1\\ -2\end{pmatrix}\lambda_2^+ \qquad .$$

with conditions:

$$\lambda_i^+, \lambda_i^- \geq 0 \quad i = 1,2 \qquad \lambda_2^+ - \lambda_2^- = \lambda_1^+ - \lambda_1^- - 1$$

Note that there is an important difference with the description based on the *signum* function:

- The λ-parametrization delivers a continuous description at the value $x = 0$.
- The continuity of this parametrization is important in some applications: It is now meaningful for instance to define the *zero of the step function.* As an example, it can be shown that neural nets with binary neurons, when implemented continuously, can never have a stable zero [15]!
- The continuity also allows for a meaningful inversion of the piecewise linear relation without artificial mathematical ambiguities.

Let's now compute the intersection of the step function with the line $\beta = \alpha$. This is equivalent with adding a new equation or premultiplying the equations with the vector $(1\ -\ 1)$. This

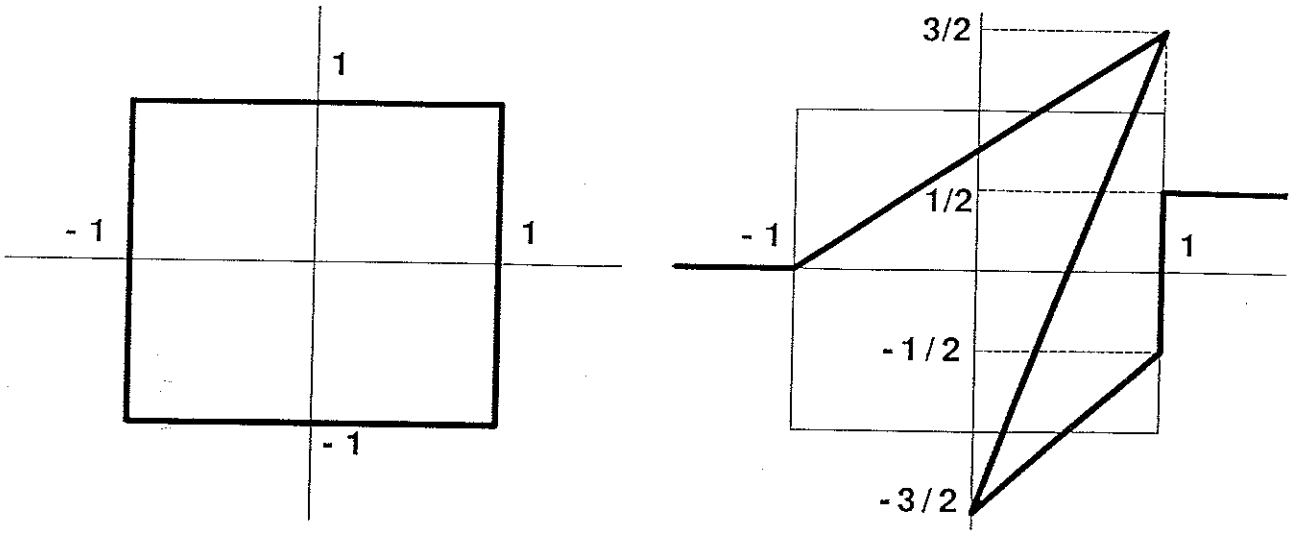


Figure 3.5: (a) Edges of a square and (b) intersection with a piecewise linear snake

then results in the (*homogenized*) GLCP:

$$\begin{pmatrix} -2 & 3 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1^+ \\ \lambda_2^+ \end{pmatrix} + \begin{pmatrix} -1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1^- \\ \lambda_2^- \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \gamma$$

with the complementarity conditions following directly from the sign decomposition (lemma 1) and $\gamma \geq 0$. Its three discrete solutions are:

	1	2	3
λ_1^+	2	1/2	0
λ_2^+	1	0	0
λ_1^-	0	0	1
λ_2^-	0	1/2	2
γ	1	1	1
λ	2	1/2	-1
α	1	0	-1
β	1	0	-1

Example 8 : The Edges of a Square
We shall derive the piecewise linear λ-parametrization of the edges of the square, depicted in figure 3.5a. Employing theorem 2, while starting in the left hand corner, results in the description :

$$\begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} 0 \\ 2 \end{pmatrix} \lambda_1^+ + \begin{pmatrix} -2 \\ -2 \end{pmatrix} \lambda_2^+ \\ \begin{pmatrix} 2 \\ -2 \end{pmatrix} \lambda_3^+ + \begin{pmatrix} 2 \\ 2 \end{pmatrix} \lambda_4^+ + \begin{pmatrix} -2 \\ 0 \end{pmatrix} \lambda_5^+$$

with complementarity conditions:

$$\lambda_2^+ - \lambda_2^- = \lambda_1^+ - \lambda_1^- - 1$$

$$\lambda_3^+ - \lambda_3^- = \lambda_2^+ - \lambda_2^- - 1$$
$$\lambda_4^+ - \lambda_4^- = \lambda_3^+ - \lambda_3^- - 1$$
$$\lambda_5^+ - \lambda_5^- = \lambda_4^+ - \lambda_4^- - 1$$

Note that λ_1^- disappeared from the equation because there is no breaking point defined at $-\infty$. Similarly, there is no breaking point at $+\infty$ so that the last correction term of theorem 2 is also lacking. Let's compute the intersection of these edges with the piecewise linear *snake* of figure 3.5b. Its piecewise linear description is given by:

$$\begin{pmatrix} \alpha_2 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 \\ 0 \end{pmatrix} \kappa_1^- + \begin{pmatrix} 2 \\ 3/2 \end{pmatrix} \kappa_1^+ + \begin{pmatrix} -3 \\ -9/2 \end{pmatrix} \kappa_2^+ \\ \begin{pmatrix} 2 \\ 4 \end{pmatrix} \kappa_3^+ + \begin{pmatrix} -1 \\ 0 \end{pmatrix} \kappa_4^+ + \begin{pmatrix} 1 \\ -1 \end{pmatrix} \kappa_5^+$$

with complementarity conditions similar as those above. The intersection of the two objects, is obtained from the additional constraints:

$$\alpha_1 = \alpha_2 \quad \beta_1 = \beta_2$$

Together with the complementarity conditions, this results in the following GLCP:

$$\begin{pmatrix} 0 & -2 & 2 & 2 & -2 & -2 & 3 & -2 & 1 & -1 \\ 2 & -2 & -2 & 2 & 0 & -3/2 & 9/2 & -4 & 0 & 1 \\ 1 & -1 & . & . & . & . & . & . & . & . \\ . & 1 & -1 & . & . & . & . & . & . & . \\ . & . & 1 & -1 & . & . & . & . & . & . \\ . & . & . & 1 & -1 & . & . & . & . & . \\ . & . & . & . & . & 1 & -1 & . & . & . \\ . & . & . & . & . & . & 1 & -1 & . & . \\ . & . & . & . & . & . & . & 1 & -1 & . \\ . & . & . & . & . & . & . & . & 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1^+ \\ \lambda_2^+ \\ \lambda_3^+ \\ \lambda_4^+ \\ \lambda_5^+ \\ \kappa_1^+ \\ \kappa_2^+ \\ \kappa_3^+ \\ \kappa_4^+ \\ \kappa_5^+ \end{pmatrix} .$$

$$+ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & . & . & . & . & . & . & . & . \\ . & -1 & 1 & . & . & . & . & . & . & . \\ . & . & -1 & 1 & . & . & . & . & . & . \\ . & . & . & -1 & 1 & . & . & . & . & . \\ . & . & . & . & . & -1 & 1 & . & . & . \\ . & . & . & . & . & . & -1 & 1 & . & . \\ . & . & . & . & . & . & . & -1 & 1 & . \\ . & . & . & . & . & . & . & . & -1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1^- \\ \lambda_2^- \\ \lambda_3^- \\ \lambda_4^- \\ \lambda_5^- \\ \kappa_1^- \\ \kappa_2^- \\ \kappa_3^- \\ \kappa_4^- \\ \kappa_5^- \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

The solutions are the following:

	1	2	3	4	5	6	7	8	9
λ_1^+	.	1	4/3	5/2	1/4	13/12	43/12	15/4	3/4
λ_2^+	.	1	1/3	3/2	.	1/12	31/12	11/4	.
λ_3^+	.	1	.	1/2	.	.	19/12	7/4	.
λ_4^+	.	1	.	.	.	.	7/12	3/4	.
λ_5^+	.	1	.	.	.	.	.	.	.
κ_1^+	.	.	2/3	.	3	14/12	22/12	5/2	4
κ_2^+	.	.	.	.	2	2/12	10/12	3/2	3
κ_3^+	.	.	.	.	1	.	.	1/2	2
κ_4^+	.	.	.	.	.	.	.	.	1
κ_5^+	.	.	.	.	.	.	.	.	.
λ_1^-	1	.	.	.	.	.	.	.	.
λ_2^-	1	.	.	.	3/4	.	.	.	1/4
λ_3^-	1	.	2/3	.	7/4	11/12	.	.	5/4
λ_4^-	1	.	5/3	1/2	11/4	23/12	.	.	9/4
λ_5^-	1	.	8/3	3/2	15/4	35/12	5/12	1/4	13/4
κ_1^-	.	.	.	.	.	.	.	.	.
κ_2^-	.	.	1/3	1	.	.	.	.	.
κ_3^-	.	.	4/3	2	.	10/12	2/12	.	.
κ_4^-	.	.	7/3	3	.	22/12	14/12	1/2	.
κ_5^-	.	.	10/3	4	1	34/12	26/12	3/2	.
γ	0	0	1	1	1	1	1	1	1
α			1/3	-1	1	5/6	1/6	1/2	1
β			1	0	-1/2	1	-1	-1	1/2

Hence, there are 9 solution vectors. The first 2 are solutions at infinity. They have no physical interpretation in the 2-dimensional (α, β) space but they are to be interpreted in the 20 - dimensional space of the $(\lambda_i^+, \lambda_i^-, \kappa_i^+, \kappa_i^-)$. The other solutions are finite solutions. One can verify that the cross-complementarity conditions are only satisfied for the pair of solutions 5-9. Hence all convex combinations of this pair of vertices are allowed. Hence, the intersection consists of 5 discrete vectors and 1 convex polytope.

The preceding example also allows to illustrate how the order, in which the equations of the GLCP are processed, may influence considerably the computational and memory requirements. Hereto consider the GLCP as presented (case 1) and another version of the same problem, obtained by reorganizing the matrices. The first two rows of the problem have been processed as the last two equations while first, row 1 - 8 have been processed (case 2). For each equation in both cases, three numbers will be given, characterizing the threefold induction of the algorithm: n_1 is the number of vertices obtained from theorem 6, chapter 4 (solving nonnegatively without redundancy reduction); n_2 is the number of extremal rays when the non-extremal ones have been omitted (theorem 8, chapter 4); n_3 is the resulting number of extremal rays that do not violate the complementarity conditions (section 3.1.3).

Case 1 :

equation	1	2	3	4	5	6	7	8	9	10
n_1	40	178	255	181	211	123	132	124	32	34
n_2	40	58	71	63	58	45	43	42	32	22
n_3	39	58	50	50	43	33	38	33	20	9

Case 2 :

equation	1	2	3	4	5	6	7	8	9	10
n_1	22	22	22	22	22	30	38	46	100	56
n_2	22	21	20	19	22	25	28	71	18	13
n_3	20	19	18	17	20	23	26	29	17	9

Observe that the vertices corresponding to n_1 and n_2 need not to be stored. However, for each of them, a test has to be performed whether they satisfy the necessary extremity and complementarity conditions. Only the vertices corresponding to n_3 are to be stored in memory. This means that for case 1, both the number of tests (computational requirements) and the number of vertices to be memorized, are much larger than in case 2. Hence, this example illustrates clearly how the number of floating point operations and the necessary memory may be influenced by the order of processing the equations. However, further research is needed to find the appropriate optimizing strategy and corresponding rules.

3.4.3 Solving geometrical problems : Divide and Conquer !

This section shows how to describe any geometrical object (in any number of dimensions!), including non-convex objects. The only condition is that it must be possible to find a parametrization with p finite vertices v^i , $i = 1, \ldots, p$ and q vertices w^j , $j = 1, \ldots, q$ at infinity. The systematic procedure to describe these objects is the following:

1. Each vector x that 'belongs' to the object , can be parametrized as :

$$x = \sum_{i=1}^{p} v^i \kappa_i + \sum_{j=1}^{q} w^j \lambda_j$$

   where κ_i , $\lambda_j \geq 0$ and $\sum_{i=1}^{p} \kappa_i = 1$.

2. Divide the object in polyhedral cones (for the vertices at infinity) and in polytopes (for the finite vertices).

3. Subdivide each polyhedral cone in sub-cones with maximally d extremal rays, where d is the dimension of the affine hull of the cone. Subdivide each polytope in d-simplices, where d is the dimension of the affine hull of the polytope.

4. Add complementarity conditions that express that a vector can belong only to one of these sub-objects. There are as many complementarity equations as there are sub-objects in step 3. For each sub-object, define two index sets

$$\mathcal{I}_{kl} = \{k\ ,\ l\ \mid v^k\ ,\ w^l \in \text{subobject}\ \} \quad \bar{\mathcal{I}}_{ij} = \{i\ ,\ j\ \mid v^i\ ,\ w^j \notin \text{subobject}\ \}$$

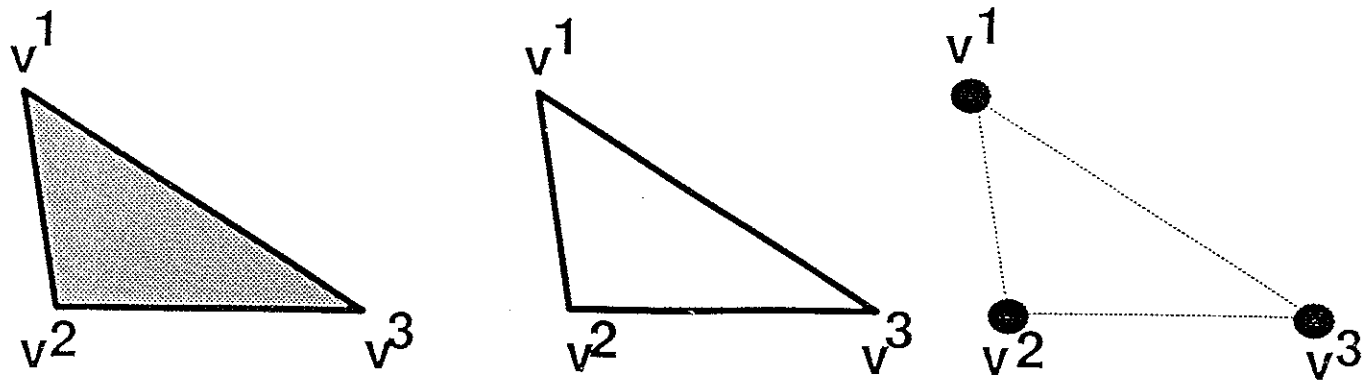


Figure 3.6: Three different objects based upon a triangle

The complementarity condition corresponding to a sub-object with finite vertices v^k and vertices w^l at infinity looks as follows:

$$\prod_{\mathcal{I}_{kl}}(\kappa_k\lambda_l)\sum_{\mathcal{I}_{ij}}(\kappa_i+\lambda_j)=0$$

Observe that the complementarity methodology may cause redundant complementarity conditions, for instance terms that appear twice or more in the conditions. However, since the complementarity may be investigated using Boolean algebra only and applying logical **AND** and **OR**, one could exploit existing logical minimisation techniques in order to find the minimal number of necessary complementarity conditions. Instead of formalizing too much the preceding recipe, we prefer to clarify its precise contents via some (didactical) examples, such that the reader can find out for himself how this parametrization works.

Example 9 : A triangle
Consider 3 vectors v^1 , v^2 , v^3 in 2 dimensions (figure 3.6). The equations

$$v=v^1\kappa_1+v^2\kappa_2+v^3\kappa_3 \quad \kappa_i\geq 0 \quad i=1,2,3 \quad \kappa_1+\kappa_2+\kappa_3=1$$

describe all vectors v that belong to the *full closed* triangle, which is the polytope (in 2 dimensions it is a simplex) that is the convex hull of v^1 , v^2 , v^3. By adding appropriate complementarity conditions, one can find the descriptions of the geometrical objects that consist of the edges only or of the vertices only:

object	complementarity	figure
full triangle	none	3.5a
three edges	$\kappa_1\kappa_2\kappa_3=0$	3.5b
vertices only	$\kappa_1\kappa_2+\kappa_2\kappa_3+\kappa_1\kappa_3=0$	3.5c

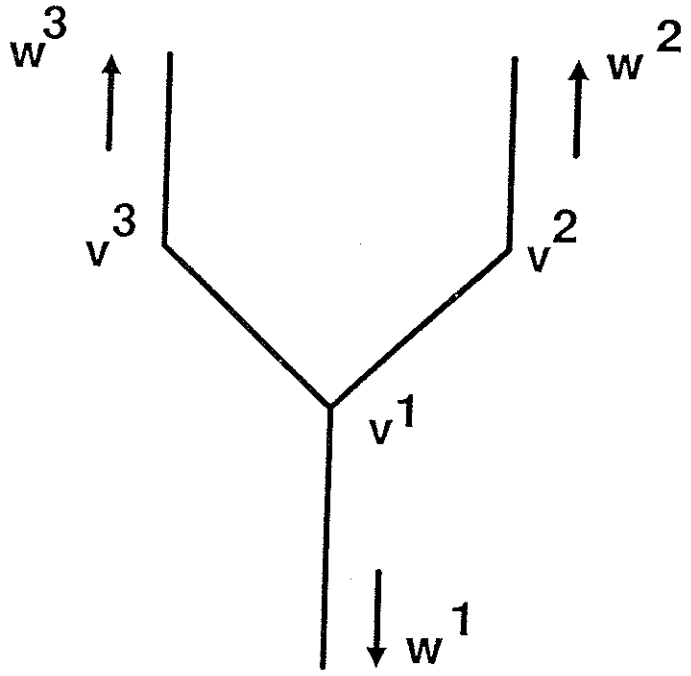


Figure 3.7: An infinite candelabrum

Observe that the polytopes of step 2 of the recipe are the edges themselves. The complementarity condition then expresses the fact that a vector can only belong to one of them at the same time. For the case of the vertices only, the polytopes reduce to one vertex.

Example 10 : An infinite candelabrum
Consider the object depicted in figure 3.7 with three finite vertices v^1 , v^2 , v^3 and three vertices at infinity w^1 , w^2 , w^3. Every vector x on the arms of this object, is expressible as:

$$x = v^1\kappa_1 + v^2\kappa_2 + v^3\kappa_3 + w^1\lambda_1 + w^2\lambda_2 + w^3\lambda_3$$
$$\kappa_1 + \kappa_2 + \kappa_3 = 1 \qquad \kappa_1 , \kappa_2 , \kappa_3 , \lambda_1 , \lambda_2, \lambda_3 \geq 0$$

with 5 complementarity conditions, since the object consists of 2 polytopes and 3 polyhedral cones.

$$\kappa_1\lambda_1(\kappa_2 + \kappa_3 + \lambda_2 + \lambda_3) = 0$$
$$\kappa_1\kappa_3(\kappa_2 + \lambda_1 + \lambda_2 + \lambda_3) = 0$$
$$\kappa_1\kappa_2(\kappa_3 + \lambda_1 + \lambda_2 + \lambda_3) = 0$$
$$\kappa_2\lambda_2(\kappa_1 + \kappa_3 + \lambda_1 + \lambda_3) = 0$$
$$\kappa_3\lambda_3(\kappa_1 + \kappa_2 + \lambda_1 + \lambda_2) = 0$$

The necessity of the sub-division into sub-objects , is highlighted in the following example :

Example 11 : A non-convex object
Consider the object depicted in figure 3.8. It can be described by 8 finite vertices with nonnegative coefficients κ_i , $i = 1, \ldots, 8$. Suppose we would *not* subdivide the square 1-2-3-4 in two triangles, as required by the recipe. Then, its complementarity condition would look like:

$$\kappa_1\kappa_2\kappa_3\kappa_4(\ \kappa_5 + \kappa_6 + \kappa_7 + \kappa_8) = 0$$

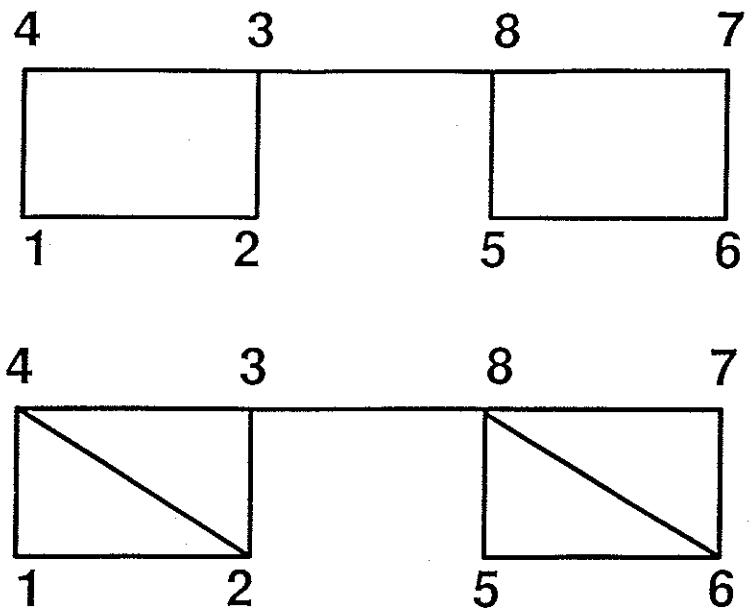


Figure 3.8: A non-convex object

Now consider a vector x in the lower triangular part of the square 1-2-3-4 :

$$x = v^1\kappa_1 + v^2\kappa_2 + v^3\kappa_3$$

Then, since $\kappa_4 = 0$, the complementarity condition is satisfied even if none of the $\kappa_5, \kappa_6, \kappa_7, \kappa_8$ equals zero, which allows for a convex combination of for instance vertex v^1 with vertex v^8. The correct way to describe the fact that a vector x belongs to the square 1-2-3-4 is therefore to specify also to which half (the upper or the lower triangular part) it belongs. Hence the necessary subdivisions and necessary complementarity conditions:

$$\begin{aligned}
\kappa_1\kappa_2\kappa_4(\kappa_3 + \kappa_5 + \kappa_6 + \kappa_7 + \kappa_8) &= 0 \\
\kappa_2\kappa_3\kappa_4(\kappa_1 + \kappa_5 + \kappa_6 + \kappa_7 + \kappa_8) &= 0 \\
\kappa_5\kappa_6\kappa_7(\kappa_1 + \kappa_2 + \kappa_3 + \kappa_4 + \kappa_8) &= 0 \\
\kappa_6\kappa_7\kappa_8(\kappa_1 + \kappa_2 + \kappa_3 + \kappa_4 + \kappa_5) &= 0
\end{aligned}$$

Example 12 : Intersection of two pyramids
The two pyramids, depicted in figure 3.9a and 3.9b can be parametrized as the convex hull of their 4 vertices :

$$\begin{pmatrix} \alpha_1 \\ \beta_1 \\ \gamma_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \kappa_1^1 \\ \kappa_2^1 \\ \kappa_1^3 \\ \kappa_4^1 \end{pmatrix} \quad \sum_i^4 \kappa_i^1 = 1 \quad \kappa_i^1 \geq 0 \;\; i = 1, \ldots, 4$$

$$\begin{pmatrix} \alpha_2 \\ \beta_2 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & -2 \\ 1 & 0 & -1 & 1 \\ 0 & -2 & 1 & 2 \end{pmatrix} \begin{pmatrix} \kappa_1^2 \\ \kappa_2^2 \\ \kappa_3^2 \\ \kappa_4^2 \end{pmatrix} \quad \sum_i^4 \kappa_i^2 = 1 \quad \kappa_i^2 \geq 0 \;\; i = 1, \ldots, 4$$

The intersection can be computed by equating $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2$, $\gamma_1 = \gamma_2$. If they are

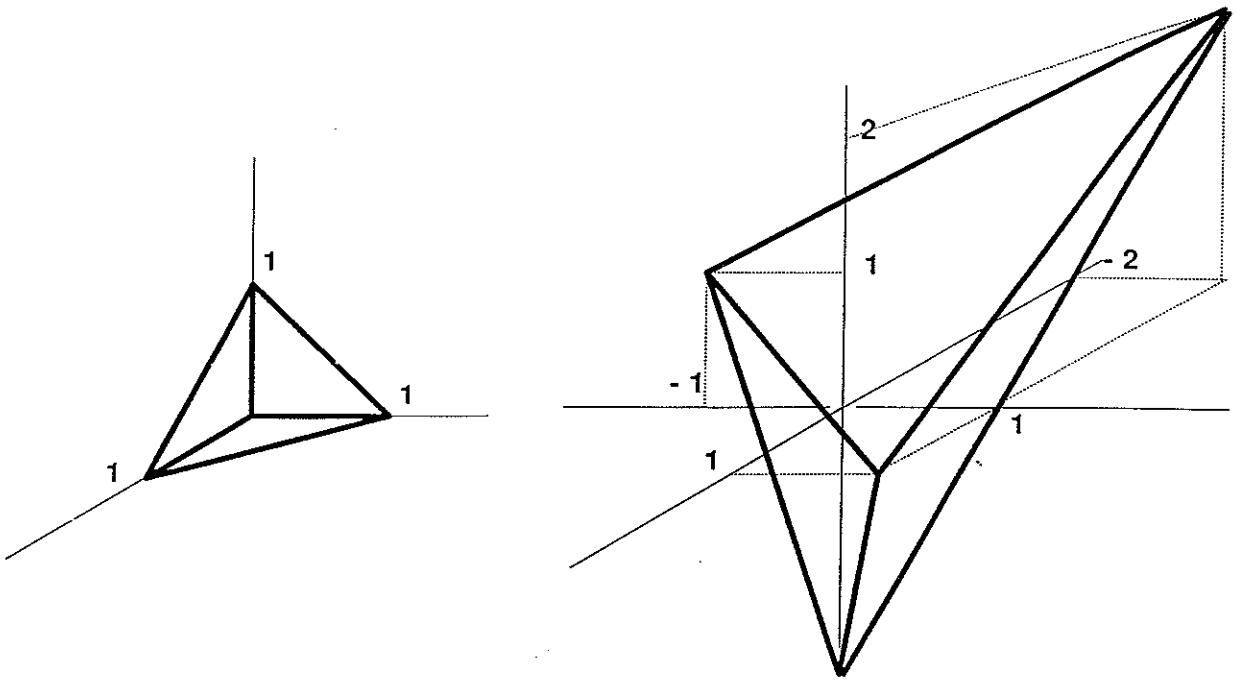


Figure 3.9: Two pyramides

'full' tetrahedrons, there are no complementarity conditions and the problem reduces simply to solve nonnegatively a set of linear equations. All of the solutions are finite:

	1	2	3	4	5	6	7	8	9
κ_1^1	0	1	1/6	0	1/4	0	0	3/5	0
κ_2^1	1/2	0	0	0	0	1/5	0	2/5	5/8
κ_3^1	0	0	0	1/5	3/4	4/5	9/11	0	3/8
κ_4^1	1/2	0	5/6	4/5	0	0	2/11	0	0
κ_1^2	1/2	4/17	2/6	2/5	2/4	3/5	6/11	2/5	5/8
κ_2^2	0	5/17	0	0	1/4	1/5	2/11	1/5	1/8
κ_3^2	1/2	6/17	3/6	2/5	0	0	0	2/5	2/8
κ_4^2	0	2/17	1/6	1/5	1/4	1/5	3/11	0	0

The solutions can obviously be read off immediately from rows 2,3 and 4 of this table. The resulting polytope is depicted in figure 3.10.a. If only the *edges* of the first tetrahedron are considered, then simply add the necessary complementarity conditions and track the solutions of the table that satisfy them together with the cross-complementary pairs, in order to find the appropriate intersection. The necessary complementarity conditions are :

$$\kappa_1^1\kappa_2^1\kappa_3^1 = 0$$
$$\kappa_1^1\kappa_2^1\kappa_4^1 = 0$$
$$\kappa_1^1\kappa_3^1\kappa_4^1 = 0$$
$$\kappa_2^1\kappa_3^1\kappa_4^1 = 0$$

These conditions have already undergone a logical minimisation. They are equivalent with the requirement that a solution must have at least two zeros among the first 4 rows of the table. Obviously all solutions satisfy this requirement. One can verify that the solution pairs 2-3, 2-5, 2-8, 4-7, 6-9 are cross-complementary. Hence the solution set consists of 5 polytopes (line segments) and 1 discrete solution (the first one). The solution is depicted in figure 3.10.b. As

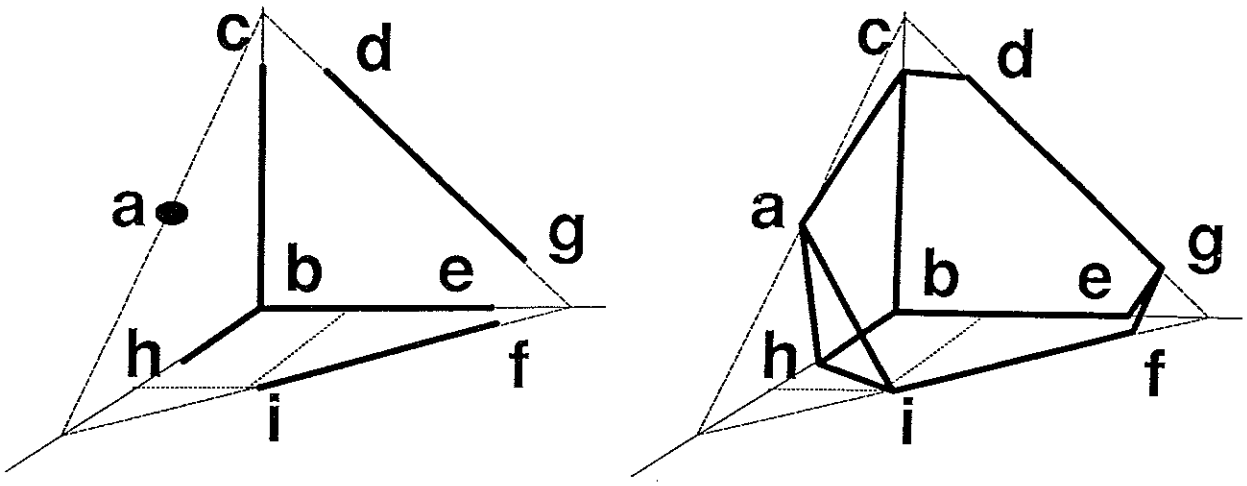


Figure 3.10: Intersection of 2 tetrahedrons with (a) and without (b) complementarity conditions

a minor observation, mind that the zero appearing as the first vertex of the first pyramide, should be treated as all other vectors and by no way deserves a privileged treatment. We mention this fact explicitly since it could be inviting to omit the first row of the solution table, since it contains the parameter κ_1^1, that takes care of the contribution of the first vertex in the convex combinations. Since this vertex equals the zero vector, this has no numerical effect. Yet it is important from the conceptual point of view !

Example 13 : A chair
Denote the 10 vertices of the chair depicted in figure 3.11 by v^i , $i = 1, \ldots, 10$. The polytopes are here the 4 legs which are the line segments 1-5, 2-6, 3-7, 4-8, the square 5-6-7-8 and the square 5-8-9-10. Every vector x of the object, satisfies the following equations:

$$\forall\ x \in \text{ chair } : \quad x = \sum_{i=1}^{10} v^i \kappa_i \quad \kappa_i \geq 0 \quad \sum_{i=1}^{10} \kappa_i = 1$$

with 8 complementarity conditions:

$$\begin{aligned}
\kappa_1\kappa_5(\kappa_2 + \kappa_3 + \kappa_4 + \kappa_6 + \kappa_7 + \kappa_8 + \kappa_9 + \kappa_{10}) &= 0 \\
\kappa_2\kappa_6(\kappa_1 + \kappa_3 + \kappa_4 + \kappa_5 + \kappa_7 + \kappa_8 + \kappa_9 + \kappa_{10}) &= 0 \\
\kappa_3\kappa_7(\kappa_1 + \kappa_2 + \kappa_4 + \kappa_5 + \kappa_6 + \kappa_8 + \kappa_9 + \kappa_{10}) &= 0 \\
\kappa_4\kappa_8(\kappa_1 + \kappa_2 + \kappa_3 + \kappa_5 + \kappa_6 + \kappa_7 + \kappa_9 + \kappa_{10}) &= 0 \\
\kappa_5\kappa_6\kappa_7(\kappa_1 + \kappa_2 + \kappa_3 + \kappa_4 + \kappa_8 + \kappa_9 + \kappa_{10}) &= 0 \\
\kappa_5\kappa_7\kappa_8(\kappa_1 + \kappa_2 + \kappa_3 + \kappa_4 + \kappa_6 + \kappa_9 + \kappa_{10}) &= 0 \\
\kappa_5\kappa_8\kappa_9(\kappa_1 + \kappa_2 + \kappa_3 + \kappa_4 + \kappa_6 + \kappa_7 + \kappa_{10}) &= 0 \\
\kappa_8\kappa_9\kappa_{10}(\kappa_1 + \kappa_2 + \kappa_3 + \kappa_4 + \kappa_5 + \kappa_6 + \kappa_7) &= 0
\end{aligned}$$

The first four conditions concern the legs of the chair. The next 2 arise from the fact that the sitting square is subdivided into 2-dimensional simplices because its affine hull is two-

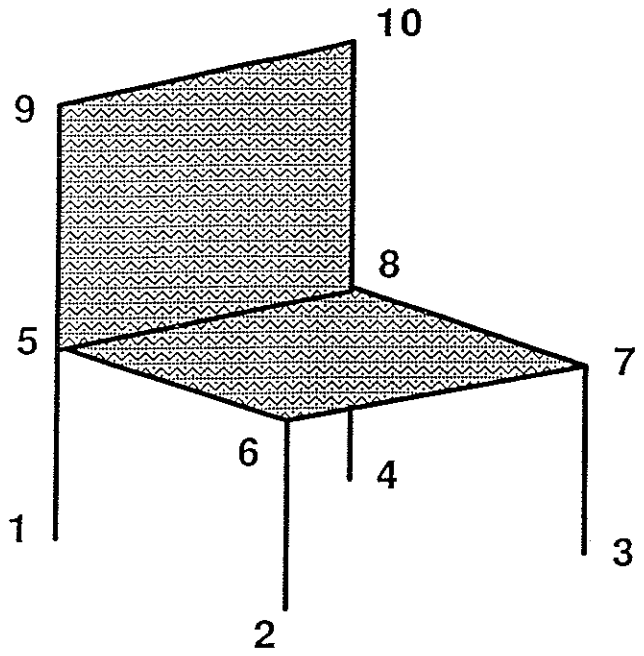


Figure 3.11: A chair

dimensional. The last 2 conditions express similar complementarity conditions.

In order to finish these really tiring examples and at the same time, conclude this section, let me offer you some mathematical recreation. If you want a good laugh, simply add two additional complementarity conditions to your friend's chair :

$$\kappa_5\kappa_7 = 0 \qquad \kappa_6\kappa_8 = 0$$

Of course , be aware that adding $\kappa_4\kappa_8 = 0$ could break his neck!

3.5 Mathematical Programming

The general nonlinear programming problem is the following. Minimize a scalar function

$$\varphi(x_1,\ x_2,\ \ldots,\ x_q) \tag{3.13}$$

subject to the constraints

$$f_i(x_1,\ x_2,\ \ldots,\ x_q) \geq 0 \tag{3.14}$$

where $i = 1,\ldots,p$ and p and q are two independent integers. The well known Lagrange multiplier approach for solving this problem consists of defining a Lagrange function:

$$L(x,\lambda) = \varphi(x) + \sum_{j=1}^{p} \lambda_j f_j(x)$$

where the real constants are called Lagrange multipliers. If the program has a solution x^*,i.e.

$$\min\varphi(x) = \varphi(x^*)$$

with

$$f_j(x^*) \geq 0, \qquad j = 1,2,\ldots,p$$

then the following conditions must hold:

$$\frac{\partial \varphi}{\partial x_i}(x^*) + \sum_{j=1}^{p} \lambda_j^* \frac{\partial f_j}{\partial x_l}(x^*) = 0 \tag{3.15}$$

$$f_j(x^*) \geq 0 \qquad \lambda_j^* \leq 0 \tag{3.16}$$

$$(\lambda_j^*) f_j(x^*) = 0 \tag{3.17}$$

where $l = 1, 2, \ldots, q$ and $j = 1, 2, \ldots, p$. Here, $\varphi(.)$ and $f_j(.)$ are assumed to be differentiable at x^*. Moreover, the constraints (3.14) are assumed to satisfy some regularity conditions , usually referred to as constraint qualifications in the literature [3]. Equations (3.15) to (3.17) are called the *Kuhn-Tucker conditions.* In the general case, the Kuhn-Tucker condition are *necessary* conditions only. However, when $\varphi(.)$ is convex and $f_i(.)$ concave, we have what is called a *convex programming problem.* It has the most convenient property that local optimality implies global optimality [35, p.15].
It will now be demonstrated how the linear and quadratic programming problem can be considered as an LCP via the Kuhn-Tucker conditons. The idea is to convert the optimization problem into an equation with additional conditions, so that it becomes a Generalized LCP. Let's first point out that it is *not* our intention to claim that the equivalence to be presented also leads to an algorithm, that could challenge the existing ones, which are known to be very efficient, especially those for linear programming such as e.g. the *simplex* algorithm. Ever since its discovery in 1948, it has been the most frequently used algorithm. Even this very efficient algorithm is now challenged by the *Karmarkar* algorithm, despite the hesitation of the scientific community to recognize this due to the lack of scientific openness of its inventors (a most regrettable fact of course !). We shall only establish some interesting conceptual links, although of course it may not be excluded a priori that this may lead to new efficient algorithms.

The data (A, D, b, c) of a quadratic programming problem with p constraints and q variables consists of a $p \times q$ matrix A, a $q \times q$ matrix D (which in most of the cases is symmetric nonnegative definite), a p-vector b and a q-vector c :
Find the q-vector x that minimizes the scalar function:

$$\varphi(x) = c^t x + 1/2 x^t D x$$

subject to

$$Ax \geq b \quad x \geq 0$$

Introduce two Lagrange multiplier vectors u_1 $(p \times 1)$ and v_1 $(q \times 1)$ and a $p \times 1$ vector $w_1 = Ax - b$ (the reason for the index *1* will soon be clarified. Applying the Kuhn-Tucker conditions results in :

$$c + Dx + A^t u_1 + v_1 = 0 \quad w_1 = Ax - b \geq 0 \quad u_1, v_1 \leq 0$$

$$(u_1^t v_1^t) \begin{pmatrix} w_1 \\ x \end{pmatrix} = 0$$

These equations are nothing else than a Linear Complementarity Problem:

$$\begin{pmatrix} -v_1 \\ w_1 \end{pmatrix} = \begin{pmatrix} D & -A^t \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ -u_1 \end{pmatrix} + \begin{pmatrix} c \\ -b \end{pmatrix}$$

subject to the nonnegativity and complementarity conditions:

$$-u_1, -v_1, w_1, x \geq 0 \qquad v_1^t x + w_1^t u = 0$$

Remarks on the Quadratic Programming Problem

- If D is symmetric nonnegative definite, then the condition for optimality is necessary *and* sufficient, and *all* solutions of the GLCP will be solutions of the minimisation problem. For more general D, the GLCP solver of section 3.3 will find *all* minima, local and global, including saddle points.

- Some electrical circuit implementations of the Kuhn-Tucker conditions are derived in [3]. We mention explicitly this link because it might be interesting taking into consideration the results of the next section, where it is shown that Chua's canonical piecewise linear models are equivalent with a GLCP.

- An appealing *genericity* result is the following, the proof of which can be found in [46]. The probability $\alpha_D(p,q)$ that the quadratic programming problem with given positive definite matrix D, p constraints and q variables indeed possesses a finite optimal solution, where Gaussian measures are considered on the data is given by:

$$\alpha_D(p,q) = \frac{1}{2^{p+q}} \left[\sum_{i=0}^{n} \binom{p+q}{i} \right]$$

- An example of a linear quadratic programming problem is linear least squares estimation with inequality constraints.

$$\min_x \|Ax - b\|_2$$

subject to $u \leq Cx \leq v$ with A an $m \times n$ and C a $p \times n$ matrix. Of course, from a numerical point of view, one could object against the explicit formation of the Gramian A^tA. However, we believe that in our algorithmic GLCP approach this sin against one of the first commands of numerical linear algebra can be avoided, but this conjecture is subject to further investigation. In any case, numerical reliable algorithms to solve the constrained linear least squares problem are extensively treated in [2], where also important special cases such as nonnegativity constraints LLS and the least distance problem are discussed. It is interesting to note that quadratic programming also arises in the context of adaptive control with constraints on the input signals [18].

Remarks on the Linear Programming Problem

- The Linear Programming Problem is obtained by simply setting $D = 0$ in the above derivation.

- Consider the so called *dual* linear program: Maximize the scalar function $\phi(y) = b^ty$ subject to $A^ty \leq c$, $y \geq 0$. This problem is the same as minimizing $-\phi(y)$ subject to $-A^ty \geq -c$, $y \geq 0$, so we can apply the above translation of the problem into a GLCP. Hereto define $w_2 = -A^ty + c$ and introduce two Lagrange multiplier vectors u_2 $(q \times 1)$ and v_2 $(p \times 1)$. The solution y follows from the GLCP :

$$\begin{pmatrix} -v_2 \\ w_2 \end{pmatrix} = \begin{pmatrix} 0 & A^t \\ -A^t & 0 \end{pmatrix} \begin{pmatrix} y \\ -u_2 \end{pmatrix} + \begin{pmatrix} -b \\ c \end{pmatrix}$$

Identifying the vectors of the *primal* and the *dual* problem with each other results in:

$$\begin{aligned} -v_1 &= w_2 \qquad x = -u_2 \\ -v_2 &= w_1 \qquad y = -u_1 \end{aligned}$$

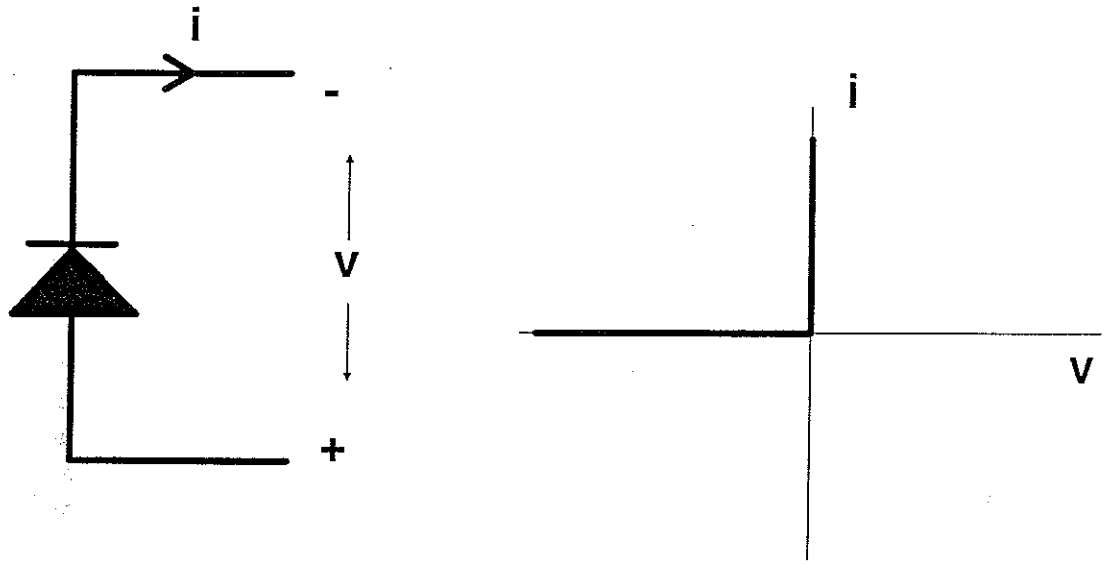


Figure 3.12: Voltage/current characteristic of an ideal diode

Moreover, we have the following :
Proposition : If there is a feasible vector x for the primal Linear Programming Problem and a feasible y for the dual such that $c^t x = y^t b$, then x minimizes $x^t c$ and y maximizes $b^t y$.
Proof : see e.g. [35, p.69].
Hence the Lagrange multipliers of both the primal and the dual linear programmimg problem have a meaningful (and very useful) interpretation: The Lagrange multiplier u_1 represents the optimal solution of the dual problem while the Lagrange multiplier u_2 of the dual problem represents the optimal solution of the primal problem. The Lagrange multipliers v_1 and v_2 represent the excess by which each inequality in the dual, resp. primal problem is satisfied. Classically, they are called the *slack variables* and the associated complementarity conditions are called the *complementary slackness* conditions [35].

3.6 Piecewise Linear Resistive Networks

In order to establish a piecewise linear behavior, an electrical network should contain *piecewise linear elements* besides all other kinds of linear static elements like resistors and independent or controlled linear sources. The ideal diode will be used as the most primitive piecewise linear element while all other piecewise linear components will and can be constructed from a combination of these ideal diodes and standard linear elements. The ideal diode is a two-pole element with a voltage/current characteristic as in figure 3.12. Its voltage v and the current i satisfy:

$$v \geq 0, \quad i \geq 0, \quad v\, i = 0$$

All ideal diodes can be extracted from any piecewise linear network and the remaining network is then a linear lumped memoryless one, containing resistors and all types of fixed or controlled linear sources. Let's assume that the network contains n ideal diodes and denote the port

voltages and currents of the linear n-port by:

$$v^t = (v_1,\ v_2,\ \ldots, v_n\ ) \qquad i^t = (i_1,\ i_2,\ \ldots, i_n\ )$$

Then the network can be described by a so-called constraint matrix description [31] of the form:

$$Mv + Ni = b \tag{3.18}$$

where M, N are both $m \times n$ matrices and $m \leq n$. Typically, the vector b is related to the current and voltage sources in the circuit. Together with the nonnegativity conditions $v, i \geq 0$ and the complementarity condition $v^t i = 0$, equation (3.18) represent a GLCP. Hence, the GLCP arises naturally in the context of network descriptions of electrical circuits. Classically, with square matrices M and N, the way to proceed is to reorder the columns of M and N, and correspondingly the variables contained in v and i until one gets a GLCP with a nonsingular leading matrix M'. By a matrix inversion, one then arrives at the 'conventional' LCP formulation, which is then solved by one of the algorithms of section 3.2.1 [38] [41] [43]. Another approach is to employ Chua's piecewise linear canonical approach and use Katzenelson's algorithm for the solution of sets of piecewise linear equations. [4] [5] [6] [7] [8] [9] [27]. The disadvantages of these approaches, to be contrasted with the possibilities delivered by our GLCP based approach, are :

1. The matrices may be ill-conditioned or even singular (e.g. from a singular perturbation analysis), making inversion difficult or even impossible (see example 5, section 3.3.1).

2. The matrix inversion requires an amount of flops which is proportional to the third power of its dimension. This represents a problem of computational complexity for large nets.

3. The analysis of degenerated nets, characterized by $m < n$ may represent some interest from the theoretical point of view. The classical approach excludes a priori these networks. Also the case of 'autonomous' networks, with $b = 0$ may provide additional insight. However, the classical approaches require a new computation for this case, while our approach delivers this solution at once.

4. As will be demonstrated soon, these remarks are *not* inherent to the problem, but to the attempt to try to *reformulate* it as a conventional LCP instead of solving it as a GLCP. The main reason for this was the lack of a reliable algorithm to solve *any* GLCP, besides of course the appropriate problem formulation as provided by the GLCP and the piecewise linear parametrizations.

5. The most popular algorithms for the solution of an LCP as well as Katzenelson's algorithm, are homotopy methods that generate a path in the solution space, leading from an initial point to *one* solution of the problem. Determining the complete solution set would require trying all possible initial points. In for instance the determination of driving-point and transfer characteristics, unconnected parts of the characteristic cannot be determined in this way, unless a point is given on each part.

However, observe that Chua's canonical model can be converted into a GLCP, as will now be demonstrated: The canonical piecewise linear approach proposed in [4] [5] [6] [7], reduces to finding the zeros of the function :

$$f(x) = a + Bx + C \mid D^t x - e \mid = 0$$

where D^t is an $m \times n$ matrix with $m > n$. Partition the matrix D^t in an upper $n \times n$ matrix D_1^t and a lower $(m-n) \times n$ matrix D_2^t. Under mild conditions, the matrix D_1^t can be assumed to be non-singular. Set $(D^t x - e) = y$ with obvious partitioning of $y^t = (y_1^t \;\; y_2^t)$. Using the sign decomposition, one finds:

$$\left((-BD_1^{-t} - C_1) \quad -C_2 \right) \begin{pmatrix} y_1^- \\ y_2^- \end{pmatrix} + \left((BD_1^{-t} + C_1) \quad C_2 \right) \begin{pmatrix} y_1^+ \\ y_2^+ \end{pmatrix} = -(a + BD_1^{-t} e)$$

Together with the complementarity conditions of the sign decomposition, this is a GLCP.

We shall now propose a structured approach to solve piecewise linear networks. It consists of three steps:

1. First, the *parametrized* $i - v$-characteristic of every piecewise linear component of the network is derived using any of the three parametrizations for piecewise linear models of section 3.4.

2. The network's topology induces additional equations that, together with the piecewise linear descriptions of the components and the corresponding complementarity conditions, generate a rectangular GLCP.

3. The GLCP is solved. Cross-complementarity conditions are checked in order to determine the polyhedral cones that are solutions. If there are no cones in the solution set (no cross-complementarity satisfied) then there are a finite number of solutions.

The most essential features of this recipe are :

- The network description (step 1 and 2) takes on the form $Mv + Ni = b$ with complementarity conditions for i and v. M and N may be singular and rectangular.

- Even if there are an infinite number of solutions, they will all be found at once, including all solutions for $b = 0$.

- The algorithm is non-iterative nor path following, requiring no convergence test or starting strategies whatsoever.

Let's now present some spectacular examples of electrical networks consisting of devices with piecewise linear i-v characteristics.

Example 14: The Butterfly
Consider the following series connection of two piecewise linear tunneldiodes with the given i - v characteristics. The piecewise linear description of the first resistor is:

$$\begin{pmatrix} i_1 \\ v_1 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} + \begin{pmatrix} -3 & 0 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \gamma_1^- \\ \gamma_2^- \end{pmatrix} + \begin{pmatrix} -2.5 & 3.5 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} \gamma_1^+ \\ \gamma_2^+ \end{pmatrix}$$

with complementarity condition :

$$1 = (1 \;\; -1) \begin{pmatrix} \gamma_1^+ \\ \gamma_1^- \end{pmatrix} + (-1 \;\; 1) \begin{pmatrix} \gamma_2^+ \\ \gamma_2^- \end{pmatrix}$$

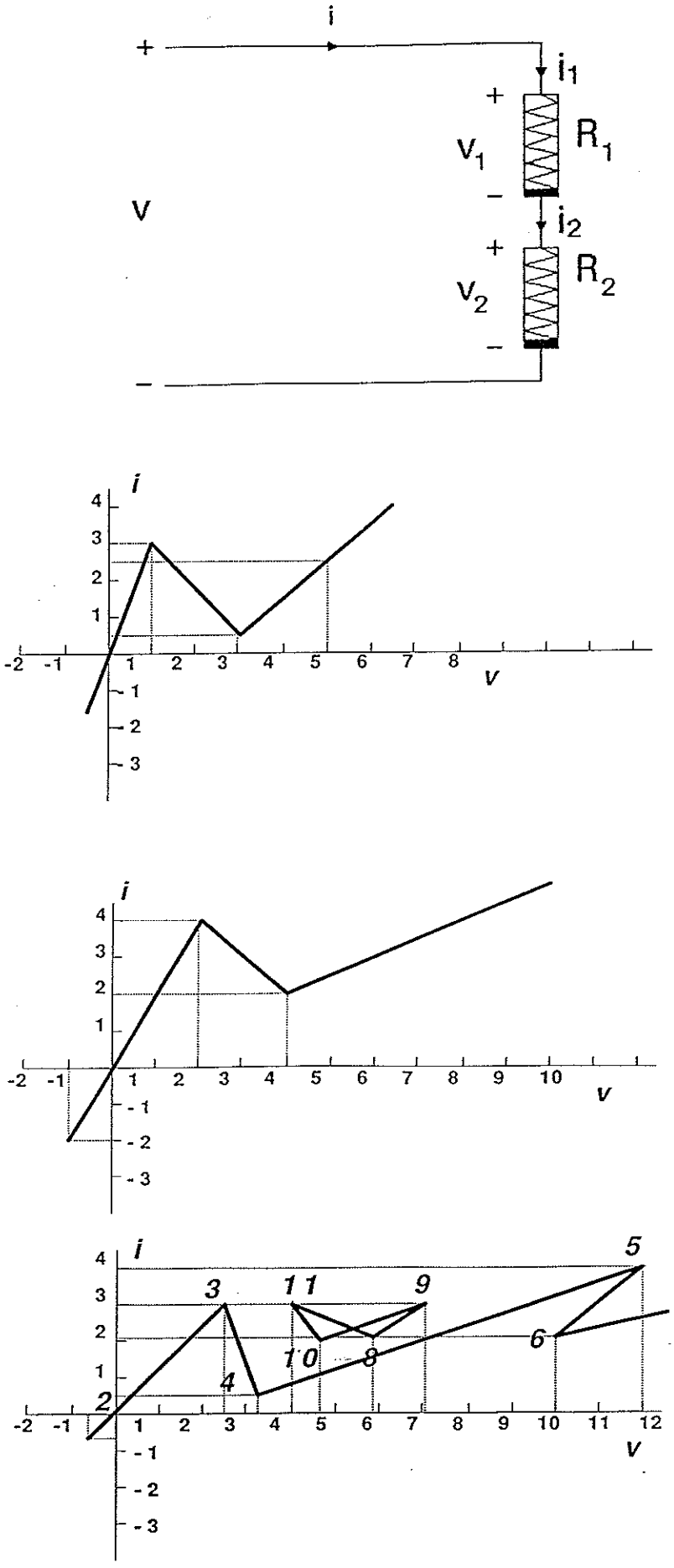


Figure 3.13: Series connection of piecewise linear diodes

In a similar way, the piecewise linear description for the second device is derived :

$$\begin{pmatrix} i_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix} + \begin{pmatrix} -4 & 0 \\ -2 & 0 \end{pmatrix} \begin{pmatrix} \gamma_3^- \\ \gamma_4^- \end{pmatrix} + \begin{pmatrix} -2 & 3 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} \gamma_3^+ \\ \gamma_4^+ \end{pmatrix}$$

with complementarity condition :

$$1 = (1 \ -1) \begin{pmatrix} \gamma_3^- \\ \gamma_3^+ \end{pmatrix} + (-1 \ \ 1) \begin{pmatrix} \gamma_4^- \\ \gamma_4^+ \end{pmatrix}$$

Now, joining the pieces together can be done by adding the constraint matrix description, which finds its origin in the series connection of the diodes:

$$\begin{pmatrix} 1 & 0 & -1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} v \\ i \\ v_1 \\ i_1 \\ v_2 \\ i_2 \end{pmatrix} = 0$$

By employing the sign decomposition, one gets the GLCP :

$$\begin{pmatrix} 0 & -1 & 3 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 2 & 0 \\ 0 & -1 & 0 & 0 & 4 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} i^- \\ v^- \\ \gamma_1^- \\ \gamma_2^- \\ \gamma_3^- \\ \gamma_4^- \end{pmatrix} + \begin{pmatrix} 0 & 1 & 5/2 & -7/2 & 0 & 0 \\ 1 & 0 & -2 & 0 & -2 & 0 \\ 0 & 1 & 0 & 0 & 2 & -3 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} i^+ \\ v^+ \\ \gamma_1^+ \\ \gamma_2^+ \\ \gamma_3^+ \\ \gamma_4^+ \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 4 \\ 1 \\ 1 \end{pmatrix} \alpha \tag{3.19}$$

subject to the obvious complementarity conditions. With our GLCP algorithm, one can find all solutions :

number	1	2	3	4	5	6	7	8	9	10	11
i^-	0.77	0	0	0	0	0	0	0	0	0	0
v^-	0.64	0	0	0	0	0	0	0	0	0	0
γ_1^-	0.26	1	0	0	0	0	0	0	1	0.33	0
γ_2^-	0.19	2	1	0	0	0	0	0.6	0	1.33	1
γ_3^-	0.19	1	0.25	0.88	0	0	0	0	0	0	0
γ_4^-	0	2	1.25	1.88	1	0	0	0	0	0	0.5
i^+	0	0	3	0.5	4	2	0.29	2	3	2	3
v^+	0	0	2.5	3.25	12	10	1.48	5.8	7	4.67	4
γ_1^+	0	0	0	1	4.5	2.5	0.29	0.4	0	0	0
γ_2^+	0	0	0	0	3.5	1.5	0.29	0	0	0	0
γ_3^+	0	0	1/2	0	0	1	0.29	1	2	1	0.5
γ_4^+	0	0	0	0	0	0	0.29	0	1	0	0
α	0	1	1	1	1	1	0	1	1	1	1
i	-0.64	0	3	0.5	4	2	0.29	2	3	2	3
v	-0.77	0	2.5	3.25	12	10	1.48	5.8	7	4.67	4

Verifying cross-complementarity conditions allows to trace the complete driving point plot which consists of 2 disjoint pieces : 1-2-3-4-5-6-7 and 8-9-10-11-8. The solution is depicted in figure 3.13.

Example 15

Consider the piecewise linear network depicted in figure 3.14. The current source i is current dependent. The piecewise linear description of the first resistor is straightforward :

$$\begin{pmatrix} i_1 \\ v_1 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \gamma_1^- \\ \gamma_2^- \end{pmatrix} + \begin{pmatrix} 2 & -2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \gamma_1^+ \\ \gamma_2^+ \end{pmatrix}$$

with the additional condition:

$$\gamma_1^+ - \gamma_1^- - \gamma_2^+ + \gamma_2^- = 1$$

The second resistor has the following piecewise linear description:

$$\begin{pmatrix} i_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix} + \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \gamma_3^- \\ \gamma_4^- \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} \gamma_3^+ \\ \gamma_4^+ \end{pmatrix}$$

with the complementarity condition originating from the sign decomposition:

$$\gamma_3^+ - \gamma_3^- - \gamma_4^+ + \gamma_4^- = 1$$

The network's topology induces the following constraints:

$$v = v_2 \qquad i = i_1 + i_2$$

Combining the piecewise linear description of the two resistors and the topology constraints, one arrives at the following GLCP formulation, which is a again a rectangular one :

$$\begin{pmatrix} 1 & 0 & -2 & 2 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 & -2 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} i^+ \\ v^+ \\ \gamma_1^+ \\ \gamma_2^+ \\ \gamma_3^+ \\ \gamma_4^+ \end{pmatrix} + \begin{pmatrix} -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} i^- \\ v^- \\ \gamma_1^- \\ \gamma_2^- \\ \gamma_3^- \\ \gamma_4^- \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \alpha \tag{3.20}$$

This equation is to be completed with the obvious complementarity conditions . As our GLCP algorithm reveals, the number of solutions is infinite!

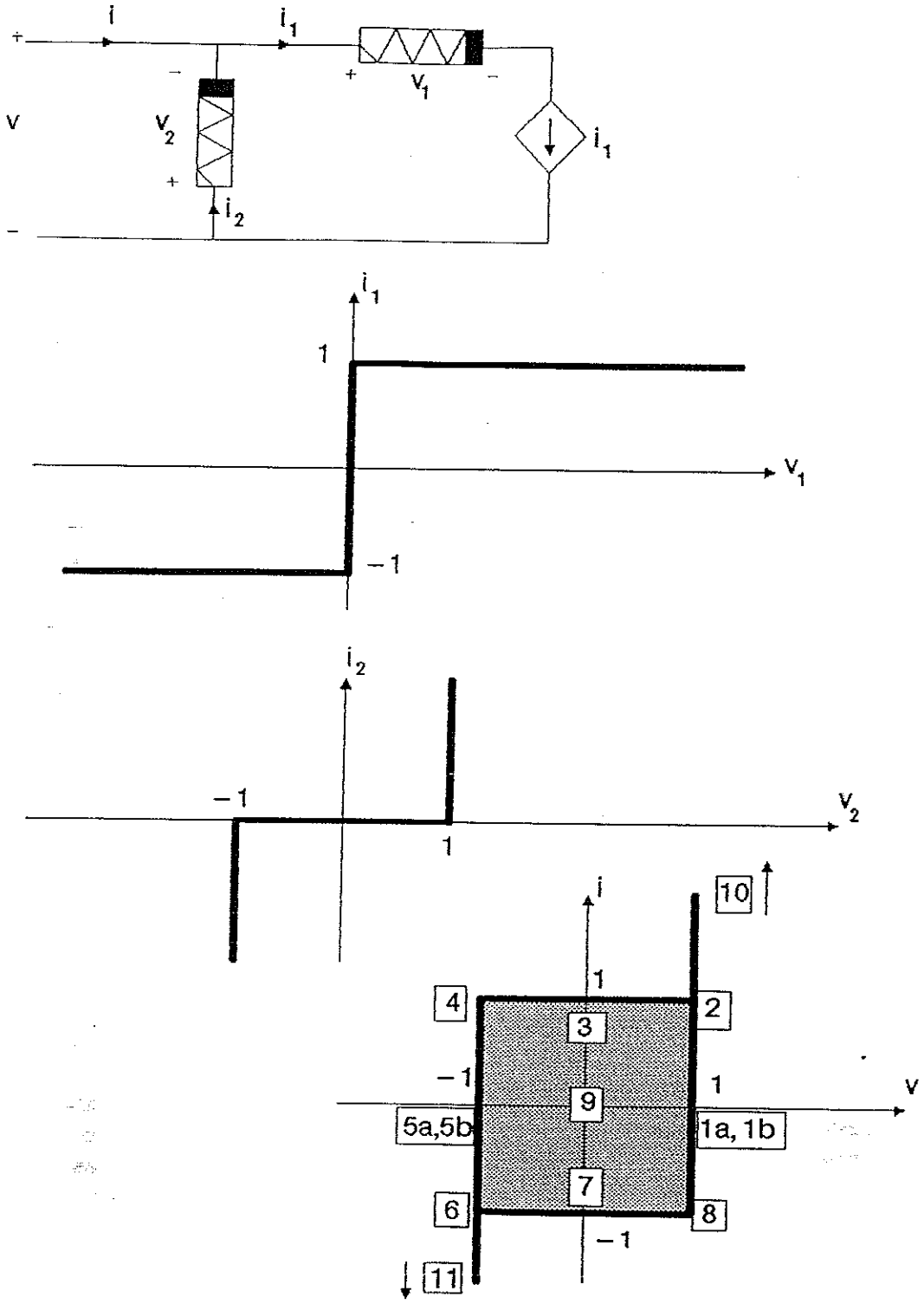


Figure 3.14: A simple *(idealized) piecewise linear circuit*

number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
i^+	0	0	1	0	1	0	0	0	1	1	0	0	0	0	0
v^+	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0
γ_1^+	1	0	0	0	1	0	0	1/2	1	1	1	0	0	1/2	1/2
γ_2^+	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
γ_3^+	0	0	0	1	0	0	0	0	1	1/2	2	1	1/2	1	1/2
γ_4^+	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
i^-	0	0	0	1	0	1	0	0	0	0	0	1	1	0	0
v^-	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0
γ_1^-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
γ_2^-	0	1	0	0	0	1	1	1/2	0	0	0	1	1	1/2	1/2
γ_3^-	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
γ_4^-	0	0	1	0	1	1	2	1	0	1/2	0	0	1/2	0	1/2
α	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
i	0	0	1	-1	1	-1	0	0	1	1	0	-1	-1	0	0
v	0	0	0	0	1	1	1	1	-1	0	-1	-1	0	-1	0

Note that solutions 1, 2, 3, 4 are solutions at infinity, hence these $i - v$ pairs are directions. The cross-complementary pairs are : (1-3), (1-4), (1-5), (1-9), (1-10), (1-11), (2-3), (2-4), (2-5), (2-6), (2-7), (2-8), (2-9), (2-10), (2-11), (2-12), (2-13), (2-14), (2-15), (3-5), (3-7), (3-8), (4-11), (4-12), (4-14), (5-7), (5-15), (6-7), (6-8), (6-12), (6-13), (6-15), (7-8), (8-10), (8-13), (9-10), (9-11), (9-14), (9-15), (10-14), (10-15), (11-12), (11-14), (12-13), (12-14), (12-15) (13-14), (13-15), (14-15). The solution set is depicted in the figure 3.14. The arms towards infinity are caused by the positive combinations of the cross-complementarity pairs : (3-5), (3-7), (3-8) for $+\infty$ and (4-11), (4-12), (4-14) for $-\infty$.

3.7 Neural Networks and the GLCP

A neural network is a massive parallel array of simple computational units (called neurons), linked together by so called synapses, that models some of the functionality of the human brain and attempts to capture some of its computational strenghts. Essential features of a neural net are its massiveness, the non-linearities in the neurons, the training facility ,the iterative way of operation (which can be synchronous or asynchronous) and its convergence properties. The set of invariant states of a network constitutes the memory of the neural net. Training is nothing else than applying certain strategies (mostly heuristic) in order to impress as many as possible 'good' invariant states (i.e. robust against perturbations, noise insensitive,...). In a lot of cases, one is also interested in the *Information Storage Capacity*, the maximal number of invariant states that can be present in a neural net.

The main results of this section are the explicit demonstration of the *equivalence between the invariant states of a neural net and a certain Generalized Linear Complementarity Problem.* Moreover, it will be shown how the main result applies for both continuous as discrete time neural nets, with possibly different types of neurons, as long as the neuronal input - output characteristics are piecewise linear functions. Another result is *the computation of all invariant states that share a prespecified amount of partial information*, which is higly significant for the analysis of the *associative memory* capacities of neural nets.

3.7.1 Mathematical Models of Neural Network

The basic neural net model consists of n neurons which are connected to each other by synapses. The state of the neural net at time k is given by the vector v:

$$v(k)^t = \left(v^1(k)\ \ v^2(k) \ldots v^n(k)\right)$$

where $v^i(k)$ is the state of the i-th neuron. The strength of the 'synaptic' interaction between the neurons is described by a $n \times n$ matrix T of which the element T_{ij} models the strength of the *directed* interaction from the j-th neuron to the i-th one: These elements may be positive (excitatory) and negative (inhibitory) and any neuron may therefore tend to turn any other neuron either "on" or "off" respectively. There is also a direct input s^i to neuron i. This can be used to force certain neurons to take on a certain value. The next state of each neuron is updated according to the total input that flows into it and is decided by a threshold function $f^i(.)$, which can be different for each neuron. We shall confine our attention to the case of piecewise linear describable non-linearities . The fact that each neuron may have a different piecewise linear input - output description allows to concentrate part of the memory in the morphology and parameters of the functions. This is indeed the case in a lot of neural networks, where the synaptic weights together with the parameters of the threshold-decision function constitute the memorized patterns, that were impressed during a learning period. A general model for a *continuous time* neural net, is given by the following equation [17] [28] :

$$dv(t)/dt = -Av(t) + B\mathcal{F}(Tv(t) + s) \tag{3.21}$$

In this equation, A, B and T are $n \times n$ matrices while v and s are $n \times 1$ vectors, containing the states and the constant inputs to each neuron. $\mathcal{F}(.)$ is the n-vector of threshold functions. For *discrete* time neural nets, the following model is used [21] [45] :

$$Av(k+1) = B\mathcal{F}(Tv(k) + s) \tag{3.22}$$

where usually one will find $A = I_n$ in literature. $\mathcal{F}$ denotes an elementwise, decoupled, piecewise linear vector function. However, the matrix A is inserted in the discrete time case, because of a unifying observation, which will now be discussed. As is easily seen, the invariant states v of both continuous and discrete time neural nets are described by the equation:

$$Av = B\mathcal{F}(Tv + s) \tag{3.23}$$

The set of invariant states of a certain network will be called henceforth the *invariant set* of the neural net. Note that the invariant set is independent of the mode of operation of the updating strategy (synchronous or asynchronous). However, it is only a weak requirement for fundamental memories to belong to the invariant set, as described in [21]. Although the mode of operation of the neural net (synchronously or asynchronously updating) has no effect on the invariant set, it may influence the reachability of the invariant states. This is the so called question of *recoverability*. However, we shall restrict our attention strictly to the characterization of *all* invariant states of a neural net with a prescribed structure, without looking at the dynamics and the convergences properties, which are nevertheless important features to be investigated. The invariant set constitutes the *memory* of the neural net and can be formed via training and learning. Adaptation or learning is a major focus of neural net research. Given a set of vectors z_k, $k = 1, \ldots, p$ to be stored in the system. The synaptic matrix T is defined by applying a 'learning' rule which impresses these vectors as the

invariant states of the system. In some neural nets, the parameters of the piecewise linear neuronal decision functions are also adaptable and constitute part of the memory. Training reduces to finding the matrix T and the decision function parameters such that the patterns z_k are invariant states. It is now expected that , if the initial state of the system is not one of its invariant states, it will be dynamically attracted to the invariant state that is most similar to it. There exist a lot of training strategies : For a survey of training for different kinds of neural nets, including training a Hopfield Net , a Hamming net, Carpenter-Grossberg training, Kohonen training, the back propagation algorithm, one may find [45] to contain a nice survey. Several other training strategies and results such as adaptive α and δ updating, training along a scheme closely following Hebb's observations regarding parameter updating in biological networks , spectral training, training via outer-product algorithms and subspace projection training algorithms are described in [17] [21] [26] [44].

The following questions are of central importance:

Given a certain neural net structure i.e. a synaptic weighting matrix T , an input vector s , the matrices A and B and the vector of threshold functions $\mathcal{F}$:

1.
   - What are *all* invariant states of the neural net ? The most obvious way to determine the invariant set of a neural net would be by exhaustive trial with different initial guess vectors. However, no explicit results exist that relate the net structure to its number of invariant points. Hence one never knows whether *all* invariant states have been found or not.
   - Are there any spurious (undesired) invariant states, for which the neural net was not intentionally trained.
   - What is the maximal number of invariant states. This is the problem of the Information Storage Capacity. For instance, it is well known [26] [21] that a Hopfield net with n neurons, can store $O(n)$ arbitrarily chosen patterns while it has a number of spurious stable states, that grows exponentially with increasing n.

2. What are the robustness properties of the invariant points ? One can distinguish between two robustness problems :
   - What about the sensitivity of the invariant states if the original matrix T and other parameters describing the net are perturbed (finite precision, component tolerances, implementation faults, etc ...).
   - What if the applied input pattern (the initial state vector $v(0)$) is perturbed : Will the neural net iteration still converge to the most similar invariant states. This is the so called question of the characterization of the "zones of attraction". What is the geometrical nature of these zones of attraction ?

3.7.2 Assessing the invariant set of a neural net

In this section, it will be shown how *all* invariant states of a neural net with and without partial information can be computed from the solution to a GLCP.

Theorem 3 The Invariant Set of a neural net

- *Assume that the neural net contains n neurons. The synaptic weights are contained in the $n \times n$ matrix T. The direct inputs are contained in the n-vector s and the invariant states v satisfy the equation:*

$$Av = B\mathcal{F}(Tv + s) \tag{3.24}$$

where A, B are $n \times n$ matrices and $\mathcal{F}$ symbolizes the neuronal decision funcions.

- *Each neuronal decision function f^i is described as a piecewise linear relation between the variables α^i and β^i. For neuron i there are $k^i + 2$ knots $(\mu_j^i \ \ \nu_j^i)$, $j = 0, \ldots, k^i + 1$. The equation then reads:*

$$\begin{aligned}\begin{pmatrix} \alpha^i \\ \beta^i \end{pmatrix} &= \begin{pmatrix} \mu_1^i \\ \nu_1^i \end{pmatrix} + \begin{pmatrix} \mu_0^i \\ \nu_0^i \end{pmatrix} \lambda_1^{i-} + \begin{pmatrix} \mu_2^i - \nu_1^i \\ \beta_2^i - \beta_1^i \end{pmatrix} \lambda_1^{i+} \\ &+ \sum_{j=3}^{k^i} \begin{pmatrix} \mu_j^i - 2\mu_{j-1}^i + \mu_{j-2}^i \\ \nu_j^i - 2\nu_{j-1}^i + \nu_{j-2}^i \end{pmatrix} \lambda_{j-1}^{i+} \\ &+ \begin{pmatrix} \mu_{k^i+1}^i - \mu_{k^i}^i + \mu_{k^i-1}^i \\ \nu_{k^i+1}^i - \nu_{k^i}^i + \nu_{k^i-1}^i \end{pmatrix} \lambda_{k^i}^{i+}\end{aligned} \tag{3.25}$$

subject to the equalities and complementarity conditions :

$$L^{i+} z^{i+} + L^{i-} z^{i-} = e^i \quad (z^{i+})^t z^{i-} = 0 \quad z^{i+}, z^{i-} \geq 0 \tag{3.26}$$

Here L^{i+} and L^{i-} are $(k^i - 1) \times k^i$ matrices and z^{i+}, z^{i-} are $k^i \times 1$ vectors containing the λ_j^{i+}, resp. λ_j^{i-} parameters. The $(k^i - 1) \times 1$ vector e^i is defined as $e_j^i = j$.

- *Aggregate the descriptions of the n neurons into the following equation:*

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} c \\ d \end{pmatrix} + \begin{pmatrix} P_1 \\ P_2 \end{pmatrix} z^+ + \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} z^- \tag{3.27}$$

The $n \times 1$ vectors a, b, c, d are defined as $a_i = \alpha^i$, $b_i = \beta^i$, $c_i = \mu_1^i$, $d_i = \nu_1^i$. The matrices P_1, P_2, Q_1, Q_2 are $n \times (\sum_{i=1}^n k^i)$ matrices constructed in an obvious way from the equations. The vectors z^+, z^- are constructed as the concatenation of the vectors z^{i+} resp. z^{i-}.

- *The conditions are summarized in:*

$$L^+ z^+ + L^- z^- = e \quad (z^+)^t z^- = 0 \quad z^+, \ z^- \geq 0 \tag{3.28}$$

where L^+, L^- are $(\sum_{i=1}^n (k^i - 1)) \times (\sum_{i=1}^n k^i)$ matrices.

- *Then the invariant set of the neural net equals the solution set of the Generalized Linear Complementarity Problem:*

$$\begin{pmatrix} P_1 & -T \\ BP_2 & -A \\ L^+ & 0 \end{pmatrix} \begin{pmatrix} z^+ \\ v^+ \end{pmatrix} + \begin{pmatrix} Q_1 & T \\ BQ_2 & A \\ L^- & 0 \end{pmatrix} \begin{pmatrix} z^- \\ v^- \end{pmatrix} = \begin{pmatrix} s - c \\ -Bd \\ e \end{pmatrix} \tag{3.29}$$

with nonnegativity and complementarity conditions:

$$\begin{aligned} z^+, \ z^-, \ v^+, \ v^- \ &\geq 0 \\ (z^+)^t z^- = 0 \quad (v^+)^t v^- &= 0 \end{aligned} \tag{3.30}$$

Here $v = v^+ - v^-$ is the sign decomposition of any invariant state v.

Proof : The piecewise linear description of the neural net is given by:

$$\begin{pmatrix} a \\ b \\ e \end{pmatrix} = \begin{pmatrix} c \\ d \\ 0 \end{pmatrix} + \begin{pmatrix} P_1 \\ P_2 \\ L^+ \end{pmatrix} z^+ + \begin{pmatrix} Q_1 \\ Q_2 \\ L^- \end{pmatrix} z^-$$

Obviously, $a = Tv + s$, $b = \mathcal{F}(a)$ and $Av = Bb$. The proof then follows from the premultiplication of equation (3.29) with the matrix

$$\begin{pmatrix} I_n & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & I_{n_1} \end{pmatrix}$$

where $n_1 = \sum_{i=1}^{n}(k^i - 1)$ and from the sign decomposition of $v = v^+ - v^-$ □

A particularly interesting application for collective decision circuits such as neural nets, is the concept of *associative memory.* It will be shown how the preceding result can easily be adapted to the case where a priori some partial information is available about the invariant states. When some partial information is specified a priori, it is a natural question to ask what are the invariant states that share this prespecified piece of information. In order to formalize this idea, suppose that the states are partitioned as :

$$v = \begin{pmatrix} v_I \\ v_{II} \end{pmatrix}$$

where v_I is a $n_I \times 1$ vector of known partial information and v_{II} is the undetermined $(n-n_I)\times 1$ part of the state. Partition also the matrices T and A accordingly as:

$$A = \begin{pmatrix} A_I & A_{II} \end{pmatrix} \qquad T = \begin{pmatrix} T_I & T_{II} \end{pmatrix}$$

where A_I, T_I are $n \times n_I$ and A_{II}, T_{II} are $n \times (n - n_I)$ matrices. Then we have the following theorem:

Theorem 4 Invariant Set with Partial Information
Consider the neural net described in theorem 1 and the partioning in known and unknown partial states. The invariant set of this neural net with partially fixed components contained in v_I is the solution to the Generalized Linear Complementarity Problem:

$$\begin{pmatrix} P_1 & -T_{II} \\ BP_2 & -A_{II} \\ L^+ & 0 \end{pmatrix} \begin{pmatrix} z^+ \\ v_{II}^+ \end{pmatrix} + \begin{pmatrix} Q_1 & T_{II} \\ BQ_2 & A_{II} \\ L^- & 0 \end{pmatrix} \begin{pmatrix} z^- \\ v_{II}^- \end{pmatrix} = \begin{pmatrix} s - c + T_I v_I \\ -Bd + A_I v_I \\ e \end{pmatrix} \tag{3.31}$$

with nonnegativity and complementarity conditions:

$$z^+, \quad z^-, \quad v_{II}^+, \quad v_{II}^- \geq 0 \tag{3.32}$$

$$(z^+)^t z^- = 0 \quad (v_{II}^+)^t v_{II}^- = 0 \tag{3.33}$$

Here $v_{II} = v_{II}^+ - v_{II}^-$ is the sign decomposition of the partial state v_{II}.

Proof : Follows directly from the partitioning and theorem 3. □

The relationship between the invariant set of a neural net and the GLCP certainly suggests that the characterization of the properties of neural nets (at least of their invariant points) may be possible in terms of the properties of the (G)LCP :

- The LCP is sometimes called the 'Fundamental Problem of Mathematical Programming', a characterization which it owes to the fact that almost all mathematical programming problems are equivalent to a (G)LCP, such as linear and quadratic optimization, ...As an interesting connection, the Hopfield neural net may be mentioned: Convergence here is described in terms of an energy function [23] [40], which however in [28] was corrected and analyzed in terms of the so-called *cocontent function.* A lot is known about the relation between piecewise linear resistive networks and the GLCP (section 3.6). Together with the results of [28], this certainly suggests an implementation strategy for neural nets .

- An extensive literature exists that aims at characterizing, with varying degrees of success, the number of solutions of the (conventional) LCP in terms of properties of the matrices in the problem formulation (section 3.2.2). This relation may allow to obtain more insight in the Information Storage Capacity of neural nets, which is currently a subject of intensive research. [21].

- There exist some results about the robustness properties of the LCP [25]. Possibly, the robustness analysis of neural nets may proceed via robustness analysis of the LCP and its generalization to GLCP.

- Stability analysis of neural nets with piecewise linear decision functions, may take profit from the fact that, around an invariant point the net will behave as a linear system, so that generically all existing sensitivity results of ordinary linear autonomous systems apply for neural nets. For some interesting results about the stability properties of the invariant states, the reader is referred to [16].

- It is expected that the initial state converges to the stable state that is most similar to it. Hence, a kind of similarity measure is needed. More than probably, one can gain considerably insight by exploiting the geometrical piecewise linear properties of the generalized Linear Complementarity Problem ! More specifically, it is conjectured that the geometrical insight provided by the GLCP, may help in characterizing the so called 'zones of attraction' of a neural net.

A question which deserves further investigation is the question of computational efficiency of the algorithm to solve the GLCP required by theorem 2 and 3.

- We only know of the algorithm described in [12] [13] [14] and in this chapter in order to solve the Generalized Linear Complementarity Problem as required by theorem 3 and 4. Indeed, in both theorems, the LCP problems are *rectangular* which makes them untractable by conventional LCP - solvers. The matrix dimensions of M and N are, for theorem 2 :

$$(2n + \sum_{i=1}^{n}(k^i - 1)) \times ((\sum_{i=1}^{n} k^i) + n)$$

while for theorem 3 they are:

$$[2n + \sum_{i=1}^{n}(k^i - 1)] \times [(\sum_{i=1}^{n} k^i) + (n - n_I)]$$

where n_I is the dimension of the vector with known partial information. The fact that possibly more efficient algorithms could be developed, is mainly a matter of computational complexity and does not influence the conceptual results expressed in theorems 3 and 4.

- As can be seen from the examples in section 6, the matrices involved in the GLCP of theorem 3 and 4 contain a lot of zeros. In this context, the use of *sparse matrix techniques* could be worth considering.

- Moreover, in a lot of practical cases, the structure of neural nets induces structural properties in the matrix T. For the Hopfield neural net, T is symmetric with zero diagonal entries [45]. It seems that biological neural nets exhibit a band structure in the matrix T, which would be caused by the fact that only neighbouring neurons communicate during the iteration proces . Possibly these properties could be exploited in optimizing algorithms for the corresponding GLCP.

- Another important practical simplification results from the use of *binary* neural nets, consisting of neurons with only two possible values (e.g. the hard limiter) .

- As can be observed from the examples, the equations that correspond to the complementarity conditions are highly structured and the corresponding matrices are sparse. These facts should be exploited in a fast solution scheme.

3.7.3 An example

Consider the discrete time neural net, depicted in figure 3.15. It consists of 4 neurons , all of which have a different neuronal decision function. The neural net is described by the following net equation:

$$v_{k+1} = \mathcal{F}(Tv_k)$$

The state vector of the neural net has 4 components and the synaptic weight matrix T is given as:

$$T = \begin{pmatrix} 2 & 0 & 0 & -4 \\ 0 & 2 & 4 & 0 \\ 0 & 4 & 8 & 0 \\ -4 & 0 & 0 & 8 \end{pmatrix}$$

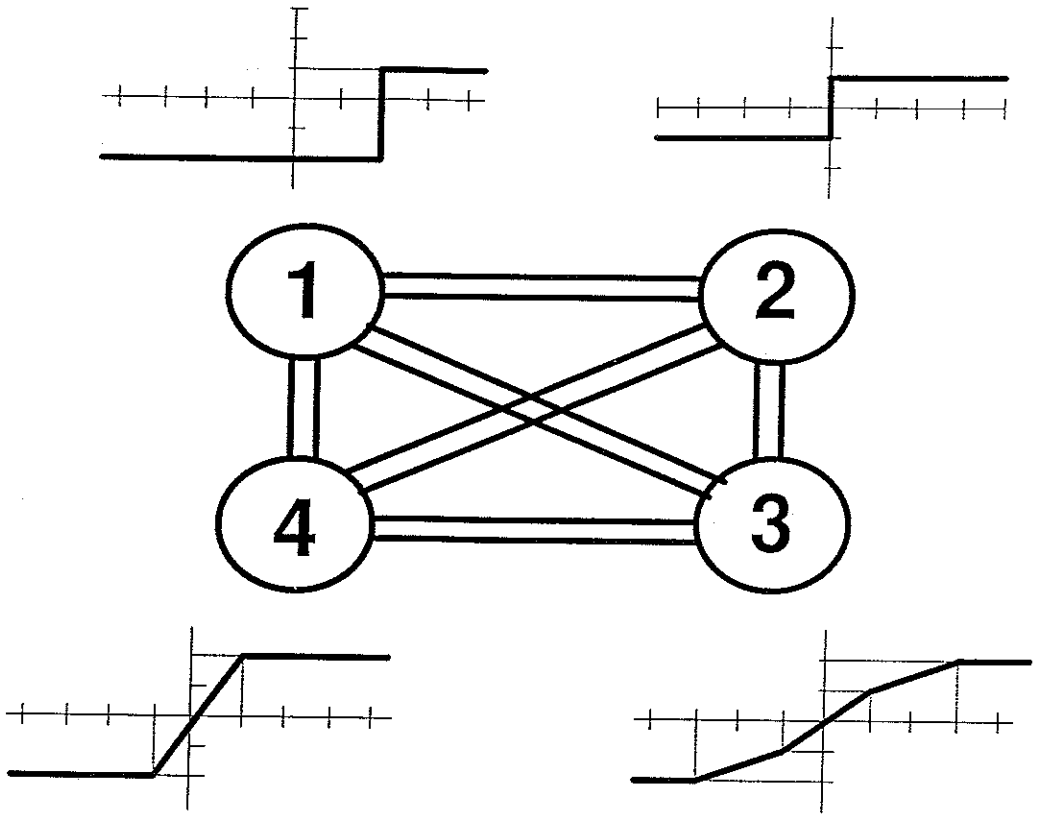


Figure 3.15: Neural Net with 4 different neurons

Compared to theorem 2, we have $A = B = I_4, s = 0$. Applying the parametrization described in section 3.4.2., one finds the piecewise linear description for this specific net:

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ 1 \\ 1 \\ 1 \\ 2 \\ 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ -3 \\ -1 \\ -2 \\ -1 \\ -2 \\ -2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} . & 1 & . & . & . & . & . & . & . & . \\ . & . & . & 1 & . & . & . & . & . & . \\ . & . & . & . & 2 & . & . & -1 & . & . \\ . & . & . & . & . & . & . & . & 2 & -1 \\ 3 & -3 & . & . & . & . & . & . & . & . \\ . & . & 2 & -2 & . & . & . & . & . & . \\ . & . & . & . & 1 & 1 & -1 & -1 & . & . \\ . & . & . & . & . & . & . & . & 4 & -4 \\ 1 & -1 & . & . & . & . & . & . & . & . \\ . & . & 1 & -1 & . & . & . & . & . & . \\ . & . & . & . & 1 & -1 & . & . & . & . \\ . & . & . & . & 1 & . & -1 & . & . & . \\ . & . & . & . & 1 & . & . & -1 & . & . \\ . & . & . & . & . & . & . & . & 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1^{1+} \\ \lambda_2^{1+} \\ \lambda_1^{2+} \\ \lambda_2^{2+} \\ \lambda_1^{3+} \\ \lambda_2^{3+} \\ \lambda_3^{3+} \\ \lambda_4^{3+} \\ \lambda_1^{4+} \\ \lambda_2^{4+} \end{pmatrix}$$

$$+\begin{pmatrix}
-1 & . & . & . & . & . & . & . & . & . \\
. & . & -1 & . & . & . & . & . & . & . \\
. & . & . & . & -1 & . & . & . & . & . \\
. & . & . & . & . & . & . & . & -1 & . \\
. & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . \\
-1 & 1 & . & . & . & . & . & . & . & . \\
. & . & -1 & 1 & . & . & . & . & . & . \\
. & . & . & . & -1 & 1 & . & . & . & . \\
. & . & . & . & -1 & . & 1 & . & . & . \\
. & . & . & . & -1 & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & -1 & 1
\end{pmatrix}
\begin{pmatrix}
\lambda_1^{1-} \\ \lambda_2^{1-} \\ \lambda_1^{2-} \\ \lambda_2^{2-} \\ \lambda_1^{3-} \\ \lambda_2^{3-} \\ \lambda_3^{3-} \\ \lambda_4^{3-} \\ \lambda_1^{4-} \\ \lambda_2^{4-}
\end{pmatrix}$$

subject to the necessary nonnegativity and complementarity conditions. The solution of the corresponding GLCP is :

number	1	2	3	4	5	6	7	8	9
λ_1^{1+}	.	.	.	9	.	9	.	9	.
λ_2^{1+}	.	.	.	8	.	8	.	8	.
λ_1^{2+}	0.5	11	.	0.5	0.5	11	11	.	.
λ_2^{2+}	.	10	.	.	1.5	10	10	.	.
λ_1^{3+}	1.5	20	.	1.5	0.5	20	20	.	.
λ_2^{3+}	0.5	19	.	0.5	.	19	19	.	.
λ_3^{3+}	.	18	.	.	.	18	18	.	.
λ_4^{3+}	.	17	.	.	.	17	17	.	.
λ_1^{4+}	24	24	24	.	.	.	0.2333	.	0.2333
λ_2^{4+}	23	23	23	.	.	.	.	.	.
v^{1+}	.	.	.	1	.	1	.	1	.
v^{2+}	.	1	.	.	.	1	1	.	.
v^{3+}	.	2	.	.	.	2	2	.	.
v^{4+}	2	2	2	.	.	.	.	.	.
λ_1^{1-}	14	14	14	.	1.7333	1	1.7333	.	1.7333
λ_2^{1-}	15	15	15	.	2.7333	1	2.7333	.	2.7333
λ_1^{2-}	.	.	10	.	.	2	.	10	10
λ_2^{2-}	0.5	.	11	0.5	0.5	.	.	11	11
λ_1^{3-}	.	.	17	.	.	.	.	17	17
λ_2^{3-}	.	.	18	.	.	.	.	18	18
λ_3^{3-}	0.5	.	19	0.5	0.5	.	.	19	19
λ_4^{3-}	1.5	.	20	1.5	1.5	.	.	20	20
λ_1^{4-}	.	.	.	19	.	19	.	19	.
λ_2^{4-}	.	.	.	20	0.7666	20	.	20	0.7666
v^{1-}	2	2	2	.	2	.	2	.	2
v^{2-}	.	.	1	.	.	.	.	1	1
v^{3-}	.	.	2	.	.	.	.	2	2
v^{4-}	.	.	.	2	1.0666	2	1.0666	2	1.0666

Hence the invariant states are :

number	1	2	3	4	5	6	7	8	9
v^1	-2	-2	-2	1	-2	1	-2	1	-2
v^2	.	1	-1	.	.	1	1	-1	-1
v^3	.	2	-2	.	.	2	2	-2	-2
v^4	2	2	2	-2	-1.066	-2	-1.066	-2	-1.066

It can be verified [16] by Lyapunov's first method, that only invariant states 2, 3, 6 and 8 are stable.

3.8 General Conclusions

In this chapter, a new class of linear complementarity problems has been introduced. The most essential features are the fact that the complementarity conditions are very general and that s ingular matrices and solutions at infinity can be included without complication. An algorithm has been proposed. It is based on a threefold inductive approach : Solving recursively a set

of linear equations in a nonnegative way, elimination of redundancy and verification of the complementarity in each recursion. A convenient feature is that *all* solutions are found as the union of polyhedral cones and polytopes. Numerous examples and applications have been given in order to demonstrate the usefulness of the GLCP in the solution of geometrical problems, modelling systems with piecewise linear descriptions and the computation of the invariant states of a neural net. However, as is always the case in research, the number of new questions grows exponentially as a function of the number of answers. Open problems with respect to the GLCP are the computational complexity and the derivation of efficient computational strategies and a deeper analysis of genericity and robustness issues. Other items demanding further research concern the interesting and exciting relations between the GLCP, optimization theory and the analysis and design of piecewise linear electrical circuits and neural networks. It is to be hoped that this will reveal important contributions towards the practical implementation of neural nets. As a matter of fact, many more results have been obtained since this chapter received its final shape. [16] [42] ... Therefore, let's conclude this chapter with turning a blind eye to Pierre de Fermat and the mystery surrounding his famous Last Theorem: *This thesis is too small to contain them all.*

Bibliography

[1] Berman A., Plemmons R.J.*Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York, 1979.

[2] Björck A. *Least Squares Methods, Handbook of Numerical Analysis, Vol.1 : Solution of equations in $\mathcal{R}^n$.* Elsevier/North Holland, 1987.

[3] Chua L.O., Lin P.M. *Nonlinear Programming without Computation.* IEEE Trans. Circuits and Systems, Vol.CAS-31, no.2, pp.182-188, February 1987.

[4] Chua L.O., Ying R.L.P. *Canonical Piecewise Linear Analysis.* IEEE Trans. Circuits and Systems, Vol. CAS-30, pp.125-140, March 1983.

[5] Chua L.O., Deng A.-C. *Canonical Piecewise Linear Analysis : Part II - Tracing Driving Point and Transfer Characteristics.* IEEE Trans. Circuits and Systems, Vol. CAS-32, pp.417-444, May 1985.

[6] Chua L.O., Deng A.-C. *Canonical Piecewise Linear Analysis: Generalized Breakpoint Hopping Algorithm.* Int. J. Circuit Theory Appl., Vol.14, pp.35-52, 1986.

[7] Chua L.O., Lin P.M. *Computer Aided Analysis of Electronic Circuits: Algorithms and Computational Techniques.* Englewood Cliffs, NJ : Prentice Hall, 1975.

[8] Chua L.O., Ying R.L.P. *Finding All Solutions of Piecewise Linear Circuits.* Int. J. Circuit Theory Appl., Vol.10, pp.201-229, 1982.

[9] Chua L.O., Matsumoto T., Ichiraku S. *Geometric Properties of Resistive Nonlinear N-Ports: Transversality, Structural stability, Reciprocity and anti-reciprocity.* IEEE Trans. Circuits and Systems, Vol. CAS-27, July 1980.

[10] Cottle R.W., Giannessi F., Lions J.-L. *Variational Inequalities and Complementarity Problems, Theory and Applications.* John-Wiley & Sons, New York,

[11] Dantzig G.B., Cottle R.W. *Positive (Semi-) Definite Programming.* in Nonlinear Programming (J.Abadie, Ed.), North-Holland, Amsterdam, 1967, pp.55-73.

[12] De Moor B. , Vandewalle J. *All nonnegative solutions to sets of linear equations and the linear complementarity problem.* Proc. of the International Symposium on Circuits and Systems

[13] De Moor B., Vandenberghe L., Vandewalle J. *A new approach to the analysis of piecewise linear resistive circuits.* Internal Report, ESAT-SISTA, Katholieke Universiteit Leuven, January 1988.

[14] De Moor B., Vandenberghe L., Vandewalle J. *The Generalized Linear Complementarity Problem and an algorithm to find all solutions.* Submitted to Mathematical Programming.

[15] De Moor B., Vandenberghe L., Vandewalle J. *Computing all invariant states of a neural net : The generalized linear complementarity problem.* Submitted.

[16] De Moor B., Vandenberghe L., Vandewalle J. *Stability Analysis of the Invariant States of Neural Nets.* Internal Report, ESAT, Katholieke Universiteit Leuven, February 1988.

[17] Denker J.S. *Neural network models of learning and adaptation.* Physica , 22D, pp.216-232, North-Holland, Amsterdam.

[18] Dion J.M., Dugard L., Nguyen M.T. *Multivariable adaptive control with input output constraints.* Proc. of the 26th conference on Decisions and Control, pp.1233, Los Angeles, December 1987.

[19] Eaves B.C. *The linear complementarity problem.* Management Science, 1971, Vol.17, no.9, pp.612-634.

[20] Eaves B.C., Gould F.J., Peitgen H.O, Todd M.J. (editors). *Homotopy methods and global convergence.* Nato Conference Series, Series II, Systems Science, Plenum Press, New York, 1983.

[21] McEliece R.J., Posner E.C., Rodemich E.R., Venkatesh S.S. *The capacity of the Hopfield associative memory.* IEEE Trans. Information Theory , Vol.IT-33, no.4, July 1987.

[22] Fiedler M., Ptak V. *Some generalisations of positive definiteness and monotonocity.* Num. Math., Vol.9, pp.163-172, Dec.1966.

[23] Hopfield J.J. *Neural Networks and physical systems with emergent collective computational abilities.* Proc.Natl.Acad.Sci.USA, Vol.79,pp.2554-2558, April 1982.

[24] Isac G. *Complementarity Problem and Coincidence equations on convex cones.* Bolletino U.M.I. (6) 5-B (1986), 925-943.

[25] Jansen M.J.M., S.H.Tijs. *Robustness and nondegenerateness for linear complementarity problems.* Mathematical Programming 37 , North - Holland pp.309-317, 1987.

[26] Kam M., Cheng R., Guez A. *A binary neural network which emulates some properties of biological memories.* IEEE 9th Annual Conference of the Engineering in Medicine and Biology Society, p.1354-1356.

[27] Kang S.M., Chua L.O. *A global representation of multidimensional piecewise linear functions with linear partitions.* IEEE Trans. Circuits and Systems, Vol.CAS-25, pp.938-940, Nov. 1978.

[28] Kennedy M.P., L.O. Chua. *Unifying the Tank and Hopfield Linear Programming Circuit and the Canonical Nonlinear Programming Circuit of Chua and Lin.* IEEE Trans. Circuits and Systems, Vol.CAS-34, no.2, February 1987.

[29] Lemke C.E. *On complementary pivot theory.* In Mathematics of the Decision Sciences (G.B.Dantzig and A.F. Veinott, Jr. Eds.). American Mathematical Society, New York,1968.

[30] Lemke C.E. *A survey of Complementarity Theory.* In 'Variational Inequalities and Complementarity Problems, Theory and Applications".(Eds. R.W.Cottle, F.Giannessi, J.-L. Lions) John Wiley and Sons Ltd. Chapter 15, p.213-239.

[31] Lin P.M. *Formulation of hybrid matrices for linear multiports containing controlled sources.* IEEE Trans. Circuits and Systems, Vol.CAS-21, pp.169-175, March 1974.

[32] Lippmann R.P. *An introduction to computing with neural nets.* IEEE ASSP Magazine April 1987.

[33] Murty K.G. *On a characterization of P-matrices.* Technical Report 69-20, Department of Industry and Engineering, University of Michigan, Ann Arbor, Mich, May 1969.

[34] Murty K.G. *On the number of solutions to the complementarity problem and the spanning properties of complementary cones.* Linear Algebra and its Applications 5 (1972) 65-108.

[35] Papadimitriou C.H., Steiglitz K. *Combinatorial Optimization, Algorithms and Complexity.* Prentice Hall Inc., Englewood Cliffs, New Jersey, 1982.

[36] Saigal R., Simon C. *Generic properties of the complementarity problem.* Mathematical Programming 4 , North - Holland Publishing Company, 324-335, 1973.

[37] Saylor J.M., Stork D.G. *Neural networks for decision tree searches.* IEEE 9th Annual Conference of the Engineering in Medicine and Biology Society, 1987. p.1366-1367.

[38] Stevens S.N., Lin P.M. *Analysis of piecewise linear resistive networks using complementary pivot theory.* IEEE Trans. Circuits and Systems, Vol. CAS-28, pp.429-441, May 1981.

[39] Tan Shaohua. *A unified approach for analysis and control of multivariable non-causal systems.* Doctoral Thesis, Fac. Applied Sciences, Electrical Engineering Department, Katholieke Universeit Leuven, October 1987.

[40] Tank D.W., Hopfield J.J. *Simple "Neural" Optimization Networks : An A/D converter, Signal Decision Circuit and a Linear Programming Circuit.* IEEE Trans. Circuits and Systems, Vol. CAS-33, No.5, may 1986.

[41] van Bokhoven W.M.G. *Piecewise Linear Modelling and Analysis.* Kluwer Technische Boeken B.V. - Deventer - Antwerpen, 1981.

[42] Vandenberghe L., De Moor B., Vandewalle J. *The Generalized Linear Complementarity Problem applied to Complete Analysis of Resistive Piecewise Linear Circuits.* Submitted to IEEE Trans. Circuits and Systems.

[43] Van Eijndhoven J.T.J. *Solving the Linear Complementarity Problem in circuit simulation.* SIAM J. Control and Optimization, Vol.24, No.5, September 1986.

[44] Venkatesh S.S. *Computation with neural networks.* IEEE 9th Annual Conference of the Engineering in Medicine and Biology Society , p.1364-1365 1987.

[45] Lippmann R.P. An introduction to computing with neural nets. IEEE ASSP Magazine april 1987.

[46] Wan Y. *On the average speed of Lemke's algorithm for quadratic programming.* Mathematical Programming 35 (1986) 236 -246.

[47] Widrow B., Winter R.G. *Neural Nets for Adaptive Filtering and Adaptive Pattern Recognition.* Computer, March 1988, pp.25-39.

Chapter 4

Oriented Energy and Oriented Signal-to-Signal Ratio Concepts in the Analysis of Vector Sequences and Time Series.

4.1 Introduction

In a wide variety of systems and signal processing applications, vector sequences are measured or computed. Such a situation naturally arises whenever multivariable signals are measured in time, at fixed locations, in a measurement set up. For the analysis of such data sequences, a wide variety of multivariate analysis tools are available. The underlying theme of much multivariate analysis is simplification and explanation of the observed phenomena. In this chapter, the problem of analysis of one or two $m \times n$ data matrices A and B is addressed. Usually $n \gg m$, where m denotes the number of measured channels while n denotes the number of measurements. It has occurred to some researchers in both signal processing and control systems that the singular value decomposition of matrices formed from observed data could be used to improve methods of signal parameter estimation and system identification. However, rationales for these methods have been very heuristic and in almost all cases are based upon the well-posedness of the algorithm, in casu the singular value decomposition (SVD). One purpose of this chapter is to present a more convincing framework which is intended both to unify existing techniques and widen the area of applications.

The results obtained in this work, bear a lot of similarity with existing (statistical) techniques, such as principal component analysis, factor analysis, analysis of variance etc... Principal component analysis, originating in some work by Karl Pearson around the turn of the century and further developed in the 1930's by Harold Hotelling, consists of finding an orthogonal transformation of the original - stochastic - variables to a new set of uncorrelated variables, which are derived in non-increasing order of importance. These so-called principal components are linear combinations of the original variables and it is the analysist's hope that the first few components will account for most of the variation in the original data so that the effective dimensionality of the data can be reduced [3] [9]. The concept of *oriented energy* defined and studied in this work, is closely related to principal component analysis. The concept of

oriented signal-to-signal ratio is closely related to factor-analysis like methods (in 'modern' approaches so-called subspace methods) in which the used metric is imposed by the noise covariance matrix, acting as a prewhitener of the measurements via a so-called Mahalanobis transformation.

It will be shown how the framework of oriented signal-to-signal ratio provides a rationale for linear modeling problems where:

1. the complexity of the model is the rank of a certain matrix. Hence, the decision for the complexity essentially reduces to the meaningful determination of the rank of certain (prewhitened) matrices.
2. the model parameters are linked in one way or another to the subspaces and their properties, that are associated to the determined rank.

Several new aspects are emphasised throughout this chapter.

- A general framework is derived, explicitly based upon the properties of the singular and generalized singular value decomposition. No statistical a priori assumptions (about e.g. probability distributions) are imposed. However, when such a priori information is available, it can be taken into account.
- The conceptual derivations based upon the (generalized) singular value decomposition are at the same time constructive: The very use of these factorizations delivers algorithms, that may be implemented in numerically robust and reliable software.
- Whereas the singular value decomposition is one of the tools in the analysis of a single vector sequence, it will be shown how the generalized singular value decomposition is the technique to be used in analyzing the relation of two vector sequences.

Several examples will be presented that lend themselves to a translation and interpretation in the novel framework : total linear least squares with specified admissible complexity or tolerated misfit, high resolution location of narrowband sources, separation of maternal from fetal ECG and linear dynamical realization theory.

This chapter is organized as follows: In section 4.2, the basic definitions and theorems of the concept of oriented energy and signal-to-signal ratio are defined and derived. The numerical tool to analyse the spatial activity of one vector sequence is the singular value decomposition as is demonstrated in section 4.3. When two vector sequences are to be studied relatively to each other, the generalized singular value decomposition applies. It is shown in section 4.4 how there exists a strong similarity between the singular value decomposition for the analysis of one vector sequence and the generalized singular value decomposition for the analysis of two vectorsequences. In section 4.5, the results are illustrated with some clarifying examples. The conclusions can be found in section 4.6.

4.2 Oriented energy and oriented signal-to-signal ratio concepts of a set of vectors

In this section, the basic definitions of oriented energy are given. The column vectors of an $m \times n$ matrix A are considered to form an indexed set of m-vectors, denoted by $\{a_k\}, k =$

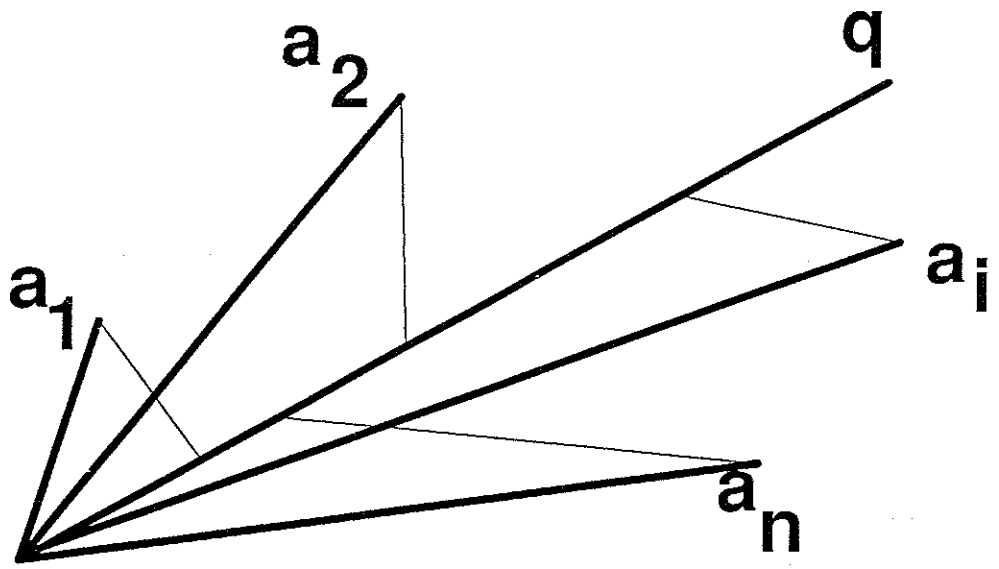


Figure 4.1: Illustration of oriented energy measurement

$1, \ldots, n$. An m-vector q and the direction it represents in a vector space, are used as synonyms.

Definition 1 Energy of a vector sequence.
Consider a sequence of m-vectors $\{a_k\}, k = 1, \ldots, n$ and associated $m \times n$ matrix A. Its total energy $E[A]$ is defined via the Frobeniusnorm of the $m \times n$ matrix A:

$$E[A] = \|A\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2$$

Definition 2 Oriented energy.
Let A be a $m \times n$ matrix and denote its n columnvectors as $a_k, k = 1, \ldots, n$ (n is possibly infinite). For the indexed vectorset $\{a_k\}$ of m-vectors $a_k \in \mathcal{R}^m$ and for any unit vector $q \in \mathcal{R}^m$ the energy E_q, measured in the direction q, is defined as:

$$E_q[A] = \sum_{k=1}^{n} (q^t \cdot a_k)^2$$

The energy E_Q measured in a subspace $Q \subset \mathcal{R}^m$, is defined as:

$$E_Q[A] = \sum_{k=1}^{n} \|P_Q(a_k)\|^2$$

where $P_Q(a_k)$ denotes the orthogonal projection of a_k into the subspace Q and $\|\cdot\|$ denotes the Euclidean norm.

A geometric visualisation is represented in fig.4.1.
Of course, the summations require l^2-type convergence conditions on the set $\{a_k\}$ when n is infinite. In words, the oriented energy of a vector sequence $\{a_k\}$, measured in the direction

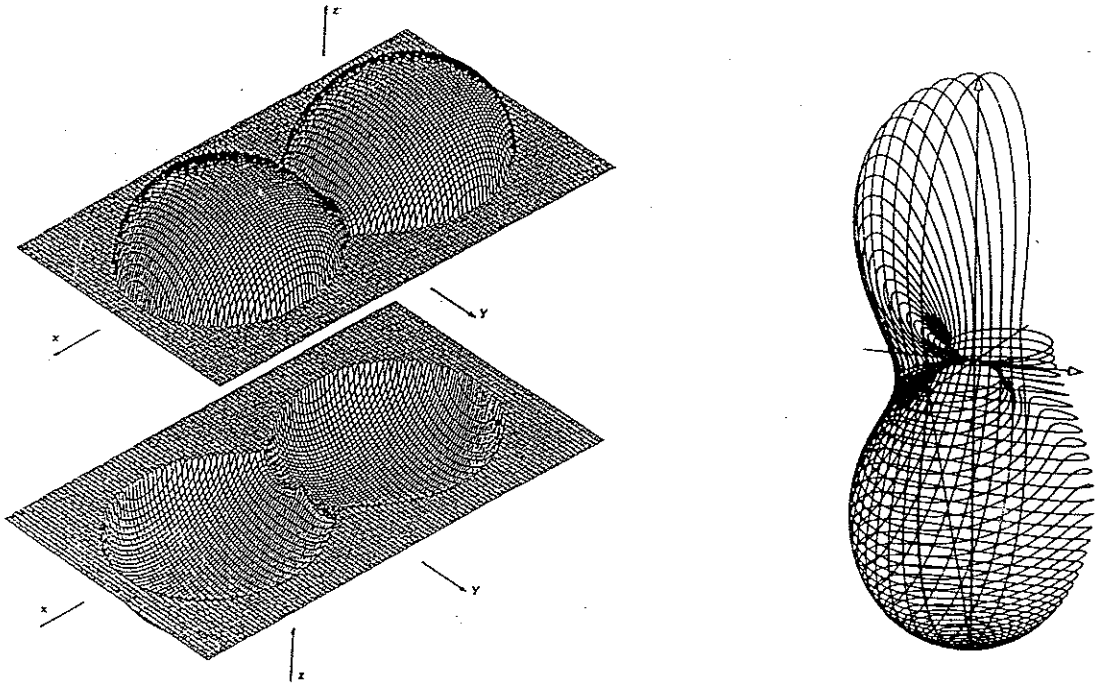


Figure 4.2: Oriented energy distribution in 3 dimensions

q (subspace Q) is nothing else than the energy of the signal, orthogonally projected on the vector q (subspace Q).

In order to sharpen the geometric intuition on the above interpretations, the oriented energy distribution and the square root of this distribution is shown in fig.4.2.a. for a 3-vector sequence which is nearly singular. The physical significance of the definition becomes clear, when one considers the unit vector q (subspace Q) to be variable and to 'sense' in all directions of the vectorspace $\mathcal{R}^m$. For a sequence of 3-vectors, we have plotted in each 'sensing' direction q of the 3-dimensional vectorspace, a vector of length E_q. All points lie on a smooth surface. In fig.4.3.b. quarter of the surface has been cut out to show the typical ('sharp') narrowing that may occur in some directions. In three dimensions, the surfaces of oriented energy exhibit one direction of maximal oriented energy, one of minimal energy an a direction with a saddle point. This can readily be generalized to more dimensions.

Let us now consider two vector sequences $\{a_k\}$ and $\{b_k\}$:

Definition 3 Oriented signal-to-signal ratio
The oriented signal-to-signal ratio $R_q[A,B]$ for two sets of m-vectors $\{a_k\}$ and $\{b_k\}, k = 1,\ldots,n$ (n possibly infinite), measured in the direction of a unit vector $q \in \mathcal{R}^m$, is defined by:

$$R_q[A,B] = \frac{E_q[A]}{E_q[B]}$$

More generally, the oriented signal-to-signal ratio $R_Q[A,B]$ for two vector sequences $\{a_k\}$ and $\{b_k\}$, measured in a subspace $Q \subset R^m$, is defined as:

$$R_Q[A,B] = \frac{E_Q[A]}{E_Q[B]}$$

In words, the signal-to-signal ratio of two vector sequences, measured in a direction q or subspace Q, is simply the ratio of the two oriented energies of the involved signal sequences in that direction or subspace. Note that when B is rankdeficient, there exist directions in which the oriented energy $E_q[B] = 0$, possibly making the signal-to-signal ratio infinite if $E_q[A] \neq 0$.

The oriented energy distribution shows of course a more than coincidental relationship with the ellipsoid, described by the nonnegative definite quadratic form of AA^t:

Theorem 1 *Consider the $m \times n$ matrix A and the vector sequence $\{a_k\}$ of its column vectors. Then, any m-vector r of the ellipsoid $\{r \mid r^t AA^t \cdot r = 1\}$, associated with the quadratic form of the matrix AA^t and the oriented energy of the vector sequence $\{a_k\}$ in the direction $r/\|r\|$ are related by:*

$$\|r\|^2 E_{r/\|r\|}[A] = 1$$

Proof: Trivial □

In words, the energy distribution of a sequence of vectors $\{a_k\}$ can be constructed from the ellipsoid of the quadratic form of AA^t by scaling any vector r on the ellipsoid until its length is $1/\|r\|^2$. Hence, the matrix AA^t characterizes the quadratic form ellipsoid as well as the oriented energy distribution. This correspondance implies that the oriented energy is everywhere continuous on the unit sphere and also everywhere differentiable. The theorem then implies that *directions of extremal energy coincide with the principal axes of the ellipsoid, hence are orthogonal.* This observation really invites to use the oriented energy concept in the analysis of the spatial activity of vector signals.

The importance of the new notion of oriented energy, with its close relation to the 'classical' quadratic form, will follow from both the conceptual as the numerical arguments developed in the remainder of this chapter. It will be demonstrated how the oriented energy concept is indeed a powerful tool to separate signals from different sources, to 'filter' signals from noise and to select subspaces of maximal signal activity and integrity.

An important observation is that the range of values of the oriented energy distribution of a vector sequence, depends upon the *choice of basis* to which the vector coordinates refer. Moreover, the directions of extremal oriented energy are not preserved and the shape and values of the oriented energy distribution may totally change under *non-orthonormal* basis transformations. This indicates that the notion of oriented energy is a workable concept only if the choice of basis for the representation of the vector set $\{a_k\}$ is fixed by external or physical arguments. A similar observation holds for the oriented signal-to-signal ratio $R_q[A,B]$ of two vector sequences $\{a_k\}$ and $\{b_k\}$ in a fixed direction q. However, the range of the values of $R_q[A,B]$ considered over all unit directions q, is independent of the choice of basis. This implies that they have a wider physical significance. This important invariance property is stated more precisely as follows.

Theorem 2 Invariance property of the signal-to-signal ratio
Consider 2 sequences of m-vectors $\{a_k\}, \{b_k\}, k = 1, \ldots, n$. *For every unit vector* $q \in R^m$ *and for every non-singular* $m \times m$ *matrix* T *that transforms* $\{a_k\}, \{b_k\}$ *into* $\{Ta_k\}, \{Tb_k\}$ *there exists an associated vector* q' *such that*

$$R_q[A, B] = R_{q'}[TA, TB]$$

Proof: Verify that the theorem is satisfied for $q' = \frac{(T^{-1})^t q}{\|(T^{-1})^t q\|}$. □

The message of theorem 2 (derived in [15]) is the following: Although the measurements of oriented energy ratios in a fixed direction depend on the choice of basis, the existence of a ratio with a certain value is independent of the chosen basis. Stated otherwise : if the oriented signal-to-signal ratio has a certain value with respect to a certain basis, it will have the same value in some direction for all possible choices of basis. Note that theorem 2 is constructive in the sense that it allows to compute this specific direction in the new basis.

4.3 The oriented energy concept and the singular value decomposition

In section 4.2, attention was paid to the basic concepts and properties of the oriented energy distribution. In this section, the tools will be studied which allow to characterize numerically the oriented energy concept. In section 4.3.1, results about the singular value decomposition are summarized. More conceptual relations between the SVD and the oriented energy properties of a vector sequence are established in section 4.3.2. In section 4.3.3, some numerical considerations are discussed.

4.3.1 The singular value decomposition (SVD)

For conceptual, numerical, algebraic and computational reasons, the singular value decomposition (SVD) is receiving more and more attention [7]. The SVD for real matrices is based upon the following theorem [7] which we name after its most important contributors:

Theorem 3 The Autonne-Eckart-Young theorem *(restricted to real matrices)*
For any real $m \times n$ *matrix* A, *there exist a real factorization :*

$$\underset{m \times n}{A} = \underset{m \times m}{U} \cdot \underset{m \times n}{S} \cdot \underset{n \times n}{V^t}$$

in which the matrices U *and* V *are real orthonormal, and the matrix* S *is real pseudo-diagonal with nonnegative diagonal elements.*

The diagonal entries σ_i of S are called the singular values of the matrix A. It is assumed that they are sorted in non-increasing order of magnitude. The set of singular values $\{\sigma_i\}$ is called the singular spectrum of the matrix A. The columns $u_i(v_i)$ of $U(V)$ are called the left (right) singular vectors of the matrix A. The space $S_U^r = \text{span}_{\text{col}}[u_1, \ldots, u_r]$ is called the r-th left principal subspace. In a similar way, the r-th right singular subspace is defined. The triple (u_i, σ_i, v_i) is called the i-th singular triplet of the matrix A. Note

that the singular value decomposition of a real matrix is not unique. However, the singular values are uniquely determined. If the non-zero singular values are distinct, the corresponding singular vectors are unique up to the sign. If r singular values (zero or non-zero) coincide, the corresponding singular vectors are arbitrary as long as they generate an orthonormal basis for the corresponding r-dimensional subspace which is unique. Proofs of the above classical existence and uniqueness theorems are found in [7] and the references therein. Some more properties of the singular value decomposition are mentioned here without proof.

Lemma 1 *The number of singular values, different from zero, equals the algebraic rank of the matrix A.*

In fact, the SVD is one of the most reliable tools to estimate in a numerically sound way the algebraic rank of a matrix.

Lemma 2 Dyadic decomposition
Via the SVD , any matrix A can be written as the sum of $r = \mathrm{rank}(A)$ rank one matrices :

$$A = \sum_{i=1}^{r} u_i \cdot \sigma_i \cdot v_i^t$$

where (u_i, σ_i, v_i^t) is the i-th singular triplet of the matrix A.

Lemma 3 Frobenius norm of $m \times n$ matrix A of rank r

$$\|A\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 = \sum_{k=1}^{r} \sigma_k^2$$

where the σ_k are the singular values of A.

In words, the total energy in a vector sequence $\{a_k\}$ with associated matrix A as defined in definition 1, is equal to the energy in the singular spectrum.

The smallest non-zero singular value corresponds to the distance in Frobeniusnorm, of the matrix to the closest matrix of lower rank. This well known property makes the SVD attractive for approximation and data reduction purposes. Observe that this result is quite remarkable: Since, *generically*, in the space of all $m \times n$ matrices, a matrix is of full rank (in other words, the set of rank deficient matrices has *Lebesgue measure zero*), it follows that the smallest singular value measures the distance from a matrix to the set of rank deficient matrices without the need to compute explicitly this set of rank deficient matrices.

There exists an important well-known relation between the singular value decomposition and the eigenvalue decomposition:

Lemma 4 *Let the $m \times n$ matrix A have an SVD as in theorem 3. Then the columns of U are the eigenvectors of the Grammian AA^t. The columns of V are the eigenvectors of the Grammian A^tA. The positive real number σ_i is a non-zero singular value of A iff σ_i^2 is a non-zero eigenvalue of both AA^t and A^tA.*

This relation with the eigenvalue decomposition, allows to generalize existing results on the eigenvalue decomposition into similar results for the singular value decomposition.

4.3.2 Conceptual relations between SVD and oriented energy

We are now in the position to establish the link between the singular value decomposition and the concept of oriented energy distribution.

Define the unit ball UB in $\mathcal{R}^m$ as $\text{UB} = \{q \in \mathcal{R}^m \mid \|q\|_2 = 1\}$.

Theorem 4 *Consider a sequence of m-vectors* $\{a_k\}, k = 1, \ldots, n$ *and the associated* $m \times n$ *matrix A with SVD as defined in theorem 3 with* $n \geq m$. *Then:*

$$E_{u_i}[A] = \sigma_i^2$$

$\forall q \in \text{UB}$: *if* $q = \sum_{i=1}^{m} \gamma_i \cdot u_i$, *then*

$$E_q[A] = \sum_{i=1}^{m} \gamma_i^2 \cdot \sigma_i^2$$

where UB *is the unit ball.*

Proof: Trivial from theorem 3. □

In words, the oriented energy measured in the direction of the i-th left singular vector of the matrix A, is equal to the i-th singular value squared. The energy in an arbitrary direction q can be reconstructed additively as a sum of 'orthogonal' oriented energies associated to the left singular directions, as soon as the coordinates γ_i of the vector q with respect to the left singular vectors are known. If the matrix A is rankdeficient, then there exist directions in $\mathcal{R}^m$ that contain no energy at all. It should be observed that the singular values and vectors are generally critically dependent *upon the scales used to measure the variables.* This scaling could be the result of data acquisition requirements such as amplification, A/D conversion etc... Hence, additional physical motivitation is required to choose those scalings for the different measurement channels for which it is meaningful to 'compare' via the oriented energy concept. In the sequel it is assumed that this question has been resolved in advance.

With the aid of theorem 4, one can easily obtain, using the SVD, the directions and spaces of extremal energy, as follows.

Corollary 1 *Under the assumptions of theorem 4:*

1. $\max_{q \in \text{UB}} E_q[A] = E_{u_1}[A] = \sigma_1^2$
2. $\min_{q \in \text{UB}} E_q[A] = E_{u_m}[A] = \sigma_m^2$
3. $\max_{Q^r \subset \mathcal{R}^m} E_{Q^r}[A] = E_{S_U^r}[A] = \sum_{i=1}^{r} \sigma_i^2$
4. $\min_{Q^r \subset \mathcal{R}^m} E_{Q^r}[A] = E_{(S_U^{m-r})^\perp}[A] = \sum_{i=m-r+1}^{m} \sigma_i^2$
5. $\max_{Q^r \subset \mathcal{R}^m} \{\min_{q \in Q^r} E_q[A]\} = \min_{q \in S_U^r} E_q[A] = \sigma_r^2$
6. $\min_{Q^r \subset \mathcal{R}^m} \{\max_{q \in Q^r} E_q[A]\} = \max_{q \in (S_U^{m-r})^\perp} E_q[A] = \sigma_{m-r+1}^2$

where 'max' and 'min' denote operators, maximizing or minimizing over all r-dimensional subspaces Q^r *of the ambient range space* $\mathcal{R}^m$. S_U^r *is the r-dimensional principal subspace of the matrix A while* $(S_U^{m-r})^\perp$ *denotes the r-dimensional orthogonal complement of* S_U^{m-r}.

Proof: The first 4 properties follow immediately from the SVD and from theorem 3 and from theorem 4. The last 2 properties are nothing else then the classical Courant-Fischer minimax and maximin characterizations of the eigenvalues. [7] □

In words, the first 2 properties relate the SVD to the mimima and maxima of the oriented energy distribution. In fact, it can be shown that extrema occur at each left singular direction. The r-th principal subspace S_U^r is, among all r-dimensional subspaces of $\mathcal{R}^m$, the one that senses a maximal oriented energy. The orthogonal decomposition of the energy via the singular value decomposition is canonical in the sense that it allows to find subspaces of dimension r where the sequence has minimal and maximal energy. This decomposition of the ambient space, as a direct sum of a space of maximal and minimal energy for a given vector sequence, leads to very interesting rank considerations, which will be exploited furtheron. The last 2 properties characterize the min-max properties of the vectorsequence if this is restricted to p-dimensional subspaces. For a fixed subspace Q^r, the minimum energy is achieved for a certain direction q. When all minima are considered for all possible r-dimensional subspaces Q^r, then there is at least one maximum. This maximum of all minima can be interpreted as the best of all worst cases. Its algorithmic computation and the determination of the corresponding maximizing subspace, is closely related to the singular value decomposition.

In signal processing, one often encounters long sequences of m-vectors a_k. This means that the corresponding $m \times n$ matrix will be largely overdetermined with many more columns n than rows $m : n \gg m$. The singular value decomposition allows to compact such sequences into sequences with equivalent oriented energy properties:

Theorem 5 *Consider a sequence of m-vectors $\{a_k\}, k = 1, \ldots, n$ and the associated $m \times n$ matrix A with SVD as defined in theorem 3 with $n \geq m$. Then the sequence of m-vectors*

$$\{u_1\sigma_1, u_2\sigma_2, \ldots, u_m\sigma_m\}$$

has the same oriented energy distribution as that of $\{a_k\}$.

Proof: Straightforward. □

One of the main applications of this theorem concerns vector stochastic processes: Ergodic vector stochastic processes can be characterized by an equivalent vector signal $U\Sigma$, closely related to the second-order joint moment matrix of the process. This implies that from the point of view of oriented energy, the sequence should not be known by an actual time realization. Only the knowledge of the equivalent sequence with identical oriented energy distribution is required.

Definition 4 Isotropy
The oriented energy distribution of a sequence of m-vectors $\{a_k\}$ will be called isotropic if the singular values of the corresponding matrix A are all equal to each other.

4.3.3 Numerical considerations.

The practical value of theorem 5 is that it allows to compact a large amount of data without algebraic or numerical degeneracies. In this context, one should clearly distinguish theorem

4 from the compaction obtained by the computation of AA^t described in theorem 1, and this especially concerning the numerical caveats. Indeed, all properties have been stated in terms of energies. This implies the summing of squares, which, in the presence of numerical round off errors caused by the limited machine precision, can lead to numerical disasters. The strength of the approach is that all computations can be performed without explicitly using these squares. The singular value decomposition obtains the decomposition of the vector sequence as a sum of orthogonal dyadic terms, weighted with singular values that can be computed within full machine precision. Numerically stable and reliable algorithms for the singular value decomposition are by now well known [7], fully tested and documented and available in reliable standard softwarepackages (e.g. Matlab, Eispack,NAG,...)

Another crucial issue concerns the computational cost of the singular value decomposition. Typically, in signal processing applications, the number of vectors n in a m-vectorsequence $\{a_k\}$ is much larger than the number of components m. This implies that the associated $m \times n$ matrix A is largely overdetermined ($n \gg m$) and hence the SVD of a 'very rectangular' matrix is required. Fortunately, there exists a simple 'trick' which allows to stably compute such SVD's. Without going into technical detail, it suffices to mention that first the R-Q factorization of the overdetermined matrix is computed, followed by the SVD of the lower triangular matrix R. This results a considerable computational saving. Moreoever, since it is 'easy' and 'cheap' to compute a rank one update of the R-Q factorization (necessary when one extra column is added), this opens interesting perspectives for an algorithm that is adaptive in the number of measurements (updating and downdating strategies).

4.4 Signal- to-signal ratios and the generalized singular value decomposition.

While in the previous section, the link between the oriented energy distribution of one vector sequence with the SVD of the associated matrix was established, in this section the relation of the signal-to-signal ratio of two vector sequences with the generalized singular value decomposition will be studied. The use of the generalized singular value decomposition allows to develop a highly instrumental parallellism between the concept of oriented energy distribution and the signal-to-signal ratio of two vector sequences.

In section 4.4.1, the theorem stating the existence of the generalized singular value decomposition is given together with its main properties. In section 4.4.2., it is shown how to apply the GSVD in order to compute the maximal minimal signal-to-signal ratio of two vectorsequences. In section 4.4.3., attention is paid to the numerical implications of the GSVD.

4.4.1 The Generalized Singular Value Decomposition [GSVD]

Theorem 6 The Generalized Singular Value Decomposition.
Let A be a $m \times n (n \geq m)$ and B a $m \times p$ matrix, then there exist orthonormal matrices U $(n \times n)$ and V $(p \times p)$ and a non-singular $m \times m$ matrix X such that

$$\begin{aligned} A &= X^{-t} D_A U^t \\ B &= X^{-t} D_B V^t \end{aligned}$$

where

$$D_A = \mathrm{diag}(\alpha_1, \ldots, \alpha_m), \quad \alpha_i \geq 0,$$

is a rectangular diagonal $m \times n$ matrix, and

$$D_B = \mathrm{diag}(\beta_1, \ldots, \beta_q) \quad , \beta_i \geq 0, \quad q = \min(m, p),$$

is a rectangular diagonal $m \times p$ matrix and

$$\beta_1 \geq \ldots \geq \beta_r > \beta_{r+1} == \beta_q = 0, \qquad r = \mathrm{rank}(B)$$

Proof: see e.g. [7] □

The elements of the set $\sigma(A, B) = \{\alpha_1/\beta_1, \ldots, \alpha_r/\beta_r\}$ are referred to as the generalized singular values of A and B. The theorem is a generalization of the SVD since $\sigma(A, B)$ equals the singular spectrum of the matrix A if $B = I_m$. In this paper, the case where $n > m$ and $p > m$ will be of interest. In that case, note that if U and V are partitioned as $U = [U_A \;\; U_2]$ and $V = [U_B \;\; V_2]$ where U_A, U_2, U_B, V_2 are $m \times n, m \times (n-m), p \times m, p \times (p-m)$ matrices, the GSVD of A and B can be written as

$$\begin{aligned} A &= X^{-t} D_A U_A^t \\ B &= X^{-t} D_B U_B^t \end{aligned}$$

where D_A and D_B are now square diagonal. This notation will be used from here on. There exists an intimate theoretical link between the generalized singular value decomposition of the matrix pair $[A, B]$ and the generalized symmetric eigenvalue problem:

Lemma 5 *: Let A and B be as in theorem 6. Then the generalized singular values are the square roots of the generalized eigenvalues γ of the symmetric eigenvalue problem :*

$$AA^t x = BB^t x \gamma$$

The matrix X of theorem 6 contains the generalized eigenvectors and diagonalizes simultaneously AA^t and BB^t. Moreover, $\gamma_i = (\alpha_i/\beta_i)^2$.

Another consequence is the relation between the generalized singular value decomposition and the product of a matrix with the pseudo-inverse of another:

Lemma 6 *Let A and B have a GSVD as in theorem 6. Then, an SVD of $B^+ A$ is given by:*

$$B^+ A = U_B D_B^+ D_A U_A^t$$

Proof: Straightforward. □

For other interesting properties, the reader is referred to [7].

4.4.2 Conceptual relations between the signal-to-signal ratio and GSVD

In this section, it will be demonstrated how the Generalized Singular Value Decomposition allows to characterize the signal-to-signal ratio of two given sequences of m-vectors $\{a_k\}, \{b_k\}, k = 1, \ldots, n$ with associated $m \times n$ matrices A and B. It is assumed that $n > m$ as is the case in signal processing applications. Often, one of the signals, say $\{a_k\}$ can be considered as the desired signal while the second one $\{b_k\}$ represents the undesired signal that in some sense corrupts the desired one. The problem of interest then clearly consists in separating the desired part from the undesired signal. In quite some applications, examples of which will be presented in section 4.5, the associated $m \times n$ matrix A is rankdeficient: $\text{rank}(A) < m$. In this case, one has a physical motivation to restrict the attention to r-dimensional subspaces of the ambient space (a recent rationale is developed in [14]; the literature covering such rank-decision tests is enormous, including maximum likelihood eigenvalue ratio tests [2], Akaike's Information Criterion ([1]), Rissanen's Minimum Description Length Criterion [11], Willems' recent results on complexity/misfit approximate modeling [22] etc....) Once the rank r has been fixed, the question of interest is to find the optimal r-dimensional subspace, in which the desired signal sequence $\{a_k\}$ can be optimally distinguished from the corrupting sequence $\{b_k\}$. It will be shown that this is equivalent to determining the maximal minimal signal-to-signal ratio of the two vectorsequences. In order to avoid unnecessary complication, caused by possible rank deficiency of B, it is assumed from now on that B is of full row rank i.e. $\text{rank}(B) = m$. If B is rank deficient with $\text{rank}(B) = r < m$, the vector sequence $\{b_k\}$ has no energy in the orthogonal complement of the r-th left principal subspace of B. For every direction q in this orthogonal complement, the signal-to-signal ratio $R_q[A, B]$ is infinite if $E_q[A] \neq 0$. Such directions can easily be dealt with in advance by some kind of deflation-orthogonalization procedure applied to the column spaces of the matrices A and B. In a signal processing context however, the full row rank situation is the generic one. For these reasons, the possible rank deficiency of B will not be considered in detail in our further discussion. Two elements will be used : The invariance of the signal-to signal ratio distribution under non-singular transformations (theorem 2) and the generalized singular value decomposition of the pair of $m \times n$ matrices A and B.

Theorem 7 *Given two sequences of m-vectors* $\{a_k\}$ *and* $\{b_k\}, k = 1, \ldots, n$ *with associated* $m \times n$ *matrices* A *and* B $(n > m)$ *where* $\text{rank}(B) = m$. *Consider the GSVD of* A *and* B*:*

$$A = X^{-t} D_A U_A^t$$
$$B = X^{-t} D_B U_B^t$$

Define the linear transformation:

$$T = D_B^{-1} X^t$$

and the transformed vector sequence $\{c_k\}$ *via* $c_k = Ta_k$ *with associated* $m \times n$ *matrix* $C = TA$. *Then:*

$$E_{q'}[C] = R_q[A, B]$$

where

$$q' = \frac{(T^{-1})^t q}{\|(T^{-1})^t q\|}$$

Proof: This theorem is an immediate consequence of the invariance theorem 2 for the signal-to-signal ratio.

$$\begin{aligned} R_q[A,B] &= R_{q'}[TA,TB] \\ &= R_{q'}[D_B^{-1}D_AU_A^t,U_B^t] \\ &= E_{q'}[D_B^{-1}D_AU_A^t] \end{aligned}$$

because the oriented energy distribution of the sequence U_B^t is isotropic : $E_{q'}[U_B^t] = 1$. □

Theorem 7 links the signal-to-signal ratio with the oriented energy distribution in the following way. A linear transformation T transforms the vectorsequence $\{b_k\}$ into an isotropic sequence with unit energy distribution. By the invariance theorem 2, it is guaranteed that the signal-to-signal ratio distribution is preserved. Moreover, all information on the signal-to-signal ratio is now available in the oriented energy distribution of the transformed sequence $TA = D_B^{-1}D_AU_A^t$. The sequence TB has become isotropic. Hence, the linear transformation T could be considered as some kind of 'whitening' operator, an idea which is commonly applied in statistics (in e.g. minimum variance and Markov-type estimators) [8]. But, the most important observation is that, by the very choice of the linear transformation T as $T = D_B^{-1}X^t$, the resulting sequence $D_B^{-1}D_AU_A^t$ has precisely the form of a singular value decomposition, in which the left singular matrix equals the identity matrix. This allows to adapt theorem 4 and corollary 1 directly to the properties of the signal-to-signal ratios in a straightforward way:

Corollary 2 *Consider the GSVD of the matrix pair A and B. Assume that the generalized singular values are ordered in non-increasing order of magnitude:*

$$(\alpha_i/\beta_i) \geq (\alpha_{i+1}/\beta_{i+1})$$

Denote by t_i the i-th column of the matrix $T^t = XD_B^{-1}$ and by x_i the i-th column of the matrix X. (Clearly $t_i = x_i/\beta_i$). UB *is the unit ball. Under the assumptions and notations of theorem 7:*

1. *For $q = t_i/\|t_i\|$, $R_q[A,B] = (\alpha_i/\beta_i)^2$*
2. *If $q = \sum_{i=1}^m \gamma_i t_i$, then*
$$R_q[A,B] = \frac{\sum_{i=1}^m (\gamma_i\alpha_i/\beta_i)^2}{\sum_{i=1}^m \gamma_i^2}$$
3. $\max_{q\in \mathrm{UB}} R_q[A,B] = R_{t_1/\|t_1\|}[A,B] = (\alpha_1/\beta_1)^2$
4. $\min_{q\in \mathrm{UB}} = R_q[A,B] = R_{t_m/\|t_m\|}[A,B] = (\alpha_m/\beta_m)^2$

Proof: The first 2 properties follow by straighforward substitution while 3 and 4 are special cases of property 2 and of the extremal relationships between oriented energy and the singular value decomposition (theorem 4 and corollary 1). □

First note that $t_i/\|t_i\| = x_i/\|x_i\|$. The GSVD not only provides the extrema of the signal-to-signal ratio but also the directions in which those extrema occur: These are simply the columns of the matrix X, which, from lemma 5 are the generalized eigenvectors of the matrix

pair (A^tA, B^tB). Hence the extreme directions of oriented signal-to-signal ratio need not to be orthogonal. The minimal and maximal signal-to-signal ratios and direction can be found in properties (3) and (4). If the coordinates of a vector q are known with respect to the basis generated by the columns of the matrix T, then the signal-to-signal ratio follows immediately from the generalized singular values (property (2)) Let us now consider some optimal and worst case signal-to-signal ratios.

Definition 5 Maximal minimal and minimal maximal signal-to-signal ratio
The maximal minimal signal-to-signal ratio of two m-vector sequences $\{a_k\}$ *and* $\{b_k\}$ $k = 1, \ldots, n$ *over all r-dimensional subspaces* $(r \leq m < n)$, *denoted by* $\mathrm{MmR}[A, B, r]$, *is defined as:*

$$\mathrm{MmR}[A, B, r] = \max_{Q^r \subset \mathcal{R}^m} \min_{q \in Q^r} R_q[A, B]$$

The minimal maximal signal-to-signal ratio, $\mathrm{mMR}[A, B, r]$ *over all r-dimensional subspaces, is defined as:*

$$\mathrm{mMR}[A, B, r] = \min_{Q \subset \mathcal{R}^m} \max_{q \in Q^r} R_q[A, B]$$

The idea behind these definitions is the following: For a given subspace Q^r, there is a certain direction q in Q^r for which the signal-to-signal ratio of the two vector sequences $\{a_k\}$ and $\{b_k\}$ is minimal. This direction corresponds to the worst direction q in the subspace Q^r in the sense that the energy of A is difficult to distinguish from the energy of B. This worst case of course depends upon the subspace Q_r. Among all r-dimensional subspaces, there must be at least one subspace where the worst case is better than all other worst cases. Hence the maximal minimal signal-to-signal ratio is in some sense the best of all worst cases : In the corresponding maximizing subspace, it can be guaranteed for all directions that the energy of A is at least $\mathrm{MmR}[A, B, r]$ times larger than that of B. A similar explanation can be derived for the minimal maximal signal-to-signal ratio: it is the worst of all best cases considered over the r-dimensional subspaces. Note that 3 elements are involved in the definition of $\mathrm{MmR}[A, B, r]$ (*resp.* $\mathrm{mMR}[A, B, r]$):

- Some motivation must be available to determine a suitable r.
- In each possible r-dimensional subspace, there is a worst (best) direction q that minimizes (maximizes) $R_q[A, B]$
- That r-dimensional subspace is selected in the ambient space $\mathcal{R}^m$ where the worst (best) case is best (worst).

The results of theorem 7 and corollary 2, lead immediately to a computational procedure to compute $\mathrm{MmR}[A, B, r]$ and $\mathrm{mMR}[A, B, r]$, based upon the GSVD of the matrix pair A, B and the fact that the signal-to-signal ratio of two signals is invariant under linear transformations (invariance theorem 2).

Corollary 3 *Consider two m-vector sequences* $\{a_k\}, \{b_k\}$ $k = 1, \ldots, n$, *associated* $m \times n$ *matrices* A *and* B *and integer number* $0 < r \leq m < n$. *Consider the GSVD of* A *and* B:

$$A = X^{-t} D_A U_A^t$$
$$B = X^{-t} D_B U_B^t$$

Let (α_i/β_i) *be the generalized singular values of* A *and* B, *arranged in non-increasing order. Denote by* x_i *the i-th column vector of* X. *Then :*

$\mathrm{MmR}[A,B,r] = \alpha_r/\beta_r$ *The corresponding subspace is generated by the first r column vectors* $[x_1 \ldots x_r]$ *of* X.

$\mathrm{mMR}[A,B,r] = \alpha_{m-r+1}/\beta_{m-r+1}$ *The corresponding subspace is generated by the last r column vectors* $[x_{m-r+1} \ldots x_m]$ *of* X.

Proof: Define the linear transformation $T = D_B^{-1}X^t$. Use the invariance theorem 2 and theorem 7, relating the signal-to-signal ratio to the oriented energy, in order to find that:

$$\mathrm{MmR}[A,B,r] = \max_{(T^{-1})^t Q^r \subset \mathcal{R}^m} \min_{q' \in (T^{-1})^t Q^r} E_{q'}[D_B^{-1}D_A U_A^t, U_B^t]$$

and

$$\mathrm{MmR}[A,B,r] = \min_{(T^{-1})^t Q^r \subset \mathcal{R}^m} \min_{q' \in (T^{-1})^t Q^r} E_{q'}[D_B^{-1}D_A U_A^t, U_B^t]$$

From the Courant-Fischer minimax and maximin characterization properties (Corollary 1), the result follows after back transformation with $T^{-1} = X^{-t}D_B$. □

4.4.3 Numerical considerations

Given an $m \times n$ matrix A with $n \geq m$ and a $p \times n$ matrix B, it can be proved that there exists a non-singular $n \times n$ matrix X such that both $X(AA^t)X^t$ and $X(BB^t)X^t$ are diagonal [7, p. 314]. The great value of the GSVD is that these diagonalizations can be achieved without forming the Grammians AA^t and BB^t, hence avoiding the numerically dangerous implicit squaring, which can lead to a loss of accuracy caused by the limited machine precision. As an example, consider the following GSVD:

$$A = \begin{bmatrix} 1 & \mu & 0 \\ 1 & 0 & \mu \end{bmatrix} \qquad B = \begin{bmatrix} \epsilon & 1 & -1 \\ \epsilon & 1 & 1 \end{bmatrix}$$

with $\mu \geq \epsilon_m \geq \mu^2$ and $\epsilon \geq \epsilon_m \geq \epsilon^2$ where ϵ_m is the machine precision. By some straightforward calculations, one can show that the generalized singular values of the matrix pair $[A,B]$ are given by :

$$\begin{array}{ll} \alpha_1 = (\frac{2+\mu^2}{4+\mu^2+2\epsilon^2})^{1/2} & \beta_1 = (\frac{2(1+\epsilon^2)}{4+\mu^2+2\epsilon^2})^{1/2} \\ \alpha_2 = (\frac{\mu^2}{2+\mu^2})^{1/2} & \beta_2 = (\frac{2}{2+\mu^2})^{1/2} \end{array}$$

However, when the Grammian products $A.A^t$ and $B.B^t$ are explicitely computed as a first step in the computation of the GSVD, already a lot of information is lost, and the result will even depend upon the ordering of the terms in the computation of the inner product:

$$A \cdot A^t = \begin{bmatrix} 1+\mu^2 & 1 \\ 1 & 1+\mu^2 \end{bmatrix} \overset{\approx}{finite\ machine\ precision} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

For the inner products in BB^t, one has to compute the inner product

$$\begin{array}{lll} [\epsilon \ 1 \ 1].[\epsilon \ 1 \ -1]^t & = (\epsilon^2+1)-1 \approx 0 & (case1) \\ & = \epsilon^2 + (1-1) \approx \epsilon^2 & (case2) \end{array}$$

Then :

$$case\ 1 : BB^t \approx \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad case\ 2 : BB^t \approx \begin{bmatrix} 2 & \epsilon^2 \\ \epsilon^2 & 2 \end{bmatrix}$$

This leads to the generalized singular values, both for case 1 and case 2:

$$\begin{array}{ll} \alpha_1 = \sqrt{2}/2 & \beta_1 = \sqrt{2}/2 \\ \alpha_2 = 0 & \beta_2 = 1 \end{array}$$

As observed in [7], the proof of the GSVD theorem, which makes use of the C-S decomposition, is constructive since it can be shown how to stably compute the C-S decomposition [19]. Another possible implementation is considered in [10].

The GSVD provides a structured algorithm to analyse the oriented signal-to-signal distribution of two vector sequences. It computes directly the several extrema (the generalized singular values) but also the corresponding extremal directions (the columns of the matrix X). Of course, any strategy that first orthonormalises the vector sequence $\{b_k\}$ via a linear transformation T and then considers the oriented energy distribution of the matrix TA will work for well conditioned vectorsequences $\{b_k\}$. As an example, consider first the R-Q factorization of the matrix B, define $T = R^{-1}$ and then study the oriented energy distribution of $R^{-1}A$ by computing its singular value decomposition. This requires a R-Q factorization, a matrix inversion and a singular value decomposition. Moreover, several additional matrix matrix multiplications are necessary in the backsubstitution. The big advantage of the GSVD is that it replaces these three algorithms and matrix multiplications by one, which is numerically reliable and can more easily handle the near-singularity case, where B is 'almost' rankdeficient.

4.5 Applications and examples

In this section, several examples are presented in order to illustrate the practical significance of the above derived framework of oriented energy and oriented signal-to-signal ratio. In section 4.5.1., the oriented energy distribution and the technique of prewithening are considered. In section 4.5.2., the concept of total linear least squares is formalized using the complexity/misfit approximative modeling framework of [22] and the oriented energy concept. In section 4.5.3., it is shown how a lot of factor analysis like modeling problems lend themselves very naturally to a formulation in terms of oriented signal-to-signal ratios.

4.5.1 The oriented energy distribution of stochastic vector sequences.

Consider a stochastic process, consisting of a m-vector sequence $\{b_k\}, k = 1, \ldots, n$. The process is assumed to be ergodic and the elements b_{ij} of the associated matrix B are independenlty distributed. Under these assumptions, the Grammian BB^t/n is an estimate for the second-order joint moment matrix of the vector stochastic process and for increasing n, it will tend to become a symmetric Toeplitz matrix: By applying theorem 5, one can replace the actual time realization contained in the matrix B, by a sequence of m m-vectors having the same oriented energy distribution. As a special case, assume that the components of the vector process are independently and identically distributed with first and second order moments m_1 and m_2^2. Then the covariance matrix will have, at least *asymptotically*, the

following specific symmetric Toeplitz structure :

$$\underset{m \times m}{BB^t} \approx n \begin{bmatrix} m_2^2 & m_1^2 & m_1^2 & \dots & m_1^2 \\ m_1^2 & m_2^2 & m_1^2 & \dots & m_1^2 \\ \dots & \dots & \dots & \dots & \dots \\ m_1^2 & m_1^2 & m_1^2 & \dots & m_2^2 \end{bmatrix}$$

The nice fact about this matrix is that its eigenstructure is straightforward to compute: There are $m-1$ eigenvalues $n(m_2^2 - m_1^2)$. There is one largest eigenvalue $\gamma_1 = n[(m-1)m_1^2 + m_2^2]$ with corresponding eigenvector $u_1 = 1/\sqrt{m}(1\ 11)^t$. Via the equivalence theorem 5, a compact presentation of a stochastic sequence with the above mentioned characteristics, is the vector-sequence:

$$[u_1\sigma_1 \quad u_2\sigma_2 u_m\sigma_m]$$

where $u_1 = 1/\sqrt{m}(1\ 11)^t$ and the vectors $u_j, j = 2,\dots,m$ are an arbitrary orthonormal set of vectors orthogonal to u_1. Moreover, $\sigma_1 = \sqrt{n[(m-1)m_1^2 + m_2^2]}$ and $\sigma_j = \sqrt{n(m_2^2 - m_1^2)}$, $j = 2,\dots,m$. In words, the oriented energy distribution of a stochastic sequence which is ergodic and which has identically independently distributed elemenents, is asymptotically isotropic except for one principal direction along the direction (1 1 ... 1) in the first orthant in which the energy is larger. Clearly, it is in this direction that this stochastic disturbance sequence will have the largest corrupting influence on any 'exact' signal sequence. A special case is of course obtained if the first moment $m_1 = 0$. In this case, the second order joint moment matrix reduces to a diagonal matrix and the oriented energy distribution is isotropic. This is the case in a lot of engineering applications, where it is assumed that the noisy vector sequence consists of independent identically normally distributed zero mean random variables. This assumption is quite natural if the central limit theorem is invoked to argue that the macroscopic effect of noise is due to the superposition of a lot of independent microscopic causes, and if all offsets are eliminated a priori.

Now consider the situation of an 'exact' m-vector signal contained in the $m \times n$ matrix A, which is of rank $r < m < n$ and assume that only the $m \times n$ matrix $C = A + B$ is observable, where B is some stochastic noisy sequence, with a priori known (for instance from experiments) second order statistics, that are summarized in an equivalent m-vector sequence of m vectors, that are the columns of the matrix $U_b.S_b$, where U_b is $m \times m$ orthonormal and S_b is $m \times m$ positive definite diagonal. Moreover, assume that the row spaces of A and B are orthogonal (which under mild conditions is the case for large overdetermination n/m). Then, it is not difficult to see that the generalized singular value decomposition of the matrix pair $[A, U_bS_b]$ or equivalently, via lemma 5 the singular value decomposition of the matrix $S_b^{-1}U_b^tA$ will provide the subspaces of maximal minimal signal-to-signal ratio, which are the subspaces in which the vector sequence A can best be distinguished (in the average) from the perturbing influence of the fuzzy sequence B, in terms of spatial oriented energy distribution. The generalized singular values are appropriate tools to make meaningful decisions for the correct dimension, because the transformation $S_b^{-1}U_b^t$ has (under mild conditions) caused a noise threshold in the singular values equal to 1. It can be shown in a straightforward way that this technique is nothing else than the 'classical' prewhitening technique, in which the data are transformed via a so-called Mahalanobis transformation [12] in such a way that the noise covariance matrix equals the identity matrix. Hereto, assume that Σ_c is the measurement

sample covariance matrix and that Σ_b is the noise covariance matrix. Then, the problem to be solved is the generalized symmetric eigenvalue problem

$$\det(\Sigma_c - \gamma.\Sigma_b) = 0$$

The whitening transformation then consists in converting this expression into the eigenvalue problem, under the assumption that Σ_b is non-singular :

$$\det(\Sigma_b^{-\frac{1}{2}}.\Sigma_c.\Sigma_b^{-\frac{1}{2}} - \gamma I) = 0$$

where $\Sigma_b^{\frac{1}{2}}$ is any symmetric square root of Σ_b. Via lemma 5 this establishes of course the link with the oriented signal-to-signal ratio framework and the generalized singular value decomposition. In a certain sense, these are even more general, since even (almost) rank deficient Σ_b can be allowed without numerical complications. Another example is the use of GSVD in prewhitening the data for the estimation of parameters in a general Gauss-Markov linear model [8]. We shall come back to this prewhitening technique in our discussion on the 'lever' theorem in chapter 5.

4.5.2 Total Linear Least Squares

In [22] a conceptual framework is developed in which the modeling problem is translated into an approximation context based upon the paradigm of low complexity and high accuracy models. The key concepts in this approach are the complexity of a model and the misfit between a model and the observations. Approximate modeling then consists of implementing the principle that either the desired optimal model is the least complex one in a given model class which approximates the observed data up to a preassigned tolerated misfit, or that it is the most accurate model within a preassigned tolerated complexity level. A particularly simple example is the total linear least squares approach [7,18] which consists of fitting a linear subspace to a finite number of points. Consider an $m \times n$ matrix A ($n \gg m$) containing n measurements on a m-vector signal. Denote by a_i its i-th column. Let Q^r be a r-dimensional subspace of $\mathcal{R}^m$ then, the complexity is defined as :

$$c(r) : Q^r \rightarrow C = [0,1] : c(r) = \dim(Q^r)/m = r/m$$

Suppose that we are looking for linear relations among the m measurement channels of the form $x^t.A = 0$ Define the error between the data and the law $x^t.A = 0$ as:

$$d(A,x) = \frac{\sqrt{[\frac{1}{n}\sum_{i=1}^{n}(x^t.a_i)^2]}}{\|x\|}$$

and the misfit associated with the r-dimensional subspace Q^r as:

$$\epsilon(A,Q^r) = \max_{x \perp Q^r} d(A,x)$$

Then, we have the following theorem [22] :

Theorem 8 *Let $\frac{1}{\sqrt{n}}A = U.\Sigma.V^t$ be the SVD of the $m \times n$ matrix A of rank s $(s \leq m < n)$ with singular values $\sigma_1 \geq\geq \sigma_s > 0$ and left singular vectors $u_i, i = 1,\ldots,m$. The unique optimal approximate model Q^r with complexity $c(Q^r) = \frac{r}{m}$ and misfit $\epsilon(A,Q^r) = \sigma_{r+1}$ is an r-dimensional subspace where :*

- *If* c_{adm} *is the maximal admissible complexity, then :*
  - *if* $\text{int}[m.c_{\text{adm}}] = 0$, $r = 0$ *and* $Q^r = 0$.
  - *if* $\text{int}[m.c_{\text{adm}}] \geq s, r = s, Q^r = \text{span}_{\text{col}}[A]$
  - *if* $\sigma_k > \sigma_{\text{int}[m.c_{\text{adm}}]+1}, r = k, Q^r = S_U^k$
- *If* ϵ_{tol} *is the maximal tolerated misfit, then :*
  - *if* $\epsilon_{\text{tol}} \geq \sigma_1, r = 0$ *and* $Q^r = 0$
  - *if* $\epsilon_{\text{tol}} < \sigma_s, r = s, Q^r = \text{span}_{\text{col}}[A]$
  - *if* $\sigma_k > \epsilon_{\text{tol}} \geq \sigma_{k+1}, r = k$ *and* $Q^r = S_U^k$

Proof : see [22] □

In this framework of approximate modeling, the appropriate rank r is determined from either an a priori fixed admissible complexity or a maximal tolerable misfit. Observe that these concepts readily reduce to the framework of oriented energy in that :

$$\begin{aligned} [d(A,x)]^2 &= E_x[A] \\ [\epsilon(A,Q^r)]^2 &= \min_{Q^r} \max_{q \in Q^r} E_q[A] \end{aligned}$$

where $p = m - r$. Hence the misfit is nothing else than a subspace of minimal maximal oriented energy . The authors conjecture that also the dynamical case for the identification of state space models developed in [22] can be translated in the oriented signal-to-signal ratio framework.

4.5.3 Factor-analysis like subspace methods

A lot of identification and modeling problems can be formulated in a factor-analysis like framework:

Given noisy measurements of an m-vector process which can be modeled as

$$\underset{m \times 1}{x(t)} = \underset{m \times r}{Q(\theta)} \cdot \underset{r \times 1}{s(t)} + \underset{m \times 1}{n(t)} \qquad r < m$$

where $Q(\theta)$ *contains the* r *linear independent so-called factor loadings,* $s(t)$ *are the source signals and* $n(t)$ *are the corrupting noise signals. The subspace generated by the columns of* $Q(\theta)$ *is called the factor loading subspace which is in one sense or another parametrized by unknown parameters* θ*. The task is then to estimate the parameters* θ*, given a priori knowledge of the second order statistics of the measurement noise and varying degrees of knowledge concerning the sensor response function.*

Factor analysis is not that well reputed in the statistical community. The reason is that the loadings are undetermined and that only the subspace that they generate is well determined under appropriate conditions on the noise. Classically, attempts were undertaken to resolve this problem by fixing so-called structural zeroes, which corresponds to fixing a coordinate system in the loading subspace. The determination of the remaining non-zero components

however frequently leads to an ill-conditioned parameter estimation problem, which explains (at least heuristically) the bad reputation of factor analysis. However, in modern applications (so-called subspace methods), the indeterminancy is immaterial because the properties (the parameters θ) are in a one-to-one correspondence with the generated subspace : Any set of vectors that generates a basis for this subspace, will allow to determine the parameters θ : a precise choice of a basis in that space is not important.

We shall now present 3 applications : high resolution spectral analysis and sensor array processing techniques, the separation of fetal ECG and maternal ECG and realization of dynamical systems.

High resolution sensor array processing.

As is derived in [12,13] a factor analysis model can be used to model the arrival of narrowband signals impinging on an array consisting of sensor pairs that are separated by a fixed distance δ. With m sensor pairs and when the signals from sensor pair i are $x_i(t)$ and $y_i(t)$, the following model is appropriate when $d < m$ narrowband sources $s_j(t)$ are present [12,13] :

$$\underset{m\times 1}{x(t)} = \underset{m\times d}{A} \cdot \underset{d\times 1}{s(t)} + \underset{m\times 1}{n_x(t)}$$

$$\underset{m\times 1}{y(t)} = \underset{m\times d}{A} \cdot \underset{d\times d}{\Omega} \cdot \underset{d\times 1}{s(t)} + \underset{m\times 1}{n_y(t)}$$

The d-dimensional columnspace of A is generated by the so-called steering vectors while the matrix Ω is a complex diagonal shift matrix that contains information on the phase shifts between the pairs of sensors from which the direction of arrival can be estimated. Schematically, this can be achieved as follows : Define $z(t)$ as

$$z(t) = \begin{bmatrix} A \\ A.\Omega \end{bmatrix} .s(t) + n_z(t)$$

Now store n consecutive samples $z(i), i = 1, \ldots, n$ in a $2m \times n$ matrix Z, n consecutive samples $s(i)$ in a $d \times n$ matrix S and the noise in a $2m \times n$ matrix N

$$Z = \begin{bmatrix} A \\ A.\Omega \end{bmatrix} .S + N$$

If the second order noise statistics of $n_z(t)$ are known, one can obtain an equivalent m-vector sequence of m vectors $U_m.\Sigma_m$ from the SVD of the sample covariance $N.N^t$. Using the insights of [12,13], one can then prove that the best approximation for the loading subspace follows from the maximal minimal signal-to-signal ratio and corresponding subspace that can be estimated from the GSVD of the matrix pair $[Z, U_m\Sigma_m]$. The number of sources can be estimated from the generalized singular values via rank determination tests [1,2,11,12,14] while the imposed shift-structure of the corresponding subspace of maximal minimal signal-to-signal ratio can be exploited to determine the angles of arrival of the signals, hence the location of the sources.

Realization of dynamical systems.

Realization theory of dynamical systems reduces to the determination of the matrices of a state space model for a linear finite dimensional systems starting from (possibly noisy) measurements of its Markov parameters. As it is well known, the relation between the state space model with states x_k, inputs u_k and outputs y_k

$$\begin{array}{ccccccccc} x_{k+1} & = & A & . & x_k & + & B & . & u_k \\ n \times 1 & & n \times n & & n \times 1 & & n \times m & & \times 1 \\ y_k & = & C & . & x_k & & & & \\ l \times 1 & & l \times n & & n \times 1 & & & & \end{array}$$

and its Markov - parameters, is $H_k = C.A^{k-1}.B$. Classically, the realization of the model from the H_k proceeds by the following algorithm (there are several variants [4] and references therein) :

- Construct a sufficiently large block Hankel matrix with the H_k.
- Determine its rank via SVD. The rank decision results in an estimate of the minimal state dimension n
- The matrices A, B , C are then realized up to a similarity transformation:
  - B and C can be read off from certain block- and column rows in the SVD
  - The matrix A follows from the shift invariant structure of the column space of the block Hankel matrix.

When the measurements are noisy, the following novel realization framework, based upon the oriented signal-to-signal ratio concept, is appropriate to determine an estimate of the matrix A up to a similarity transformation.

Construct a sufficiently large block Hankel matrix H and partition it in two blocks H_1 and H_2

$$H = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix}$$

The reader may now wish to verify that the best approximation for the appropriate shift invariant subspace follows from the GSVD of the matrix pair $[H_1, H_2]$. The generalized singular values allow to estimate the minimal order n. The corresponding subspace of maximal minimal signal to signal ratio contains information on the minimal poles of the system via an imposed shift structure [4].

The separation of fetal ECG from maternal ECG.

In this biomedical application, the cutaneous measurements (typically 6 to 9 channels) are contained in a vector $m(t)$ which is modeled as :

$$m(t) = T.s(t) + n(t)$$

where the signal $s(t)$ corresponds to the sources (electrical activity of the heart of mother and fetus), $n(t)$ is the noise, and the columns of T are the so-called lead vectors (typically 2 for

the fetus, 3 for the mother)[17].

Three methods can be analysed in our oriented signal-to-signal ratio distribution framework.

1. **Singular Value Decomposition:** Under certain conditions (orthogonality of source signals, certain statistical conditions on the noise, placement of electrodes), it can be verified that one singular value decomposition suffices to determine the factor loading subspace generated by the lead vectors of the fetus. This allows to project the measurements into this subspace, hence eliminating almost completely the maternal ECG. The conceptual framework is provided by the oriented energy distribution of the vector signal $m(t)$ [16] [17].

2. **Generalized Singular Value Decomposition:** In [20], the same problem is solved using an approach that can be interpreted in the oriented signal-to-signal framework. By visual inspection, two matrices A and B are constructed from the measurements $m(t)$. A window in time is selected visually so that A contains only fetal ECG complexes. Another window is chosen so that B contains only maternal ECG complexes. The loading subspace generated by the fetal lead vectors is then nothing else than the subspace of maximal minimal oriented signal-to-signal ratio. This subspace and its dimension can be computed from the GSVD of the matrix pair $[A, B]$.

3. **Generalized Singular Value Decomposition, method 2:** While in the method described above, the matrices A and B are constructed by visual inspection, one could also envisage to implement the following strategy. Construct the matrix A containing only measurements of the mother heart. This can be done by considering only measurements that are derived from electrodes nearby the mother heart. The matrix B contains measurements of both fetal and maternal ECG. Recall that the oriented signal-to-signal ratio concepts allow to extract from a certain vector sequence that information which does not belong to another sequence (maximal minimal and minimal maximal signal to signal ratio). Hence, from the generalized singular value decomposition of the matrix pair (A, B) one could derive everything what does belong to the matrix B but not to the matrix A. Obviously, this is the information concerning the fetal ECG:

4.6 Conclusions

Two important concepts have been defined : The oriented energy distribution of a vector sequence and the oriented signal-to-signal ratio of two vector sequences. For the former, the singular value decomposition is the appropriate quantification tool while for the latter the general singular value decomposition applies. Both allow a numerically robust implementation. Conceptually important properties have been analysed. With some clarifying examples, the practical significance of the framework in the formalization of so-called subspace methods has been demonstrated. These notions of oriented energy and oriented signal-to-signal ratio distribution will return frequently in the sequel of this dissertation.

Bibliography

[1] Akaike H. *Information theory and an extension of the maximum likelihood principle.* In Proc. 2nd Int.Symp.Inform.Theory, pp.267-281, 1973.

[2] Anderson T.W. *Asymptotic theory for principal component analysis.* Ann. Math. Statist., 34:122-148, 1963.

[3] Chatfield C. , Collins A. *Introduction to multivariate analysis.* Chapman and Hall Ltd., London 1980.

[4] De Moor B., Vandewalle J. *Non-conventional matrix calculus in the analysis of rank deficient Hankel matrices of finite dimensions.* Systems and Control Letters, 9, pp.401-410, 1987

[5] Vandewalle J., De Moor B. *A wide variety of applications of singular value decomposition in identification and signal processing.* Proc. of the Workshop on Singular Value Decomposition and Signal Processing, September 21-23, 1987, Les Houches France. Published by Elsevier Science Publishers B.V., North-Holland (Eds. Ed Deprettere) (in press).

[6] De Moor B., Vandewalle J., Staar J. *Oriented Energy and Oriented Signal-to-Signal Ratio Concepts in the Analysis of Vector Sequences and Time Series.* Proc. Workshop on SVD and Signal Processing, September 21-23, 1987, Les Houches, France. Published by Elsevier Science Publishers B.V., North-Holland (Eds. Ed Deprettere) (in press).

[7] Golub G., Van Loan C. *Matrix Computations.* Johns Hopkins University Press, North Oxford Academic, 1983.

[8] Hammarling S. *The Numerical Solution of the General Gauss-Markov Linear Model.* NAG-Technical report TR2/85, October 1985.

[9] Jolliffe I.T. *Principal Component Analysis.* Springer Series in Statistics. Springer Verlag New York, 1986.

[10] Paige C.C. *Computing the generalized singular value decomposition.* SIAM J. Sci. Statist. Comput., 7, pp. 1126-1146, 1986.

[11] Rissanen J. *Modeling by shortest data description.* Automatica, 14:465-471, 1978.

[12] Roy R. *ESPRIT: Estimation of Signal Parameters via Rotational Invariant Techniques.* Ph.D. Dissertation, Stanford University, August 1987.

[13] Roy R., Paulraj A., Kailath T. *Estimation of Signal Parameters via Rotational Invariance Techniques - ESPRIT.* In Proc. IEEE ICASSP pp.2495-2498, Tokyo, Japan, 1986.

[14] Scharf L.L., Tufts D.W. *Rank reduction for modeling stationary signals.* IEEE Trans. on Acoustics, Speech and Signal Processing, Vol.ASSP-35,no.3, march 1987.

[15] Staar Jan. *Concepts for reliable modeling of linear systems with application to on-line identification of multivariable state space descriptions.* PhD. thesis, Esat - Katholieke Universiteit Leuven, 1982.

[16] Vanderschoot J., Vandewalle J., Janssen J., Sansen W., Vantrappen G. *Extraction of weak bioelectrical signals by means of SVD.* Proc. of 6th Int. Conf. on Analysis & Optimization of Systems, Nice, June 1984, Springer Verlag, Berlin, pp.434-448.

[17] Vanderschoot J., Callaerts D., Sansen W., Vandewalle J., Vantrappen G., Janssens J. *Two methods for optimal MECG elimination and FECG detection from skin electrode signals* IEEE Trans. on Biomedical Engineering, Vol. BME-34, No. 3, pp. 233-243, March 1987.

[18] Van Huffel S., Vandewalle J. *The total least squares technique: Computation, properties and applications.* Proc. of the workshop on SVD and Signal Processing, September 21-23 1987, Les Houches, France. Published by Elsevier Science Publishers B.V., North-Holland. (Eds. Ed Deprettere) (in press).

[19] Van Loan C. *Computing the C-S and the Generalized Singular Value Decompositions.* Numerische Mathematik, 46:479-491,1985.

[20] Van Oosterom A. Alsters J. *Removing the maternel component in the Fetal ECG using singular value decomposition.* Electrocardiography '83, I. Rattkay-Nedecky and P.MacFarlane, Eds. Amsterdam, The Netherlands. Excerpta Medica, 1984, pp. 171-176.

[21] Wilkinson J. *The Algebraic Eigenvalue Problem.* Oxford University Press, 1965.

[22] Willems J.C. *From time series to linear systems. Part I - II - III.* Automatica. Part I: vol.22, no.5,pp. 561-580, 1986. Part II:vol.22, no.6, pp.675-694, 1986. Part III: Vol.23, no.1, pp.87-115, 1987.