Determination of Design of Optimal Actuator Location
Based on Control Energy

by

Afrouz Farshad Mehr

Thesis submitted for the degree of
Doctor of Philosophy
in
Control Engineering

January 2018
Systems and Control Research Centre
School of Engineering and Mathematical Sciences
City University of London, Northampton Square, London, EC1V 0HB
Copyright Statement

The copyright of this thesis rests with author. No quotation from it should be published without author’s prior consent and information derived from it should be acknowledged and cited properly.
I declare that this thesis was composed by myself and that this work has not been submitted for any other degree or professional qualification.
Abstract

The thesis deals with the selection of the sets of inputs and outputs using the energy properties of the controllability and observability of a system and aims to define input and output structures which require minimization of the energy for control and state reconstruction. Such a study explores the energy dimension of the properties of controllability and observability, develops computations for the controllability and observability Gramians for stable and unstable systems and examines measures of the degree of controllability and observability properties using SVD (Singular Value Decomposition) of Gramians to compute the maximal and minimal energy requirements. These characterize the relative degree of controllability and observability under conditions where the available energy is constrained. The notion of energy surfaces in the state space is introduced and this enables the characterization of restricted notions of controllability and observability when the available energy is bounded. The maximal and minimal energy requirements for different input vectors is demonstrated and this provides the basis for the development of strategies and methodologies for selection of systems of inputs and outputs to minimize the energy required for control, respectively state reconstruction. These results enable the development of input, output structure selection methodology using a novel optimization method. This thesis contributes in the further development of the area of systems, or global instrumentation, developed so far based on the assignment of structural characteristics by incorporating the role of energy requirements. The research provides energy based tools for the selection of input and outputs schemes with a main criterion the minimization of the energy required for control and observation and thus provide an alternative approach based on quantitative system properties in characterizing control and state observation as functions of given sets of inputs and output sets. The methodologies developed may be used as design tools where apart from energy requirements other design criteria may be also incorporated for the selection of inputs and outputs. The methodology that is used is based on linear systems theory and tools from numerical linear algebra. The solution to the problems considered here is an integral part of the effort to develop an integrated approach to control and global process instrumentation.
Acknowledgements

I would like to express my sincere gratitude to my advisor Professor Nicos Karcanias for his guidance, advice, help, support, and patience throughout this work. Prof Karcanias has left a deep impression on me, because of his professionalism, pursuit of excellence and dedication to all his work over the years. What I learned from him will always remain with me to help me throughout my professional life. Without his support, this thesis would never have emerged. I am very thankful to him for making this journey possible.

I would like to gratefully acknowledge the enthusiastic help and support of professor George Halikias during this work.

Sincere thanks go to the members of my committee for their patience and support in overcoming numerous obstacles I have been facing through my research.

I am also grateful to the following university staffs: Dr. Youyou Yan, Mrs. Nathalie Chatelain and Dr. Weiping Wu for their helps.

My deepest gratitude and much love goes to my Maman and Baba, Afsaneh and Ebrahim, my brothers, Omid and Arash and my beloved sister Afshan. You have always given me more support and love than I could have asked for, while I was thousands of miles away when you were going through difficult times. I am indebted, forever grateful, and so lucky to have you all!

Finally, there are my friends. We were not only able to support each other by deliberating over our problems and findings, but also happily by talking about things other than just our researches.
خداً را با سی شاکر که از روی کرم پدر و مادری فداکار نصیب ساخته‌تا در سایه
درخت پر بار وجود‌شان پیاسایم و از ریشه آنها شاخ و برگ گیرم و از سایه وجود‌شان
در راه کسب علم و دانش تلاش نمایم.

والدینی که بودنشان تاج افتخاری است بر سرم و نامشان دلیلی است بر بودن چرا
که این دو وجود پس از پروردگار ماهی هستی ام بوده اند دستم را گرفتن و راه رفت‌ن
را در این وادی زندگی پر از فرآ و نشیب آمختند.

تقدیم به پدر و مادر عزیز و مهربان
که در سختی‌ها و دشواری‌های زندگی همواره یاوری فداکار
و پشتیبانی محکم و مطمئن برام بوده‌اند.
## Table of Contents

**List of Tables** ........................................................................................................................................... ix
**List of Figures** .......................................................................................................................................... x

### Chapter 1: Introduction ............................................................................................................................. 1

### Chapter 2: Literature Review

- 2.1 Introduction ............................................................................................................................................. 6
- 2.2 Input/Output Structure Selection .......................................................................................................... 6
- 2.3 Problem Statement: ............................................................................................................................... 9
- 2.4 Different Cost Functions: ....................................................................................................................... 10
- 2.5 Summary .................................................................................................................................................. 24

### Chapter 3: Overview of Controllability Gramian and its Computation

- 3.1 Introduction ........................................................................................................................................... 25
- 3.2 Reachability and Reachability Gramian ............................................................................................... 26
- 3.3 Controllability Gramian and its Computation ....................................................................................... 28
  - 3.3.1 A Novel Approach for Calculation of Controllability Gramian ...................................................... 30
  - 3.3.2 Controllability Gramian Assignment with Respect to Control Variables .................................. 35
- 3.4 Summary ................................................................................................................................................. 40

### Chapter 4: Minimum Input Energy

- 4.1 Introduction ........................................................................................................................................... 41
- 4.2 Minimum Input Energy ........................................................................................................................ 42
- 4.3 Influence of Different Terms on the Value of Minimum Input Energy ............................................. 43
  - 4.3.1 Terminal Time $t_f$ ......................................................................................................................... 44
    - 4.3.1.1 Stable System ............................................................................................................................ 45
    - 4.3.1.2 Unstable System ....................................................................................................................... 49
Chapter 5: Controllability, Energy for Unstable Systems and Measures of Degree of Controllability

5.1 Introduction...........................................................................................................................................61
5.2 Minimum Energy for Unstable Systems ..............................................................................................62
  5.2.1 Decomposition of Controllability Gramian in LTI Systems with Stable and Anti-Stable Modes ..........................................................................................................................63
  5.2.2 Minimum Input Energy for LTI System with Anti-Stable Part ......................................................68
5.3 Energy Levels for LTI Systems with Anti-Stable Modes .................................................................72
5.4 The Energy and Relative Degree of Controllability ........................................................................80
5.5 Degree of Controllability ..................................................................................................................87
5.6 Degree of Disturbance Rejection ......................................................................................................91
5.7 Summary...............................................................................................................................................100

Chapter 6: Controllability Gramian and Energy Calculations of the Stable LTI System in Canonical Forms

6.1 Introduction...........................................................................................................................................101
6.2 Controllability Gramian Computation in Controller Canonical Form .............................................102
6.3 Inverse of Controllability Gramian and Energy Calculation in Controller Canonical Form 116
6.4 Controllability Gramian Computation in Diagonal Canonical Form ..............................................128
6.5 Summary............................................................................................................................................131

Chapter 7: Optimal Selection of Input Structure, Minimizing the Average of Input Energy

7.1 Introduction...........................................................................................................................................132
7.2 Optimal Selection of Input to Minimize the Average of Minimum Input Energy ................................133
  7.2.1 Single Input Matrix Structure Selection Problem .....................................................................135
  7.2.1.a Optimal Input Vector Selection Considering the Finite Terminal Time 136
7.2.1.b Optimal Input Vector Selection for the Case of Stable System .......... 139
7.2.1.c Optimal Input Vector in The Case That $A$ Is a Normal Matrix .......... 141
7.2.1.d Optimal Input Vector in The Case That $A$ Is Stable and Normal .......... 143
7.2.2 Multi-Input Matrix Structure Selection .............................................. 145
7.2.2.a (Part I) Input Matrix Selection Within a Finite Terminal Time .......... 146
7.2.2.a (Part II) Input Matrix Selection Within a Finite Terminal Time .......... 148
7.2.2.a (Part III) Input Matrix Selection Within a Finite Terminal Time .......... 149
7.2.2.b Optimal Multi-Input Selection in a Stable System ............................. 162
7.2.2.c Optimal Input Matrix Selection with Condition On 2Norm of Input Matrix $B$ 163
7.3 Summary:........................................................................................................ 166

Chapter 8: Kamineh Algorithm Optimization

8.1 Introduction ........................................................................................................ 168
8.2 Kamineh Optimization Algorithm (KA) ............................................................ 169
  8.2.1 Inspiration ........................................................................................................ 169
  8.2.2 Mathematical Model and Algorithm ............................................................... 175
8.3 Experimental Results and Discussion ................................................................. 181
  8.3.1 Benchmark Functions ..................................................................................... 182
    8.3.1.1 Unimodal Test Functions ......................................................................... 184
    8.3.1.2 Multimodal Test Functions ..................................................................... 190
  8.3.2 Analysis of KA Optimizer ............................................................................... 202
  8.3.3 Optimization Problems Using the KA Algorithm ......................................... 203
8.4 Summary:............................................................................................................ 230

Chapter 9: Conclusions & Future Research

9.1 Summary and Contribution of Thesis ................................................................. 232
9.2 Future Directions ............................................................................................... 234

References ............................................................................................................... 236
Appendix .................................................................................................................... 261
List of Tables:

Table (5.1): Results of Example (5.8) ................................................................. 91
Table (5.2): Distance to uncontrollability for AMB system in Example (5.4) .................. 91
Table (5.3): Distance to uncontrollability for seesaw-inverted pendulum system in Example (5.5) 91
Table (5.4): DODR results of Example (5.10) ......................................................... 98
Table (5.5): DODR results of Example (5.11) ......................................................... 99
Table (7.1): optimum input vector in single input cases, which minimizes the average of $E_{\min}$ 166
Table (7.2): optimum input matrix in multi-input cases, which minimizes the average of $E_{\min}$ 167
Table (8.1): Unimodal test functions ...................................................................... 183
Table (8.2): Multimodal test functions with variable dimensions .............................. 183
Table (8.3): Fixed-dimension multimodal test functions ........................................... 184
Table (8.4): Results of unimodal test functions ....................................................... 186
Table (8.5): P-values produced by Wilcoxon’s test comparing KA versus GA, PSO, FA, GSA and CAB over all runs for unimodal test functions .............................................. 186
Table (8.6): Results of high-dimensional multimodal test functions .......................... 191
Table (8.7): Results of fixed-dimensional multimodal test functions .......................... 192
Table (8.8): P-values produced by Wilcoxon’s test comparing KA versus GA, PSO, FA, GSA and CAB over all runs for high-dimensional multimodal test functions .............................................. 193
Table (8.9): P-values produced by Wilcoxon’s test comparing KA versus GA, PSO, FA, GSA and CAB over all runs for fixed-dimensional multimodal test functions .............................................. 193
Table (8.10): The numerical results of Example (8.1) .............................................. 204
Table (8.11): The numerical results of Example (8.2) .............................................. 206
Table (8.12): The numerical results of Example (8.3) .............................................. 207
Table (8.13): Numerical results for worst-case optimization case of Example (8.4) ....... 211
Table (8.14): Numerical results for average performance optimization case of Example (8.4) 212
Table (8.15): Numerical results for Example (8.6) .................................................. 222
Table (8.16): The eigenvalues of the controllability Gramians in Example (8.8) .......... 228
List of Figures:

Figure (4.1): Controllability and the energy required to transfer states from the origin to terminal states $x_f, \|x_f\| = R$ ..............................................................42

Figure (4.2): Mass-spring-damper system ..........................................................46

Figure (4.3): $E_{\text{min}}$ as a function of variable terminal time where $t_f \in [0.5, 10]$ ..........47

Figure (4.4): $E_{\text{min}}$ as a function of terminal time variable, $0 < t_f \leq \infty$ .......................48

Figure (4.5): The schematic picture of Furuta pendulum .................................................49

Figure (4.6): $E_{\text{min}}$ as a function of final time variable, $0 < t_f < \infty$ ..............................51

Figure (4.7): Energy levels considering $E_{\text{min}}^{\text{min}}$ ....................................................55

Figure (4.8): Energy levels for furuta pendulum considering $E_{\text{min}}^{\text{min}}$ .............................55

Figure (4.9): Energy levels considering $E_{\text{min}}^{\text{max}}$ ..................................................57

Figure (4.10): Energy levels for furuta pendulum considering $E_{\text{min}}^{\text{max}}$ ...........................57

Figure (4.11): Minimum input energy as a function of the variables $t_f, x_f$ .......................58

Figure (4.12): Reachability area changes based on different terminal times for 2 dimensional Example model in (4.16) .................................................................59

Figure (5.1): Minimum energy levels with respect to the norm of initial and terminal states (Example (5.1)) ............................................................................................................73

Figure (5.2): Minimum energy levels of Example (5.2) .....................................................74

Figure (5.3): Max and Min energy levels’ directions in Example (5.2) ............................75

Figure (5.4): Minimum energy levels of Example (5.3) .....................................................76

Figure (5.5): Max and Min energy levels’ directions in Example (5.3) ............................77

Figure (5.6): Pure energy values taken by the unstable subsystem when $R_0 = 1$ and $R_f = 0$ in Example (5.2) ............................................................................................................................78

Figure (5.7): Minimum energy levels when the system steers from the worst initial direction $R_0 (= 100, 1000)$ and moves toward the final states located on the sphere with the radius $R_f = 1$ in different directions in Example (5.2) ................................................78

Figure (5.8): Pure energy values taken by the stable subsystem when $R_f = 1$ and $R_0 = 0$ in Example (5.3) ............................................................................................................................79
Figure (5.9): Minimum energy levels when the system moves from various initial directions $R_{i} = 1$ toward the final states located on the sphere with the radius $R_{f} (= 100,1000)$ in the worst direction in Example (5.3) .............................................................................................................. 79
Figure (5.10): Schematic of the AMB system .................................................................................................................. 81
Figure (5.11): The schematic of seesaw-pendulum process ............................................................................................... 85
Figure (5.12): Two-CSTR Process ........................................................................................................................................ 93
Figure (5.13): Model of bus suspension system (matched disturbance) .............................................................................. 98
Figure (5.14): Model of bus suspension system (unmatched disturbance) .......................................................................... 99
Figure (7.1): trace($W_{c}(0,1)$) with respect to different values of $b$ ............................................................................. 139
Figure (7.2): The optimal input vectors and the spectral subspace ................................................................................... 150
Figure (7.3): $b_{2}$ must be located on the boundary of the cone with centre $b_{1}$ ................................................................. 154
Figure (7.4): Vector $b_{j}$ lies on the intersection of the two cones centered at vectors $b_{1}$ and $b_{2}$ ... 154
Figure (7.5): trace($W_{c}(0,1)$) according to the variable input matrix B in the interval $[-1:0.2:1]^{2}$ ................................................................. 160
Figure (7.6): trace($W_{c}(0,1)$) respect to the variable input matrix B in the interval $[-1:0.1:1]^{2}$ ......................................................... 160
Figure (7.7): trace($W_{c}(0,1)$) based on the angle $\varepsilon$ between $b_{1}, b_{2}$ (top), and trace($W_{c}(0,1)$) as a function of $\alpha_{1}^{2}, \alpha_{2}^{2}$ (bottom) .............................................................................................................................. 161
Figure (8.1): Steady states of logistic equation (8.2) ........................................................................................................ 171
Figure (8.2): Graphical resolution for various values of $r$ ................................................................................................ 173
Figure (8.3): Schematic representation of the bifurcation diagram .................................................................................. 174
Figure (8.4): Schematic representation of the KA optimization method ........................................................................ 175
Figure (8.5): Change of parameters over the course of iterations ..................................................................................... 179
Figure (8.6): The movement of a solution around the destination via change of the parameters over the iterations ................................................................. 180
Figure (8.7): Typical two-dimensional form of unimodal benchmark functions .......................................................... 183
Figure (8.8): Typical two-dimensional form of high-dimensional multimodal benchmark functions ......................................................... 184
Figure (8.9): Typical two-dimensional form of fixed-dimensional multimodal benchmark functions ......................................................... 184
Figure (8.10): Benchmark function F1-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history .................................................................................. 187
Figure (8.11): Benchmark function F2-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................187
Figure (8.12): Benchmark function F3-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................188
Figure (8.13): Benchmark function F4-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................188
Figure (8.14): Benchmark function F5-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................189
Figure (8.15): Benchmark function F6-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................189
Figure (8.16): Benchmark function F7-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................190
Figure (8.17): Benchmark function F8-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................194
Figure (8.18): Benchmark function F9-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................194
Figure (8.19): Benchmark function F10-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................195
Figure (8.20): Benchmark function F11-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................195
Figure (8.21): Benchmark function F12-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................196
Figure (8.22): Benchmark function F13-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................196
Figure (8.23): Benchmark function F14-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................197
Figure (8.24): Benchmark function F15-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................197
Figure (8.25): Benchmark function F16-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................198
Figure (8.26): Benchmark function F17-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................198
Figure (8.27): Benchmark function F18-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................199
Figure (8.28): Benchmark function F19-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history.................................................................199
Figure (8.29): Benchmark function F20-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history..............................................................200
Figure (8.30): Benchmark function F21-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history..............................................................200
Figure (8.31): Benchmark function F22-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history..............................................................201
Figure (8.32): Benchmark function F23-The convergence rate, average fitness of all individuals, trajectory in first dimension and the fitness history..............................................................201
Figure (8.33): Convergence curve of Example (8.1)...204
Figure (8.34): A tension/compression spring.................................................................204
Figure (8.35): A pressure vessel, and its design variables..................................................205
Figure (8.36): Convergence curve of KA in Example (8.2)................................................205
Figure (8.37): Convergence curve of KA in Example (8.3)................................................208
Figure (8.38): Mass-Spring-Damper system......................................................................208
Figure (8.39): Block diagram of the mass-spring-damper system with uncertainty..............209
Figure (8.40): The augmented system $G_{aug}$ ......................................................................209
Figure (8.41): Three-story building model with supplemental passive viscous dampers............213
Figure (8.42): The best objective values of Example (8.5) using KA optimization...............215
Figure (8.43): Top-floor displacement response of the 3-storey building.............................216
Figure (8.44): Floors’ maximum displacement (One damper).............................................216
Figure (8.45): Inter-story displacement response of the floors for one damper in different locations .........................................................................................................................217
Figure (8.46): Acceleration of the floors using two dampers ...............................................217
Figure (8.47): Floors’ maximum displacement (Two dampers)............................................218
Figure (8.48): Inter-story displacement response of the floors for two dampers in different locations .........................................................................................................................218
Figure (8.49): Acceleration of the floors using two dampers ...............................................219
Figure (8.50): Floors’ maximum displacement (Three dampers).........................................219
Figure (8.51): Inter-story displacement response of the floors for three dampers in different locations .........................................................................................................................220
Figure (8.52): Acceleration of the floors using three dampers.............................................220
Figure (8.53): The input and state of the system with the actuator set $s_{actuator} = \{1, 2, 3, 4\}$ ...222
Figure (8.54): The input and state of the system with the actuator set $s_{actuator} = \{1, 2, 3\}$ .....223
Figure (8.55): The input and state of the system with the actuator set $s_{actuator} = \{2, 3, 4\}$ .....223
Figure (8.56): The input and state of the system with the actuator set $s_{\text{actuator}} = \{1, 4\}$ ........223
Figure (8.57): The open-loop system of Example (8.7) ................................................................. 224
Figure (8.58): The convergence curve of Example (8.7) ................................................................ 224
Figure (8.59): The graph of cost function of Example (8.7) ................................................................ 225
Figure (8.60): Local minimum of Example (8.7) ............................................................. 225
Figure (8.61): Control force input needed to control the system when using the optimal actuation .......................................................... 226
Figure (8.62): Control force input needed to control the system when using the input vector $b^i, i = 1, 2, 3, 4$ ........................................................................................................ 226
Figure (8.63): The convergence curve of the KA algorithm in Example (8.8) ....................... 229
Chapter 1: Introduction

The instrumentation of a process, that is the selection of measurement variables (outputs) and actuation variables (inputs) has a “micro” (local), as well as a “macro” (global) aspect. The “micro” role of instrumentation has been well developed (Finkelstein & Grattan, 1994) and deals with the problem of measurement, or implementation of action upon given physical variables; instrumentation theory and practice deals almost exclusively with the latter problems. The “macro” aspects (Karcanias, 1994) of instrumentation stem from that designing an instrumentation scheme for a given process (classification and selection of input and output variables) expresses the attempt of the “observer” (designer) to build bridges with the “internal mechanism” of the process in order to observe it and/or act upon it. What is considered as the final system, on which Control Systems Design is to be performed, is the object obtained by the interaction of the “internal mechanism” and the specification of the overall instrumentation scheme. Difficulties in control of the final system may be assessed in terms of certain structural characteristics of the final system model (MacFarlane & Karcanias, 1976), (Kouvaritakis & MacFarlane, 1976) and the shaping of the degree of the presence of certain system properties, such as controllability and observability. These structural characteristics are formed through various stages, where the design goes through; however, the process of formation of such structural characteristics, as well as the link between their types, values and nature to control problems is not yet well understood. From the systems viewpoint, global instrumentation is seen as a model structure shaping, design stage as far as the characteristic of the final model is concerned. Given
that the structure of the model determines, in a sense, what can be achieved under compensation, global instrumentation is intimately linked to control design. The problems of control design and overall selection of input, output schemes for a process, referred to here as Global Process Instrumentation (GPI) (Karcanias, 1994), are strongly interrelated and this has been specially recognized in the Process Control area, where issues of selection of input, output schemes have been considered within the area of control structure selection (Morari, et al., 1980), (Morari & Stephanopoulos, 1980), (Morari & Stephanopoulos, 1980), (Govind & G.J. Powers, 1982), (Georgiou & Floudas, 1989). So far, however, there has been no systematic attempt to develop a unifying framework for selection of systems of measurement and actuation variables for processes, where the model structure shaping role of Global Instrumentation is the central feature. The overall research activity here reflects the view that instrumentation and control cannot be seen as independent activities, but as interrelated tasks with an integrated methodology.

The problem of selection of input, output schemes for a process, is part of the overall design of the process, which is of cascade nature and has as main stages, the Process Synthesis, Global Process instrumentation (GPI) and Control System design. It has been argued (Karcanias, 1994) that there is a correspondence between the successive design stage decisions of the cascade design process and the evaluation of structural characteristics of process models. Global Instrumentation plays a crucial role in the shaping of structural characteristics, as well as the boundary values of related system properties. This has the advantage that very frequently many degrees of freedom are available, which may be used for design purposes. The central characteristic of this approach is that we view GPI as a process of shaping further the inherited structure from the process synthesis stage. This expresses a fundamental property in the structure evolution during the design (Karcanias, 2008). Note that the term structure is viewed here as a linear graph and/or as system invariants of the underlined model.

The problem of selection of sets of inputs and outputs has been based so far on the shaping of structural invariants (Karcanias & Vafiadis, 2002), such as the set of invariant zeros (MacFarlane & Karcanias, 1976), (Kouvaritakis & MacFarlane, 1976), (Rosenbrock & Rower, 1970), (Karcanias & Giannakopoulos, 1989), (Karcanias, 1994), (Karcanias, 1996), (Karcanias & Vafiadis, 2002), (Karcanias & Vafiadis, 2002), (Leventides & Karcanias, 2008), (Georgiou & Floudas, 1990). Shaping properties such as controllability and observability, goes beyond the shaping of structural invariants. Structural invariants (Karcanias & Vafiadis, 2002), (Karcanias, 2002) determine the shape of system properties, but do not provide the measure of presence of such properties in a system. In this thesis, we will use energy considerations to evaluate the role of selection of systems of inputs and respectively outputs on the energy requirements for state control and state reconstruction. The study of required energy for control and state reconstruction has been a standard theme in the study of linear systems (Kailath, 1980), (Skelton, 1988). However, this has not being used for evaluation and selection of systems of inputs (location of actuators) and systems of outputs (location of sensors) so far. Such a study is intimately related to the study of degree of controllability and observability of a linear system (Moore, 1981), (Arbel, 1981), or alternatively measuring the distance of a system from the set of uncontrollable, respectively unobservable systems.
The objectives of this thesis are:

**Objectives:**

(i) Study the energy dependency of the properties of controllability and observability

(ii) Define the relative degree and presence of controllability and observability under conditions where the available energy is constrained.

(iii) Develop strategies and methodologies for selection of systems of inputs and outputs to minimize the energy required for control, respectively state reconstruction.

(iv) Evaluate the effect of systems of inputs and outputs on various other system properties.

This thesis contributes to the further development of the area by clarifying the role of energy requirements in achieving control and state observation as functions of given sets of inputs and outputs and thus provide energy based tools for the selection of input and outputs schemes with a main criterion the minimization of the energy required for control and observation. The methodologies developed may be used as design tools where apart from energy requirements other design criteria may be also incorporated for the selection of inputs and outputs. The methodology that is used is based on Linear Systems theory (Kailath, 1980), (Karcanias, 2002), (Karcanias & Vafiadis, 2002) and tools from Numerical Linear Algebra. The solution to the problems considered here is an integral part of the effort to develop an integrated approach to Control and Global Process Instrumentation (Karcanias, 1994).

The main achievements of the thesis are in the following areas:

- Provide a literature review for link of energy to controllability and observability and a review of methodologies for selection of inputs and outputs.
- Perform an overview of controllability/observability Gramians, and methodologies for their computation.
- Introduce the minimum energy and define the constant energy surfaces linked to controllability and observability.
- Develop computations of Energy for unstable systems and measures of degree of controllability/observability.
- Develop a selection methodology of system inputs, and thus of the input matrix $B$, to minimize the average required energy for controllability, respectively observability.
- Present a metaheuristic optimization method for the optimal selection of input/output system structure.

The thesis is structured as follows:

---

1 Note: (Kalman, 1959) has pointed out that observability and controllability are duals of each other. We can observe some similarities between optimal actuator model and optimal observation model because of this duality. In order to avoid repetition in this study, what is said about actuator selection can be applied to the concept of sensor selection with the proper interpretation.
Chapter 1
This is an introduction chapter, which provides a brief overview of the problem and summarizes the contents of each of the following chapters.

Chapter 2
Describes a review of the literature showing the importance of the input/output selection in control system design. It reviews the current methodologies and criteria for selection of inputs/outputs and states the open challenges.

Chapter 3
The concepts of reachability and reachability gramian are introduced for linear continuous time-invariant systems. The system properties of controllability and controllability Gramian and a novel approach of calculation of controllability gramian are studied in this chapters. The problem of controllability Gramian assignment is investigated.

Chapter 4
The chapter develops the fundamentals for selection of input (and by duality output) structure based on the energy type criteria, and provides answer to the question of the required energy for the transfer of the origin to a given state in the state space within some given final time when the input structure is given. It also introduces first the problem of minimum input energy and then discusses the important related factors which can influence its value.

Further, this chapter presents the definition of the energy levels in the state space characterizing the maximum of minimum input energy required for transfer to a given distance. This reveals the presence of the energy stratification of the state space.

Chapter 5
The chapter discusses the energy calculation for unstable systems and the impact of actuator selection on various issues such as the degree of controllability, and disturbance rejection. Chapter 5 also presents the generalized energy stratification of the state space for possibly unstable systems.

A method of actuator selection is also presented considering the degree of controllability and the degree of disturbance rejection.

The discussed methods which are proposed in chapter 5 are general methods in the sense that they support both stable and unstable systems.
Chapter 6

The chapter considers the case where the system matrix $A$ is Hurwitz and has certain canonical form structures. This chapter presents some Theorems based on which the controllability Gramian of the continuous-time linear time invariant system can be constructed directly based on the coefficients of the characteristic polynomial, along with the value of the trace and some upper bounds for the maximum eigenvalue of the controllability. Furthermore, the chapter investigates how the value of the energy could be derived based on the characteristic polynomial of the system.

Chapter 7

This chapter proposes a strategy for the selection of the proper input matrix $B$, based on the average of the minimum input energy. This is formulated and solved as an optimization problem for different cases.

First the problem of finding the best single input structure to minimize the energy requirements is discussed and a solution for a general system with a finite terminal time $t_f$ is proposed. Then, the same optimization problem is discussed for the case of stable system with infinite terminal time. This is followed by the input structure selection for the normal system where $AA^* = A^*A$. In the last part of this chapter, the case of multi-input systems subject to different possible conditions are investigated.

The work here provides a new approach to those in existing literature, which considers the problem of input structure selection over a binary set, i.e. where the input matrix $B$ can only take the values $\{0,1\}$ or can be chosen among the given sets. The current approach developed in this chapter does not rely on such restrictive assumptions.

Chapter 8

The chapter develops a new metaheuristic optimization method based on the logistic equations. Superior performance in terms of exploration, exploitation, and convergence is demonstrated relative to other state-of-the-art methods. These properties make the proposed algorithm powerful and capable of solving complex high-dimensional multi-parameter problems, such as optimal actuator and sensor placement.

Chapter 9

This chapter reviews the contribution of this thesis and draws some conclusions. Directions of further work are discussed.
Chapter 2: Literature Review

2.1. Introduction

In this chapter, we review previous works directly relevant to the problem of sensor and actuator selection. In section 2.3, the problem of optimal actuation in continuous linear time-invariant systems is described. The dual problem can also be formulated to find the optimal location of the sensors. With increasing focus on the particular problem of energy optimization, in section 2.4, we consider some of the available methods on input/output selection (IO) and we review the indexes and the criteria specialized for optimal actuator/sensor placement in continuous linear time-invariant systems. Finally, in section 2.5, we summarize this chapter.

2.2. Input/Output Structure Selection

44 years ago (Foss, 1973) challenged the process control research community as he made the observation that in many areas, application was ahead of theory. He introduced control structure design and the gap between theory and applications in this important area:

“The central issue to be resolved by the new theories are the determination of the control system structure. Which variables should be measured, which inputs should be manipulated and which links
should be made between the two sets. . . . The gap is present indeed, but contrary to the views of many, it is the theoretician who must close it.”

A similar observation on the need of control structure design was made by (Findeisen, et al., 1980) (p. 10). Control structure design determines which variables to control, which variables to measure, which inputs to manipulate and which links that should be made between them. (Morari, et al., 1980), (Morari & Stephanopoulos, 1980), (Morari & Stephanopoulos, 1980) introduced new and exciting ideas and theories in the area of control structure design. In the late 1980s (Nett, 1989), (Minto & Nett, 1989) presented a number of lectures about the selection and partitioning of measurements and manipulations for the control of complex systems based on his experience on aero-engine control at General Electric.

A good review of the literature on control structure design can be found in (Larsson & Skogestad, 2000).

While the area of control structure design has received some interest in the literature, e.g. (Georgiou & Floudas, 1989), (Govind & Powers, 1982), (Karcanas, 1996), (Morari & Stephanopoulos, 1980), (Rijnsdorp, 1991), (Skogestad & Postlethwaite, 1996), (Skogsted, 2000) and references therein, the gap still remained, and still does to some extent today. Control structure design cannot be compared to the enormous amount of work on controller design, although it is probably the most important in practice. It involves a number of key sub-problems (Karcanas, 1996) which are:

I. The classification of process variables into potential inputs, outputs and referred to as Model Orientation Problem (MOP)

II. Specification of effective sets of inputs, outputs on an oriented model and referred to as Model Projection Problem (MPP)

III. The selection of a control configuration (i.e. CC), which decides on the way we couple effective inputs and outputs for control design purposes (also called the measurement/manipulation partitioning or input/output pairing).

The above sub-problems have received attention in the literature, see e.g. (Wal & Jager, 2001), (Padula & Kincaid, 1999) and references therein. (Chen, 2002) and (Assali, 2008) summarized different methods for the above sub-problems. Furthermore, (Reinschke, 1988), (Lin, 1974), (Siljak, 1991), (Siljak, 1978), (Murota, 2009) and references therein discussed different solution criteria and applications of the IO and CC selection problems. However, most of the attention so far has been focused on the last sub-problem, i.e. control configuration, when heuristics and diagnostic indicators have been used. For the first two sub-problems, which are the focus of this thesis less attention has been given, especially from the Control Theory viewpoint, with the exception of the work in (Karcanas, 1996), (Georgiou & Floudas, 1989), (Karcanas & Giannakopoulos, 1989), (Morari & Stephanopoulos, 1980) on some specific problems.

In control structure design, the determination of the adequate number, place and sensor-actuator type is the input/output selection (i.e. IO selection) process. The system outputs are the set of the measured variables and the controlled variables. The system inputs are the set of manipulated inputs and the exogenous inputs, such as disturbances, sensor noise and reference inputs or the set-points.
In the control structure selection framework and in this thesis outputs are the measured variables and inputs are the manipulated variables.

The Importance of the IO structure selection lies in the fact that it may affect many properties of the system such as zeros locations, cost, maintenance, reliability, controllability, observability and complexity.

A formal definition for the IO selection is given by (Wal & Jager, 2001):

"Select suitable variables $u$ to be manipulated by the controller and suitable variables $y$ to be supplied to the controller"

Qualitative rules for IO selection problem could be found in (Seborg, et al., 1989):

" i) Control the outputs that are not self-regulating, ii) Control the outputs that have favourable dynamic and static characteristics, iii) Select inputs that have large effects on the outputs, iv) Select inputs that rapidly affect the controlled variables."

Over the last few decades different performance criteria and measurements have been proposed for the selection of actuation variables (inputs) and sensor locations (outputs) (Georges, 1995), (Marx, et al., 2002), (Marx, 2003), (Marx, et al., 2004), (Singh & Hahn, 2005), (Singh & Hahn, 2006). A survey on IO selection is given by (Wal & Jager, 2001). They list several criteria for evaluation of control structure design methods: generality, applicable to nonlinear control systems, controller-independent, direct, quantitative, efficient, effective, simple and theoretically well developed. Ideally, an IO selection method should satisfy all these criteria. After reviewing they conclude that such a method does not exist.

Some discussions on IO selection in the process industry are given in (Morari, 1982), (Shinskey, 1988), (Stephanopoulos, 1984) and (Balchen & Mummé, 1988).

A literature search for optimal actuator or optimal sensor placement methods yields a wide range of publications from different engineering disciplines. Some of these references describe small optimization problems and employ manual optimization techniques or intuitive placement recipes rather than systematic optimization methods. Other references discuss challenging numerical optimization problems and most often use genetic algorithms as the optimization method.

In this thesis, we study the mathematical strategies to determine the optimal actuator locations. A new metaheuristic optimization algorithm is also developed to solve the challenging actuation problems in the case of large scale systems. By duality, the solution methodology may be readily extended to the problem of optimal sensor selection. To verify the results, wherever applicable, numerical examples are designed and the results are compared with those available in the literature.
2.3. Problem Statement:

As it is discussed in the previous sub-section, the location of an actuator/sensor has a tremendous impact on the performance of the controlled system. Misplaced actuators can lead to lack of controllability (Kumar & Narayanan, 2007). The actuators/sensors should therefore be located at positions that optimize certain performance objectives.

In aerospace engineering, the actuators at optimal locations reduce the vibrations (Mehrabian & Yousefi-Koma, 2007). The optimal placement of actuators/sensors is essential for effective control of structural vibration and acoustic noise (Pulthasthan & Pota, 2008). In acoustic problems, an arbitrarily placed actuator can actually increase the sound field locally (Fahroo & M.A. Demetriou, 2000). Civil structures such as high-rise buildings and suspension bridges are designed to protect against earthquake excitation through the placement of actuators at appropriate heights (Abdullah, et al., 2001). The energy consumption is a concern in various control applications, e.g. vibration control of smart structures, chemical process control (Arbel, 1981), (Peng, et al., 2005) showed that the optimal placement of actuators/sensors improves the performance of the control system significantly and at the same time the energy consumption is minimized. In (Leleu, et al., 2001), (Sung, 2002), (Yue, et al., 2008), (Jha & Inman, 2003), (Hac & Liu, 1993) improved performance of vibration control is observed when actuators/sensors are placed at optimal locations. In (Antoniades & Christo, 2002), an integrated feedback design problem and optimal actuator placement minimizing the control energy was investigated. It is shown in (Morris, 1998) that the optimal actuator placement can reduce noise in a duct. In systems modelled by partial differential equations, optimal location calculations are performed on approximated problems, although the state space for the full model is infinite dimensional. The theory that guarantees optimality of the cost and existence of the optimal actuator location for these models has not been developed in its entirety (Hebrard & Henrot, 2005).

The aim of this thesis is to develop a mathematical framework for calculating optimal actuator locations with regard to the energy-based criteria, and to develop a new evolutionary algorithm to solve the combinatorial problem of optimal actuator/sensor placement in the challenging real world applications.

Consider the following linear time-invariant system in \( \mathbb{R}^n \):

\[
\begin{align*}
\dot{x} &= Ax(t) + Bu(t) \\
y(t) &= Cx(t) + Du(t)
\end{align*}
\]

where \( x \) denotes the state of the system, \( A \) denotes the system dynamics and \( u(t) \) is the control applied to the system as a function of time. The effect of control on the state of the system is described by the input matrix \( B \).

In many control systems, the location of actuators/sensors can often be chosen. These locations should be selected in order to optimize the performance criterion of interest.

Consider the situation where there are \( m \) actuators with locations that could be varied over some compact set, call it \( \Omega \subset \mathbb{R}^n \). Parametrize the actuator locations by \( a \) and denote the dependence of the corresponding input matrix with respect to the actuator location by \( B_a \). Note that \( a \) is a vector of length \( m \) with components in \( \Omega \) so that \( a \) varies over a space denoted by \( \Omega^m \). Based on the designer’s
interest on the choice of a desired performance measure, a suitable cost function $J_a$ that depends on the actuator location is formulated.

**Definition (2.1):** The optimal cost $J_{opt}$ over all possible locations is defined as,

$$J_{opt} = \inf_{a \in \Omega^m} J_a$$ (2.2)

Also, if it exists, the location $a_{opt} \in \Omega^m$ that satisfies

$$a_{opt} = \arg \inf_{a \in \Omega^m} J_a$$ (2.3)

is called the optimal actuator location.

In the rest of this chapter, the background of the formulations of the cost function $J_a$ with regard to the energy-based strategies used in diverse applications will be discussed.

### 2.4. Different Cost Functions:

Previous studies have used various measures and criteria in order to increase the effectiveness of sensors and actuators and to reduce their numbers and sizes while some specific variables are optimized (Wal & Jager, 2001), (Frecker, 2003), (Padula & Kincaid, 1999). The most commonly used criteria in literature are:

- **Accessibility**

It is a qualitative technique for IO selection based on cause and effect graphs, which can be generated for linear and nonlinear systems. Vertices of the graph represent system variables, disturbances, and control signals. Directed edges depict relations between various variables (Kościelny, et al., 2017). The key idea for IO selection is that a causal path must exist between the manipulated and the controlled variables on the one hand and the measured and the controlled variables on the other hand. A large number of candidate IO sets may be termed viable if they are only assessed for accessibility. So, additional criteria should be invoked (Wal & Jager, 2001).

Accessibility may be chosen as a tool for actuator/sensor placement because of the following features:

I. Faults can be included directly in the model as additional vertices.

II. The graph model is a simple, intuitive, and easily understandable way of describing a process. Mathematical description is not needed, only basic knowledge about the physics of a process. This is an important feature because actuator/sensor placement analysis should be performed at the design stage and a detailed process model can be unavailable.

(Lambert, 1977) used fault-trees to analyse the location of sensors. Even though this work was the first step toward the design of sensor locations based on a diagnostic observability criterion, it had drawbacks such as: i) inability to handle cycles, and ii) the development of a fault tree is in itself an error-prone and time-consuming process. To solve the problem of observability, based on the process graph, (Ali & Narsimhan, 1993), (Ali & Narsimhan, 1995) have presented sensor placement strategies
for linear and bilinear processes (Ali & Narasimhan, 1996). They introduced the concept of reliability and the degree of redundancy (spatial redundancy: measuring more variables than the minimum required to ensure observability, so that there are multiple ways of estimating a variable (Mah, 1990) and hardware redundancy: measuring a variable using more than one sensor) for sensor placement in linear systems. There are also other methods based on linear and bilinear mathematical models for sensor location (Madron & Veverka, 1992). They proposed two objective functions for optimization: overall measurement cost and overall precision of a system and utilized Gauss-Jordan elimination to identify a minimum set of variables that need to be measured in order to observe all important variables. Their work is essentially an extension of the work of (Vaclavek & Loucka, 1976), (Raghuraj, et al., 1999), (Bhushan & Rengaswamy, 2000), (Bhushan & Rengaswamy, 2000b), (Khemliche, et al., 2006), (Bhushan, et al., 2008), (Yang, et al., 2009) also used qualitative (cause-effect) graph theoretic approaches to find the optimal sensor locations in a process.

- **Linear Quadratic Control**

Linear quadratic control is a well-known closed-loop strategy for controller design that minimizes energy of both the control signal and the measured signal. The control is calculated by minimizing a quadratic cost function with penalty on both the state of the system and the control input. This objective is concerned with controlling the initial condition to 0 and disturbances are neglected.

Consider the system (2.1) on $\mathbb{R}^n$. In the case of a single control $u$, the linear-quadratic (LQ) controller design objective is to find a control $u(t)$ to minimize the quadratic cost functional $J(u, x_0)$ over infinite-time interval, for a given initial state $x_0$:

$$J(u, x_0) = \int_0^\infty \langle x(t), Qx(t) \rangle + \langle u(t), Ru(t) \rangle \, dt$$

(2.4)

where $R > 0$ weights the control cost, $Q$ is a self-adjoint positive semi-definite matrix weighting the state. The linear quadratic control problem is to minimize the cost function (2.4) over all possible controls:

$$\min_{u \in L_2(0, \infty, \mathbb{R}^m)} J(u, x_0)$$

(2.5)

The control that achieves this minimum, $u_{opt}$, is often called the linear quadratic optimal control.

**Definition (2.2):** The pair $(A, B)$ is stabilizable if there exists $K \in \mathbb{R}^{m \times n}$ so that $A - BK$ is Hurwitz.

**Definition (2.3):** The pair $(A, C)$ is detectable if there exists $F \in \mathbb{R}^{n \times l}$ so that $A - FC$ is Hurwitz.

**Theorem (2.1):** (Morris, 2001) Consider the system given by (2.1). Assume the case that there are $m$ actuators with locations that could be varied over the compact set $\Omega$. If $(A, B, Q^{1/2})$ is both stabilizable and detectable. Then the infinite-horizon optimization problem (2.5) has a minimum for every given initial condition $x_0$. Furthermore, there exists a symmetric, semi-positive definite matrix, $P_a \in \mathbb{R}^{n \times n}$, such that:

$$\min_{u \in L_2(0, \infty, \mathbb{R}^m)} J_a(u, x_0) = J_a(u_{opt}, x_0) = x_0^T P_a x_0$$

(2.6)
where $P_a$ is the unique solution of the Algebraic Riccati Equation (ARE):

$$A^TP_a + P_aA - P_aB R^{-1}B^TP_a + Q = 0$$

(2.7)

The optimal control $u_{opt}$ is as follows:

$$u_{opt} = -R^{-1}B^TP_a x(t)$$

(2.8)

and the corresponding optimal state feedback, $K = -R^{-1}B^TP_a$ is stabilizing, i.e. $A - BK$ is Hurwitz.

The proof is given in (Morris, 2011), see Theorem (5.12) and Theorem (5.16). According to the above Theorem, for a particular initial condition $x_0$, the optimal actuator location problem can be formulated as follows:

$$\inf_{a \in \Omega^a} x_0^TP_a x_0$$

(2.9)

The linear quadratic index has been studied in various literature as an objective function to optimise actuator location and feedback gain, (Demetriou, 2000), (Chen, et al., 2012), (Li & Huang, 2013).

The optimal linear-quadratic cost at a given actuator location $a$, which is given by (2.9) depends on the initial condition $x_0$. Researchers have used a number of techniques in the past to remove this dependency. The first approach is to consider the worst case initial condition (Devasia, et al., 1993), (Morris, 2011). The cost function is:

$$\max_{|x_0| = 1} \min_{u \in L_2(0,\infty;\mathbb{R}^n)} J_a(u, x_0) = \max_{|x_0| = 1} (x_0^TP_a x_0) = \lambda_{\text{max}}(P_a)$$

(2.10)

Then the optimal actuator selection problem can be formulated as:

$$\inf_{a \in \Omega^a} \lambda_{\text{max}}(P_a)$$

(2.11)

Another method is to view the initial condition as a random vector with zero mean and unity variance (Morris, 2011), (Geromel, 1989). The expected optimal linear-quadratic cost is:

$$E \left[ \min_{u \in L_2(0,\infty;\mathbb{R}^n)} J_a(u, x_0) \right] = E \left[ x_0^TP_a x_0 \right] = \text{trace}(P_a)$$

(2.12)

Thus, the problem of optimal actuator location can be defined as:

$$\inf_{a \in \Omega^a} \text{trace}(P_a)$$

(2.13)

Another approach is to average the cost over a set of linearly independent initial states (Antoniades & Christo, 2002). The effect of a disturbance with fixed frequency content (for instance, a single white noise disturbance) is considered in (Morris & Demetriou, 2010). This leads to a $H_2$ control problem and if the spatial distribution of the disturbance is unknown, then the cost function (2.10) is used to calculate the optimal actuator location. Linear quadratic control is a popular choice, since the controller is designed simultaneously with the optimal actuator location. A methodology based on the minimisation of the optimal linear quadratic control index as an objective function was proposed by (Devasia, et al., 1992). This objective function was applied to a simply-supported beam in order to optimise the size and location of actuators using a simple search method. They reported that this objective function achieves stability of closed-loop control system and allows a designer to choose various values of weighted matrices of optimal linear quadratic index in order to optimise actuator location and vibration reduction. (Kondoh, et al., 1990) proposed an objective function based on the minimisation of the linear quadratic index to optimise sensor and actuator location and feedback gain.
for a cantilever beam. The authors reported that the optimal linear quadratic control has a clear physical meaning and the flexibility to allow the implementation of varying weighted matrices. (Sung, 2002) used the linear quadratic index with the worst case initial condition for placing actuators on a simply supported beam with a moving mass. In (Fahroo & M.A. Demetriou, 2000), the linear quadratic cost function with a random vector uniformly distributed initial condition was minimized for finding the best locations of actuators to reduce the interior noise in an acoustic cavity. To determine the most effective placement and design of an active control system, (Levine & Athans, 1970), (Abdullah, et al., 2001) used the cost function (2.12) to consider both the building response and the control effort. (Kumar & Narayanan, 2007), (Kumar & Narayanan, 2008) considered the optimal placement of collocated piezoelectric actuator/sensor pairs on flexible beams using a model-based linear quadratic regulator (LQR) controller.

• $H_\infty$ and $H_2$ Norms

Noise in the environment and disturbances, which are often unknown, will influence the performance of the control system. Therefore, in many situations, the problem is to reduce the effect of disturbances on the performance. The control system is now described by:

$$\begin{cases}
\dot{x} = Ax(t) + Bu(t) + B_d d(t) \\
y(t) = Cx(t) + Du(t)
\end{cases}, \quad x \in \mathbb{R}^n, A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}, C \in \mathbb{R}^{l \times n}, D \in \mathbb{R}^{l \times m}, B_d \in \mathbb{R}^{m \times k}$$

(2.14)

where $d(t)$ is the exogenous input or disturbance. $H_\infty$ and $H_2$ norm functions are a measure of open and closed loop system frequency response to an external disturbance. These criteria have been used as objective functions to optimise the location of sensors and actuators. The problem is to find a controller $u(t)$ to minimize:

$$\sqrt{\int_0^\infty \|y(t)\|^2 dt}$$

(2.15)

Since the $L_2$ norm of $y$ equals the $H_2$ norm of the Laplace transform of $y$, this is known as an $H_2$ controller design problem. The $H_2$ and $H_\infty$ cost (Geromel, 1989) are closely related to the linear quadratic criteria examined in the previous subsection. (Leonides, 2012)

The usual orthogonality hypotheses $C^T D = 0$ and $D^T D = I$ are assumed in order to simplify the subsequent equations. The full column rank of matrix $D$ ensures a non-singular map $y$ on the control $u(t)$. It is assumed that all states of the system are available for measurement. This system (2.14) is a special form of the generalised plant configuration, known as the full information problem.

Assume that $d(s)$ and $y(s)$ be the Laplace transforms of the disturbance and the output respectively. Then, minimizing the effect of all disturbances on the output is equivalent to minimizing the size of the transfer function:

$$G_y d = \frac{y(s)}{d(s)}$$

(2.16)

The notation $H_\infty$ indicates the Hardy space of all functions $G(s)$, which are analytic in the right-half plane $\text{Re}(s) > 0$ and for which:

$$\sup_{\omega} \lim_{\alpha \to 0} |G(x + j\omega)| < \infty$$

(2.17)

Thus, $H_\infty$ norm is defined as:
\[
\|G(j\omega)\|_\infty = \sup_{\omega} \sigma(G(j\omega))
\]  
(2.18)

where \(\sigma(G(j\omega))\) denotes the maximum singular value of the transfer function \(G(j\omega)\). The details of calculating \(H_\infty\) norm can be found in many literature, see e.g. (Morris, 2001), (Zhou & Doyle, 1996).

Let \(G\) be the transfer function of the system (2.14):
\[
G(s) = C(sI - A)^{-1}[B \quad B_d] + [D \quad 0]
\]  
(2.19)

With state feedback control:
\[
u = -Kx(t)
\]  
(2.20)

The closed loop transfer function from disturbance to output is:
\[
G_{yd}(s) = (C - DK)(sI - (A - BK))^{-1}B_d
\]  
(2.21)

**Definition (2.4):** (Doyle, et al., 1989) The controlled system will have input \(d\) (disturbance) and output \(y\) (measurement). The fixed attenuation \(H_\infty\) control problem for attenuation \(\gamma\) of (2.14) is to construct a stabilizing controller with transfer function \(G(s)\) so that the closed loop system \(G_{yd}\) with \(u(s) = G(s)x(s)\) is \(L_2\) stable and satisfies the bound:
\[
\|G_{yd}\|_\infty < \gamma
\]  
(2.22)

Even for stabilizable systems, the fixed attenuation problem cannot be solved for every attenuation since such a controller may not exist. However, if it is solvable, as in the case of linear quadratic control, the control law can be chosen to be constant state feedback. (Özdemir, 2003)

**Definition (2.5):** The optimal \(H_\infty\) control problem for (2.14) with full-information is to find:
\[
\hat{\gamma} = \inf \gamma
\]  
(2.23)

over all \(\gamma\) for which the fixed attenuation problem is solvable. The infimum \(\hat{\gamma}\) is called the optimal \(H_\infty\) attenuation, which can be calculated using a bisection-type algorithm.

The \(H_\infty\) cost for a particular actuator location \(a\) is \(\hat{\gamma}_a\). The value of \(\hat{\gamma}_a\) provides the best possible attenuation of the worst-case disturbance for the system at the actuator location \(a\).

Another approach for placing actuators is to use the \(H_\infty\) norm of the closed-loop system, which has some advantages:
- The effect of the worst-case disturbance on the output of the system is minimized
- with optimal actuator location, the corresponding optimal state feedback controller is simultaneously designed
- The closed loop system is stabilized

(Gawronski, 1998) discusses the \(H_\infty\) criteria in more details.

Consider the performance index:
\[
J(u, d; x_o) = \|y\|_{L_2(0,\infty; R^l)} - \gamma^2 \|d\|_{L_2(0,\infty; R^l)}
\]  
(2.24)
subject to (2.14) for some $\gamma > 0$. Calculating a controller that achieves the given $H_\infty$ attenuation bound is equivalent to solving the optimization below (Morris, 2001):

$$\max_{d \in \mathbb{R}^n} \min_{u \in \mathbb{R}^m} J(u, d; x_0)$$

(2.25)

Consider the first term of the performance index (2.24). The $L_2$ norm of $y$ equals the $H_2$ norm of the Laplace transform of $y$, this is known as an $H_2$ controller design problem. The $H_2$ optimal control is the state feedback (Zhou, et al., 1997):

$$u(t) = -B^*_a P_a x(t)$$

(2.26)

Subject to stabilizability of $(A, B_a)$. $P_a$ is the unique solution of the Algebraic Riccati Equation (2.7).

Therefore, the $H_2$ optimal actuator location problem is to find the actuator location $a$ that minimizes the corresponding cost (Morris, et al., 2015):

$$\sqrt{\text{trace}(B^*_d P_d B_d)}$$

(2.27)

Mathematically, the problem is identical to that of minimizing the LQ cost when the initial condition is random with covariance $B^*_d B_d$. If $B_d$ is a continuous function of $a$, both the optimal actuator location problem with a fixed disturbance location, and the problem where the disturbance is unknown, lead to well-posed optimization problems. Since the optimal cost relies on the norm of the solution to Algebraic Riccati Equation (2.7), well-posedness of this problem follows using techniques and results similar to the linear quadratic case (Morris, et al., 2015), (Morris & Yang, 2015).

In a fluid application, (Chen & Rowley, 2011) used $H_2$ criteria for optimal actuator and sensor placement with respect to the disturbance and with respect to each other, which is shown to have a significant effect on performance. In (Ambrosio, et al., 2012), an $H_2$ criterion based on a modal approximation of a structure is applied. They have considered the neglected modes to accomplish spill-over attenuation. In (Guney & Eskinat, 2008), sensor/actuator pair placement on a simply supported beam and $H_\infty$ control with an impulse disturbance was performed. The control signal is shown, but no comparisons were made between locations.

(Liu, et al., 2006), (Hiramoto, et al., 2000), (Arabyan & Chemishkian, 1998), (Chemishkian & Arabyan, 1999) have investigated the optimal locations of sensor and actuator under closed loop control using optimisation algorithms to find the optimal combinations of sensors and actuators. In (Hiramoto, et al., 2000), the product of a frequency weighting term that represents the design specification and the closed-loop transfer function was used for optimal actuator placement.

In (Raja & Narayanan, 2009) $H_\infty$ cost function was used for locating actuators in the vibration control of tensegrity structures. The results showed that the displacement is less with $H_\infty$ criteria than the $H_2$ performance index. (Silva, et al., 2006) used an $H_\infty$ approach to place actuators on a plate.

(Sweeney, et al., 2005), (Demetriou & Grigoriadist, 2004) have used an analytical expression to compute the upper bound on the $H_\infty$ norm of the controlled system was used to optimize the actuator locations. Using this analytical approach, it was shown in (Demetriou & Grigoriadist, 2007) that the resulting optimal actuator location exhibits spatial robustness. Calculation of optimal actuator locations with spatially varying disturbances was addressed in (Demetriou, 2004).
• **Degree of Controllability/Observability and Gramians**

Controllability refers to the property of being able to steer the state of a dynamical system from any starting point to any terminal point by means of appropriate inputs. Observability is concerned with whether without knowing the initial state, one can determine the state of a system given the input and the output. Controllability and observability of a system, which depend primarily on actuator and sensor location, will have a major influence on the efficiency of the control system and the control effort required to satisfy design requirements.

**Definition (2.6):** Consider the following linear time-invariant system:

\[
\begin{aligned}
\dot{x} &= Ax(t) + B_a u(t), \\
y &= C_s x(t) + Du(t),
\end{aligned}
\]

\[A \in \mathbb{R}^{n \times n}, B_a \in \mathbb{R}^{n \times m}, C_s \in \mathbb{R}^{l \times n}, D \in \mathbb{R}^{l \times m}\] (2.28)

\(B_a\) denotes a family of input operators that are compact and continuous with respect to actuator location \(a\). This system is said to be controllable at the actuator location \(a\) if, for any initial state \(x_0\) and any final state \(x_f\), there exists an input \(u(t)\) such that the integral curve \(x(t)\) generated by \(u(t)\) with \(x(0) = x_0\), satisfies \(x(t_f) = x_f\) in a finite time \(t_f > 0\). Otherwise, the system or the pair \((A, B_a)\) is said to be uncontrollable at the actuator location \(a\).

**Definition (2.7):** Consider the linear time-invariant system in (2.28), \(C_s\) denotes a family of output operators that are compact and continuous with respect to sensor location \(s\). This system is called observable at the sensor location \(s\) if for any finite time \(t_f > 0\), the initial state \(x(0) = x_0\) can be determined from the time history of the input \(u(t)\) and the output \(y(t)\) in the interval \([0, t_f]\). Otherwise, the system or the pair \((A, C_s)\) is said to be unobservable at the sensor location \(s\).

Since the performance of the controlled system is highly related to the actuator/sensor locations (Fahroo, 1995), (Morris, 1998), many authors optimized actuator locations based on maximization of observability and controllability (Liu & Hu, 2010), (Peng, et al., 2005), (Qiu, et al., 2007), (Zhang & Erdman, 2006).

A general rule for IO selection method would be to reject candidate IO sets for which \((A, B_a)\) is uncontrollable or \((A, C_s)\) is unobservable. The problem of associating physically meaningful measures of quality with the notions of controllability and observability was discussed a little by (Brown, 1966) and (Monzuncio, 1967). Later, (Johnson, 1969) after a comprehensive review of the problem maximized one special scalar measure of the quality of complete controllability and complete observability for a class of linear dynamical systems. It is shown in (Olshevsky, 2014) that the minimal controllability problem is NP-hard; indeed, this paper utilizes a simple greedy heuristic which sequentially picks variables to maximize the rank increase of the controllability matrix. (Václavek & Loučka, 1976) proposed a sensor selection on steady-state systems using graph theory to ensure observability of...
variables. (Caruso, et al., 2003) studied the optimal placement of actuators and sensors for a collocated flexible plate structure. For a given number of available piezoelectric patches of fixed dimensions to be bonded to the plate, the optimal locations were obtained by maximizing the modal controllability and observability of the structure. Various simple tests for this qualitative measure are available (Zhou, et al., 1997). According to Kalman rank condition, the system given in (2.28) is observable and controllable if the controllability and observability matrices (2.29) are full rank.

\[
C = \begin{bmatrix}
B & AB & A^2B & \ldots & A^{n-1}B
\end{bmatrix} \\
O = \begin{bmatrix}
C & AC & A^2C & \ldots & A^{n-1}C
\end{bmatrix}^T
\]

A large number of studies has considered classical binary controllability metrics based on Kalman rank (Liu, et al., 2011), (Rajapakse, et al., 2011), (Cowan, et al., 2012), (Nepusz & Vicsek, 2012), (Wang, et al., 2012), (Olshevsky, 2014), (Pequito, et al., 2016). (Morari & Stephanopoulos, 1980) proposed structural state controllability and observability as the selection criteria. The controllability and observability concepts has been strengthened by the fact that they are robust properties in linear constant systems (Lee & Markus, 1967). That is, the set of all controllable pairs \((A, B)\) is open and dense in the space of all such pairs. This fact was exploited by (Lin, 1974) in formulating his concept of structural controllability, which states that all uncontrollable systems structurally equivalent to a structurally controllable system are a typical. Not only is this concept consistent with physical reality in the sense that values of system parameters are never known precisely, but it is also helpful in testing the properties of controllability using only binary computations. It is a well-known fact that testing controllability is a difficult numerical problem (Nour-Eldin, 1987), and structural controllability offers a computable alternative especially for systems of high dimension. Regarding the duality, the same facts hold for observability.

Definition (2.8): An \(n \times n\) matrix \(\tilde{M} = (\tilde{m}_{ij})\) is said to be a structured matrix if its elements \(\tilde{m}_{ij}\) are either fixed zeros or independent free parameters. For example, a \(2 \times 3\) structured matrix is:

\[
\tilde{M} = \begin{bmatrix}
\times & \times & 0 \\
\times & 0 & \times
\end{bmatrix}
\]

To relate a numerical \(n \times n\) matrix \(M = (m_{ij})\) to a structured matrix \(\tilde{M} = (\tilde{m}_{ij})\) we define \(n = \{1, 2, \ldots, n\}, m = \{1, 2, \ldots, m\}\) and state the following:

Definition (2.9): A numerical matrix \(M\) is said to be admissible with respect to a structured matrix \(\tilde{M}\), that is, \(M \in \tilde{M}\), if and only if \(\tilde{m}_{ij} = 0\) implies \(m_{ij} = 0\) for all \(i \in n\) and \(j \in m\).

\[
\tilde{y} = A\tilde{x} + B\tilde{u} \\
\tilde{y} = C\tilde{x}
\]

A structured system \(\tilde{S} = (\tilde{A}, \tilde{B}, \tilde{C})\) so that \((A, B, C) \in (\tilde{A}, \tilde{B}, \tilde{C})\). Structural controllability of the system \(\tilde{S}\) is defined via the pair \((\tilde{A}, \tilde{B})\) as in (Lin, 1974).
**Definition (2.10):** A pair of matrices \((A, B)\) is said to be structurally controllable if there exists a controllable pair \((A, B)\) such that \((A, B) \in (\tilde{A}, \tilde{B})\).

**Definition (2.11):** A pair of matrices \((A, C)\) is said to be structurally observable if there exists an observable pair \((A, C)\) such that \((A, C) \in (\tilde{A}, \tilde{C})\).

In (Commault & Dion, 2013) and references therein, given the structure of a linear-time invariant system and a possible collection of inputs, the objective is to determine the minimum subset of inputs, which yields the structural controllability. (Commault & Dion, 2013) proposes solution methods to the above constrained minimal input selection problem rely on a two-step optimization procedure, which, in general, leads to suboptimal solutions. Determining feasible solutions to the constrained minimal input selection problem has been a major focus of (Dion, et al., 2003), (Murota, 2009). Similar work is presented in (Boukobza & Hamelin, 2011), where the analysis of IO selection is investigated via ensuring structural observability for linear systems in descriptor form. The problem of identifying the optimal input/output, which yields the structural controllability/observability was also considered in (Liu, et al., 2011), (Dion, et al., 2002).

(Liu, et al., 2011) states that the minimum number of controlling agents required to achieve structural controllability is related to the number of right unmatched vertices of an associated bipartite matching problem. However, these criteria have some disadvantages. They are binary qualitative concepts and tell only whether a set of actuators/sensors make the system controllable/observable or not, regardless of other considerations such as the energy required to actually drive the system around the state. Therefore, a measure of the degree of controllability/observability has to be defined to help choose the best IO set. This measure should have the following attributes (Abdel-Mooty & Roorda, 1994):

- It must vanish for uncontrollable/unobservable case
- It must indicate the effectiveness of the selected set of actuators/sensors, which is the ability of the actuators to induce the desired control effect, i.e. enhance the controllability properties, and the ability of the sensors to enhance the detection, i.e. improve the observability properties.
- It must indicate the control/observe cost as a measure of the effort made to achieve the required control/detect level
- It must reflect the control/observe objective

**Definition (2.12):** A recovery region in the state space was defined as the region that includes all of the initial conditions (or disturbed states) that can be returned to the origin in a finite time using the bounded control forces. The degree of controllability (DOC) was defined as a scalar measure of the size of the recovery region.

In other words, the degree of controllability will be determined by the minimum distance from the origin to a state that cannot be brought to the origin in a finite time \(t_f\). More loosely, it is the minimum initial condition disturbance from which the system cannot recover in \(t_f\).
**Theorem (2.2):** (Morris, 2001) The pair \((A, B_a)\) defined in (2.1) is controllable at \(a\) if and only if 
\[ W_c(a, 0, t_f) \] 
is positive definite for all \(t_f > 0\).

where:
\[
W_c(a, 0, t_f) = \int_0^{t_f} e^{At_f} B_a^T e^{At} d\tau
\]

The proof is given as Theorem (2.5) in (Morris, 2001). To address the degree of controllability/observability issue various quantitative measures have been proposed in literature. Among the first attempts was the work of (Kalman, et al., 1963), in which a symmetric controllability matrix for time-varying linear systems was defined. That study further defined the determinant and the trace of controllability Gramian as scalar measures of the controllability, see e.g. (Leleu, et al., 2001). (Muller & Weber, 1972), (Arbel, 1981) added the minimum eigenvalue of the controllability Gramian as a third measure. Controllability and observability Gramian have some advantages when compared to controllability/observability matrices. The superior numerical properties of Gramians are due to their symmetry, semi-definiteness and compactness. An interpretation of the Gramians is the geometry of principal components of an ideal sensor/actuator configuration. The controllability Gramian quantifies the energy required to move the system around the state space. Relevant results are also found in (Pasqualetti, et al., 2014), where the authors study the controllability of a system with respect to the smallest eigenvalue of the controllability Gramian, and they derive a lower bound on the number of actuators so that this eigenvalue is lower bounded by a fixed value. Nonetheless, they do not provide an algorithm to identify the actuators that achieve this value. They propose a heuristic actuator placement procedure that does not constrain the number of available actuators and does not optimize their control energy objective. A small eigenvalue of the controllability Gramian matrix would lead to at least one state requiring very high control effort. This implies that all the eigenvalues of the controllability Gramian matrix should be as large as possible. (Georges, 1995) extended this idea to nonlinear systems. (Hac & Liu, 1993), (Qiu, et al., 2007) developed an optimization method for finding the optimal locations of piezoelectric actuators based on the degree of controllability using the Gramian matrix as follows:

\[
\max_{a \in \Omega} \rho = \max_{a \in \Omega} \sum_{i=1}^{n} \lambda_i \left( \prod_{i=1}^{m} \lambda_i \right)
\]

where \(m\) is the number of actuators with locations that could be varied over some compact set \(\Omega \subset \mathbb{R}^n\), \(\lambda_i, i = 1, ... , n\) denotes the \(i\)-th eigenvalue of the controllability Gramian. The summation term in (2.32) is the trace of the Gramian matrix. To ensure that all the eigenvalues of the Gramian are high, the geometric mean of all the eigenvalues is included in the objective function. The optimal locations of multiple actuators calculated using (2.32) suppressed the vibrations in a plate (Peng, et al., 2005). An adaptive feedforward controller was designed in (Peng, et al., 2005) after calculating the optimal actuator location.

A term related to the standard deviation of Gramian eigenvalue, \(\sigma^{-1}(\lambda_i)\), was often included in the performance index (2.32) as a product term, see e.g., (Jha & Inman, 2003), (Pulthasthant & Pota, 2008), where the objective function is defined as:
\[ \rho = \left( \sum_{i=1}^{n} \lambda_i \right) \left( \sqrt[n]{\prod_{i=1}^{n} \lambda_i} \right) / \sigma(\lambda_i) \]  

(2.33)

Here \( \lambda_i \) could be the \( i \)th eigenvalue of the controllability gramian in the case of actuator locations or the observability gramian when sensor positions are considered. Factor \( \left( \sqrt[n]{\prod_{i=1}^{n} \lambda_i} \right) / \sigma(\lambda_i) \) in equation (2.5) are used to normalize the factor \( \left( \sum_{i=1}^{n} \lambda_i \right) \) which is directly proportional to the total exchanged energy between the plant and actuators or sensors. The term \( \left( \sqrt[n]{\prod_{i=1}^{n} \lambda_i} \right) \) is the geometric mean of the ellipsoid axes length, \( \sigma(\lambda_i) \) penalizes locations where there is a poorly controllable or observable state hidden by a highly controllable or observable state or there exist both very high and very small eigenvalues. Optimal sensor and actuator location is obtained at the position where the objective function \( \rho \) in equation (2.5) is a maximum.

(Muller & Weber, 1972), (Dochain, et al., 1997), (Waldraff, et al., 1998), (Van den Berg, et al., 2000) presented a sensor placement technique based on a suitable norm of the observability Gramian or the observability matrix. For optimal actuator placement (Choe & Baruh, 1992) minimizes the objective functions based on the entries of the actuator influence matrix, which gives general measures of controllability. (Devasia, et al., 1993), (Jha & Inman, 2003), (Bruant & Proslier, 2005) proposed the maximization of a controllability/observability criterion using the Gramian matrices as well. (Wang & Wang, 2001) suggested the maximization of the control forces transmitted by the actuators to the structure. (Dhuri & Seshu, 2006) proposed a modal controllability index based on the same singular value analysis of the control vector. In an attempt to define a degree of controllability, (Cheng & Pantelides, 1988) suggested a weighted sum of the squares of the modal displacements of seismic buildings at the actuator position, each multiplied by the maximum modal response spectrum value for the design earthquake. However, this criterion does not satisfy the basic requirement that the DOC vanishes when the system becomes uncontrollable.

For actuator placement (Kim & Junkinsf, 1991) introduced a combination of the squares of Hamdan and Nayfeh’s modal controllability measures weighted by the respective modes’ contributions to a quadratic output cost function. (Bagajewicz & Sanchez, 1999) presented a sensor placement technique with the goal of achieving a certain degree of observability or redundancy for a variable in a system. They introduced the degree of estimability of a variable by merging the concept of degree of redundancy for measurements and degree of observability for unmeasured variables into one single property. (Bruant, et al., 2010) used two modified optimization criteria, ensuring good observability or controllability of the structure, and considering residual modes to limit the spill-over effect. They considered two optimization variables for each piezoelectric device: the location of its centre and its orientation. They implied Genetic algorithms to find the optimal configurations. (Han & Lee, 1999) also used genetic algorithm (GA) to determine the sensor and actuator locations with the consideration of controllability, observability, and spill-over prevention. However, in their work, the dynamic characteristics of both rectangular plate and piezoelectric sensors/actuators were not derived explicitly.
Lim (1992) proposed a method based on the combined degree of controllability and observability to select the most suitable set of actuator and sensor locations, which is capable of simultaneously controlling and observing the system to a high degree. The controllability Gramian was computed and the controllable subspace for each actuator location was derived. The observability Gramian was also calculated and the observable subspace for each sensor location was determined. An intersection subspace was formed from the controllable and observable subspaces of each actuator/sensor pair. The proposed optimization cost functional did not incorporate closed-loop control; therefore, any control method could be applied after the actuator positions are determined. Additionally, the approach was computationally efficient because the problem size and design space are proportional to the number of actuators, unlike other methods where the design space is factorially related to the number of actuators. One disadvantage of the method is that a priori knowledge of the system is required.

The use of Gramians as quantitative metrics of controllability in networks is studied in (Rajapakse, et al., 2011), (Yan, et al., 2012), (Sun & Motter, 2013), (Tang, et al., 2012), (Pasqualetti, et al., 2014). Other important studies of controllability measures in networks include (Olshevsky, 2014), (Sorrentino, et al., 2007), (Rahmani, et al., 2009). They have studied the problem of leader selection in networks with consensus dynamics, in which a set of leader states are selected to act as control inputs to the system. Controllability Gramian based approach is an open-loop strategy. Calculation of optimal actuator location does not simultaneously provide a controller. For example, in (Pulthasthan & Pota, 2008), the design of optimal control and the optimal actuator location were treated as two separate problems which is cumbersome.

**Definition (2.13):** A set function $f : 2^V \rightarrow \mathbb{R}$ is called submodular if for all subsets $A \subseteq B \subseteq V$ and all elements $s \notin B$, it holds that:

$$f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$$

or equivalently, if for all subsets:

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$$

A set function is called super-modular if the reversed inequalities in (2.34) and (2.35) hold and is called modular if (2.34) and (2.35) hold with equality.

The authors in (Summers & Lygeros, 2014) showed that the mapping from possible actuator placements to the trace of the controllability Gramian is a modular set function and therefore a simple optimization can be implemented to compute the metric individually for all possible actuator placement combination, sorts the outcome, and selects the globally optimal subset that minimizes the Gramian energy metric (Summers, 2016). Several classes of Gramian metrics are shown to be submodular in (Summers, et al., 2016), (Summers & Lygeros, 2014), (Cortesi, et al., 2014), (Tzoumas, et al., 2016):

- $-\text{trace}(W_c^{-1})$
- $\log|W_c|$
- $\text{rank}(W_c)$

In addition, many other problems featuring super-modularity or sub-modularity are discussed in (Pasqualetti, et al., 2014), (Clark, et al., 2014), (Summers, et al., 2015), (Bushnell, et al., 2014), (Yan, et
al., 2015), (Shames & Summers, 2015). Although optimization of submodular functions is difficult, submodularity allows for an approximation guarantee if one uses a simple greedy heuristic for their optimization (Nemhauser, et al., 1978).

The approaches presented in (Summers & Lygeros, 2014) assume that the system is initially stable with strict applications to linear time-invariant systems with the controllability Gramian and its derivatives as the control metric.

(Juang & Rodriguez, 1979) regarded the framework of the optimal control theory for systems subjected to initial disturbances and defined the cost function that is to be minimized as an indication, which relies on the controllability and disturbance sensitivity Gramians, so closed-form solution can be found. Many studies to qualify controllability of a system with external disturbances have been carried out (Stanley, et al., 1985), (Morari, 1983), (Morari, et al., 1985), (Shimizu & Matsubara, 1985), (Skogestad & Morari, 1987), (Hovd & Skogestad, 1992), (Luyben, 1988), (Cao & Rossiter, 1996), (Cao, et al., 1997), (Mirza & Niekerk, 1999). However, these works are mainly defined in frequency domain and hard to be applied to a system with unmatched disturbance on output, though they have an advantage to calculating specific frequency range. Naturally DOC for disturbance rejection in time domain is defined by (Kang & Park, 2009). (Lee & Park, 2014) extended this idea to unstable systems. The optimal actuator placement in controlling large structures in space subjected to initial disturbances was considered in (Viswanathan, et al., 1979), (Lindberg & Longman, n.d.), (Longman & Lindberg, 1986), (Viswanathan & Longman, 1983) and (Longman & Horta, 1989). (Viswanathan, et al., 1979) developed some numerical methods for generating the degree of controllability and evaluating the effectiveness of candidate actuator distributions. The method is shown to take on a relatively simple form when spacecraft modal coordinates are used.

(Hughes & Skelton, 1980) used the norms of the rows of the control location matrix as measures of the controllability of the individual modes. (Trajkov & Nestorvic, 2012) analyzed the placement based on controllability and observability criteria. The optimization was conducted based on H2-norm and controllability and observability gramian function which is dependent on vibrational modes. The structure model was designed using finite element and after the reduction of order process, optimization operations were done on the reduced model and the optimal place was suggested for the plate and cantilever beam.

(Vilnay, 1981) and (Ibidapo-Obe, 1985) proposed a method based on the interpretation of the functional relationship (transfer matrix/control influence matrix) between the actuators and modes of the structural system. It is shown that, from the form of the matrix, the controllability and observability of the system with respect to differing locations of the sensors and actuators can be established. In an attempt to consider the control effort in the actuator placement, (Abdei-Rohman, 1984) suggested that the optimal distribution of the sensors and the actuators is that which minimizes the observer gains and the control gains for the same level of control. However, the control gains do not represent completely the control effort, which depends greatly on the type and configuration of the control mechanism. (Sinha, et al., 2013) derived an analytical expression for the finite time controllability and observability Gramian for advection PDE. The selection criteria for the optimal location of actuators and sensors was proposed based on the maximization of Gramians. (Safizadeh, et al., 2010) studied the best position of piezoelectric actuator for active vibration control using controllability Gramian performance index and genetic algorithm. Their proposed method optimizes the controllable performance index in order to minimize the output energy of actuator for limited modes of frequency. In this method, the main responsibility is system’s controllability and expressing...
an optimum control input so that by applying forces on this optimal place of structure, system, can be damped. (Yang & Chen, 2010), (Yang & Chen, 2010 b) studied the optimal placement of piezoelectric sensor and actuator on a plate by increasing the controlling performance of system. In their work, the performances of two algorithms are used and compared to achieve better optimization result among variables of vibration response deformation precision, control energy and number of actuators. The utilized algorithms are GATSP (Genetic for the TSP) algorithm and HTTSP (Hopfield-Tank for the TSP) algorithm.
2.5. Summary

In this chapter, a brief survey based on some most popular optimization strategies that were used extensively in engineering literature for actuator/sensor placement is done. An overview of the problem of optimal actuator/sensor selection is presented and some energy-based criteria are described.

The originality of this thesis is to use the reviewed concepts within a systematic framework in order to select the best actuator/sensor location with which the optimum performance is achieved while the required control energy is minimized. A new evolutionary algorithm is developed and merged within the proposed methodologies for the purpose of the control energy optimisation for each actuator/sensor set. The next chapter discusses the controllability Gramian approach in more detail.
Chapter 3: Overview of Controllability Gramian and its Computation

3.1. Introduction

The principal goals of this chapter are to introduce the system properties of controllability and controllability Gramian (and of reachability, reachability Gramian), which play a central role in the study of energy consumption and input/output structure selection, topics that will be studied in the following chapters.

Controllability refers to the ability to manipulate the state by applying appropriate inputs (in particular, by steering the state vector from an initial vector value to a final vector value in finite time). Such is the case, for example, in satellite attitude control, where the satellite must change its orientation by manipulating its thrusters.

In Section 3.2, the concepts of reachability and reachability Gramian are introduced for linear continuous time-invariant systems.

Controllability and controllability Gramians are treated in detail in Section 3.3, where a novel approach of calculation of controllability Gramian is presented.
In the last section, we introduce a new approach for selecting input matrix $B$, which can assign a desired controllability Gramian to the LTI system.

### 3.2. Reachability and Reachability Gramian

Consider the LTI state space system:

\[
\dot{x} = Ax + Bu \\
y = Cx + Du
\]  

(3.1)

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$ and $u(t) \in \mathbb{R}^m$, Then the state at time $t_f$ is given by:

\[
x(t_f) = \Phi(t_f, t_0)x_0 + \int_{t_0}^{t_f} \Phi(t_f, \tau)Bu(\tau)d\tau
\]

(3.2)

where $\Phi(t, \tau)$ is the state transition matrix of the system, and $x(t_0) = x_0$ is the initial state.

In the time-invariant case considered here the state transition matrix can be defined as:

\[
\Phi(t, \tau) = \Phi(t - \tau, 0) = e^{A(t-\tau)}
\]

(3.3)

Here we are interested in using the input to transfer the state from $x_0$ to some other value $x(t_f) = x_f$ at some finite time $t_0 \leq t_f$. Note that here because of time invariance, only the difference $t_f - t_0 = T$, is relevant (rather than the individual times $t_0$ and $t_f$) and we can always take $t_0 = 0$ and $t_f = T$.

Considering (3.3), equation (3.2) can be written as:

\[
x_f = e^{AT}x_0 + \int_0^T e^{A(T-\tau)}Bu(\tau)d\tau
\]

(3.4)

Clearly, there exists $u(t), 0 \leq t \leq T$, that satisfies (3.4) if and only if such transfer of the state is possible. Also, if we define $\dot{x}_f = x_f - e^{AT}x_0$, it follows that the control input $u(t)$, which transfers the states from $x_0$ at time 0, to $x_f$ at time $T$ will also cause the state to reach $\dot{x}_f$ at $T$, starting from the origin at time zero.

**Definition (3.1):** A state $x_f$ is reachable if there exists an input $u(t), 0 \leq t \leq T$, that transfers the states $x(t)$ from the origin at 0 to $x_f$ over some finite time $t_f = T$. A reachable state $x_f$ is sometimes also called controllable from the origin (Antsaklis & Michel, 2007).

**Definition (3.2):** The set of all reachable states $\mathcal{R}_f$ is the reachable subspace of the LTI system (3.1), or of the pair $(A, B)$. Note that the set of all reachable states $x_f$ contains the origin and constitutes a linear subspace of the state space $(\mathbb{R}^n, \mathbb{R})$ (Antsaklis & Michel, 2007).
**Definition (3.3):** The LTI system (3.1), or the pair \((A, B)\) is completely reachable if every state is reachable, i.e., if \(\mathcal{R}_r = \mathbb{R}^n\) (Antsaklis & Michel, 2007).

**Definition (3.4):** For the linear continuous-time invariant system (3.1), the \(n \times n\) reachability Gramian is (Lewis, et al., 2012) (Antsaklis & Michel, 2007):

\[
W_r(0,T) = \int_0^T e^{(T-\tau)A}BB^*e^{(T-\tau)A'}\,d\tau 
\tag{3.5}
\]

**Lemma (3.1):** (Antsaklis & Michel, 2007) \(W_r\) is a symmetric, positive semidefinite matrix, for every \(T > 0\); i.e., \(W_r = W_r^*\) and \(W_r \geq 0\).

One can also show that the reachable subspace of LTI system (3.1) is exactly the range of the reachability Gramian \(W_r\). This is described as Theorem 3.1, but before that it is worth mentioning that the range of \(W_r(0,T)\), is independent of \(T\), it is the same for any finite \(T > 0\), and in particular, it is equal to the range of the controllability matrix \(C\), it is stated in Corollary (3.1) below:

**Corollary (3.1):** (Antsaklis & Michel, 2007) \(\mathcal{R}(W_r(0,T)) = \mathcal{R}(C)\) for every \(T > 0\).

(Note: In discrete case, the reachable (and/or controllable) subspace does not depend on the finite time only when the interval has length larger than or equal to \(n\) time steps, when \(0 < T < n\), we have \(\mathcal{R}(W_r(0,T)) \subset \mathcal{R}(C)\) where \(C = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}\) (Hespanha, 2009).)

**Theorem (3.1):** (Antsaklis & Michel, 2007) Consider the LTI system (3.1), and let the initial condition \(x(0) = 0\). There exists an input \(u\) that transfers the state \(x_0 = 0\) to \(x_f\) in finite time \(T > 0\) if and only if \(x_f \in \mathcal{R}(C)\), or equivalently, if and only if \(x_f \in \mathcal{R}(W_r(0,T))\). Thus, the reachable subspace \(\mathcal{R}_r = \mathcal{R}(C) = \mathcal{R}(W_r(0,T))\). Furthermore, an appropriate \(u\) that accomplishes this transfer in time \(T\) is given by:

\[
u(t) = B^*e^{(T-t)A'}\eta_t \tag{3.6}
\]

with \(\eta_t\) such that \(W_r(0,T)\eta_t = x_f\) and \(0 \leq t \leq T\).

Note that here there is no restriction on time \(T\), other than it has to be finite, so \(T\) can be as small as we wish; i.e., the transfer can in theory be accomplished in arbitrarily short time, though we will discuss later that this may be infeasible in practice as it may require control signals of arbitrarily large
magnitude or energy. Thus, it has its own costs, leading to a trade-off due to some physical restrictions, which appears in practical cases.

Next we introduce Corollary (3.2), which is a well-known result that has been used explicitly or implicitly by many authors (Farina & Rinaldi, 2011) (Antsaklis & Michel, 2007) (Guzzella, 2011) (Hespanha, 2009).

**Corollary (3.2):** The LTI system (3.1), or the pair \((A, B)\) is completely reachable, if and only if:

\[
\text{Rank } C = n \tag{3.7}
\]

or equivalently, if and only if:

\[
\text{Rank } W_r(0,T) = n \tag{3.8}
\]

for any finite \(T > 0\).

**Lemma (3.2):** (Fabrizio L. Cortesi, Tyler H. Summers, and John Lygero, 2014) Let the LTI system (3.1) or pair \((A, B)\) be completely reachable. Then, there exists an input that will transfer any state \(x_0\) to any other state \(x_f\) in some finite time \(T > 0\). One such input is given by:

\[
u(t) = B^* e^{(T-t)A} W_r^{-1}(0,T) [x_f - e^{AT} x_0] , \quad 0 \leq t \leq T \tag{3.9}
\]

In fact, this input control is not unique; there exist many various control inputs \(u(t)\) transferring the state \(x_0\) to the state \(x_f\), in time \(T\), but it can be shown that the control input \(u(t)\) given by (3.9), accomplishes this transfer while expending a minimum amount of energy, i.e. it minimizes the cost function \(\int_0^T \|u(\tau)\|^2 d\tau\), where \(\|u(t)\|^2 = u^*(t)u(t)\) denotes the Euclidean norm of \(u(t)\).

This result is discussed in full in subsequent chapters, where the effect of the terminal time \(T\) and the terminal state \(x_f\) on the value of the cost function will also be investigated.

### 3.3. Controllability Gramian and its Computation

**Definition (3.5):** A state \(x_0\) is controllable if there exists an input \(u(t)\), which transfers the LTI system (3.1) from \(x_0\) at \(t=0\) to the origin in some finite time \(T > 0\) (Mellodge & Kachroo, 2008) (Lewis, et al., 2003).

**Definition (3.6):** The set of all controllable states \(\mathcal{X}_c\) is the controllable subspace of the LTI system (3.1) or of the pair \((A, B)\) (Antsaklis & Michel, 2007).
Definition (3.7): The continuous linear time invariant system (3.1), or the pair \((A, B)\) is completely state controllable if every state is controllable, i.e., if \(\mathcal{R}_c = \mathbb{R}^n\), i.e. there exists an input \(u(t)\) that can transfer the states of the system from an arbitrary initial state \(x_0\) to a final state in a finite interval of time. This definition was given first by Kalman (Kalman, et al., 1963) and has been used by many authors; i.e. (Mellodge & Kachroo, 2008) (Owens, 2015).

We now establish the relationship between reachability and controllability for the LTI system (3.1). In view of (3.4), \(x_0\) is controllable when there exists \(u(t), 0 \leq t \leq T\), so that:

\[-e^{AT}x_0 = \int_0^T e^{(T-\tau)A} Bu(\tau) d\tau\]  

(3.10)

Or when \(e^{AT}x_0 \in \mathcal{R}(W_x(0,T))\), or according to Corollary (3.1),

\[e^{AT}x_0 \in \mathcal{R}(C)\]  

(3.11)

Recall that \(x_f\) is reachable when \(x_f \in \mathcal{R}(C)\). In fact, in the continuous-time case there is no distinction between reachability and controllability, in contrast to discrete-time case, since the state transition matrix is \(e^{AT}\) (rather than \(A^k\)) and is always invertible.

Lemma (3.3): If \(x \in \mathcal{R}(C)\), then \(Ax \in \mathcal{R}(C)\); i.e., the reachable subspace \(\mathcal{R}_r = \mathcal{R}(C)\) is A-invariant.

The proof is given in (Levine, 1996).

Corollary (3.3): the reachability subspace \(\mathcal{R}_r\) is the smallest A-invariant subspace containing \(\text{Im}[B]\).

For the proof see (Bittanti & Colaneri, 2009), p. 114.

Theorem (3.2): Consider the LTI system (3.1),

(i) A state \(x\) is reachable if and only if it is controllable.

(ii) \(\mathcal{R}_r = \mathcal{R}_c\).

(iii) The system (3.1), or the pair \((A, B)\), is completely state reachable if and only if it is completely state controllable.

For the proof see Theorem (5.20) in (Antsaklis & Michel, 2007). The following Theorem explains the connection of reachability with controllability in the LTI system. Note: Although this Theorem always holds for continuous LTI systems in the discrete time case the Theorem is only valid when the system matrix \(A\) is non-singular.

Definition (3.8): (Antsaklis & Michel, 2007) The controllability Gramian in the linear continuous time-invariant system (3.1) is \(n \times n\) matrix:
\[ W_c(0,T) = \int_0^T e^{AT} B B^T e^{A^T \tau} d\tau \]  \hspace{1cm} (3.12)

It can be verified directly that:

\[ W_c(0,T) = -e^{TA} W_c(0,T) e^{TA} \]  \hspace{1cm} (3.13)

The controllability Gramian is widely used to check the controllability of the linear dynamical systems.

\textbf{Proposition (3.1):} The LTI system (3.1) is controllable:

(i) If and only if $\text{rank}(W_c(0,T)) = n$ for some finite $T > 0$

(ii) If and only if $\text{rank}(C) = n$, where $C = [B, AB, ..., A^{n-1}B]$

(iii) If and only if $\text{rank}([\lambda_i I - A; B]) = n$, where $\lambda_i, i = 1, 2, ..., n$, is an eigenvalue of $A$.

The proof is given in (Klamka, 2009).

Controllability Gramian is related to the energy necessary to transfer the initial state to one final state (Moore, 1981). The Gramians are widely studied in the process of model order reduction (Moore, 1981), (S. Gugercin and A.C. Antoulas, 2003), (Antoulas, 2004).

In the next section, we introduce a novel approach for the computation of controllability Gramian, which presents many advantages not only in decreasing the complexity of the calculation but also in investigating the effect of model uncertainty or actuator noise. It also helps in input energy optimization problem when we are selecting the input matrix $B$.

\subsection*{3.3.1. A Novel Approach for Calculation of Controllability Gramian}

In this part, we introduce a new approach of calculating the controllability Gramian matrix $W_c(0,t_f)$.

The key innovative idea of this method is to split the controllability Gramian $W_c(0,t_f)$ into two distinct independent parts: one part deals with the input matrix $B$ and the other part is a function of the system matrix $A$. We will show that in the presence of uncertainty, and perturbation, this method helps to examine the effects of the matrices $A$ and $B$ on controllability Gramian (3.12).

By this method, the controllability Gramian can be calculated faster and easier, however it is noteworthy that in the case of large scale systems or multi input systems, this method may result in some large matrices while it still preserves its computation advantages. It is shown later that in the problem of input selection or finding the optimal placement of the actuators we are usually concerned with an eigenvalue optimization problem or a problem of maximization of the trace of controllability Gramian. This means that in most practical cases we do not need to calculate the entire controllability Gramian. Thus, especially in the case of large scale systems our method has significant advantages. This method has been developed by applying the concept of quadratic form matrices and sum of

**Note:** We denote by \( p(x) := p[x_1, \ldots, x_n] \) the ring of polynomials in \( n \) variables with coefficients in the field \( \mathbb{R} \).

**Proposition (3.2):** (Parrilo, 2003), (Parrilo, 2000) A multivariate polynomial \( p(x) \in \mathbb{R} \) in \( n \) variables and of degree 2d is a sum of squares (SOS) if and only if there exists a real symmetric positive semidefinite matrix \( Q \) such that:

\[
p(x) = z^T Q z
\]

where \( z \) is the vector of monomials of degree up to d, we call matrix \( Q \) the Gram matrix.

\[
Z = [1, x_1, x_2, \ldots, x_n, x_1 x_2, \ldots, x_n^d]^T
\]

Given a Gram matrix \( Q \), of \( \text{rank}(Q) = t \), we can construct polynomials \( h_1, h_2, \ldots, h_t \), such that:

\[
p(x) = \sum_{i=1}^{t} h_i^2
\]


The construction is given in the Appendix. We also refer reader to (Powers & Wörmann, 1998).

Thus, to find a representation of \( p(x) \) as a sum of squares, we just need to find a matrix \( Q \), which satisfies the above relation.

As the controllability Gramian definition (3.8) suggests, the most difficult part of the related computation, which gives rise to complexity, is related to the computation of the integral, in (3.12). thus, the simpler the matrix inside is to integrate the easier and faster are the calculations.

Inspired from quadratic form associated matrices, we introduce Theorem (3.3) below. First, we give two definitions and an intermediate research.

**Definition (3.9):** (Lam, 2005) Let \( A \) be an \( n \times n \) symmetric matrix with real entries, and let \( x \) be an \( n \) vector over the field \( \mathbb{R} \). The mapping \( Q : x \rightarrow x^T A x \) is the quadratic form defined by \( A \).

\[
Q = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j
\]

According to the definition above for any multivariate polynomial \( p(x) \in F \) in \( n \) variables and of degree 2d, there exist an associated symmetric matrix \( A \in F^{n \times n} \), with the entries:

\[
a_{ij} = \frac{\text{coefficient}(x_i x_j) + \text{coefficient}(x_j x_i)}{2}, 1 \leq i, j \leq n, i \neq j
\]

\[
a_{ii} = \text{coefficient}(x_i^2), 1 \leq i \leq n
\]
Now regarding Proposition (3.2), we define Lemma (3.4) below whose proof is omitted.

**Lemma (3.4):** A polynomial \( p(x) \in F \) is a sum of squares (SOS) polynomial if and only if its associated matrix \( A \in F^{n \times n} \) is a positive semi definite matrix.

**Definition (3.10):** (Loan, 2000) For any matrix \( A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{p \times q} \) the Kronecker product (i.e., the direct product or tensor product), denoted as \( A \otimes B \), is defined by:

\[
A \otimes B = \begin{bmatrix}
a_{11}B & a_{12}B & \cdots & a_{1m}B \\
a_{21}B & a_{22}B & \cdots & a_{2m}B \\
\vdots & \vdots & & \vdots \\
a_{n1}B & a_{n2}B & \cdots & a_{nm}B 
\end{bmatrix} \in \mathbb{R}^{np \times mq}
\]  

(3.17)

**Remark (3.1):** (Loan, 2000) The following basic properties easily follow:

i. \( (A \otimes B)^T = A^T \otimes B^T \)

ii. \( (A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \)

iii. \( (A \otimes B)(C \otimes D) = (AC \otimes BD) \)

iv. \( \text{trace}(A \otimes B) = \text{trace}(A) \text{trace}(B) \)

**Theorem (3.3):** For \( n \) dimensional LTI system (3.1) with \( m \) inputs, the controllability Gramian

\[
W_c(0,t_f) = \int_0^{t_f} e^{At}BB^Te^{A^Tt}d\tau ,
\]

can also be written as:

\[
W_c(0,t_f) = (I_n \otimes \Upsilon)^T Q(I_n \otimes \Upsilon)
\]

(3.18)

Where \( \Upsilon \) is the vectorised input matrix and \( Q \in \mathbb{R}^{m \times n \times m} \) is a real symmetric matrix.

**Proof:**

For simplicity and without loss of generality we consider the case of a 3-dimensional LTI system in (3.1), with two inputs. Let:

\[
e^{At} = \begin{bmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{bmatrix}
\]

(3.19)

According to (3.12), we can define the controllability Gramian as:
Clearly, we can simplify (3.20) as:

\[
(W_c(0,t_f))_{col,row} = \int_0^{t_f} \sum_{i=1}^{n} a_{col,i} \sum_{j=1}^{3} a_{row,j} B_{ij} d\tau \quad , \quad B_p = \sum_{i=1}^{2} b_{ij} b_{di} \quad , \quad col,row = 1,2,3
\]

(3.21)

In the same way, for a general \( n \) dimensional LTI system with \( m \) inputs the controllability Gramian matrix could be defined as:

\[
(W_c(0,t_f))_{col,row} = \int_0^{t_f} \sum_{i=1}^{n} a_{col,i} \sum_{j=1}^{n} a_{row,j} B_{ij} d\tau \quad , \quad B_p = \sum_{i=1}^{m} b_{ij} b_{di} \quad , \quad col,row = 1,\ldots,n
\]

(3.22)

Now using this fact that the \( b \)'s are independent monomials with respect to \( \tau \), one can readily consider \((W_c(0,t_f))_{col,row}, col,row = 1,\ldots,n\), as a quadratic polynomial, which can be written as:

\[
(W_c(0,t_f))_{col,row} = [b_{j1},\ldots,b_{jm}] \int_0^{t_f} \sum_{i=1}^{n} a_{col,i} \sum_{j=1}^{n} a_{row,j} \begin{bmatrix} b_{i1} \\
\vdots \\
 b_{im} \end{bmatrix} d\tau 
\]

\[ \quad , \quad col,row = 1,\ldots,n \]

(3.23)

Then by defining the monomial matrix \( Y = [b_{j1},b_{j2},\ldots,b_{j3},b_{j4},b_{j5},\ldots,b_{jn}]^T \) the controllability Gramian matrix could be defined as:

\[
(W_c(0,t_f))_{col,row} = Y^n \int_0^{t_f} \sum_{i=1}^{n} a_{col,i} \sum_{j=1}^{n} a_{row,j} d\tau T
\]

(3.24)

and the Theorem (3.3) is proved.
Example (3.1): Assume an LTI system as:

\[
A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\]

A practical example of this system is the linearized pitch motion equation of a rigid spacecraft or a lunch vehicle \( I \ddot{\theta} = -TL \sin \delta \) where \( T \) is the rocket thrust, \( L \) is the distance between the center of mass and thrust, \( \delta \) is the gimbal angle, \( \theta \) is the pitch angle and \( I \) is the pitch moment of inertia.

Consider the transfer of the states from the origin to a sphere of the radius one and center the origin within one second, i.e. \( t_f = 1 \). In this case, according to Theorem (3.3) we have:

\[
W_c(0, t_f) = \int_0^{t_f} e^{A \tau} BB^*e^{A^* \tau} d\tau
\]

Let us define:

\[
e^{A \tau} = e^{\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \tau} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix}
\]

Then we have:

\[
W_c(0, 1) = \int_0^1 \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} d\tau
\]

\[
= \int_0^1 \begin{bmatrix} a_{11} b_{11}^2 + a_{12} b_{11} b_{21} + a_{21} b_{11} b_{21} + a_{22} b_{21}^2 \\ a_{21} b_{11}^2 + a_{22} b_{11} b_{21} + a_{22} b_{22} b_{21} + a_{22} b_{21}^2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} d\tau
\]

Thus:

\[
W_c(0, 1) = \int_0^1 \begin{bmatrix} a_{11} b_{11}^2 + a_{12} b_{11} b_{21} + a_{12} a_{11} b_{11} b_{21} + a_{12} a_{12} b_{11} b_{21} + a_{22} b_{21}^2 \\ a_{21} b_{11}^2 + a_{22} b_{11} b_{21} + a_{22} a_{11} b_{11} b_{21} + a_{22} a_{12} b_{11} b_{21} + a_{22} b_{22} b_{21} + a_{22} b_{21}^2 \end{bmatrix} d\tau
\]

\[
= \begin{bmatrix} b_{11} & b_{21} & 0 & 0 \\ 0 & 0 & b_{11} & b_{21} \end{bmatrix} \int_0^1 \begin{bmatrix} a_{11}^2 & a_{12} a_{11} & a_{12} a_{12} \\ a_{21} a_{11} & a_{12}^2 & a_{12} a_{22} \\ a_{21} a_{21} & a_{22} a_{11} & a_{22}^2 \\ a_{21} a_{12} & a_{22} a_{12} & a_{22} a_{22} \end{bmatrix} d\tau
\]

Therefore considering:

\[
\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = e^{\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \tau} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\]

we obtain:
\[
W_c(0,1) = \begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\int_0^1 \begin{bmatrix}
1 & \tau & 0 & 0 \\
\tau & \tau^2 & 1 & \tau \\
0 & 1 & 0 & 0 \\
0 & \tau & 0 & 1
\end{bmatrix} d\tau \begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

Now by calculating the integral we get:

\[
W_c(0,1) = \begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0.5 & 0 & 0 \\
0.5 & 0.333 & 1 & 0.5 \\
0 & 1 & 0 & 0 \\
0 & 0.5 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix} = \begin{bmatrix}
0.333 & 0.5 \\
0.5 & 1
\end{bmatrix}
\]

**Definition (3.11):** A symmetric polynomial matrix \(P(x) \in \mathbb{R}[x]^{m \times m}, x \in \mathbb{R}^n\) is SOS matrix if there exists a polynomial matrix \(M(x) \in \mathbb{R}[x]^{s \times m}\) for some \(s \in \mathbb{N}\), such that \(P(x) = M^T(x)M(x)\).

If \(m = 1\) we have a standard SOS polynomial. Also, when \(P\) is a constant matrix, then the condition simply states that \(P\) is positive semidefinite matrix. Thus, SOS matrices generalize both positive semidefinite (constant) matrices and SOS polynomials. (Blekherman, et al., 2013)

**Remark (3.2):** The controllability Gramian \(W_c(0, t_f)\) is a SOS matrix in input matrix \(B\) and hence can be written as \(W_c(0, t_f) = Q^T(x)Q(x), x = [1; \text{vec}(B_{n,m})] \in \mathbb{R}^{(mn+1)c}\).

In the sequel, we will see that remark (3.2) simplifies energy minimization and optimal input selection achieved by maximizing the eigenvalues of the controllability Gramian, which are generally nonlinear functions of its entries.

### 3.3.2. Controllability Gramian Assignment with Respect to Control Variables

One motivating factor behind this section is the fact that many system properties are related to the controllability Gramian. Moreover, the eigenvalues and their corresponding eigenvectors of the controllability Gramian determine the speed and direction of energy dissipation, a problem, which will be studied in more detail in the following chapters. The basic idea behind this part is to find the optimal control input matrix that assigns a specific controllability Gramian to a LTI system.

In particular, we propose a solution to this assignability problem and specify the best possible set of input matrices \(B\) to steer the system to a target state, under energy limitations imposed by a given controllability Gramian.

**Theorem (3.4):** In the LTI system (3.1) a specified controllability matrix \(W_c(0, t_f)\) is assignable if and only if there exist an input matrix \(B\) such that:
\[ \begin{align*}
\Phi & = \int_0^T \left( e^{At} \otimes e^{At} \right) d\tau = \int_0^T \left( e^{(A \otimes A)t} \right) d\tau \\
\end{align*} \tag{3.25} \]

Where \( m \) is the number of control inputs.

**Proof:**

Consider the general definition of controllability Gramian defined in this chapter:

\[ W_c(0,t_f) = \int_0^t e^{At} BB^T e^{A^T \tau} d\tau \]

Now based on the Kronecker properties (Loan, 2000) (Jain, 1989) (Petersen & Pedersen, 2012) (Brookes, 2004) we have:

\[ \text{vec}(A_{nxm} X_{npx} B_{pexo}) = (B^T \otimes A)_{qxpm} \text{vec}(X)_{pmo} \] \tag{3.26} \]

Then we obtain:

\[ \begin{align*}
\Phi & = \int_0^T \left( e^{At} \otimes e^{At} \right) d\tau \\
\end{align*} \tag{3.27} \]

and the proof is completed.

An important consequence of the above Theorem is that it relates the problem of controllability Gramian assignment to a linear system equations where the matrix of coefficients is itself a Kronecker product. Note that the linear system \( Ax = b \) is efficiently solvable if \( A \) is a Kronecker product.

For the \( n \) dimensional system \( e^{At} \in \mathbb{R}^{nxn} \), and \( \text{vec}(BB^T) \) can be obtained in \( O(n^3) \) flops via the LU factorization of \( e^{At} \). Without the exploitation of structure, an \( n^2 \times n^2 \) system would normally require \( O(n^6) \) flops to solve (Loan, 2000).

Now considering the fact that both matrices \( W_c(0,t_f) \) and \( BB^T \) are symmetric, one may readily simplify (3.27) as:

\[ \int_0^T \left( e^{At} \otimes e^{At} \right) d\tau P_{BB^T} \text{vec}_{\text{sym}}(BB^T) = P_{W_c(0,t_f)} \text{vec}_{\text{sym}}(W_c(0,t_f)) \] \tag{3.28} \]

where \( P_{BB^T} \in \{0,1\}^{n^2 \times n(n+1)/2}, P_{W_c(0,t_f)} \in \{0,1\}^{n^2 \times n(n+1)/2} \) denote appropriate permutation matrices.
are obtained by stacking the sub-diagonal columns of $BB^T$ and $W_c(0,t_f)$ respectively.

This results in a system of linear equations with the special coefficient matrix:

$$
\int_0^{t_f} (e^{At} \otimes e^{B^T \tau}) d\tau P_{BB^T}^{-1}
$$

consisting of $\frac{n(n+1)}{2}$ equations and $\frac{n(n+1)}{2}$ variables, which can be solved almost twice as fast as a system without the symmetry property.

Assume that we are given a designed controllability Gramian according to the energy restrictions on the system. Then the assignability problem is defined as selecting the proper set of input matrices $B$ such that (3.28) is satisfied, which assigns the given controllability Gramian to the system. To do so we just need easily to solve the linear equations based on $B$.

The difficulty of this approach is that the desired Gramian may not be assignable, i.e. there may be no solution to the linear equations. In this case, one must choose an input matrix $B$ from the set which will closely approximate the desired Gramian. One approach is to minimize the matrix norm of the difference between the desired controllability Gramian and the assignable approximation to this matrix. According to (Wicks & Decarlo, 1990) it is not clear, however, that minimizing some norm of the error matrix implies a good approximation to the eigenvalues and eigenvectors of the desired matrix, nor is it clear that a large error necessarily means a bad approximation. An alternative method could be to approximate the controllability Gramian using least squares by minimizing the Frobenius norm of the error matrix over our variables, which are the elements of $vec_{sym}(BB^T)$.

The following example clarifies the method.

**Example (3.2):** Consider the single input system below:

$$
A = \begin{bmatrix} -2 & -1 \\ 2 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 2 \\ -1 \end{bmatrix}
$$

using Lyapunov equations or through the method proposed in this chapter we obtain

$$
W_c(0,\infty) = \begin{bmatrix} 1.1250 & -0.2500 \\ -0.2500 & 0.7500 \end{bmatrix}.
$$

Now assume that we are looking for the input matrix $B = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}$ such that it assigns the controllability Gramian $W_c(0,\infty) = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$ to the system.

According to the above Theorem we have:
\[ \int_0^t \left( e^{At} \otimes e^{At} \right) d \tau P_{bb^2} \text{vec}_{\text{sym}}(BB^T) = P_{w(0,t_f)} \text{vec}_{\text{sym}}(W_c(0,t_f)) \]

\[\begin{bmatrix}
1/4 & 0 & 0 & 1/8 \\
0 & 1/4 & -1/4 & -1/4 \\
0 & -1/4 & 1/4 & -1/4 \\
1/2 & 1/2 & 1/2 & 3/4
\end{bmatrix} \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix} = \begin{bmatrix}
W_1 \\
W_2 \\
W_3
\end{bmatrix}
\]

As explained above, due to symmetry the 2nd and 3rd rows are identical and one of them can be removed. Then the set of all assignable controllability Gramians in this system is determined by \( S_w \) and clearly the given controllability Gramian is not assignable to this system. This we must choose a matrix from the set \( S_w \) which will closely approximate the desired controllability Gramian.

\[ S_w = \{ W_1, W_2, W_3 : B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \in \mathbb{R}^{2\times1} \} \]

Minimizing the Frobenius norm of the difference between the desired matrix and the assignable approximation to this matrix results in:

\[ B = \begin{bmatrix} 3.1779 \\ -1.2195 \end{bmatrix}, \quad \bar{W}_c = \begin{bmatrix} 2.7106 & -0.3718 \\ -0.3718 & 2.2894 \end{bmatrix} \]

Note that minimizing the spectral norm in this case also leads to the same result.

The eigenvalues of approximated controllability Gramian are: \( \lambda_1 = 2.9257, \lambda_2 = 2.0743 \) which are close to the eigenvalues of the desired Gramian.

\[ \square \]

Next consider a similar example to the multi input case.

**Example (3.3):** Consider the 3-dimensional system with 2-inputs:

\[ A = \begin{bmatrix} 0 & 0 & 2 \\
0 & 0 & 4 \\
0 & 3 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \]

The eigenvalues of this system are: \( \lambda_1 = -3, \lambda_2 = 0, \lambda_3 = 4 \). Hence the system is unstable and \( W_c(0,t_f) \) becomes unbounded as \( t_f \rightarrow \infty \) so the Lyapunov equations do not produce proper solution for the controllability Gramian. Hence, we apply the proposed method for \( t_f = 1 \). In this case:

\[ W_c(0,1) = \begin{bmatrix}
21.7711 & 69.3133 & 34.2409 \\
21.3257 & 34.2409 & 67.8167
\end{bmatrix} \]

with the eigenvalues: \( \lambda_1 = 15.7912, \lambda_2 = 34.3159, \lambda_3 = 113.4868 \).
Next, we define the problem of finding the input matrix:

\[
B = \begin{bmatrix}
    b_{11} & b_{12} \\
    b_{21} & b_{22} \\
    b_{31} & b_{32}
\end{bmatrix} \in \mathbb{R}^{3 \times 2}
\]

which assigns the controllability Gramian:

\[
W_c(0,1) = \begin{bmatrix}
    3 & 0 & 0 \\
    0 & 2 & 0 \\
    0 & 0 & 1
\end{bmatrix}.
\]

Taking a similar approach to the single input case, it easily follows that the given controllability Gramian is not in the set of all assignable Gramians \( S_W \) and therefore we need to use an approximation.

By minimizing the spectral norm, we obtain:

\[
B = \begin{bmatrix}
    -1.5844 & -0.0243 \\
    0.0004 & 2.3652 \\
    -0.0235 & -1.6908
\end{bmatrix}, \quad \tilde{W}_c = \begin{bmatrix}
    3.1329 & -0.2907 & 0.3252 \\
    -0.2907 & 2.1759 & 0.2827 \\
    0.3252 & 0.2827 & 1.1245
\end{bmatrix}
\]

The eigenvalues of the approximating matrix are: \( \lambda_1 = 3.2422, \lambda_2 = 2.2082, \lambda_3 = 0.9829 \)

If the Frobenius norm of the difference matrix is minimized, we get the optimal input matrix as:

\[
B = \begin{bmatrix}
    1.7100 & -0.4690 \\
    0.0018 & -2.4668 \\
    0 & 1.7750
\end{bmatrix}
\]

The corresponding approximate Gramian and its eigenvalues are:

\[
\tilde{W}_c = \begin{bmatrix}
    3.0019 & 0.0015 & 0.0165 \\
    0.0015 & 2.0198 & -0.0022 \\
    0.0165 & -0.0022 & 0.9671
\end{bmatrix}
\]

\( \lambda_1 = 3.0020, \lambda_2 = 2.0198, \lambda_3 = 0.9670 \)
3.4. Summary

In this chapter, we defined the reachability/controllability Gramian for continuous-time LTI system and introduced some of their fundamental properties of them.

Quadratic forms and sum of square polynomials were introduced as well, which led to a new approach for controllability Gramian calculation. In contrast to Lyapunov equations, our method does not imply any restriction on the stability of the system, and this is a noticeable advantage.

Another advantage of our method is that it divides the controllability Gramian matrix into two distinct parts: one part linked to the system matrix $A$ and the other part is a tensor product of the vectorised input matrix $B$. Some applications of this method can be demonstrated in perturbation analysis and robust control design. Though the eigenvalues of a matrix are nonlinear functions of its entries it will be attempted to optimize the eigenvalues of the controllability matrix over the elements of the input matrix.

In the last part of this chapter, the controllability Gramian assignment problem was solved by introducing a new approach for selecting the proper input matrix $B$. This approach can be used for any stable/unstable system with one or more control inputs.

In control design, it is very important to select the proper inputs, to construct an effective control system and to satisfy various constraints and limitations. The controllability assignment is a core problem, which can be considered wherever we have constraints on the input energy, or when the robustness properties of linear system need to be considered.

In the next chapter, the minimum of input energy is introduced as a quadratic form of the inverse of the controllability Gramian. Then we will discuss how the terminal time and the final states may affect the value of the energy. The various energy levels will be depicted in the state space.
Chapter 4: Minimum Input Energy

4.1. Introduction

In this chapter, we develop the fundamentals for selection of input (and by duality output) structure based on the energy type criteria. The fundamental question is the evaluation of the required energy for the transfer of the origin to a given state in the state space within some given final time when the input structure is given. The overall aim is to use these preliminary results to examine in the following parts of the thesis the selection of the input structure under the minimum energy requirement for the worst state. Thus, we introduce first the problem of minimum input energy and then we will discuss the important related factors which can influence the value of this minimum energy. These results lead to the definition of the energy levels in the state space which can be used as an instrument for input structure selection.

The development of the material involves the definition first of the minimal energy and this follows by examining the factors influencing the minimum energy requirements. This involves the influence of the terminal state factor, as well as the role of the required final time that has to be specified for the state transfer. The computations of energy are based on the controllability Gramian and thus we examine separately the case of stable and unstable systems. The essence of our investigation is comparing the minimum of minimum energy \( E_{min} \) transferring the states of a given system from the
origin to a terminal state $x_f$, such that $\|x_f\| = R (= 1)$ within a fixed terminal time $t_f$. In fact, we are interested to determine the minimum input energy required to transfer the states of the LTI controllable system from the origin to the states defined by the distance $R$ away from the origin. We compare the maximum of minimum energy $E_{\text{min}}^{\text{max}}$, required to transfer the states of a given system from the origin to a terminal state $x_f$, $\|x_f\| = R$, where $R$ is variable (within a fixed terminal time $t_f$), that can guaranty to reach a state defined $R$ away from the origin, ignoring the direction of the specific states. This leads to the definition of energy levels in the state space characterizing the maximum of minimum input energy required for transfer to a given distance. This reveals the presence of the energy stratification of the state space.

### 4.2. Minimum Input Energy

Consider a controllable SISO LTI system, we are interested in transferring the states from the origin to some point on the surface of a sphere with the radius $R$. The state-space realization of the system is:

$$\begin{align*}
\dot{x} &= Ax + Bu \\
y &= Cx + Du
\end{align*}$$

(A $\in \mathbb{R}^{n \times n}$, B $\in \mathbb{R}^{n \times m}$, C $\in \mathbb{R}^{1 \times n}$, D $\in \mathbb{R}^{1 \times m}$) (4.1)

The energy dissipated by the controller for such transfer is given as the cost function below:

$$E = \|u(t)\|_{[0,t_f]}^2 = \int_0^{t_f} u^T(t)u(t)dt$$

(4.2)

We are interested to minimize this cost function, so that the states transfer to the desired terminal, within the defined control time, by expending the least possible input energy.

![Figure 4.1](image)

As stated in literature, (VanderVelde & Carignan, 1982), (Georges, 1995), (Toan & Georges, 2007) the minimum input energy to transfer the states for the system (4.1) from the origin to the terminal state $x_f$, such that $\|x_f\| = R (= 1)$, is given as:

$$E_{\text{min}} = \min \|u(t)\|_{[0,t_f]}^2 = \min \int_0^{t_f} u^T(t)u(t)dt = x_{fW_{c}^{-1}}^T(0,t_f)x_f$$

(4.3)
Note that the system is assumed controllable and so the indicated inverse of the controllability Gramian is well-defined. The corresponding (unique) optimal control input is:

\[ u_{E_{\text{min}}} (t) = u_0(t) = B^T e^{A^T (t_f - t)} W_c^{-1} (0, t_f) x_f \]  

(4.4)

These results are defined in the Theorem (4.1) below (Fairman, 1998):

**Theorem (4.1):** In the controllable LTI system (4.1), the minimum energy required to transfer the initial states from the origin to the terminal states located on a sphere around the origin with the radius \( R \) within the terminal time \( t_f \) is \( E_{\text{min}} = x_f^T W_c^{-1} (0, t_f) x_f \) and the required control input for this optimal transfer is:

\[ u_{E_{\text{min}}} (t) = u_0(t) = B^T e^{A^T (t_f - t)} W_c^{-1} (0, t_f) x_f . \]

The proof can be found in (Fairman, 1998). In different stages of control design, the concept of minimum input energy, may play an important role, e.g., it is important for an airplane or a satellite to complete a defined journey using the minimum fuel, in a smart house minimization of input energy as the cost function usually defines the aim of the design, etc.

As the Figure (4.1) illustrates, to transfer the states of a controllable LTI system from the origin to a terminal point on the surface of a sphere with the radius \( R \), within a terminal time \( t_f \) there are many different control inputs in various directions, the Theorem above states that only one of these inputs in a unique direction leads to the minimum energy dissipation. In control design and input selection, it is also important to investigate the factors, which can influence the value of minimum energy corresponding to this optimal direction. This will be discussed in detail in a subsequent section.

### 4.3. Influence of Different Terms on the Value of Minimum Input Energy

In this part, some significant variables in the value of the minimum input energy are investigated.

Considering the definition of controllability Gramian in (4.5):

\[ W_c (0, t_f) = \int_0^{t_f} e^{A \tau} B B^T e^{A^T \tau} d\tau \]  

(4.5)

As equation (4.3) suggests, in an LTI system, when the matrices A and B are fixed and the system is controllable, two important factors may change the value of minimum input energy:

i. The terminal time: \( t_f \)

ii. The terminal state: \( x_f \)

It is, of course, clear that if \( W_c (0, t_f) \) is nearly singular then a very large energy input is needed to reach certain states aligned with the eigenspace of the controllability Gramian corresponding to its "small" eigenvalues (in this case, the system is hardly controllable.)
### 4.3.1 Terminal Time $t_f$

In this part, for the fixed given LTI system \((4.1)\), we consider various transition times to transfer the initial states from the origin to the fixed terminal states \(\hat{x}_f\) such that \(\|\hat{x}_f\| = R (= 1)\), using the minimum energy. In other words, we define the minimum input energy as a function of the variable terminal time.

The results show that the value of the minimum input energy decreases as the terminal time increases, which means that if we give more time to the system it can reach the final states with less input energy requirement.

It is presented in the following Theorem.

**Theorem (4.2):** In the controllable LTI system \((4.1)\), where the system is fixed, i.e. \(A\) and \(B\) are invariable matrices, the minimum energy required to transfer the states from the origin to the defined terminal states \(\hat{x}_f, \|\hat{x}_f\| = R (= 1)\), decreases as the terminal time \(t_f\) increases, however the rate of this change is not fixed and could be shown as a concave up decreasing graph i.e. the change of the required energy decreases over the terminal time.

**Proof:**

As stated in \((4.3)\) the minimum input energy for the LTI controllable system \((4.1)\) is defined as follows:

\[ E_{\text{min}} = \hat{x}_f^T W_c^{-1}(0,t_f) \hat{x}_f \]

By differentiating both sides of the equation above with respect to the terminal time, we get:

\[
\frac{\partial E_{\text{min}}}{\partial t_f} = \hat{x}_f^T \frac{\partial W_c^{-1}(0,t_f)}{\partial t_f} \hat{x}_f
\]  \hspace{1cm} (4.6)

Then using identity:

\[
\frac{\partial A^{-1}(x)}{\partial x} = -A^{-1} \frac{\partial A(x)}{\partial x} A^{-1}
\]  \hspace{1cm} (4.7)

Equation (4.6) gives:

\[
\frac{\partial E_{\text{min}}}{\partial t_f} = -\hat{x}_f^T(t) W_c^{-1}(0,t_f) \frac{\partial W_c(0,t_f)}{\partial t_f} W_c^{-1}(0,t_f) \hat{x}_f(t)
\]  \hspace{1cm} (4.8)

Hence the minimum input energy decreases as the final time increases.

In other words, the set of states reachable with a cost less than 1 (or any given value) grows as \(t_f\) increases.
In this part, we may consider two different cases: In the first case, all eigenvalues of matrix $A$ are located in the open left half plane (OLHP), which means the LTI system is asymptotically stable, while in the second case an unstable system is considered.

We are interested to know how the value of minimum energy is affected by the terminal time $t_f$, in each of the above cases.

### 4.3.1.1 The Stable System

For the stable case, we have the following result (Yoshizawa, 2012):

**Theorem (4.3):** Consider the fixed controllable LTI system (4.1). If the system is asymptotically stable, then we can increase the terminal time to $\infty$, i.e. $t_f \to \infty$, in this case, the controllability Gramian could be defined as the unique solution of the Lyapunov equation (4.9), and the minimum energy required to move the states from the origin to the defined terminal states $x_f$, $\|x_f\| = R (= 1)$ is also minimized.

$$AW_c(0, \infty) + W_c(0, \infty)A^T = -BB^T$$  \hspace{1cm} (4.9)

**Proof:**

For an asymptotically stable system, whose state matrix $A$ has all its eigenvalues in the OLHP, then, $e^{At} \to 0$ as $t \to \infty$. Thus, there are constants, $M, \sigma > 0$ such that:

$$\|e^{At}\| = \|e^{At}\| \leq Me^{-\sigma t}, \forall t > 0$$  \hspace{1cm} (4.10)

Then:

$$\|W_c(0, t_f)\| = \|\int_0^{t_f} e^{At}BB^Te^{At^T}d\tau\| \leq \int_0^{t_f} \|e^{At}\| \|BB^T\| \|e^{At^T}\|d\tau$$

$$\leq M^2 \|BB^T\| \int_0^{t_f} e^{-2\sigma t}d\tau = M^2 \|BB^T\|(1 - e^{-2\sigma t_f})/2\sigma$$  \hspace{1cm} (4.11)

$$\leq \frac{1}{2\sigma} M^2 \|BB^T\|$$

This implies:

$$\lim_{t_f \to \infty} W_c(0, t_f) = W_c(0, \infty)$$  \hspace{1cm} (4.12)

and the controllability Gramian matrix $W_c(0, \infty)$ can be obtained as the steady state solution of the equation below:
By setting $\dot{W}_c(0,t_f) = 0$, we then have that the controllability Gramian matrix $W_c(0,\infty)$ satisfies the matrix Lyapunov equation:

\[
AW_c(0,\infty) + W_c(0,\infty)A^T = -BB^T
\]

(4.14)

which means that for a stable system the minimum input energy can be calculated in the case $t_f \to \infty$ via the controllability Gramian which in this case is the unique symmetric positive-definite solution of the algebraic Lyapunov equation (4.14). However, in a stable system we have $W_c^{-1}(0,t_f) > 0$, which implies we cannot get anywhere for free.

**Example (4.1):** An example of a classical linear system is the mass-spring-damper system. Newton’s laws of motion for an object moving horizontally on a plane and attached to a wall via a spring and damper gives:

\[
m\ddot{y}(t) = u(t) - b\dot{y}(t) - ky(t)
\]

(4.15)

where $y(t)$ is the position of the center of mass of the object, $\dot{y}(t)$ its velocity, $\ddot{y}(t)$ its acceleration, $u(t)$ is the applied force. Further $b$ is the viscous friction coefficient, $k$ is the spring constant and $m$ is the mass of the moving object. Assume the origin as the initial condition, i.e. $y(t) = \dot{y}(t) = 0$, if we define the target states as $y(t_f) = 1$, $\dot{y}(t_f) = 0$. We are interested in the minimum energy dissipated by the system considering various terminal times.

![Mass-spring-damper system](image.png)

Figure (4.2): Mass-spring-damper system

The state equation is:

\[
\begin{align*}
\dot{x} &= \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix} x + \begin{bmatrix} 0 \\ 1/m \end{bmatrix} u \\
y &= \begin{bmatrix} 1 & 0 \end{bmatrix} x
\end{align*}
\]

(4.16)
where \( x_1(t) \) represents the position of the object, and \( x_2(t) \) its velocity.

For this example, the controllability test shows that such LTI system is controllable for all \( b \) and \( m \):

\[
\begin{bmatrix} B \\ AB \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & -\frac{k}{m} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & -\frac{b}{m} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{m} \\ 1 & -\frac{b}{m^2} \end{bmatrix}
\] (4.17)

Clearly, the system is stable for all values of \( b \), \( k \) and \( m \), since all eigenvalues of \( A \) are located in OLHP:

\[
\det(\lambda I - A) = \det \begin{bmatrix} \lambda & -1 \\ \frac{k}{m} & \frac{b}{m} \end{bmatrix} = \lambda^2 + \lambda \frac{b}{m} - \frac{k}{m} = 0
\] (4.18)

\[
\Rightarrow \lambda = \text{real}(\frac{-b}{2m} \pm \sqrt{\frac{(b}{2m})^2 + \frac{k}{m}}) < 0
\]

Suppose that \( m=1 \text{kg}, k=1 \text{N/m} \) and \( b=0.2 \text{Ns/m} \). Considering a range of terminal times \( 0.5 \leq t_f \leq 10 \), the corresponding value of the minimum input energy was calculated and is illustrated in Figure (4.3). The Figure demonstrates that the minimum energy required to move the mass from the initial condition to the terminal state \( y(t_f) = 1, \dot{y}(t_f) = 0 \) decreases from 91.6790 to 0.4694, as the terminal time \( t_f \) increases from 0.5 to 10.

![Minimum input energy for different terminal time](image)

Figure (4.3): \( E_{\text{min}} \) as a function of variable terminal time where \( t_f \in [0.5, 10] \)
Figure (4.4) shows that the minimum energy converges as $t_f \to \infty$. The limiting energy can be calculated via the steady-state controllability Gramian which is the solution of the algebraic Lyapunov equation (4.14).

$$AW(0, \infty) + W_0(0, \infty)A^T = -bb^T$$

$$W_0(0, \infty) = \int_0^\infty e^{A\tau}bb^Te^{A^T}\tau = \begin{bmatrix} 2.5 & 0 \\ 0 & 2.5 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 0 & 1 \\ -1 & -0.2 \end{bmatrix} \begin{bmatrix} 2.5 & 0 \\ 0 & 2.5 \end{bmatrix} + \begin{bmatrix} 2.5 & 0 \\ 0 & 2.5 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & -0.2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$$

(4.19)

$$E_{\text{min}} = x_f^TW_0^{-1}(0, \infty)x_f = \begin{bmatrix} 1 & 0 \\ 0 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 0.4$$

In fact, 0.4 is the minimum possible energy, by which this system can move from the origin to $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and is achieved when $t_f \to \infty$.

All the results of this example, supports Theorem (4.3).

Figure (4.4): $E_{\text{min}}$ as a function of terminal time variable, $0 < t_f \leq \infty$
4.3.1.2 Unstable System

If the system is unstable \( e^{At}B \) becomes unbounded as \( t = t_f \to \infty \). Thus, the controllability Gramian \( W_c(0, \infty) \) also becomes unbounded which means that for some \( x_f \), \( \lim_{t_f \to \infty} E_{\min} \to 0 \). This means that some directions do not require any cost of control. In other words, for an unstable system, inverse of controllability Gramian \( W_c^{-1}(0, t_f) \) can have a nonzero null space \( W_c^{-1}(0, t_f)x_f = 0 \) for some \( x_f \neq 0 \), which means that the system can get to \( x_f \) using arbitrarily small input energy. The input just gives a little kick to the initial state and then giving the system enough time, instability carries it out to the desired final state \( x_f \).

Example (4.2): As an example, assume the linear model of a Furuta pendulum (Guzman, et al., 2016), described below:

![Figure (4.5): The schematic picture of Furuta pendulum](image)

The model equations are:

\[
\begin{align*}
\mathbf{x}'(t) &= \begin{bmatrix}
0 & 1 & 0 & 0 \\
\frac{-gm^2l_1^2}{I_0(J_1 + m_1l_1^2)} & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
\frac{(I_0 + m_1l_1^2)m_1l_1g}{I_0(J_1 + m_1l_1^2)} & 0 & 0 & 0 \\
\end{bmatrix} \mathbf{x}(t) + \begin{bmatrix}
0 \\
\frac{(J_1 + m_1l_1^2)}{I_0(J_1 + m_1l_1^2) + J_1l_1^2} \\
0 \\
\frac{m_1l_1g}{I_0(J_1 + m_1l_1^2) + J_1l_1^2} \\
\end{bmatrix} u(t) \\
\end{align*}
\]

where, \( x_1 = \theta \) is the angle of the pendulum, \( x_2 = \dot{\theta} \) is the angular velocity of the pendulum, \( x_3 = \phi \) is the angle of the arm carrying the pendulum, and \( x_4 = \dot{\phi} \) is the angular velocity of the carrying arm. Now, by defining:
The linear model of the system reduces to:

\[
\alpha = \frac{-gm_L^2 I_L^2}{I_0(J_1 + m_L I_L^2) + J m_L I_0^2}, \quad \gamma = \frac{(I_0 + m_L I_L^2)m_L g}{I_0(J_1 + m_L I_L^2) + J m_L I_0^2}
\]

\[
\beta = \frac{(J_1 + m_L I_L^2)}{I_0(J_1 + m_L I_L^2) + J m_L I_0^2}, \quad \epsilon = \frac{-m_L}{I_0(J_1 + m_L I_L^2) + J m_L I_0^2}
\]

The linear model of the system reduces to:

\[
\dot{x} = \begin{bmatrix}
0 & 1 & 0 & 0 \\
\alpha & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
\gamma & 0 & 0 & 0 \\
\end{bmatrix} x + \begin{bmatrix}
0 \\
\beta \\
0 \\
\epsilon \\
\end{bmatrix} u
\]

Assume the system is moving from the initial states \([0 \ 0 \ 0 \ 0]^T\) to \([0 \ 0 \ 1 \ 0]^T\).

Figure (4.6) illustrates the changes of the minimum input energy for various terminal times \(0 < t_f < \infty\) by substituting the following values:

\[
\dot{x} = \begin{bmatrix}
0 & 1 & 0 & 0 \\
31.32 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
-0.5584 & 0 & 0 & 0 \\
\end{bmatrix} x + \begin{bmatrix}
0 \\
-71.23 \\
0 \\
191.2 \\
\end{bmatrix} u
\]

The system above is clearly an unstable, controllable system.

As in can be seen in Figure (4.6), the minimum energy is decreasing as the terminal time increases however because of instability, the value of the minimum energy is changing really close to zero, even if the terminal time is small. It supports the results discussed in this section as for unstable systems.
4.3.2 Terminal State $x_f$

In this part, we assume a fixed given LTI system described in (4.1) and we are interested in transferring the initial states from the origin to the variable terminal states $x_f$ such that ($\|x_f\| = R$ (= 1)), within a fixed terminal time $t_f=T$.

Here the terminal state $x_f$ is assumed a variable point located on the surface of a sphere centered at the origin with radius $R$, and we are interested to find the direction for which the corresponding minimum input energy is minimized or maximized.

The results show that the minimum value of the minimum input energy happens in the direction of an eigenvector corresponding to the maximum eigenvalue of the controllability Gramian, and the maximum of minimum input energy is in the direction of an eigenvector corresponding to the minimum eigenvalue of controllability Gramian. First consider the definitions below:

Definition (4.1): If $x$ is a nonzero vector in $\mathbb{R}^n$ and $A$ is a $n \times n$ dimensional matrix, the Rayleigh quotient of $x$ with respect to $A$ is defined as (Simovici, 2012):

$$RQ_A(x) = \frac{x^T A x}{x^T x}$$

Figure (4.6): $E_{\min}$ as a function of final time variable, $0 < t_f < \infty$
**Definition (4.2):** A linear subspace $V \subseteq \mathbb{R}^n$ is called an invariant subspace of $A$ if it satisfies $Ax \in V$ for $\forall x \in V$ (Simovici, 2012).

We can now prove the following Lemma (Prasolov, 1994):

**Lemma (4.1):** If $A$ is a real symmetric matrix and $V$ is an invariant subspace of $A$, then there is some $x \in V$ such that $RQ_A(x) = \inf\{RQ_A(y), y \in V\}$. Any $x \in V$ that minimizes $RQ_A(x)$ is an eigenvector of $A$, and the value $RQ_A(x)$ is the corresponding eigenvalue.

**Proof:**

If $x$ is a vector and $r$ is a nonzero scalar, then $RQ_A(x) = RQ_A(rx)$, hence every value attained in $V$ by the function $RQ_A(x)$ is attained on the unit sphere $S(V) = \{x \in V, \|x\| = 1\}$.

The function $RQ_A(x)$ is a continuous function on $S(V)$, and $S(V)$ is compact (closed and bounded) so by Weistrass Theorem (Gould, 1957) $RQ_A(x)$ attains its minimum value at some unit vector $x \in S(V)$. Using this quotient rule we can compute the gradient as:

$$\nabla RQ_A(x) = \frac{2Ax - 2(x^T Ax)x}{(x^T x)^2}$$

(4.25)

At the vector $x \in S(V)$ where $RQ_A(x)$ attains its minimum value in $V$, this gradient vector must be orthogonal to $V$, otherwise, the value of $RQ_A(x)$ would decrease as we move away from $x$ in the direction of any $y \in V$ that has negative dot product with $\nabla RQ_A(x)$.

On the other hand, our assumption that $V$ is an invariant subspace of $A$ implies that the right side of (4.25) belongs to $V$. The only way that $\nabla RQ_A(x)$ could be orthogonal to $V$ while also belongs to $V$ is if the zero vector, hence:

$$\nabla RQ_A(x) = 0 = 2Ax - 2(x^T Ax)x \Rightarrow Ax = (x^T Ax)x \Rightarrow x^T Ax = \lambda$$

(4.26)

so the Lemma is proved.

Now considering the Lemma above, equation (4.3) along with the fact that the inverse of controllability Gramian $W_c^{-1}(0, t_f)$ is a symmetric matrix, by defining:

$$S(V) = \{x \in V, \|x\| = 1\}$$

(4.27)

it can be concluded that

$$RQ_A(x_f) = x_f^T W_c^{-1}(0, t_f)x_f$$

(4.28)
is minimized (maximized) in the direction corresponding to the smallest (largest) eigenvalue of the inverse of controllability Gramian $W_c^{-1}(0,t_f)$.

The results stated above are summarized as shown below:

**Remark (4.1):** The following relations hold true:

$$
\min_{x(t) = x_f} \|u(t)\|_{[0,t_f]} = \min_{x(t) = x_f} \|x^T W_c^{-1}(0,t_f) x_f\| = \lambda_{\min}(W_c^{-1}(0,t_f)) = \frac{1}{\lambda_{\max}(W_c(0,t_f))}
$$

\( \Rightarrow E_{\min} = \lambda_{\min}(W_c^{-1}(0,t_f)) = \frac{1}{\lambda_{\max}(W_c(0,t_f))} \) \hspace{1cm} (4.29)

and

$$
\max_{x(t) = x_f} \|u(t)\|_{[0,t_f]} = \max_{x(t) = x_f} \|x^T W_c^{-1}(0,t_f) x_f\| = \lambda_{\max}(W_c^{-1}(0,t_f)) = \frac{1}{\lambda_{\min}(W_c(0,t_f))}
$$

\( \Rightarrow E_{\max} = \lambda_{\max}(W_c^{-1}(0,t_f)) = \frac{1}{\lambda_{\min}(W_c(0,t_f))} \) \hspace{1cm} (4.30)

Now, based on the results above, in the next section two sub-problems could be defined that is:

i. Energy Levels in the best case, considering $E_{\min}$

ii. Energy Levels in the worst case, considering $E_{\max}$

### 4.3.2.1 Energy Levels in the Best Case

Here, we are comparing the minimum of minimum energy $E_{\min}$ transferring the states of a given system from the origin to the terminal states $x_f$ (within a fixed terminal time $t_f$) such that $\|x_f\| = R$ and $R$ is variable.

In other words, in this part we are interested to know how much could be the minimum input energy which transfers the states of the LTI controllable system from the origin to the states defined $R$ away from the origin, according to the Theorem below, this transfer should be done in the direction of the eigenvector corresponding to the minimum eigenvalue of $W_c^{-1}(0,t_f)$ (i.e. the maximum eigenvalue of $W_c(0,t_f)$).

**Theorem (4.4):** (Prasolov, 1994) A quadratic form can always be reduced to the form:

$$
q(x_1, \ldots, x_n) = \lambda_1 x_1^2 + \cdots + \lambda_n x_n^2
$$

\( \square \)
According to the Theorem above (so called Lagrange Theorem), we can define the minimum input energy (4.3) as:

\[
E_{\text{min}} = x_f^T W_c^{-1}(0,t_f) x_f = \sum_{i=1}^{n} \lambda_i(W_c^{-1}(0,t_f)) x_i^2
\]

That is:

\[
E_{\text{min}} = x_f^T W_c^{-1}(0,t_f) x_f = \sum_{i=1}^{n} \frac{1}{\lambda_i(W_c(0,t_f))} x_i^2 = \frac{1}{\lambda_1(W_c(0,t_f))} x_1^2 + \cdots + \frac{1}{\lambda_n(W_c(0,t_f))} x_n^2
\]

Then it can be easily inferred from (4.33) that for a given fixed system we have:

**Proposition (4.1):** The optimal minimum input energy, which describes the minimum energy level for \(\|x_f\| = R\) is defined along the eigenvector corresponding to the maximum eigenvalue of controllability Gramian \(W_c(0,t_f)\) and its value is:

\[
E_{\text{min}}^{\text{opt}} = \min_{\|x_f\|^2 = R} \{ x_f^T W_c^{-1}(0,t_f) x_f \} = \frac{1}{\lambda_{\text{max}}(W_c(0,t_f))} R^2
\]

Figure (4.7) shows the levels of minimum input energy for the LTI moving object described in example (4.1), for different values of \(R\), where \(\|x_f\|^2 = R\).

In each case the corresponding direction, in which the minimum of minimum energy happens is plotted as well, which is the same as the direction of the eigenvector corresponding to the maximum eigenvalue of the controllability Gramian \(W_c(0,t_f)\).
In Figure (4.8), the levels for the minimum of the minimum input energy is shown for furuta pendulum 3-dimensional system defined in example (4.2).
4.3.2.2 Energy Levels in the Worst Case

In this part, we compare the maximum of minimum energy $E_{\min}^{\max}$ to transfer the states of a given controllable LTI system from the origin to the terminal states $x_f, \|x_f\| = R$, where $R$ is variable (within a fixed terminal time $t_f$), it can guarantee if you want to reach a state defined $R$ away from the origin, ignoring the direction of transferring the states, the energy, which you need would be equal or less than a specific value, which is corresponding to the energy level defined for the maximum of minimum input energy.

This would be exactly the same as the solution proposed in previous part.

**Remark (4.2):** The maximum of minimum energy for transferring the initial states of controllable LTI system (4.1) from the origin to $x_f, \|x_f\| = R$ could be defined in the direction of eigenvector corresponding to the smallest eigenvalue of controllability Gramian $W_c(0, t_f)$. $V_{\min} = V\left(\lambda_{\min}\left(W_c(0, t_f)\right)\right)$:

$$E_{\min}^{\max} = \max_{\|x\| = R} \{x^TW_c^{-1}(0, t_f)x\} = \frac{1}{\lambda_{\min}(W_c(0, t_f))} R^2$$

(4.35)

The above represents the energy stratification of the state space, which is important in input selection and optimal actuator placement. By optimizing the energy levels of controllable LTI system (4.1), the optimum positions of actuators could be defined.

Figure (4.9) illustrates the energy levels for the maximum of minimum input energy for 2-dimensional mass-spring system defined in example (4.1)

The direction, in which these energy levels are calculated, is shown as well, that is the same as the direction of the eigenvector corresponding to the minimum eigenvalue of the controllability Gramian $W_c(0, t_f)$.
Figure (4.9): Energy levels considering $E_{\text{min}}^{\text{max}}$

Figure (4.10) demonstrates the energy levels for 3-dimensional pendulum system (see Example (4.2)) considering the maximum value of minimum energy input for the various values of $R = \|x\|^2 = 1, 2, 3$

Figure (4.10): Energy levels for Furuta pendulum considering $E_{\text{min}}^{\text{max}}$
Now, regarding the results achieved for the variable terminal times and the variable terminal states in the sequel, we define the minimum energy $E_{\text{min}}$ as a function of two variables $t_f, x_f$, to investigate how the minimum input energy $E_{\text{min}}$ varies, if we transfer the states of the LTI controllable system (4.1), from the origin to the arbitrary terminal states $x_f$, in various directions, satisfying the condition $\|x_f\| = R$ within different terminal times $t_f$.

Figure (4.11) shows the changes of minimum input energy in the LTI mass-spring system (example (4.1)), assuming $R = 1$.

The Figure shows that as the terminal time increases the effect of the direction decreases, since the cost function looks more like a circle, it means that by growing the terminal time $t_f$, the difference between the minimum and maximum eigenvalues of the controlability Gramian $W_c(0,t_f)$ decreases, i.e. the reachability sets looks more like a circle. This can be seen in Figure (4.12).

In other words, for well-conditioned controlability Gramian matrix, the maximum and minimum energy levels are almost same, which indicates regardless of the trajectory and the direction of movement, there is a minimum value for the energy dissipated to move the states from the origin to the target point $x_f$, $\|x_f\| = R$ (within a terminal time $t_f$).

![Minimum Energy Cost Function for different tf in different directions R=1](image)

**Figure (4.11): Minimum input energy as a function of the variables $t_f, x_f$**
Figure (4.12): Reachability area changes based on different terminal times for 2 dimensional example model in (4.16)
4.4 Summary

In this chapter, we discussed the fundamentals of minimum energy required for transferring the states of a controllable LTI system from the origin to some target states $x_f, \|x_f\| = R$ in the state space within the terminal time $t_f$. Terminal time and terminal states as two important elements which may affect the value of this minimal energy have also been discussed.

In particular, if controllability Gramian $W_c(0, t_f)$ is nearly singular, then large energy inputs are required to reach the target states belonging to the eigenspace of its least eigenvalue $\lambda_{\text{min}}(W_c(0, t_f))$. This motivated the use of the minimum/maximum eigenvalue of the controllability Gramian in the energy stratification of the state space, in this chapter we adopted the measure of the worst-case control effort when investigating maximum of optimum energy levels.

In the next chapter, we consider the energy calculation for unstable systems and the impact of actuator selection on various issues such as the degree of controllability and disturbance rejection.
Chapter 5: Controllability, Energy for Unstable Systems and Measures of Degree of Controllability

5.1. Introduction

In this chapter, our focus is on the energy calculation for unstable systems and the impact of actuator selection on various issues such as the degree of controllability, and disturbance rejection.

In the previous chapter, we have defined the minimum input energy and introduced the energy levels for stable systems, in this chapter we examine the case where the system is unstable, and we present the generalized energy stratification of the state space for the possibly unstable systems.

Section 5.2 and 5.3 cover the above issues, then in section 5.4 we will investigate the problem of the actuator selection with regard to minimum energy.

One of the most reliable methods for determining actuator locations is the improvement of state controllability of the process (Wal & Jager, 2001), then the problem for the actuator placement could be viewed as the problem of minimizing the input energy required to reach a given state. A deficiency
of this method appears where sufficient information for the initial and final states of the system is not available.

This problem is discussed in section 5.4 in details then optimization of the worst-case method is suggested as a solution.

In section 5.5 we presented a method of actuator selection considering the degree of controllability of linear time-invariant systems. Starting from a set of actuator locations which produce an uncontrollable system, but for which the number of actuators is sufficient to produce controllability, it will usually be the case that moving one of the actuators a distance \( \epsilon > 0 \) can produce a controllable system, no matter how small the value \( \epsilon \) (and vice versa). It is then important to select the actuators which maximize the degree of controllability.

The similar problem will be considered in section 5.6, where the system is affected by an external disturbance as well. In this case, we define the degree of disturbance rejection by considering both, the controllability Gramian of the system and the disturbance sensitivity Gramian, and we place the actuators such that the presented measure is maximized.

The methods proposed in this chapter are general methods in the sense that they support both stable and unstable systems. The techniques described are successfully verified by various examples.

### 5.2. Minimum Energy for Unstable Systems

In chapter 4 we have discussed in detail, the minimum energy dissipation needed to transfer the states of an LTI system from the origin to an arbitrary destination \( x_f \) within a finite time \( t_f \). This can be calculated as \( x_f^T W_c^{-1}(0, t_f)x_f \) where \( W_c(0, t_f) \) is the controllability Gramian. According to (Hać & Liu, 1993) in more general case where we are steering the states from any initial states \( x_0 \) to the final states \( x_f \) within the terminal time \( t_f \), the minimum energy required for this transferring could be represented as:

\[
E_{\text{min}} = (e^{At_f} x_0 - x_f)^T W_c^{-1}(0, t_f)(e^{At_f} x_0 - x_f)
\]  

(5.1)

If the system is asymptotically stable, \( e^{At_f} \rightarrow 0 \) as \( t_f \rightarrow \infty \), then we can attain the minimum energy as a function of final states only regardless of their initial location. In this case if we choose the origin as the final state \( x_f = 0 \) then \( E_{\text{min}} = 0 \), which means bringing the system to the state of equilibrium can be achieved by providing zero energy, as the asymptotically stable system always converges to the origin by itself, if the final time is sufficiently large. This implies that, if the system is asymptotically stable escaping from origin (i.e., the reachability problem) requires more energy than transferring to origin, (i.e., the controllability to 0 problem) because the stable system naturally tends to converge to zero.

Furthermore equation (5.1) states that in undamped systems (i.e. eigenvalues are located on the imaginary axes) where \( e^{At_f} \) is an orthonormal matrix, i.e. \( e^{At_f} e^{At_f^T} = I_n \), the minimum input energy required to bringing the system to the origin is \( E_{\text{min}} = x_0^T W_c^{-1}(0, t_f)x_0 \) (Arbel, 1981).
In this section, we use the statement (5.1) to determine the minimum energy for an unstable system, where the system is decomposable to a stable and anti-stable subsystem, i.e. when system matrix $A$ has no eigenvalues on the imaginary axis.

### 5.2.1. Decomposition of Controllability Gramian in LTI Systems with Stable and Anti-Stable Modes

Consider the unstable LTI system below:

$$
\begin{align*}
\dot{x} &= Ax + Bu \\
y &= Cx + Du
\end{align*}
$$

(5.2)

where $A$ has one or more eigenvalues in the right half plane (RHP) of the complex plane and it has no eigenvalues on the imaginary axis (as for the minimum input energy computation, no energy is required to change the states of undamped modes in sufficiently large time period $[0, t_f]$; thus this assumption imposes no extra restriction).

In general, there are two well-known transformation (Sivasundaram, 2004), (Demenkov & Goman, 2009), (Chen, 1995) suitable for the stable/unstable decomposition of a linear continuous-time system. The one is based on Schur transformation (Castelan, et al., 1996), (Golub & Loan, 1986) and the second one is based on the block Jordan transformation (Chen, et al., 2004). In this thesis, we will use the latter.

**Remark (5.1):** There exists a similarity transformation $A \rightarrow V^{-1}AV = \tilde{A}, V \in \mathbb{R}^{n \times n}$, for which $\tilde{A}$ is block diagonal, i.e. $V^{-1}AV = \Lambda = \begin{bmatrix} A_s & 0 \\ 0 & A_u \end{bmatrix}$, where $A_s, A_u$ are stable and anti-stable system matrices, respectively.

Then in the new coordinate frame one can rewrite the LTI system (5.2) as:

$$
\begin{align*}
\begin{bmatrix} \dot{x}_s \\ \dot{x}_u \end{bmatrix} &= \begin{bmatrix} A_s & 0 \\ 0 & A_u \end{bmatrix} \begin{bmatrix} x_s \\ x_u \end{bmatrix} + \begin{bmatrix} B_s \\ B_u \end{bmatrix} u \\
y &= [C_s \ C_u] \begin{bmatrix} x_s \\ x_u \end{bmatrix} + Du
\end{align*}
$$

(5.3)

with the appropriate dimensions for all matrices.

**Remark (5.2):** For the finite-horizon case $W_c(0, t_f)$ is always well defined as a symmetric positive semidefinite matrix (which is not the case for the infinite-horizon problem), then the controllability Gramian for the finite terminal time $W_c(0, t_f)$ could be derived from the state space realization above as:
and by the general formula for the inverse of the partitioned matrix we obtain:

**Lemma (5.1):** Consider the block partitioned controllability Gramian matrix in (5.4) is positive definite, then the inverse of controllability Gramian could be achieved as:

\[
W_c^{-1}(0,t_f) = \begin{bmatrix}
(W_s - W_{su}W^{-1}_{us}W_{su})^{-1} & -W^{-1}_{su}(W_s - W_{su}W^{-1}_{su}W_{us})^{-1} \\
-W^{-1}_{us}(W_s - W_{su}W^{-1}_{su}W_{us})^{-1} & (W_u - W_{us}W^{-1}_{su}W_{us})^{-1}
\end{bmatrix} > 0 \quad (5.5)
\]

We refer the interested readers to (Horn & Johnson, 2012) for the proof. Next, we define Schur complement Theorem for a positive definite symmetric matrix.

**Theorem (5.1):** (Fuzhen, 2005) Given any symmetric matrix

\[
A = \begin{bmatrix}
A & B \\
B^T & C
\end{bmatrix}
\]

the following conditions are equivalent:

i. \( M > 0 \) (\( M \) is positive definite)

ii. \( A > 0, C - B^TA^{-1}B > 0 \)

iii. \( C > 0, A - BC^{-1}B^T > 0 \)

The complete proof of this Theorem is given in (Fuzhen, 2005). Regarding the time dependency in (5.1) we can use the continuous-time differential Lyapunov equations (CDLE) (Davis, et al., 2010), then we get the Corollary below:

**Corollary (5.1):** (Middleton & Goodwin, 1990), (Amato, et al., 2015), (Singh, 2009) Consider the linear time-invariant continuous system (5.2) with no eigenvalues on the imaginary axis. Then the controllability Gramian satisfies the continuous-time differential Lyapunov equation

\[
\dot{W}_c(0,t_f) = AW_c(0,t_f) + W_c(0,t_f)A^T + BB^T \quad (5.6)
\]

with zero initial condition or equivalently by partitioning \( W_c(0,t_f) \):

\[
W_c(0,t_f) = \begin{bmatrix}
W_s(0,t_f) & W_{su}(0,t_f) \\
W_{us}(0,t_f) & W_u(0,t_f)
\end{bmatrix}
\]
\[
\begin{align*}
\dot{W}_s &= A_s W_s + W_s A_s^T + B_s B_s^T \\
\dot{W}_{su} &= A_s W_{su} + W_{su} A_s^T + B_s B_u^T \\
W_{su} &= A_s W_{su} + W_{su} A_s^T + B_u B_s^T \\
W_u &= A_u W_u + W_u A_u^T + B_u B_u^T
\end{align*}
\tag{5.7}
\]

For the proof see (Middleton & Goodwin, 1990), (Amato, et al., 2015), (Singh, 2009).

**Theorem (5.2):** Let (5.2) be a controllable LTI system with both stable and anti-stable parts. Then the inverse of controllability Gramian i.e. \(W_c^{-1}(0,t_f) > 0\), as is stated in Lemma (5.1), is well-defined in the limit \(t_f \to \infty\) and:

\[
W_c^{-1}(0,\infty) = \begin{bmatrix} W_c^{-1} & 0 \\ 0 & 0 \end{bmatrix}
\tag{5.8}
\]

**Proof:**

If the system is controllable, the controllability Gramian (5.4) must be positive definite \(W_c(0,t_f) > 0\) (E.Hendricks, et al., 2008) and according to Schur complement Theorem (5.1) we have:

\[
(W_u - W_{su} W_s^{-1} W_{su}) > 0 \text{ and } (W_u - W_{su} W_s^{-1} W_{su})^{-1} > 0.
\]

Now set \(Q = (W_u - W_{su} W_s^{-1} W_{su})\). Using the chain rule of differentiation, we get:

\[
\dot{Q} = (\dot{W}_u - \dot{W}_{su} W_s^{-1} W_{su} - W_{su} (W_s^{-1}) W_{su} - W_{su} W_s^{-1} W_{su})
\tag{5.9}
\]

Then considering (CDLE) equations (5.7) we obtain:

\[
\dot{Q} = A_s (W_u - W_{su} W_s^{-1} W_{su}) + (W_u - W_{su} W_s^{-1} W_{su}) A_s^T + B_u B_u^T - B_u B_s^T W_s^{-1} W_{su} - W_u W_s^{-1} B_s B_u^T + W_u W_s^{-1} B_u B_s^T W_s^{-1} W_{su}
\]

\[
= A_s (W_u - W_{su} W_s^{-1} W_{su}) + (W_u - W_{su} W_s^{-1} W_{su}) A_s^T + (W_u W_s^{-1} B_s - B_u)(W_s W_s^{-1} B_s - B_u)^T
\]

\[
\dot{Q} = A_s Q + QA_s^T + (W_u W_s^{-1} B_s - B_u)(W_s W_s^{-1} B_s - B_u)^T
\tag{5.10}
\]

Therefore, according to the differential Lyapunov equations (5.7) and with regard to the results of chapter 4, we obtain \(\dot{Q} \geq 0\).

Furthermore, using the differentiation property of non-singular matrices we have:

\[
(Q^{-1}) = -Q^{-1} \dot{Q} Q^{-1}
\tag{5.11}
\]
Then the positive definite matrices $Q$ and $Q^{-1} = (W_u - W_{us}W_{su}^{-1}W_{us})^{-1}$ yield:

$$(Q^{-1}) = -(W_u - W_{us}W_{su}^{-1}W_{us})^{-1} \hat{Q}(W_u - W_{us}W_{su}^{-1}W_{us})^{-1}$$

(5.12)

which clearly defines a negative definite matrix.

Then via the Lyapunov stability Theorem one can readily recognize that $(W_u - W_{us}W_{su}^{-1}W_{us})^{-1}$ is stable and tends to zero as $t_f \to \infty$.

To verify the stability of off-diagonal blocks in (5.5):

$$-W_{su}^{-1}W_{us}(W_u - W_{us}W_{su}^{-1}W_{us})^{-1}, -W_{su}^{-1}W_{us}(W_u - W_{us}W_{su}^{-1}W_{us})^{-1}$$

let us define matrices below:

$$\begin{align*}
P &= (W_u - W_{us}W_{su}^{-1}W_{us})^{-1} \\
Q &= W_{su}^{-1}W_{us}(W_u - W_{us}W_{su}^{-1}W_{us})^{-1}
\end{align*}$$

(5.13)

Considering Theorem (5.1) and the positive definite matrices $W_{c}(0, t_f), W_{c}^{-1}(0, t_f)$:

$$\begin{align*}
&\begin{cases}
W_{s} > 0 \\
W_{s} - W_{su}W_{us}^{-1}W_{us} > 0 \\
W_{u} > 0 \\
W_{u} - W_{us}W_{su}^{-1}W_{us} > 0
\end{cases} \\
&\begin{cases}
P > 0 \\
P - Q^{T}(W_{s} - W_{su}W_{us}^{-1}W_{us})Q > 0
\end{cases}
\end{align*}$$

(5.14)

According to part $(ii)$ we have:

$$P - Q^{T}(W_{s} - W_{su}W_{us}^{-1}W_{us})Q > 0 \Rightarrow P > Q^{T}(W_{s} - W_{su}W_{us}^{-1}W_{us})Q$$

(5.15)

Considering part $(i)$ of equation (5.14) we know that $Q^{T}(W_{s} - W_{su}W_{us}^{-1}W_{us})Q \geq 0$, then due to continuity that $t_f \to \infty$ the statement (5.15) implies that:

$$\lim_{t_f \to \infty} P \geq \lim_{t_f \to \infty} Q^{T}(W_{s} - W_{su}W_{us}^{-1}W_{us})Q \geq 0$$

(5.16)

It was also proved in the previous part that $\lim_{t_f \to \infty} P = 0$ then:

$$0 \geq \lim_{t_f \to \infty} Q^{T}(W_{s} - W_{su}W_{us}^{-1}W_{us})Q \geq 0$$

(5.17)

Hence according to the Squeeze Theorem one can readily conclude:
\[
\lim_{t_f \to \infty} Q^T (W_s - W_{su} W_u^{-1} W_{as}) Q = 0 \tag{5.18}
\]

Regarding part (ii) we have \((W_s - W_{su} W_u^{-1} W_{as}) > 0\) then equation (5.18) implies that the off-diagonal block \(Q\) tends to zero as \(t_f \to \infty\).

Therefore, with regard to the symmetry, it can readily be verified that both off-diagonal blocks in (5.5) are stable (i.e. they go to zero as \(t_f \to \infty\)), and they do not appear in the steady state form of the inverse of the controllability Gramian \(W_c^{-1}(0, \infty)\).

Now consider the first diagonal block matrix \((W_s - W_{su} W_u^{-1} W_{as})^{-1}\) in (5.5). Using the Sherman-Morrison-Woodbury formula (Press, et al., 2007) we obtain:

\[
(W_s - W_{su} W_u^{-1} W_{as})^{-1} = W_s^{-1} + W_s^{-1} W_{su} (W_u - W_{su} W_s^{-1} W_{as})^{-1} W_{as} W_s^{-1} \tag{5.19}
\]

For simplicity, we define matrices \(P, Q\) as:

\[
\begin{cases}
    P = (W_u - W_{as} W_s^{-1} W_{su})^{-1} \\
    Q = W_s^{-1} W_{su} P
\end{cases} \tag{5.20}
\]

The statement (5.19) can be written as:

\[
(W_s - W_{su} W_u^{-1} W_{as})^{-1} = W_s^{-1} + QP^{-1}Q^T \tag{5.21}
\]

Then using the results of the previous parts the second term in (5.21) implies that:

\[
\begin{cases}
    P^{-1}(0, t_f) > 0, 0 < t_f < \infty \\
    \lim_{t_f \to \infty} QP^{-1}Q^T = 0 \tag{5.22}
\end{cases}
\]

We know that the constant solution of continuous-time algebraic Lyapunov equations (CALE) is in fact the steady state solution (i.e. \(t_f \to \infty\)) of (CDLE) in (5.7) (Davis, et al., 2010). Then clearly the \(\lim_{t_f \to \infty} W_s\) is the solution of (CALE) \(A_s W_s + W_s A_s^T = -B_s B_s^T\) and part (ii) in equation (5.21) yields a finite positive definite matrix, i.e. \(0 < \lim_{t_f \to \infty} W_s^{-1}\).

Hence, we have shown that in the steady state condition as \(t_f \to \infty\) all partitions of:

\[
W_c^{-1}(0, t_f) = \begin{bmatrix}
    (W_s - W_{su} W_u^{-1} W_{as})^{-1} & -W_s^{-1} W_{su} (W_u - W_{su} W_s^{-1} W_{as})^{-1} \\
    -W_u^{-1} W_{as} (W_s - W_{su} W_u^{-1} W_{as})^{-1} & (W_u - W_{su} W_u^{-1} W_{as})^{-1}
\end{bmatrix}
\]

tend to zero except the first diagonal block, which would be equal to the positive definite matrix \(W_s^{-1}(\infty)\).

\[\square\]
The result above shows that in the steady state conditions, i.e. \( t_f \to \infty \), only the stable modes of LTI system can affect the inverse of the controllability Gramian. However, it is shown in the next section that the steady state of the minimum input energy is dependent to both stable and unstable modes of the system.

### 5.2.2. Minimum Input Energy for LTI System with Anti-Stable Part

According to (5.1), the minimum input energy is given as:

\[
E_{\min} = (e^{A_{t_f}x_0} - x_f)^T W_c^{-1}(0,t_f)(e^{A_{t_f}x_0} - x_f)
\]

Then, using Remark (5.2), we get:

\[
E_{\min} = (e^{A_{t_f}x_0} - x_f)^T \begin{bmatrix} W_s(0, t_f) & W_{s_u}(0, t_f) \\ W_{s_u}(0, t_f) & W_u(0, t_f) \end{bmatrix}^{-1} (e^{A_{t_f}x_0} - x_f)
\]

(5.23)

Next, we consider Theorem (5.3) in time domain:

**Theorem (5.3):** In the controllable LTI system (5.3) containing both stable and anti-stable parts, the minimum input energy under steady state conditions, i.e. \( t_f \to \infty \) is:

\[
E_{\min}(\infty) = \begin{bmatrix} (x_f)_s \\ (x_0)_u \end{bmatrix} \begin{bmatrix} W_s^{-1} & 0 \\ 0 & W_u^{-1} \end{bmatrix} \begin{bmatrix} (x_f)_s \\ (0, t_f) \\ (0, t_f) \end{bmatrix}
\]

(5.24)

Where:

\[
\begin{cases} 
A_s W_s + W_s A_s^T = -B_s B_s^T \\
(-A_u) W_u + W_u (-A_u^T) = -B_u B_u^T 
\end{cases}
\]

(5.25)

**Proof:**

Considering the minimum input energy equation (5.1) we have:

\[
E_{\min}(t_f) = \{x_f^TW_c^{-1}x_f - x_f^TW_c^{-1}e^{A_{t_f}x_0} - (e^{A_{t_f}x_0})^TW_c^{-1}x_f + (e^{A_{t_f}x_0})^TW_c^{-1}e^{A_{t_f}x_0} \}
\]

(5.26)

Define:

\[
\begin{cases} 
v_1 = x_f^TW_c^{-1}(0, t_f)x_f \\
v_2 = -x_f^TW_c^{-1}(0, t_f)e^{A_{t_f}x_0} \\
v_3 = -(e^{A_{t_f}x_0})^TW_c^{-1}(0, t_f)x_f \\
v_4 = (e^{A_{t_f}x_0})^TW_c^{-1}(0, t_f)e^{A_{t_f}x_0} 
\end{cases}
\]

(5.27)
Then:

\[ E_{\text{min}}(\infty) = \lim_{t_f \to \infty} (v_1 + v_2 + v_3 + v_4) \]  
\[ (5.28) \]

Thus, in the steady state:

\[
\begin{align*}
{v}_1^\infty &= \lim_{t_f \to \infty} v_1 = \lim_{t_f \to \infty} x^T f_{c}^{-1}(0, t_f) x_f \\
{v}_2^\infty &= \lim_{t_f \to \infty} v_2 = -\lim_{t_f \to \infty} x^T f_{c}^{-1}(0, t_f) e^{A_{tr}} x_0 = 0 \\
{v}_3^\infty &= \lim_{t_f \to \infty} v_3 = -\lim_{t_f \to \infty} (e^{A_{tr}} x_0)^T f_{c}^{-1}(0, t_f) x_f = 0 \\
{v}_4^\infty &= \lim_{t_f \to \infty} v_4 = \lim_{t_f \to \infty} (e^{A_{tr}} x_0)^T f_{c}^{-1}(0, t_f) e^{A_{tr}} x_0 
\end{align*}
\]  
\[ (5.29) \]

According to Theorem (5.2), part \( v_1^\infty \) represents the minimum input energy of the stable modes of the system when \( t_f \to \infty \) and is equal to \( x^T f_{c}^{-1}(0, \infty) x_0 \).

Now consider part the second term:

\[
{v}_2^\infty = \lim_{t_f \to \infty} v_2 = \lim_{t_f \to \infty} x^T f_{c}^{-1}(0, t_f) (e^{A_{tr}} x_0) = x^T (\lim_{t_f \to \infty} f_{c}^{-1}(0, t_f) e^{A_{tr}}) x_0 \Rightarrow
\]  
\[ v_2^\infty = x^T (\lim_{t_f \to \infty} \begin{bmatrix} (W_s - W_s W_u^{-1} W_s^{-1} W_u^{-1})^{-1} & -W_s^{-1} W_s (W_u - W_u W_s^{-1} W_u^{-1})^{-1} \\
-W_u^{-1} W_u (W_s - W_s W_u^{-1} W_u^{-1})^{-1} & (W_u - W_u W_s^{-1} W_u^{-1})^{-1} \end{bmatrix} e^{A_{tr}}) x_0 
\]  
\[ (5.30) \]

The above equation can be stated as:

\[
{v}_2^\infty = x^T (\lim_{t_f \to \infty} \begin{bmatrix} (W_s - W_s W_u^{-1} W_s^{-1} W_u^{-1})^{-1} & -W_s^{-1} W_s (W_u - W_u W_s^{-1} W_u^{-1})^{-1} \\
-W_u^{-1} W_u (W_s - W_s W_u^{-1} W_u^{-1})^{-1} & (W_u - W_u W_s^{-1} W_u^{-1})^{-1} \end{bmatrix} e^{A_{tr}}) x_0
\]

\[ i) \lim_{t_f \to \infty} (W_u - W_u W_s^{-1} W_u^{-1})^{-1} e^{A_{tr}} = 0 \]

\[ ii) \lim_{t_f \to \infty} W_u^{-1} W_u (W_s - W_s W_u^{-1} W_s^{-1} W_u^{-1})^{-1} e^{A_{tr}} = 0 \]

\[ iii) \lim_{t_f \to \infty} W_s^{-1} W_s (W_u - W_u W_s^{-1} W_u^{-1})^{-1} e^{A_{tr}} \]

\[ iv) \lim_{t_f \to \infty} (W_u - W_u W_s^{-1} W_u^{-1})^{-1} e^{A_{tr}} \]

\[ (5.31) \]

Then it is easy to show \( \lim_{t_f \to \infty} v_2 = 0 \).
Consider part \((iv)\), as it is described in the proof of Theorem (5.2), the rate of the divergence of exponential matrix \(e^{A_{ij}t}\) when \(t_f \to \infty\) is less than the rate of convergence of \((W_u - W_{sa}W_{su}^{-1}W_{su})^{-1}\) then \(\lim_{t_f \to \infty} (W_u - W_{sa}W_{su}^{-1}W_{su})^{-1} e^{A_{ij}t} = 0\). For the same reason, part \((iii)\) converges to zero as the final time \(t_f\) increases to infinity, i.e. \(\lim_{t_f \to \infty} W_{sa}^{-1}W_{su}(W_u - W_{sa}W_{su}^{-1}W_{su})^{-1} e^{A_{ij}t} = 0\).

Then part \(v_2\) is stable. By the same method, one can readily prove that part \(v_3\) also tends to zero as \(t_f \to \infty\).

Now let us consider the last part \(v_4\), it could itself be broken into four different statements:

\[
v_4^* = \lim_{t_f \to \infty} (e^{A_{ij}t} x_0)^T (W_c e^{-1} (0, t_f) e^{A_{ij}t} x_0)
\]

\[
\Rightarrow
\]

\[
\left\{
\begin{array}{l}
i) \lim_{t_f \to \infty} (e^{A_{ij}t} x_{03})^T (W_s - W_{sa} W_{su}^{-1}W_{su})^{-1} e^{A_{ij}t} x_{03} \\
ii) \lim_{t_f \to \infty} (e^{A_{ij}t} x_{0a})^T W_{sa}^{-1} W_{su}(W_u - W_{sa} W_{su}^{-1}W_{su})^{-1} e^{A_{ij}t} x_{0a}
\end{array}
\right.
\]

\[
(iii) \lim_{t_f \to \infty} (e^{A_{ij}t} x_{0a})^T W_{sa}^{-1}W_{su}(W_u - W_{sa} W_{su}^{-1}W_{su})^{-1} e^{A_{ij}t} x_{0a}
\]

\[
(iv) \lim_{t_f \to \infty} (e^{A_{ij}t} x_{0a})^T (W_u - W_{sa} W_{su}^{-1}W_{su})^{-1} e^{A_{ij}t} x_{0a}
\]

We have shown in Theorem (5.2) that:

\[
\lim_{t_f \to \infty} (W_s - W_{sa} W_{su}^{-1}W_{su})^{-1} = W_{sa}^{-1}
\]

\[
(5.33)
\]

Now regarding the stability of \(A\), part \((i)\) yields zero, i.e. \(\lim_{t_f \to \infty} (e^{A_{ij}t} x_{03})^T W_{sa}^{-1} (0, t_f) e^{A_{ij}t} x_{03} = 0\).

Part \((ii)\) becomes zero as well in the steady state (i.e. \(t_f \to \infty\)) since \(W_{sa}^{-1}W_{su}(W_u - W_{sa} W_{su}^{-1}W_{su})^{-1}\) is stable and the decay rate of its dynamic in the steady state condition is faster than the increment rate of the exponential term \(e^{A_{ij}t}\). In a similar way, it can be shown that part \((iii)\) goes to zero as \(t_f \to \infty\).

Lastly the block diagonal \((W_u - W_{sa} W_{su}^{-1}W_{su})^{-1}\) could be expanded based on the Sherman–Morrison–Woodbury formula (Press, et al., 2007) then we obtain:

\[
(W_u - W_{sa} W_{su}^{-1}W_{su})^{-1} = W_{sa}^{-1} + W_{sa}^{-1} W_{su}(W_s - W_{sa} W_{su}^{-1}W_{su})^{-1} W_{sa}^{-1} W_{su}^{-1}
\]

\[
(5.34)
\]

Hence with regard to the previous parts we obtain:

\[
\lim_{t_f \to \infty} (e^{A_{ij}t} x_{0a})^T (W_u^{-1}W_{sa}(W_s - W_{sa} W_{su}^{-1}W_{su})^{-1} W_{sa} W_{su}^{-1}) e^{A_{ij}t} x_{0a} = 0 \Rightarrow
\]

---

70
\[ iv = \lim_{t_f \to \infty} (e^{A_{tf}}x_{0u})^T W_u^{-1}(0, t_f)e^{A_{tf}}x_{0u} = x_{0u}^T (\lim_{t_f \to \infty} \alpha) x_{0u} \]

\[ \Rightarrow \lim_{t_f \to \infty} \alpha = \lim_{t_f \to \infty} (\int_0^{t_f} e^{-A_{tf}} e^{A_{tf}} B_u B_u^T e^{-A_{tf}} dt)^{-1} \]

\[ \Rightarrow \lim_{t_f \to \infty} \alpha = \lim_{t_f \to \infty} (\int_0^{t_f} e^{-A_{tf}} B_u B_u^T e^{-A_{tf}} dt)^{-1} \]

Then as \( t_f \to \infty \) part \( v_4^+ \) converges to \( x_{0u}^T W_u^{-1}(0, \infty) x_{0u} \), which represents the energy needed to regulate the initial condition of the anti-stable subsystem.

In fact, the Theorem above supports the results we discussed in the previous chapter stating that in the steady state condition the minimum energy in stable system is a function of terminal time and the terminal states, therefore, in the steady state, i.e. \( t_f \to \infty \) the minimum control energy is required, while in anti-stable systems we need no steering energy but an initial excitation to move the states from the origin to the terminal point in the controllable region when \( t_f \to \infty \).

As (Zhou, et al., 1999) suggests one can also define the steady state form of the controllability Gramian in the frequency domain through Parseval’s Theorem. Similar results can be derived for unstable systems also.

**Proposition (5.1):** (Zhou, et al., 1999) In the LTI system (5.2) if \( A \) is stable, we define the controllability Gramian in the frequency domain as:

\[ W_c = \frac{1}{2\pi} \int_{-\infty}^{\infty} (j\omega I - A)^{-1} B B^T (-j\omega I - A^T)^{-1} d\omega \]  

\[ (5.36) \]

**Corollary (5.2):** In LTI system (5.3) \( W_c \) defined in Proposition (5.1) can be decomposed as:

\[ W_c = \begin{bmatrix} W_s & 0 \\ 0 & W_{-u} \end{bmatrix} \]  

\[ (5.37) \]

where \( W_s, W_{-u} \) are the solutions of Lyapunov equations (5.25).

Proof is given in (Zhou, et al., 1999). The above Corollary uses the frequency domain to show that the controllability Gramian of LTI controllable systems with no eigenvalues on the imaginary axes can be decomposed into stable and anti-stable parts. This confirms that minimum input energy, which is

---

2 For the simplicity, we use the same symbol for the controllability Gramian in the frequency domain and the time domain.
defined in equation (5.1) can be derived as in (5.24). Theorem (5.3) expresses the same results in the time domain.

5.3. Energy Levels for LTI Systems with Anti-Stable Modes

In this section, we consider the minimum and maximum values of required energy \( E_{\text{min}} \), \( E_{\text{max}} \) transferring the states of the LTI system (5.2) with anti-stable modes from the initial point \( x_0 \), \( \|x_0\| = R_0 \) to the terminal states \( x_f \), \( \|x_f\| = R_f \) (within a very large terminal time \( t_f \to \infty \)) such that \( R_0, R_f \) are variables.

\[
E_{\min} = \min E_{\text{min}}, \quad E_{\max} = \max E_{\text{min}}
\]

In other words, we are interested to determine how much input energy we require to move the states of the LTI controllable system (5.2) from an arbitrary initial state a distance \( R_0 \) away from the origin to the states defined at a distance \( R_f \) away from the origin.

Assume \( x_{0u} \neq 0, x_{fu} \neq 0 \). According to Theorem (5.3) and Theorem (4.4), in the best case, where we need the minimum value of input energy, i.e. \( E_{\text{min}} \), this transfer is done such that the stable modes of the system are transferred in the direction of the eigenvector corresponding to the maximum eigenvalue of \( W_s \), and the unstable modes start their movement in the direction of the eigenvector associated with the maximum eigenvalue of \( W_u \). Similarly, in the worst case where we require the maximum energy, this transfer would be done in the direction of the eigenvectors corresponding to the minimum eigenvalues of \( W_s \) for stable part and the initial direction of the movement of anti-stable part would be along the eigenvector associated with the minimum eigenvalue of \( W_u \).

**Example (5.1):** Assume the LTI system as: \( A = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \) where the initial and final states can take values in the interval \([-1, 1]\), i.e. \( x_0 = \begin{bmatrix} -1 \leq x_{01} \leq 1 \\ -1 \leq x_{02} \leq 1 \end{bmatrix}, x_f = \begin{bmatrix} -1 \leq x_{f1} \leq 1 \\ -1 \leq x_{f2} \leq 1 \end{bmatrix} \). According to Theorem (5.3) the energy levels could be found as functions of both \( R_0 \) and \( R_f \). Figure (5.1) illustrates the energy levels for this unstable system. It is concluded that as the norm of \( x_{0u} \) or/and \( x_{fu} \) increases the value of energy input increases as well, however this increase depends on the direction of the movement. Wherever \( x_{0u} \neq 0, x_{fu} \neq 0 \) if the movement starts in the direction of eigenvector corresponding to the minimum eigenvalue of \( W_u \) for unstable modes and continues to reach the final states of stable modes in the direction of the eigenvector corresponding to the minimum eigenvalue of \( W_s \), then we will have the minimum increment, which means the maximum energy for transferring is used. If the unstable modes of system begin their movement along the direction of the eigenvector associated to the maximum eigenvalue of \( W_u \) and the stable modes of the system move to the final states in the direction of the eigenvector corresponding to the maximum eigenvalue of \( W_s \), the minimum value of energy would be required.
In addition, for \( R_0 \neq 0, R_f \neq 0 \) it is still possible that the system starts its movement in the direction such that \( x_{0u} = 0 \Rightarrow \|x_0\| = R_0 \) and moves toward the final state along the direction of \( x_{fu} = 0 \Rightarrow \|x_{fu}\| = R_f \), in this case the value of the required energy is zero.

Furthermore, as Figure (5.1) illustrates in this particular example, the value of minimum energy is more sensitive to the norm of initial states, \( R_0 \), rather than the distance of final states from the origin, \( R_f \), i.e. as \( R_f \) varies, for a fixed \( R_0 \), the value of minimum energy changes faster than the case that \( R_f \) is fixed and \( R_0 \) can change.

Figure (5.1): Minimum energy levels with respect to the norm of initial and terminal states (Example (5.1))

The importance of the shown energy levels is where we have a constraint on the value of control energy. This determines how far the system may move from the initial states defined by the distance \( R_0 \) away from the origin, i.e. \( \|x_0\| = R_0 \).

Additionally, it can be readily concluded from (5.24) that the unstable part contribution vanishes when the unstable modes of the LTI system start their movement from zero. In this special case \( V^{-1}x_0 = [x_{0s}, 0] \Rightarrow \|x_0\| = R_0 \), \( V \in \mathbb{R}^n \) is the transformation defined in Remark (5.1) and \( x_{0s} \) represents the stable modes’ initial states. Therefore, the input energy would be just a function of the stable part, and the energy levels in new coordination \((x \rightarrow V^{-1}x)\) could be determined as it is explained in chapter 4. Consider the 3-dimensional unstable system in the next example.

Example (5.2): Consider the LTI system:
The eigenvalues are $\lambda_1 = -30.1246, \lambda_2 = -12, \lambda_3 = 10.1246$. Assume we wish to move the states from the origin to a final state located on the sphere with radius $R_f (= 1, 2, 3)$. The steady state energy levels are shown in Figures below. Figure (5.2) demonstrates the value of minimum energy in the state space. In Figure (5.3) eigenvectors of $W_j$ are shown, which show the directions in which the minimum and maximum of energy levels occur.

![Figure (5.2): Minimum energy levels of Example (5.2)](image)
Conversely, if $x_{iu} \neq 0$, no energy is required to steer the stable modes of the LTI system to the origin, then if $V^{-1}x_f = [0, x_{fu}]$, $\|x_{fu}\| = R_f$, where $V \in \mathbb{R}^{n}$ is the transformation defined in Remark (5.1) and $x_{fu}$ represents the final states of the unstable subsystem, the input energy could be determined just as a function of the unstable part. Thus, the energy levels can be defined in the new coordinate frame ($x \rightarrow V^{-1}x$) as explained in chapter 4. For stable systems, however, the final states would be substitute by the initial states of unstable modes $x_{iu}$, and the inverse of controllability Gramian can modified as the inverse of controllability Gramian of the anti-stable part as it is described in Theorem (5.3), $W_{-1}$. Figures (5.4), (5.5) show the energy levels for example (5.3) in this particular case, where $R_0$ varies as the integer in the interval $[1, 3]$.

**Example (5.3):** Assume that we wish to transfer the state of the LTI system:
from an initial state on a sphere centered at the origin with the radius $R_0 (= 1, 2, 3)$ to the origin within a large terminal time $t_f \to \infty$. Figure (5.4) describes the minimum energy levels in the state space. These energy levels determine that in the worst case how much energy is needed to transfer the system from the initial states on a sphere with the radius $R_0$ to the origin. In Figure (5.5) the direction of eigenvectors of $W_{-u}$ are illustrated. As it can be seen the energy required to derive the system from an initial point decreases as $\|x_{0u}\|$ decreases. Furthermore, if $x_{0u} \neq 0$ the required energy decreases in the direction of the eigenvector corresponding to the maximum eigenvalue of $W_{-u}$ and increases along the direction of the eigenvector associated with the minimum eigenvalue of $W_{-u}$.

![Diagram](image.png)

Figure (5.4): Minimum energy levels of Example (5.3)
Furthermore, in the case that the initial state $\|x_0\| = R_0 \neq 0$ is fixed, the minimum input energy levels can be defined based on the scaled extremum eigenvalues of $W^{-1}_s$ via the norm of terminal state vector for the stable part, i.e. $\|x_f\| \leq R_f$, which are added by a constant value $\lambda^W_{\min} R_0^2$ related to the unstable part. $\lambda^W_{\min}$ represents the smallest eigenvalue of $W^{-1}_w$. Figure (5.6) and (5.7) describe the energy levels for example (5.2), in the case that $R_0$ is fixed at the values of 100 and 1000.

The importance of these energy levels is that for an upper bounded input energy they determine how far the system is guaranteed to be transferred from an initial state located on a sphere centered at the origin with radius $R_0 = cte$.
Figure (5.6): Pure energy values taken by the unstable subsystem when $R_0 = 1$ and $R_f = 0$ in Example (5.2)

Figure (5.7): Minimum energy levels when the system steers from the worst initial direction $R_0 (= 100, 1000)$ and moves toward the final states located on the sphere with the radius $R_f = 1$ in different directions in Example (5.2)
In a similar manner, if $R_f \neq 0$ is a fixed given value, the minimum input energy levels are defined as the extremum eigenvalues of $W^{-1}$ by the scaled norm of unstable modes’ initial states $\|x_{0u}\| \leq R_0$, which are increased by a constant value $\lambda_{\min}^W R_f^2$ related to the stable part.

$\lambda_{\min}^W$ defines the minimum eigenvalue of $W$. In this case, the energy levels describe the maximum distance $R_f$ the system can move away from the origin using a limited amount of input energy. Figure (5.8) shows this minimum energy value when $x_{0u} = 0$ and $R_f = 1$ in example (5.3). Figure (5.9) depicts the energy levels for two fixed values $R_f (= 100, 10000)$.

![Figure (5.8): Pure energy values taken by the stable subsystem when $R_f = 1$ and $R_0 = 0$ in Example (5.3)](image)

![Figure (5.9): Minimum energy levels when the system moves from various initial directions $R_0 = 1$ toward the final states located on the sphere with the radius $R_f (= 100, 10000)$ in the worst direction in Example (5.3)](image)
The energy levels discussed in this section imply that if the LTI system is anti-stable then the problem of transferring to zero requires more energy than the reachability problem, which deals with the problem of transferring the system from the origin \( \|x_0\| = R_0 = 0 \) to a terminal state \( \|x_f\| = R_f > 0 \), since the anti-stable systems tend to amplify the magnitude of the state.

In the case that the LTI system has a combination of stable and anti-stable parts, the selection of the initial and final states may cause simplifying or complicating the reachability problem and the controllability to 0 problem.

### 5.4. The Energy and Relative Degree of Controllability

In many practical cases, controllability should not be treated as a binary concept; that either a system is controllable or not. In many classes of control problems, it is also important to know how controllable a system is. Consider an uncontrollable system, if we start from a set including the sufficient number of actuators it will usually be the case that moving one of the actuators a small distance \( \varepsilon > 0 \) a controllable system will result. For a very small \( \varepsilon \), even though technically the system is controllable, in some sense it is almost uncontrollable (Viswanathan, et al., 1984).

According to (Burgmeier, 1992) consider the LTI system:

\[
\begin{array}{l}
\dot{x} = Ax + Bu \\
y = Cx + Du
\end{array}
\]

(5.38)

Then \( x(t) = x(t_f, t_0, x_0, u), t_f \geq t_0 \) denotes the unique solution of this system for a given initial condition \( x(t_0) = x_0 \in \mathbb{R}^n \) and \( u \in U \) where \( U \) is the set of all admissible controls for the system (5.38).

For a given non-empty set of final states \( X_f \in \mathbb{R}^n \) we define the \( X \)-controllability set of the system at time \( t_f \geq t_0 \) by \( S(t_0, t_f, U, X_f) = \{ x_0 \in \mathbb{R}^n, u \in U, x(t_f, t_0, x_0, u) \in X_f \} \).

As simple measures of the size of the \( X \)-controllability set at time \( t_f \geq t_0 \) with respect to a given point \( y \in \mathbb{R}^n \) we define two degrees of controllability and uncontrollability as:

**Definition (5.1):** (Burgmeier, 1992) in LTI system (5.38) the degree of controllability with respect to a given point \( y \in \mathbb{R}^n \) is defined as:

\[
\rho(t_0, t_f, U, X_f) = \inf\{ \|x - y\| : x \in \mathbb{R}^n, x(t_f, t_0, x_0, u) \}
\]

(5.39)

And the degree of uncontrollability could be stated as:

\[
\rho(t_0, t_f, U, X_f) = \sup\{ \|x - y\| : x \in \mathbb{R}^n, x(t_f, t_0, x_0, u) \}
\]

(5.40)
Where \( \| \| \) represents the Euclidean norm.

**Definition (5.2):** (Burgmeier, 1992) The recovery region for time \( T \) for normalized system (5.38) is the set 
\[
\mathcal{R} = \{ x_f \in X : \exists u \in U, \|u\|_{[0,t_f]} \leq 1, x(t_f, t_0, x_0, u) \}.
\]

Then, the recovery region denotes the reachability set of the system if we consider the problem of transferring the initial states \( x_0 \) to the origin, within the finite time \( t_f \) using the control input \( u \), \( \|u\| \leq 1 \).

Therefore, as explained in (Viswanathan, et al., 1984) the recovery region identifies all the final states to which at least one initial state can be transferred in an interval using bounded energy control. Also, the degree of controllability is a scalar measure of the size of the region chosen as the shortest distance from the final point to a given point \( y \in \mathbb{R}^n \), which is the desired terminal state.

Then, the control input energy can be employed as the degree of controllability to demonstrate the importance or lack of importance of each control input/actuator.

**Example (5.4):** Consider the linearized MIMO model of the active magnetic bearing (AMB) system (Noshadi, et al., 2016), (Schweitzer & Maslen, 2009), (Mazenc, et al., 2005):

AMBs can provide contactless suspension of the rotor by attractive forces produced by electromagnets. They are employed in a variety of rotating machines in place of conventional mechanical bearings.

In this example, the 18th order AMB has 4 coil currents as inputs and the outputs are defined as the position and the orientation of rotors.
Active magnetic bearing (AMB) systems are intrinsically unstable as they have some poles in the right half plane (RHP). One can easily decouple the whole model into stable and anti-stable parts using a linear transformation. In this case the eigenvalues are:

$$\lambda_k = \{-0.0001 - 1.2900i, -0.0001 + 1.2900i, -0.0001, -0.4826i, -0.0001 + 0.4826i, -0.0001 - 0.4835i, -0.0001 + 0.4835i, -0.0152 - 0.0336i, -0.0152 + 0.0336i\}$$

$$k = 1, 2, \ldots, 18$$

$$A_{18 \times 18} = \text{diag}\left[\begin{array}{ccc}
-0.0359 & 12900 & -0.55 & 4826 \\
-12900 & -0.0359 & -4826 & -0.55 \\
-120.9 & 341.8 & -111.8 & 228.2 \\
-341.8 & -120.9 & -228.2 & -111.8
\end{array}\right],
B_{18 \times 4} = \left[\begin{array}{c}
-0.0243 & 0.0176 & -0.0235 & -0.0193 \\
0.0453 & -0.0672 & 0.0486 & 0.0839 \\
0.5450 & 0.0348 & -0.5991 & 0.0686 \\
-1.4730 & 0.2672 & 1.7560 & 0.3069 \\
-0.1713 & -1.4250 & 0.2059 & -1.7710 \\
0.0820 & -0.7334 & -0.0522 & -0.8909 \\
-5.2730 & 4.2650 & -4.8140 & -3.4230 \\
-1.7190 & 3.8070 & 3.0520 & 9.5530 \\
-5.1950 & 2.4130 & -8.4050 & -3.1360 \\
-1.4100 & 1.6070 & 1.4170 & -3.6310 \\
-2.0850 & -1.2980 & -0.3873 & 4.2370 \\
-2.0320 & 5.3700 & -1.7270 & 5.5080 \\
7.4800 & -0.7283 & -5.0160 & -1.0070 \\
1.1000 & 0.9193 & 1.7110 & 1.5220 \\
8.7810 & -21.0600 & 0.5281 & 33.6300 \\
-44.7700 & -16.0000 & -53.4300 & 1.0640 \\
-50.9400 & -8.0200 & 65.9800 & 13.8900 \\
34.1700 & -55.9100 & -50.3000 & -44.6800
\end{array}\right],
C_{4 \times 18}^T = \left[\begin{array}{c}
7.2300 & -2.1590 & -1.5950 & 6.9340 \\
-5.5360 & 16.7600 & -18.1900 & -20.8700 \\
-10.4100 & 0.5644 & 16.4900 & 1.6760 \\
-5.5310 & 1.7130 & 3.7640 & 1.2470 \\
-1.2190 & -6.6440 & 1.9040 & -1.9720 \\
0.7982 & 11.4200 & -1.7800 & 14.4700 \\
-1.3220 & -15.4500 & 0.3159 & 9.0980 \\
8.9640 & -0.9097 & 3.0000 & 1.0810 \\
3.3940 & 1.7800 & 1.0590 & -15.4300 \\
-3.4690 & -17.5500 & 3.2230 & -10.5200 \\
-8.5000 & -1.9170 & 10.1500 & -1.3500 \\
0.6677 & -2.5790 & 0.3406 & 3.1050 \\
-2.1520 & -0.2261 & -1.9370 & 0.4813 \\
-1.1260 & -1.9160 & 1.4620 & -0.5211 \\
0.6849 & -1.9850 & 0.2637 & -1.0290 \\
\end{array}\right].
$$

Let us define the initial condition and the terminal state as:

$$x_0 = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0.2 & 0 & 0 & 0 \end{bmatrix}^T, x_f = \begin{bmatrix} 2 & 0 & \cdots & 0.3 & 0 & 0 & 0 \end{bmatrix}^T. $$
We can achieve the values for the degree of controllability proposed in this section by using the Lyapunov equations for the above decoupled matrices, and we get:

\[
\begin{align*}
E_{\min_1} &= E_{s[B_1]} + E_{u[B_1]} = 217.3633 + 6.9140e+04 = 6.9357e+04 \\
E_{\min_2} &= E_{s[B_2]} + E_{u[B_2]} = 119.0317 + 1.2020e+04 = 1.2139e+04 \\
E_{\min_3} &= E_{s[B_3]} + E_{u[B_3]} = 511.1147 + 1.9115e+07 = 1.9116e+07 \\
E_{\min_4} &= E_{s[B_4]} + E_{u[B_4]} = 95.7876 + 4.7137e+03 = 4.8095e+03
\end{align*}
\]

Here \( E_{\min_i}, i = 1, 2, 3, 4 \) denote the minimum input energy for \( i \)th input, and \( E_{s[B_i]} E_{u[B_i]}, i = 1, 2, 3, 4 \) are the steady state energy values stored in the stable and anti-stable subsystems related to \( i \)th input respectively.

The results show that for this initial condition and final state, it is efficient to apply the last input since it corresponds to the minimum energy.

Now let us find DOC by a different set of initial and final state:

\[
\begin{align*}
x_0 &= [1 \ 0 \ \cdots \ 0 \ 1 \ 0 \ 0 \ 0 \ 2]^T, \quad x_f = [2 \ 0 \ \cdots \ 0 \ 2 \ 0 \ 0 \ 0 \ 0.5]^T
\end{align*}
\]

In this case, we get:

\[
\begin{align*}
E_{\min_1} &= E_{s[B_1]} + E_{u[B_1]} = 1.5376e+04 + 3.1321e+08 = 3.1322e+08 \\
E_{\min_2} &= E_{s[B_2]} + E_{u[B_2]} = 7.5852e+04 + 1.2252e+08 = 1.2260e+08 \\
E_{\min_3} &= E_{s[B_3]} + E_{u[B_3]} = 7.5184e+04 + 1.4773e+08 = 1.4780e+08 \\
E_{\min_4} &= E_{s[B_4]} + E_{u[B_4]} = 9.8826e+04 + 1.4857e+08 = 1.4867e+08
\end{align*}
\]

The achieved results are different from the first case. Here the 2nd input leads to less steady state input energy and thus it is more controllable relative to the other inputs. The results also show that this approach is sensitive to the initial and final conditions, i.e. the initial and terminal states may vary the result. In this case the proper initial and terminal conditions that coincide with the control objective should always be selected (Lee & Park, 2014).

To remove this sensitivity and increase the reliability of the input selection, we can define the degree of controllability (DOC) as the worst-case metric where we evaluate the maximum of minimum energy, which specifies the largest possible amount of energy required to transfer the states from any initial point to any final states located in the sphere with the center in the origin and the radius \( R \), i.e.

\[
\|x_f\| \leq R, \|x_{u0}\| \leq R \quad \text{(without loss of generality one can simply normalize } x_0, x_f \text{.)}
\]

wherever we have not enough information of the initial and final conditions of the system the presented energy DOC measure for testing the controllability of a system is not satisfactory in the sense it may lead to wrong conclusions.

Therefore, we introduce another controllability measure, based on the maximum value of minimum energy as a metric which seems to be useful, especially in the case that we are designing the system for various initial and final conditions.
We can readily define the above energy measure as a linear combination of eigenvalues of the inverse of controllability Gramians $W_s^{-1}, W_u^{-1}$ for the stable and anti-stable subsystems respectively, where the coefficients are determined through the initial and final states.

In the worst case, according to the explanations given in section 5.3, the normalized initial condition must represent the eigenvector corresponding to the maximum eigenvalue of $W_u^{-1}$ and the terminal states would be in the same direction of the eigenvector corresponding to the maximum eigenvalue of $W_s^{-1}$. Then DOC measure could be defined as the summation of maximum eigenvalues of $W_u^{-1}$ and $W_s^{-1}$.

If we use this DOC measure in our example, we get:

$$
\begin{align*}
E_{\text{min},u}\bigg|_{\text{max}} &= E_{s(B_1)} + E_{u(B_1)} \bigg|_{\text{max}} = 3.6377e+04 + 1.5978e+08 = 1.5982e+08 \\
E_{\text{min},s}\bigg|_{\text{max}} &= E_{s(B_1)} + E_{u(B_1)} \bigg|_{\text{max}} = 1.6471e+05 + 2.9201e+09 = 2.9202e+09 \\
E_{\text{min},s}\bigg|_{\text{max}} &= E_{s(B_1)} + E_{u(B_1)} \bigg|_{\text{max}} = 9.4060e+04 + 5.5939e+08 = 5.5948e+08 \\
E_{\text{min},u}\bigg|_{\text{max}} &= E_{s(B_1)} + E_{u(B_1)} \bigg|_{\text{max}} = 6.8675e+04 + 1.7343e+10 = 1.7343e+10
\end{align*}
$$

The results show that the first input could be a reasonable choice if we are interested in the robustness of degree of controllability over the set of all possible initial and final states, though the choosing other inputs might result in a better degree of controllability for some specific initial and final conditions, they do not provide guaranteed degree of controllability, i.e. there might exist some initial or final states which lead to nearly uncontrollable system.

\textbf{Remark (5.3):} Consider the LTI system (5.30) with stable and anti-stable modes, with defined initial and final states the energy optimal degree of controllability (DOC) could be defined as the minimum input energy in the steady state defined by Theorem (5.3).

$$
E_{\text{min}}(\infty) = \left[ (x_f)^T, (x_0)^T \right] \begin{bmatrix} W_s^{-1} & 0 \\ 0 & W_u^{-1} \end{bmatrix} \begin{bmatrix} (x_f)^T \\ (x_0)^T \end{bmatrix}
$$

(5.41)

\textbf{Remark (5.4):} Consider the LTI system (5.38) with stable and anti-stable modes. The worst-case energy optimal degree of controllability (WDOC) can be defined as:

$$
E_{\text{min}}(\infty) = \left[ (V_{\text{max}}W_s^{-1})^T_s, (V_{\text{max}}W_u^{-1})^T_u \right] \begin{bmatrix} W_s^{-1} & 0 \\ 0 & W_u^{-1} \end{bmatrix} \begin{bmatrix} (V_{\text{max}}W_s^{-1})^T_s \\ (V_{\text{max}}W_u^{-1})^T_u \end{bmatrix}
$$

(5.42)

Where $V_{\text{max}}W_s^{-1}$ shows the eigenvector corresponding to the maximum eigenvalue of $W_s^{-1}$ and $V_{\text{max}}W_u^{-1}$ determines the eigenvector associated with the maximum eigenvalue of $W_u^{-1}$.
Let us investigate further the difference of two proposed DOC measures using the example of linearized seesaw-pendulum.

We refer the interested readers for details on the linearized model of seesaw-pendulum to (Subbotin, 2004), (Jirstrand & Gunnarsson, 2001).

**Example (5.5):** consider the seesaw-pendulum process in Figure (5.11).

![Figure (5.11): The schematic of seesaw-pendulum process](image)

In this system, we have an inverted pendulum coupled with a second cart which can move in parallel on top of an unstable structure - a seesaw. This system has two inputs: a force applied to the cart with the pendulum installed on top of it and a force applied to the load cart which tries to balance the seesaw structure.

Let us define the state vector as:

\[
x = \begin{bmatrix} x_1 & \theta & x_2 & \alpha & \dot{x}_1 & \dot{x}_2 & \dot{\alpha} \end{bmatrix}^T
\]

where \( x_1 \) is the position of the load cart, \( \theta \) is the angle of the seesaw structure with respect to the vertical axis, \( x_2 \) represents the position of the pendulum cart and \( \alpha \) denotes the angle of the pendulum with respect to the seesaw symmetry axis, then we can write the linearized model of the system as:

\[
\dot{x} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} x + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} [u_1, u_2]
\]

Have:
\[ A_{11} = 0_{4 \times 4} \]
\[ A_{12} = I_4 \]
\[ A_{21} = \begin{bmatrix} -ghm_j / J & g - cghm_z / J & -gh(m_z + m_p) / J & 0 \\ gm_i / J & cgm_z / J & g(m_z + m_p) / J & 0 \\ -ghm_j / J & g - cghm_z / J & -gh(m_z + m_p) / J & -gm_p / m_z, \\ -gm_i / J & -cgm_z / J & -g(m_z + m_p) / J & g(m_z + m_p) / l_p m_z \end{bmatrix} \]
\[ A_{22} = \begin{bmatrix} -\alpha(J + m_i h^2) / m_z J & 0 & -\alpha h^2 / J & 0 \\ \alpha h / J & 0 & \alpha h / J & 0 \\ -\alpha h^2 / J & 0 & -\alpha(J + m_i h^2) / J & 0 \\ -\alpha h / J & 0 & -\alpha(l_p hm_z - J) / m_z l_p & 0 \end{bmatrix} \]
\[ b_i = 0_{4 \times 2}, b_2 = \begin{bmatrix} (J + m_i h^2) / m_z J & h^2 / J \\ -h / J & -h / J \\ h^2 / J & (J + m_i h^2) / J \\ -h / J & (l_p hm_z - J) / m_z l_p \end{bmatrix} \]

Assuming all parameters are equal to 1 then we get the LTI system below:

\[ A_{8 \times 8} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & -2 & 0 & -2 & 0 & -1 & 0 \\ 1 & 1 & 2 & 0 & 1 & 0 & 1 & 0 \\ -1 & 0 & -2 & -1 & -1 & 0 & 2 & 0 \\ -1 & -1 & -2 & 2 & -1 & 0 & 0 & 0 \end{bmatrix}, B_{8 \times 2} = \begin{bmatrix} 0 & 0 \\ 2 & 1 \\ 1 & -1 \\ 1 & 2 \\ 1 & 0 \end{bmatrix} \]

This system has three anti-stable modes. Using a similarity transformation, we can split it into a 5-dimensional stable part and a 3-dimensional anti-stable subsystem.

Let us consider the first case, where the initial and final conditions are given as:

\[ x_0 = [0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]^T, \quad x_f = [0 \ 2 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]^T \]

Then the results of the first DOC are achieved as:

\[ E_{\text{min}} = E_{x[B_1]} + E_{u[B_1]} = 1.6358e+05 + 5.8621 = 1.6358e+05 \]
\[ E_{\text{min}} = E_{x[B_2]} + E_{u[B_2]} = 3.4472e+33 + 29.9975 = 3.4472e+33 \]

Recall that \( \{B_i\}, i = 1, 2 \) means the part of the \( B \) matrix related to \( i^\text{th} \) input.
As shown in the results, it is efficient to apply force to the cart with pendulum to change the angle of the seesaw structure.

Now consider the second case, where the angle of pendulum changes from 1 to 2:

\[
x_0 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}^T, \quad x_f = \begin{bmatrix} 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \end{bmatrix}^T
\]

Then, the values of the proposed measure are:

\[
\begin{align*}
\min \{B\} u \{B\} &= 1.4566 \times 10^6 \\
\max \{B\} u \{B\} &= 4.4755 \times 10^4
\end{align*}
\]

In this case, the results show that it is efficient to apply force to the load cart to change the pendulum’s angle.

Note also that DOC result of the system varies with the initial and final conditions.

While the second DOC measure, which illustrates the degree of controllability in the worst case, leads to the same values for all normalized initial and final states:

\[
\begin{align*}
\min \{B\} u \{B\} &= 1.0588 \times 10^3 \\
\max \{B\} u \{B\} &= 1.7695 \times 10^3
\end{align*}
\]

The results show that if we use the force to the load cart there exist at least one direction in which the system would be uncontrollable or at least hardly controllable, while if we select the first input (force to the cart with pendulum), even in the worst case the value of control energy does not exceed 1.0913e+03, and the system remains controllable.

From this simple example, it becomes apparent that the proposed worst case method (WDOC) is a more meaningful controllability metric rather than the steady state energy measure (EDOC), which is dependent to the initial and final conditions and may produce opposite results according to these conditions.

In summary, if we know the initial and final condition of a system we can easily use the first DOC measure to precisely determine the degree of controllability of the system, however in the case that we are considering the system with various initial and final conditions, the second DOC measure seems more reliable.

### 5.5. Degree of Controllability

In this section, a novel measure of controllability is proposed, which amounts to generalization of Kalman’s rank test to quantify the degree of controllability.

In fact, the proposed (WDOC) approach, in the previous section, determines the distance to uncontrollability, which was firstly defined by (Paige, 1981):
**Definition (5.3):** The distance to uncontrollability is the spectral norm distance of the pair \((A, B)\) from the set of all uncontrollable pairs:

\[
d(A, B) = \min_{\varepsilon \in \mathbb{C}} \{ \| E, F \| : (A + E, B + F) \text{Uncontrollable} \}
\]  

(5.43)

where \(\| \|\) denotes the spectral norm.

This definition has its root in (Hautus, 1970), (Hautus, 1969) characterization of controllability (5.44) and confirms the definition (5.1) involving the degree of controllability, i.e.

\[
\text{rank} \left( \left[ A - sI, B \right] \right) = n, \forall s \in \mathbb{C}, A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}
\]  

(5.44)

According to (Eising, 1984), (Eising, 1984) \(d(A, B)\) can be equivalently expressed as:

\[
d(A, B) = \min \sigma, \left( \left[ A - sI, B \right] \right), A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}
\]  

(5.45)

where \(\sigma_{\min}\) denotes \(n^{th}\) singular value of augmented matrix \([A - sI, B]\).

There is a considerable number of research papers which quantify the distance to uncontrollability based on the above definition (Wicks & DeCarlo, 1991), (Boley & Lu, 1986), (Eising, 1984) (Gahinet & Laub, 1992) (Tarokh, 1992). Special purpose optimization methods have been developed for calculating this distance, however the current methods usually face difficulties due to the presence of numerous local minimum points, so that the algorithm method search for all of them in order to guarantee the optimal solution (Byers, 1989). Furthermore, a good starting point is typically needed to identify the global optimum (Boley, 1987), (Byers, 1989). In addition, some of the existing methods may fail to detect nearly uncontrollable systems (Paige, 1981) (Hu & Davison, 2001).

Next, we propose a new approach based on Kalman’s controllability criteria (Ogata, 1997) to solve distance to uncontrollability problem. This new method removes many drawbacks of the current heuristic/numerical methods, and produces more accurate results. Then we will use this novel method to check the validity and accuracy of the worst-case DOC measure (WDOC), which was introduced above.

**Lemma (5.2):** (Kalman, 1963) A necessary and sufficient condition for \((A, B)\) to be controllable is:

\[
\text{rank}(C) = \text{rank} \left( \left[ \begin{array}{c|c|c|c|c} & B & AB & A^2B & \cdots & A^{n-1}B \end{array} \right] \right) = n, A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}
\]  

(5.46)

where \(C\) is called Kalman’s controllability matrix or Kalman block matrix (of size \(n \times nm\)).

For the proof see (Kalman, et al., 1963), (Ogata, 1997). The above Lemma is a useful and simple test but less effort has been spent to generalize it from classical binary rank test to a test which indicates the distance to uncontrollability. In fact, the characterization \(d(A, B)\) based on the PBH test has received more attention in literature (Eising, 1984), (Kenney & Laub, 1988). In Theorem (5.4), we propose a new way of calculating distance to uncontrollability based on Kalman’s rank condition.
**Theorem (5.4):** Consider the LTI system described by (5.38). The distance to controllability is:

\[
d(A, B) = \min_{s \in \mathbb{C}} \sigma_n(V[I_n, \Lambda - sI_n, (\Lambda - sI_n)^2 \cdots (\Lambda - sI_n)^{n-1}](I_n \otimes V^{-1}B))
\] (5.47)

where \( A = V \Lambda V^{-1} \) is the Eigen-decomposition of system matrix \( A \), and \( \sigma_n \) denotes \( n \)th singular value of modified Kalman block matrix \( \begin{bmatrix} B & (A - sI_n)B & (A - sI_n)^2 B & \cdots & (A - sI_n)^{n-1}B \end{bmatrix} \).

\[
\square
\]

**Proof:**

Consider the controllable pair \((A, B)\). According to the definition (5.3) we are looking for the minimum-norm perturbation such that \((A + \Delta A, B + \Delta B)\) is uncontrollable, so according to Kalman’s rank condition we have:

\[
\text{rank} \left( \begin{bmatrix} B + \Delta B & (A + \Delta A)(B + \Delta B) & (A + \Delta A)^2 (B + \Delta B) & \cdots & (A + \Delta A)^{n-1}(B + \Delta B) \end{bmatrix} \right) < n
\] (5.48)

Additionally, based on (Kienitz, 2012), (Eising, 1984) we know that for uncontrollable pair \((A + \Delta A, B + \Delta B)\) there exists a left-eigenvector/eigenvalue pair \((V_i, \lambda_i)\) of \((A + \Delta A)\) such that:

\[
V_i(A + \Delta A) = \lambda_i V_i, \quad V_i B = 0
\] (5.49)

Furthermore, it is easy to see that the eigenvalues of \( A - sI_n \) for any scalar \( s \) are those of \( A \) subtracted by \( s \), while the eigenvectors remain those of \( A \). Further controllability of pairs \((A, B)\) and \((A - sI_n, B)\) is equivalent, since:

\[
V_i A = \lambda_i V_i, V_i B \neq 0 \implies V_i(A + sI_n) = (\lambda_i - s)V_i, V_i B \neq 0
\] (5.50)

Kalman’s controllability matrix for the second pair gives:

\[
\begin{bmatrix} B & (A - sI_n)B & (A - sI_n)^2 B & \cdots & (A - sI_n)^{n-1}B \end{bmatrix}
\] (5.51)

Now considering the Eigen-decomposition of \( A \), i.e. \( A = V \Lambda V^{-1} \) (in the case of repeated eigenvalues the singular decomposition is used), we get:

\[
\begin{array}{l}
[B & V(\Lambda - sI_n)V^{-1}B & V(\Lambda - sI_n)^2V^{-1}B & \cdots & V(\Lambda - sI_n)^{n-1}V^{-1}B] \\
\Rightarrow [VV^{-1} & V(\Lambda - sI_n)V^{-1} & V(\Lambda - sI_n)^2V^{-1} & \cdots & V(\Lambda - sI_n)^{n-1}V^{-1}](I_n \otimes B) \\
\Rightarrow V[V^{-1} & (\Lambda - sI_n)V^{-1} & (\Lambda - sI_n)^2V^{-1} & \cdots & (\Lambda - sI_n)^{n-1}V^{-1}](I_n \otimes B) \\
\Rightarrow (V[I_n & (\Lambda - sI_n)(\Lambda - sI_n)^2 & \cdots & (\Lambda - sI_n)^{n-1})(I_n \otimes V^{-1})(I_n \otimes B)) \\
\Rightarrow (V[I_n & (\Lambda - sI_n)(\Lambda - sI_n)^2 & \cdots & (\Lambda - sI_n)^{n-1}](I_n \otimes V^{-1}B))
\end{array}
\] (5.52)

Then according to (5.50) and using definition (5.3), calculating the distance to uncontrollability reduces to finding the minimum of \( n \)th singular values of (5.52) over \( s \in \mathbb{C} \) and the proof is completed.
Now we illustrate the method introduced above with some numerical examples, each representing a
time-invariant linear system of the form (5.38). The examples show the comparable accuracy and validity of our method.
The first example is taken from (Boley & Lu, 1986):

**Example (5.6):** The LTI system is defined by:

\[
A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}
\]

It is clear by Kalman’s criterion that this system is controllable, and by applying the Theorem above we obtain \(d(A, B) = 0.6616\).

In the second example, we start with an uncontrollable system which is studied in various references (Paige, 1981), (Boley, 1987), etc. This example shows the power of our method in detecting the systems very close to uncontrollability where many existing methods may not indicate that fact (in the most cases, our approach is comparable to Eising’s method, and in many cases, it produces better results).

For this particular example with repeated eigenvalues, the best current distance to uncontrollability by staircase algorithm is 0.00087689843786 (Boley, 1987). The found solution by Eising’s method is 1.1308e-22.

**Example (5.7):** Consider the pair:

\[
A = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\]

The distance to uncontrollability achieved by the above Theorem is exactly zero, i.e. DOC=0, which is much more accurate rather than the solutions found by other approaches.

Now consider the following example, which is borrowed from (Khare, et al., 2012):

**Example (5.8):** Consider the LTI system:

\[
\dot{x} = \begin{bmatrix} -1 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -3 \end{bmatrix} x + \begin{bmatrix} 0 \\ t \\ u, t \in \mathbb{R} \end{bmatrix}
\]
In this example, the distance to uncontrollability changes as the parameter $t$ varies. Table (5.1) demonstrates the real controllability radius obtained by Theorem (5.4) and compares our method with the results achieved by Eising’s method for five different values of $t$, clearly as $t$ decreases the real controllability radius decreases too (as would be expected since the system is uncontrollable with $t = 0$). The comparison of the results in Table (5.1) confirms the accuracy of our approach.

<table>
<thead>
<tr>
<th>$t$</th>
<th>New Theorem (5.4)</th>
<th>Eising’s Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.2192</td>
<td>0.2165</td>
</tr>
<tr>
<td>2</td>
<td>0.7181</td>
<td>0.7180</td>
</tr>
<tr>
<td>1</td>
<td>0.3779</td>
<td>0.6063</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0388</td>
<td>0.645</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>3.1659e-05</td>
</tr>
</tbody>
</table>

Table (5.1): Results of Example (5.8)

Now let us investigate if the Theorem (5.4) supports the actuator selection which was suggested in Examples (5.4), (5.5) using our WDOC criterion.

Table (5.2) lists the real controllability radius of each actuator selection of AMB system in example (5.4). The results show that the first input produces more controllability. It supports the results of WDOC measure.

<table>
<thead>
<tr>
<th>Input 1</th>
<th>Input 2</th>
<th>Input 3</th>
<th>Input 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d(A, B)$ New Method</td>
<td>0.0175</td>
<td>0.0041</td>
<td>0.0107</td>
</tr>
</tbody>
</table>

Table (5.2): Distance to uncontrollability for AMB system in Example (5.4)

The real controllability radius of seesaw-inverted pendulum in example (5.5) is shown in Table (5.3). The results also confirm the input selection by WDOC.

<table>
<thead>
<tr>
<th>Input 1</th>
<th>Input 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d(A, B)$ New Method</td>
<td>0.0147</td>
</tr>
</tbody>
</table>

Table (5.3): Distance to uncontrollability for seesaw-inverted pendulum system in Example (5.5)

### 5.6. Degree of Disturbance Rejection

Consider the linear time-invariant system given in equation (5.36) with an input disturbance $d$:

$$
\begin{align*}
\dot{x} &= Ax + Bu + B_d d, \quad A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}, B_d \in \mathbb{R}^{n \times p}, C \in \mathbb{R}^{q \times n}, D \in \mathbb{R}^{q \times m} \\
y &= Cx + Du
\end{align*}
$$

(5.53)

Degree of disturbance rejection is defined as minimum energy required to steer initial state $x_i$ to the final state $x_f$ under the presence of external disturbance $d$ (Lee & Park, 2014).

The state transition equation for above system is given by (Chen, 1984):
\[ x_f = e^{At}x_0 + \int_0^t e^{A\tau}B_d d(\tau) d\tau \]  
(5.54)

(Friedland, 1986) has defined the disturbance sensitivity Gramian as:

\[ W_d(0,t_f) = \int_0^t e^{A\tau}B_d B_d^T e^{A^T\tau} d\tau \]  
(5.55)

**Lemma (5.2):** The disturbance energy of the LTI system in equation (5.36) is given by:

\[ E^d_{(0,t_f)} = (x_f - e^{At}x_0)^T W_d^{-1}(0,t_f)(x_f - e^{At}x_0) \]  
(5.56)

where \( x_0, x_f \) are initial and terminal states respectively, and \( W_d^{-1}(0,t_f) \) is the inverse of disturbance sensitivity Gramian.

\[ \square \]

**Proof:**

Suppose that \( W_d(0,t_f) \) is non-singular for some finite \( 0 \leq t_f \), then the disturbance, which takes the system from some initial state \( x_0 \) to the final state \( x_f \) is given by:

\[ d(t) = B_d^T(t)e^{A(t-t_f)}W_d^{-1}(0,t_f)(x_f - e^{At}x_0), 0 \leq t \leq t_f \]  
(5.57)

Then the disturbance input energy is defined as:

\[ E^d_{(0,t_f)} = \|d(t)\|^2_\Sigma = \int_0^{t_f} d^T(t)d(t)dt \]
\[ = \int_0^{t_f} d^T(t)B_d^T(t)e^{A^Tt}W_d^{-1}(0,t_f)(x_f - e^{At}x_0)dt \]
\[ = (x_f - e^{At}x_0)^T W_d^{-1}(0,t_f)(x_f - e^{At}x_0) \]  
(5.58)

\[ \square \]

In the case that system is transferred from the origin \( x_0 = 0 \) equation (5.56) could be simplified as:

\[ E^d_{(0,t_f)} = x_f^TW_d^{-1}(0,t_f)x_f \]  
(5.59)

From the above Lemma, it can be seen that the inverse of disturbance sensitivity Gramian \( W_d^{-1}(0,t_f) \) should be maximized to make the system maximally insensitive to the disturbance. Equivalently, \( W_d(0,t_f) \) needs to be minimized to have a minimum effect of disturbance on the system. This could be used to define an appropriate criterion as an indication of the optimal position of the actuators to reject an external disturbance acting on the system.

The degree of disturbance rejection (DODR) is the capability of the system to prevent the disturbance energy to be transferred into the state of the system. This is defined as the trace of the inverse of
controllability Gramian multiplied by the sensitivity Gramian in the steady state (Lee & Park, 2015), (Lee & Park, 2014):

\[
\rho = \lim_{{t \to \infty}} \text{trace}(W_c^{-1}(t)W_d(t))
\]  

(5.60)

Clearly the above measure has a physically meaningful value of input energy and has the potential to be modified to unstable systems by following the procedure proposed in the previous section. \( W_c, W_d \) can be obtained from the (CALE) equations:

\[
\ddot{A}W_c + W_c\ddot{A}^T = -BB^T, \quad \ddot{A}W_d + W_d\ddot{A}^T = -B_dB_d^T
\]  

(5.61)

where \( \ddot{A} \) denotes the proper related system matrix, which would be \( A \) in the stable case, and \( -A \) in anti-stable system. Although this is computationally straightforward we will show that it does not necessarily lead to the correct answer where the growth or decrease in one eigenvalue of \( W_c^{-1}W_d \) is large, relative to the increment or decrement of the other eigenvalues. We offer the following illustrative counterexample:

**Example (5.9):** Consider the linearized model of a two-CSTR process, which is schematically shown in Figure (5.12). A full description of the system and an eight-state model can be found in (Cao & Biss, 1996).

Figure (5.12): Two-CSTR Process

Let us define cooling water temperature fluctuations as the disturbance and assume that we have two options for control inputs:

i. two feed flowrates

ii. two cooling-water flowrates

\[
\begin{align*}
\dot{x} &= Ax + Bu + B_d d \\
y &= Cx + Du
\end{align*}
\]

...
If we choose the first input $B_1$, the degree of disturbance rejection would be obtained as:

$$\rho_1 = \lim_{t_f \to \infty} \text{trace}(W_c^{-1}W_d) = 17.0353$$

While by selecting the second input $B_2$, we will have:

$$\rho_2 = \lim_{t_f \to \infty} \text{trace}(W_c^{-1}W_d) = 1.0154$$

The two results yield that the control input energy of first input $B_1$ is larger than that of second input $B_2$, which suggest input $B_2$ is the best selection, while we know that $B_d$ is same in both cases then the degree of controllability could be a good indicator of the disturbance rejection capability of the system and it suggests that input $B_1$ is a better choice, because the first input causes more controllability, $DOC_1 = 0.0024, DOC_2 = 9.6533e-05$.

In following, we introduce a new DODR, based on worst case of steady state input energy. We will show that this new method can be used as a reliable criterion to measure the degree of disturbance rejection of the system, and it can easily be modified for unstable systems.

**Definition (5.4):** Consider the LTI system in equation (5.53), the degree of disturbance rejection could be defined as:

$$\rho = \frac{\text{trace}(W_c(\infty))}{1 + \text{trace}(W_c(\infty))} + \frac{1}{\text{trace}(W_d(\infty))}$$  \hspace{1cm} (5.62)
$W_c(\infty)$ denotes the controllability Gramian in the steady state situation for the system without the disturbance, i.e. $\dot{x} = Ax + Bu, d = 0, t_f \to \infty$, and $W_d(\infty)$ demonstrates the sensitivity Gramian of the system without control input when $t_f \to \infty$ i.e. $\dot{x} = Ax + B_d d, u = 0$.

Consider the continuous time invariant system in equation (5.53), we can write it as:

$$\dot{x} = Ax + \begin{bmatrix} B & B_d \end{bmatrix} \begin{bmatrix} u \\ d \end{bmatrix}$$  \hspace{1cm} (5.63)

Then with regard to the definition of controllability Gramian we obtain:

$$W_{sys}(0,t_f) = \int_0^{t_f} e^{At} \begin{bmatrix} B & B_d \end{bmatrix} \begin{bmatrix} B \\ B_d \end{bmatrix} e^{A^T \tau} d\tau$$  \hspace{1cm} (5.64)

Thus, we have:

$$\text{trace}(W_{sys}(0,t_f)) = \text{trace} \int_0^{t_f} e^{At} \begin{bmatrix} B & B_d \end{bmatrix} \begin{bmatrix} B \\ B_d \end{bmatrix} e^{A^T \tau} d\tau$$

$$\Rightarrow \text{trace}(W_{sys}(0,t_f)) = \text{trace} \int_0^{t_f} e^{At} BB^T e^{A^T \tau} d\tau + \text{trace} \int_0^{t_f} e^{At} B_d B_d^T e^{A^T \tau} d\tau$$

Now assume we define:

$$k = \text{trace}(W_{sys}(0,t_f)), a = \text{trace} \int_0^{t_f} e^{At} BB^T e^{A^T \tau} d\tau, b = \text{trace} \int_0^{t_f} e^{At} B_d B_d^T e^{A^T \tau} d\tau$$

Then (5.65) can be written as:

$$k - b = a$$

According to $a \geq \frac{a}{1 + a}$ we get:

$$k - b \geq \frac{a}{1 + a}$$  \hspace{1cm} (5.66)

One can add $0 \leq \frac{1}{b}$ to both sides of statement (5.66), then we obtain:
\[ \Rightarrow k - b + \frac{1}{b} \geq \frac{a}{1+a} + \frac{1}{b} \]
\[ \Rightarrow \frac{kb - b^2 + 1}{b} \geq \frac{a}{1+a} + \frac{1}{b} \]
\[ \Rightarrow \frac{ab + b^2 - b^2 + 1}{b} \geq \frac{a}{1+a} + \frac{1}{b} \]
\[ \Rightarrow \frac{ab + 1}{b} \geq \frac{a}{1+a} + \frac{1}{b} \geq 0 \] \hspace{1cm} (5.67)

Hence, with regard to minimum input energy optimization problem we know that the left-hand side of the statement (5.67) should be maximized that is a non-convex optimization, then using a relaxation we can optimize input energy by maximizing the lower bound \( \rho = \frac{a}{1+a} + \frac{1}{b} \):

\[ \max(a + \frac{1}{b}) \Rightarrow \max(\frac{ab + 1}{b}) \Rightarrow \max(\frac{a}{1+a} + \frac{1}{b}) \] \hspace{1cm} (5.68)

\[ \Rightarrow \rho = \frac{\text{trace}(W_c(\infty)) + 1}{1 + \text{trace}(W_c(\infty))} \] \hspace{1cm} (5.69)

Thus statement (5.62) is verified.

\( \square \)

**Remark (5.5):** In the case that \( A \) is stable we can easily use the Lyapunov equations to get DODR defined in (5.4):

\[
\begin{align*}
AW_{\text{sys}}(\infty) + W_{\text{sys}}(\infty)A^T &= -[B \quad B_d] \begin{bmatrix} B \\ B_d \end{bmatrix} \\
AW_c(\infty) + W_c(\infty)A^T &= -BB^T \\
AW_d(\infty) + W_d(\infty)A^T &= -B_dB_d^T
\end{align*}
\] \hspace{1cm} (5.70)

Furthermore, as discussed in previous sections, in the case that \( A \) is unstable, i.e. \( A \) has some eigenvalues on the right half plane (RHP), we may decompose the system into stable and anti-stable parts, and then we can easily use the Lyapunov equations to find the controllability and sensitivity Gramians of each subsystem separately:

\[
\begin{align*}
A_sW_{\text{sys},s}(\infty) + W_{\text{sys},s}(\infty)A_s^T &= -[B_s \quad B_{ds}] \begin{bmatrix} B_s \\ B_{ds} \end{bmatrix} \\
A_sW_{c,s}(\infty) + W_{c,s}(\infty)A_s^T &= -B_sB_s^T \\
A_sW_{a,s}(\infty) + W_{a,s}(\infty)A_s^T &= B_sB_s^T \\
A_sW_{d,s}(\infty) + W_{d,s}(\infty)A_s^T &= -B_{ds}B_{ds}^T \\
A_sW_{c,a}(\infty) + W_{c,a}(\infty)A_a^T &= B_uB_u^T \\
A_sW_{a,a}(\infty) + W_{a,a}(\infty)A_a^T &= B_uB_u^T \\
A_sW_{d,a}(\infty) + W_{d,a}(\infty)A_a^T &= B_{du}B_{du}^T
\end{align*}
\] \hspace{1cm} (5.71)
Therefore, our method is not computationally intensive either. It can be readily solved using the Lyapunov equations.

Now let us test the validity of the proposed DODR measure in example (5.9):

The eigenvalues of $A$ are:

\[
\lambda = \{-17.6314, -17.7915, -0.1142, -0.1283, -0.8406, -0.6678\}
\]

By choosing the first input $B_1$, we obtain
\[
\rho = 0.9699 + 444.8508 = 445.8207
\]

while the second input $B_2$ yields
\[
\rho = 0.0122 + 444.8508 = 444.8630
\]

thus the results suggest the CSTR plant with $B_1$ would be less affected by the disturbance $d$.

Now consider another example cited from (Johnston & Barton, 1985):

**Example (5.10):** Consider the linear system with three candidate inputs, $u$ described by the following state space model:

\[
\begin{align*}
\dot{x} &= Ax + Bu + B_d d \\
y &= Cx
\end{align*}
\]

\[
A = \begin{bmatrix}
-1 & -2 & 3 & 0 & 0 & 1 \\
0 & -2 & 10 & 0 & 2 & 0 \\
0 & 0 & -20 & 10 & 0 & -20 \\
0 & 0 & 0 & -3 & 0 & 0 \\
4 & 0 & 0 & 0 & -10 & 2 \\
1 & 1 & 0 & 0 & 3 & -15
\end{bmatrix}
\]

\[
B_1 = \begin{bmatrix}
1 \\
5 \\
0 \\
0 \\
2 \\
1
\end{bmatrix},
B_2 = \begin{bmatrix}
-4 \\
1 \\
5 \\
1 \\
-4 \\
-2
\end{bmatrix},
B_3 = \begin{bmatrix}
2 \\
0 \\
1 \\
0 \\
0 \\
0
\end{bmatrix},
C = \begin{bmatrix}
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

\[
d = \begin{bmatrix}
1 \\
0 \\
0 \\
0 \\
0 \\
0
\end{bmatrix}
\]

The aim is to select two inputs from the three candidates by considering the disturbance rejection properties of each input. Table (5.4) demonstrates the results of our method in measuring the system's capability in disturbance rejection. In the original work Johnston and Barton claimed that
{B_1, B_2} is the best input set. However, later (Cao, et al., 1997) through Input-Disturbance Gain Deviation (IDGD) method in the frequency domain revealed that this selection appears to be wrong and showed that the best performance is expected from selecting the input set \{B_1, B_2\}. The results of our method also support this statement, because our proposed DODR is maximized by the selection of input set \{B_1, B_2\}, which means that this approach maintains the physical meanings of DODR as the maximum energy requirement for the disturbance.

<table>
<thead>
<tr>
<th></th>
<th>{B_1, B_2}</th>
<th>{B_1, B_3}</th>
<th>{B_2, B_3}</th>
</tr>
</thead>
<tbody>
<tr>
<td>DODR New Method</td>
<td>1.3240</td>
<td>1.2624</td>
<td>1.3210</td>
</tr>
</tbody>
</table>

Table (5.4): DODR results of Example (5.10)

To evaluate the performance and usefulness of the proposed measure, let us consider another numerical example describing a continuous-time linear model of bus suspension system. In this example, we will consider the cases when the disturbance is matched and unmatched. Next, we will identify the variation of the measure according to the variation of a model parameter.

**Example (5.11):** Consider a 1/4 bus model (one of the four wheels) to simplify the problem to a one-dimensional spring-damper system:

\[
\begin{bmatrix}
0 & \frac{1}{k} & \frac{b}{m} & \frac{k}{b} & \frac{1}{m}
\\
\frac{k}{m} & \frac{1}{m} & \frac{b}{m} & \frac{k}{b} & \frac{1}{m}
\\
0 & 0 & 0 & 1 & 0
\\
\frac{k}{m} & \frac{b}{m} & \frac{2k}{m} & \frac{b}{m} & \frac{1}{m}
\\
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4 \\
x_5 \\
\end{bmatrix}
+ \begin{bmatrix}
0 \\
1 \\
0 \\
0 \\
0 \\
\end{bmatrix} u + B_u d
\]

Assume body mass \( m_1 = 2500 \text{ kg} \), suspension mass \( m_2 = 320 \text{ kg} \), spring rate \( k = 10000 \text{ N/m} \) damping constant \( b = 140000 \text{ Ns/m} \) and the deformation of the tire as the disturbance. The system is controllable and asymptotically stable.

Assume body mass \( m_1 = 2500 \text{ kg} \), suspension mass \( m_2 = 320 \text{ kg} \), spring rate \( k = 10000 \text{ N/m} \) damping constant \( b = 140000 \text{ Ns/m} \) and the deformation of the tire as the disturbance. The system is controllable and asymptotically stable.
We first study the case when the matching condition is satisfied; that is, the disturbance is in the range of input matrix B (Figure (5.13)). For the matched disturbance, \( B_d = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}^T \) then we get \( \rho = 0.9474 + 0.0555 = 1.0029 \).

Next, we will calculate the measure for the case of the unmatched disturbance depicted in Figure (5.14).

![Figure (5.14): Model of bus suspension system (unmatched disturbance)](image)

In this case, the disturbance matrix is given by \( B_d = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^T \), then the proposed DODR is \( \rho = 0.9474 + 3.3857 = 4.3331 \).

Thus, it is found that the capability of disturbance rejection of the system would be degraded when a matched disturbance exists, i.e. \( B_d = B \), since the matched disturbance can directly affect the system, then the control input energy is equal to the disturbance energy, while in unmatched case the disturbance may require larger energy to affect the system.

Now assume that the spring constant \( k \) changes from zero to infinity while the matched disturbance \( B_d = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}^T \) is applied to the system. The results from changing the spring constant \( k \) are shown in Table (5.5). The results reveal that as the spring constant becomes larger our proposed DODR measure increases. That means that the degree of disturbance rejection improves. Those results are expected from the physical understanding of the system. The small spring constant means that the masses are almost not linked to each other. Thus, the disturbance may highly affect the system. As the spring constant \( k \) increases the two masses \( m_1 \) and \( m_2 \) are connected tightly to each other. Accordingly, the DODR measure increases.

<table>
<thead>
<tr>
<th>( K )</th>
<th>DODR New Method</th>
<th>( K=0 )</th>
<th>( K=0.01 )</th>
<th>( K=0.1 )</th>
<th>( K=1 )</th>
<th>( K=10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K=100 )</td>
<td>( K=10^3 )</td>
<td>( K=10^4 )</td>
<td>( K=10^5 )</td>
<td>( K=10^6 )</td>
<td>( K=\text{inf} )</td>
<td></td>
</tr>
<tr>
<td>DODR New Method</td>
<td>1.00002446</td>
<td>1.0019</td>
<td>1.0029</td>
<td>1.2699</td>
<td>5.4894</td>
<td>21.3538</td>
</tr>
</tbody>
</table>

Table (5.5): DODR results of Example (5.11)
5.7. Summary

In this chapter, we defined the minimum energy of unstable systems in the steady state form. Two different energy-based measures were introduced to check the degree of controllability for both stable and unstable systems. According to Kalman’s controllability rank condition a new quantitative measure of DOC is also presented, which verified the validity of the energy-based controllability metrics.

In the chapter, it was explained how these measures can be employed to determine optimal actuator locations for good degree of controllability. A measure for the degree of disturbance rejection was also proposed. The approach depends on computation of the controllability Gramian and the disturbance sensitivity Gramian.

In the next chapter, we consider the stable LTI systems and propose an approach to find the value of the minimum input energy using the coefficients of the characteristic polynomial of the system. This enables to determine the trace and some upper-bounds for the maximum eigenvalue of the controllability Gramian.
Chapter 6: Controllability Gramian and Energy Calculations of the Stable LTI System in Canonical Forms

6.1. Introduction
In this chapter, we consider the case where the LTI system is stable and it has certain canonical form structure. We derive a simple structure for the controllability Gramian based on the coefficients of the characteristic polynomial of the system.

The aim of this chapter is to present the interesting links between the value of minimum input energy and the place of the eigenvalues of a stable system. Moreover, an expression for the trace of the controllability Gramian is derived as a simple function of the coefficients of the characteristic polynomial using the fact that the controllability Gramian of a stable LTI system is the solution of a Lyapunov equation.

In section 6.2 we discuss the case where the system is defined in the controller companion form. An upper bound for the maximum eigenvalue of the controllability Gramian is proposed based on the Routh Hurwitz table. The inverse controllability Gramian and the minimum input energy using the coefficients of the characteristic polynomial of a stable system in the controller companion form is discussed in section 6.3. Finally, in section 6.4 an approach is proposed to find the controllability Gramian of a diagonal LTI system.
6.2. Controllability Gramian Computation in Controller Canonical Form

The controllability Gramian of stable linear continuous time-invariant systems:

\[
\begin{align*}
\dot{x} &= Ax + Bu, & A &\in \mathbb{R}^{n \times n}, & B &\in \mathbb{R}^{n \times 1}, \\
y &= Cx + Du, & C &\in \mathbb{R}^{1 \times n}, & D &\in \mathbb{R}^{1 \times 1}
\end{align*}
\]  

(6.1)
can be described by the solution of the following Lyapunov equation:

\[
AW_c + W_c A^T = -BB^T
\]  

(6.2)

A number of methods for solving the Lyapunov equation have been presented in the literature (Barnett & Storey, 1970), (Barnett & Storey, 1968), (Wu, et al., 2006), (Zhou & Duan, 2005), (Hauksdóttir, et al., 2008), (Hauksdóttir & Sigurðsson, 2009). In this section through the Kronecker product we present a simple method for finding the controllability Gramian directly based on the coefficients of the characteristic polynomial.

Lemma (6.1): (Barnett & Storey, 1970) The Lyapunov equation (6.2) can be written as:

\[
(A \oplus A) \text{vec}(W_c) = -\text{vec}(Q)
\]  

(6.3)

where \( Q = BB^T \) and \( \text{vec}(.) \) is the vector obtained by stacking columns of matrix over each other. \({A \oplus A}\) is the Kronecker sum of matrix \( A \). \( A \) is assumed to be in controller companion form.

(Note: The Kronecker sum of two matrices is defined as: \((A_{n \times n} \oplus B_{m \times m}) = (I_m \otimes A) + (B \otimes I_n)\) (Laub, 2005))

The proof can be found in (Barnett & Storey, 1968), (Laub, 2005).

Lemma (6.2): Consider the LTI system (6.1). There is a non-singular similarity matrix \( T \) such that:

\[
\begin{align*}
T^{-1}AT &= A_c \\
T^{-1}b &= b_c
\end{align*}
\]  

(6.4)

where:

\[
A_c = \begin{bmatrix}
0_{(n-1)\times 1} & I_{(n-1)\times(n-1)} \\
-\beta_0 & -\beta_1 & \cdots & -\beta_{n-2} & -\beta_{n-1}
\end{bmatrix}, \quad b_c = \begin{bmatrix}
0 \\
0 \\
\vdots \\
1
\end{bmatrix}
\]

Matrix \( T \) can be described as:

\[
T = \begin{bmatrix}
b & Ab & A^2b & \cdots & A^{n-1}b
\end{bmatrix}
\]  

(6.5)
where $\beta_i, i = 1, \ldots, n - 1$ are the coefficients of the characteristic polynomial of the system, i.e.

$$p(s) = s^n + \beta_{n-1}s^{n-1} + \cdots + \beta_0$$ \hspace{1cm} (6.6)

**Proof:** (Kailath, 1980)

Let $T = \begin{bmatrix} t_1 & t_2 & t_3 & \cdots & t_n \end{bmatrix}$ and define the similarity transformation:

$$A \rightarrow T^{-1}AT = A_c$$

Consider the characteristic polynomial of the system:

$$|sI - A| = s^n + \beta_{n-1}s^{n-1} + \cdots + \beta_0$$ \hspace{1cm} (6.7)

According to (6.5) we obtain:

$$AT = TA_c \Rightarrow \begin{bmatrix} At_1 & At_2 & At_3 & \cdots & At_n \end{bmatrix} = \begin{bmatrix} -\beta_0 t_n & t_1 - \beta_1 t_n & t_2 - \beta_2 t_n & \cdots & t_{n-1} - \beta_{n-1} t_n \end{bmatrix}$$ \hspace{1cm} (6.8)

and:

$$b = Tb_c \Rightarrow b = T \begin{bmatrix} 0 \\ 0 \\ \vdots \\ t_n \\ 0 \\ 1 \end{bmatrix}$$ \hspace{1cm} (6.9)

Then (6.5) shows that $T$ is the product of the controllability matrix of the system and the Hankel matrix of the characteristic polynomial’s coefficients. In vector form:

$$\begin{bmatrix} b = t_n \\ Ab + \beta_{n-1}b = t_{n-1} \\ At_{n-1} + \beta_{n-2}b = t_{n-2} \Rightarrow T = \begin{bmatrix} b & Ab & A^2b & \cdots & A^{n-1}b \end{bmatrix} \begin{bmatrix} \beta_1 \beta_2 \cdots \beta_{n-1} \\ \beta_2 \beta_3 \cdots 1 \\ \vdots \vdots \vdots \vdots \\ 1 0 \cdots 0 0 \end{bmatrix} \end{bmatrix}$$ \hspace{1cm} (6.10)

which concludes the proof.

Using the similarity transformation $T$, Lemma (6.1) yields:

**Corollary (6.1):** Consider the stable LTI system (6.1). Then the Lyapunov equation (6.2) can be written as:

$$(A_c \oplus A_c)(T \otimes T)^{-1} \text{vec}(W_c) = -\text{vec}(Q_c)$$ \hspace{1cm} (6.11)

where $A_c$ and $Q_c$ are defined in (6.4).
Considering the equation above, the controllability Gramian can be written as:

\[
W_c = -\text{vec}^{-1}\left( T \otimes T \left( A_c \oplus A_c \right)^{-1} \text{vec}(Q_c) \right)
\]

(6.12)

or equivalently:

\[
W_c = TW_c^* T^T
\]

Here \(W_c^*\) defines the controllability Gramian in the controller companion form, and \(\text{vec}^{-1}\) is the inverse \(\text{vec}\) operator reshaping the vector into a n-by-n matrix.

**Note:** In the Corollary above system is stable, i.e. \(A_c\) has no eigenvalues in the open right-half plain (ORHP) and then \(\lambda_i + \lambda_j \neq 0, \forall i, j = 1, \ldots, n\) where \(\lambda_i\) defines the \(i\)th eigenvalue of \(A_c\). Hence the inverse of the Kronecker sum \((A_c \oplus A_c)\) is well-defined.

**Proposition (6.1):** Let \(A_c \in \mathbb{R}^{n \times n}\) be in controller companion form and assume that \(\lambda_i + \lambda_j \neq 0, \forall i, j = 1, \ldots, n\). Then:

\[
(A_c \oplus A_c)^{-1}_{n \times n} = \begin{bmatrix}
\tilde{O}_{11} & \tilde{O}_{12} & \cdots & \tilde{O}_{1n} \\
\tilde{O}_{21} & \tilde{O}_{22} & \cdots & \tilde{O}_{2n} \\
\vdots & \ddots & \ddots & \vdots \\
\tilde{O}_{n1} & \tilde{O}_{n2} & \cdots & \tilde{O}_{nn}
\end{bmatrix}
\]

where the blocks \(\tilde{O}_{ij}\) are defined recursively as:

\[
\tilde{O}_{11} = \left( \sum_{i=0}^{n} (-1)^{i+1} \beta_i A_c^{i-1} \right) \sum_{i=1}^{n} (-1)^{i-1} \beta_i A_c^{i-2}, \quad \tilde{O}_{21} = I - A_c \tilde{O}_{11} \rightarrow \text{first – underdiag block}
\]

\[
\tilde{O}_{12} = \left( \sum_{i=0}^{n} (-1)^{i+1} \beta_i A_c^{i-1} \right) \sum_{i=2}^{n} (-1)^{i-2} \beta_i A_c^{i-3}, \quad \tilde{O}_{22} = I - A_c \tilde{O}_{12} \rightarrow \text{first – underdiag block}
\]

\[
\tilde{O}_{13} = \left( \sum_{i=0}^{n} (-1)^{i+1} \beta_i A_c^{i-1} \right) \sum_{i=3}^{n} (-1)^{i-3} \beta_i A_c^{i-5}, \quad \tilde{O}_{23} = I - A_c \tilde{O}_{13} \rightarrow \text{diagblock}
\]

\[
\tilde{O}_{14} = \left( \sum_{i=0}^{n} (-1)^{i+1} \beta_i A_c^{i-1} \right) \sum_{i=4}^{n} (-1)^{i-4} \beta_i A_c^{i-7}, \quad \tilde{O}_{24} = I - A_c \tilde{O}_{14} \rightarrow \text{upperdiag – and – diagblocks}
\]

\[
\vdots
\]

\[
\tilde{O}_{1n} = \left( \sum_{i=0}^{n} (-1)^{i+1} \beta_i A_c^{i-1} \right) \tilde{O}_{nn} = (-A_c)^{-i} \tilde{O}_{nn}, i = 2, \ldots, n
\]

(6.13)
Proof:

From the Kronecker sum definition (Laub, 2005) we have:

\[
(A_x \oplus A_y) = ((A_x \otimes I) + (I \otimes A_y)) = \begin{bmatrix}
A_x & I & 0_{n \times n} & \cdots & \cdots & 0_{n \times n} \\
0_{n \times n} & A_x & I & 0_{n \times n} & \cdots & 0_{n \times n} \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0_{n \times n} & 0_{n \times n} & \cdots & A_x & I \\
-\beta_0 I & -\beta_1 I & \cdots & -\beta_{n-2} I & -\beta_{n-1} I + A_y \\
\end{bmatrix}
\] (6.14)

Then, we define the inverse of this matrix as:

\[
(A_x \oplus A_y)^{-1} = \begin{bmatrix}
\bar{\delta}_{11} & \bar{\delta}_{12} & \cdots & \bar{\delta}_{1n} \\
\bar{\delta}_{21} & \bar{\delta}_{22} & \cdots & \bar{\delta}_{2n} \\
\vdots & \ddots & \ddots & \vdots \\
\bar{\delta}_{n1} & \bar{\delta}_{n2} & \cdots & \bar{\delta}_{nn} \\
\end{bmatrix}
\] (6.15)

Thus:

\[
\begin{bmatrix}
A_x & I & 0_{n \times n} & \cdots & \cdots & 0_{n \times n} \\
0_{n \times n} & A_x & I & 0_{n \times n} & \cdots & 0_{n \times n} \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0_{n \times n} & 0_{n \times n} & \cdots & A_x & I \\
-\beta_0 I & -\beta_1 I & \cdots & -\beta_{n-2} I & -\beta_{n-1} I + A_y \\
\end{bmatrix} M = \begin{bmatrix}
\bar{\delta}_{11} & \bar{\delta}_{12} & \cdots & \bar{\delta}_{1n} \\
\bar{\delta}_{21} & \bar{\delta}_{22} & \cdots & \bar{\delta}_{2n} \\
\vdots & \ddots & \ddots & \vdots \\
\bar{\delta}_{n1} & \bar{\delta}_{n2} & \cdots & \bar{\delta}_{nn} \\
\end{bmatrix} \quad \text{6.16}
\]

Applying Gaussian elimination with partial pivoting to equation (6.16), we can readily obtain 
\((A_x \oplus A_y)^{-1}\) as it is described in equation (6.13).

By multiplying the first row of \(M\) by the columns of \(N\) we obtain:

\[
\begin{align*}
A_x \bar{\delta}_{11} + \bar{\delta}_{21} & = I_n \Rightarrow \bar{\delta}_{21} = I - A_x \bar{\delta}_{11} \\
A_x \bar{\delta}_{12} + \bar{\delta}_{22} & = 0_n \Rightarrow \bar{\delta}_{22} = -A_x \bar{\delta}_{12} \\
& \vdots \\
A_x \bar{\delta}_{1n} + \bar{\delta}_{2n} & = 0_n \Rightarrow \bar{\delta}_{2n} = -A_x \bar{\delta}_{1n}
\end{align*}
\] (6.17)

Take the second row of \(M\), and multiply by the columns of \(N\):

\[
\begin{align*}
A_x \bar{\delta}_{31} & = 0_n \Rightarrow \bar{\delta}_{31} = -A_x \bar{\delta}_{31} \\
A_x \bar{\delta}_{32} & = I_n \Rightarrow \bar{\delta}_{32} = I - A_x \bar{\delta}_{32} \\
& \vdots \\
A_x \bar{\delta}_{3n} & = 0 \Rightarrow \bar{\delta}_{3n} = -A_x \bar{\delta}_{3n}
\end{align*}
\]
By multiplying, in the third row we get:

\[
\begin{align*}
A_c \tilde{O}_{31} + \tilde{O}_{41} &= 0 \Rightarrow \tilde{O}_{41} = -A_c \tilde{O}_{31} = A_c^2 \tilde{O}_{21} \\
A_c \tilde{O}_{32} + \tilde{O}_{42} &= 0 \Rightarrow \tilde{O}_{42} = -A_c \tilde{O}_{32} \\
A_c \tilde{O}_{33} + \tilde{O}_{43} &= I_n \Rightarrow \tilde{O}_{43} = I - A_c \tilde{O}_{33} \\
\vdots &
\end{align*}
\]

\[
\begin{align*}
A_c \tilde{O}_{3n} + \tilde{O}_{4n} &= 0 \Rightarrow \tilde{O}_{4n} = -A_c \tilde{O}_{3n} = A_c^2 \tilde{O}_{2n}
\end{align*}
\]

Similarly, multiplication of the rest of the rows gives:

\[
\begin{align*}
\vdots &
\end{align*}
\]

\[
\begin{align*}
A_c \tilde{O}_{(n-1)1} + \tilde{O}_{n1} &= 0 \Rightarrow \tilde{O}_{n1} = -A_c \tilde{O}_{(n-1)1} \\
A_c \tilde{O}_{(n-1)2} + \tilde{O}_{n2} &= 0 \Rightarrow \tilde{O}_{n2} = -A_c \tilde{O}_{(n-1)2} \\
\vdots &
\end{align*}
\]

\[
\begin{align*}
A_c \tilde{O}_{(n-1)n} + \tilde{O}_{nn} &= I_n \Rightarrow \tilde{O}_{nn} = I - A_c \tilde{O}_{(n-1)n}
\end{align*}
\]

\[
\begin{align*}
-\beta_0 \tilde{O}_{11} - \beta_1 \tilde{O}_{21} - \cdots - \beta_{n-1} \tilde{O}_{n1} + A_c \tilde{O}_{11} &= 0 \Rightarrow \sum_{i=0}^{n} (-1)^{i+1} \beta_i A_c^{i+1} = \sum_{i=1}^{n} (-1)^{i-1} \beta_i A_c^{i-1} \\
-\beta_0 \tilde{O}_{12} - \beta_1 \tilde{O}_{22} - \cdots - \beta_{n-1} \tilde{O}_{n2} + A_c \tilde{O}_{12} &= 0 \Rightarrow \sum_{i=0}^{n} (-1)^{i+1} \beta_i A_c^{i+2} = \sum_{i=2}^{n} (-1)^{i-2} \beta_i A_c^{i-2} \\
\vdots &
\end{align*}
\]

\[
\begin{align*}
-\beta_0 \tilde{O}_{1n} - \beta_1 \tilde{O}_{2n} - \cdots - \beta_{n-1} \tilde{O}_{nn} + A_c \tilde{O}_{1n} &= I \Rightarrow \sum_{i=0}^{n} (-1)^{i+1} \beta_i A_c^{i+1} \tilde{O}_{1n} = I
\end{align*}
\]

and the proof is complete.

**Remark (6.1):** The Proposition above, allows to calculate the inverse of \(n^2\)-dimensional matrix \((A_c \oplus A_s)^{-1} \in \mathbb{R}^{n^2 \times n^2}\) by inverting the \(n\)-dimensional matrix \((\sum_{i=0}^{n} (-1)^{i+1} \beta_i A_c^{i+1})\) where

\[
|sI - A_c| = \beta_0 s^n + \beta_{n-1} s^{n-1} + \cdots + \beta_0, \beta_n = 1.
\]

**Corollary (6.2):** Consider the stable LTI system (6.4) in controller companion form. The controllability Gramian \(W_c(0, \infty)\) can be readily derived as the matrix produced by the inverse vectorization of the last column of \((A_c \oplus A_s)^{-1}\), i.e.

\[
W_c(0, \infty) = -vec^{-1} \left( \begin{bmatrix} \tilde{O}_{11}(\cdot; n) & \tilde{O}_{21}(\cdot; n) & \cdots & \tilde{O}_{nn}(\cdot; n) \end{bmatrix}^T \right) \tag{6.18}
\]

where:
\[
(A_c \oplus A_c)^{-1} = \begin{bmatrix}
\bar{O}_{11} & \bar{O}_{12} & \cdots & \bar{O}_{1n} \\
\bar{O}_{21} & \bar{O}_{22} & \cdots & \bar{O}_{2n} \\
\vdots & \ddots & \ddots & \vdots \\
\bar{O}_{n1} & \bar{O}_{n2} & \cdots & \bar{O}_{nn}
\end{bmatrix}_{n^2 \times n^2}
\]  
\tag{6.19}

and \( \bar{O}_{in}(i,n), \ i = 1, \ldots, n \) describes the last column of the block matrix \( \bar{O}_{in} \).

\[\blacksquare\]

**Remark (6.2):** Corollary (6.2) shows that the full structure of \( (A_c \oplus A_c)^{-1} \) is not needed for the calculation of the controllability Gramian \( W_c^\infty(0, \infty) \).

\[\blacksquare\]

**Remark (6.3):** Consider the controllability Gramian of the stable LTI system (6.1) in controller companion form. Then the trace of the controllability Gramian is equal to:

\[
\text{trace}(W_c^\infty(0, \infty)) = -\sum_{i=1}^{n} o_{in}(\bar{O}_{in})
\]  
\tag{6.20}

where \((A_c \oplus A_c)^{-1} > 0\) is defined in (6.19) and \( o_{in}(\bar{O}_{in}) \) denotes the element in \( i \)th row and last column of the block matrix \( \bar{O}_{in} \), for \( i = 1, \ldots, n \).

\[\blacksquare\]

**Remark (6.4):** An upper bound on the maximum eigenvalue of the controllability Gramian of the stable LTI system (6.4) in controller form is equal to:

\[
\lambda_{\max}(W_c^\infty(0, \infty)) \leq \sqrt{\text{trace}(W_c^\infty(0, \infty))^2} = \sqrt{\sigma_n^T \bar{O}_{in}^T (I + A_c^T A_c + (A_c^T)^2 A_c^2 + \cdots + (A_c^{(n-1)})^T A_c^{n-1}) \bar{O}_{in} \sigma_n}
\]  
\[\overset{\downarrow}{= \sqrt{\sigma_n^T \sum_{i=1}^{n} \bar{O}_{in}^T \bar{O}_{in} \sigma_n}}\]  
\tag{6.21}

where \( \sigma_n = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}^T \) \( \bar{O}_{in}(\cdot,n) \) represents the last column of the block matrix \( \bar{O}_{in} \) and \( \bar{O}_{in} \) denotes the block matrix in the \( i \)th row and \( n \)th column of the matrix \((A_c \oplus A_c)^{-1}, i = 1, \ldots, n \) in (6.19).

\[\blacksquare\]

**Example (6.1):** Consider the stable LTI system below:
Using the similarity matrix:

\[
T = \begin{bmatrix}
4.2601 & 2.4902 & 1.0306 \\
-0.9425 & -0.0069 & 0.3275 \\
2.3626 & 2.4050 & 0.6521
\end{bmatrix}
\]

The system can be put in controller companion form:

\[
A_c = \begin{bmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
-6.8913 & -8.2417 & -4.1328
\end{bmatrix}, B_c = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}
\]

Then according to Corollary (6.1) we obtain the controllability Gramian \(W_c(0, \infty)\) as:

\[
W_c(0, \infty) = \text{vec}^{-1}(-T \otimes T)(A_c \oplus A_c)^{-1}\text{vec}(Q_c)
\]

Then, using Proposition (6.1) we have:

\[
(A_c \oplus A_c)^{-1} = \begin{bmatrix}
\tilde{\sigma}_{11} & \tilde{\sigma}_{12} & \tilde{\sigma}_{13} \\
\tilde{\sigma}_{21} & \tilde{\sigma}_{22} & \tilde{\sigma}_{23} \\
\tilde{\sigma}_{31} & \tilde{\sigma}_{32} & \tilde{\sigma}_{33}
\end{bmatrix}, \text{vec}(Q_c)_{9x1} = [0 \ 0 \ \cdots \ 1]^T
\]

where:

\[
\tilde{\sigma}_{11} = \left(A_c^2 - \beta_2 A_c + \beta_1 I\right) \left(A_c^3 - \beta_2 A_c^2 + \beta_1 A_c - \beta_0 I\right)^{-1} \tilde{\sigma}_{21} = I - A_c \tilde{\sigma}_{11}
\]

\[
\tilde{\sigma}_{31} = -A_c \tilde{\sigma}_{21}
\]

in which:

\[
\tilde{\sigma}_{11} = \begin{bmatrix}
-0.9130 & -0.3765 & -0.0727 \\
0.5000 & -0.3146 & -0.0761 \\
0.5236 & 1.1267 & 0
\end{bmatrix}
\]

\[
\tilde{\sigma}_{21} = \begin{bmatrix}
0.5000 & 0.3146 & 0.0761 \\
-0.5236 & -0.1267 & 0.0000 \\
0.0000 & -0.5236 & -0.1267
\end{bmatrix}
\]

\[
\tilde{\sigma}_{31} = \begin{bmatrix}
0.5236 & 0.1267 & 0.0000 \\
-0.0000 & 0.5236 & 0.1267 \\
-0.8716 & -1.0432 & -0.0000
\end{bmatrix}
\]

Also:
\[
\bar{O}_{12} = (-A_x + \beta_2 I)(A_x^3 - \beta_2 A_x^2 + \beta_1 A_x - \beta_0 I)^{-1} \left\{ \bar{O}_{22} = -A_x \bar{O}_{12}, \bar{O}_{32} = I - A_x \bar{O}_{22} \right\}
\]

\[
\begin{bmatrix}
-0.3765 & -0.2074 & -0.0457 \\
0.3146 & 0.0000 & -0.0184 \\
0.1267 & 0.4662 & 0.0761 \\
\end{bmatrix}
\]

and finally:

\[
\bar{O}_{13} = (A_x^3 - \beta_2 A_x^2 + \beta_1 A_x - \beta_0 I)^{-1} \left\{ \bar{O}_{23} = -A_x \bar{O}_{13}, \bar{O}_{33} = A_x^2 \bar{O}_{13} \right\}
\]

\[
\begin{bmatrix}
-0.0727 & -0.457 & -0.0111 \\
0.0761 & 0.0184 & -0.0000 \\
0.0000 & 0.0761 & 0.0184 \\
\end{bmatrix}
\]

Hence the controllability Gramian \( W_c(0, \infty) \) is readily obtained by Corollary (6.1) as:

\[
W_c(0, \infty) = \text{vec}^{-1}(-T \otimes T)(A_x \oplus A_x)^{-1} \text{vec}(\bar{O})
\]

In this example:

\[
W_c(0, \infty) = \begin{bmatrix} 0.3138 & -0.0000 & 0.2279 \\
-0.0000 & 0.0375 & 0.0046 \\
0.2279 & 0.0046 & 0.1758 \end{bmatrix}
\]

Example (6.2): Consider the stable LTI system below:
According to Remark (6.3) the trace of the controllability Gramian is obtained as:

\[
\text{trace}(W_c((0, \infty))) = -(o_{13}(O_{13}) + o_{23}(O_{23}) + o_{33}(O_{33})) = -0.0014 - 0.2596 = 0.2610
\]

where:

\[
\begin{align*}
O_{13} &= (A_c^3 - \beta_c A_c^2 + \beta_c A_c - \beta_c I)^{-1}, \quad \begin{cases} 
O_{23} = -A_c \overline{O}_{13} \\
O_{33} = A_c^2 \overline{O}_{13}
\end{cases} \\
O_{23} &= (A_c^3 - 5.8585 A_c^2 + 185.0919 A_c - 727.8173 I)^{-1}, \quad \begin{cases} 
O_{23} = -A_c \overline{O}_{13} \\
O_{33} = A_c^2 \overline{O}_{13}
\end{cases} \\
\overline{O}_{13} &= \begin{bmatrix}
-0.0007 & -0.0001 & -0.0000 \\
0.0082 & 0.0014 & -0.0000 \\
0.0000 & 0.0082 & 0.0014
\end{bmatrix} \\
\overline{O}_{23} &= \begin{bmatrix}
-0.0007 & -0.0001 & 0.0000 \\
-0.0000 & 0.0082 & 0.0014 \\
1.0207 & 0.2596 & -0.0000 \\
\end{bmatrix} \\
\overline{O}_{33} &= \begin{bmatrix}
-1.0207 & -0.2596 & 0.0000 \\
0.0000 & 0.0082 & 0.0014 \\
-0.0000 & -1.0207 & -0.2596
\end{bmatrix}
\end{align*}
\]

Example (6.3): Consider the stable LTI system below:

\[
A_c = \begin{bmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
-6.8913 & -8.2417 & -4.1328
\end{bmatrix}, \quad B_c = \begin{bmatrix} 0 \\
0 \\
1 \end{bmatrix}
\]

For this stable controller form system, the corresponding matrices below are calculated in example (6.1):
\[
\bar{O}_{13} = \left( A_{c}^3 - 4.1328 A_{c}^2 + 8.2417 A_{c} - 6.8913 I \right)^{-1} \begin{cases}
\bar{O}_{23} = -A_{c} \bar{O}_{13} \\
\bar{O}_{31} = A_{c}^2 \bar{O}_{13}
\end{cases}
\]

Then according to Remark (6.4) an upper-bound of the maximum eigenvalue of the controllability Gramian is:

\[
\sqrt{\text{trace}(W_{c}^\infty(0, \infty))^2} = \sqrt{\sum_{i=1}^{n} (\bar{O}_{i3}(\cdot,3))^T \bar{O}_{i3}(\cdot,3)}
\]

\[
= \sqrt{\begin{bmatrix}
-0.0111 & 0 & 0.0184 \\
0 & -0.0111 & 0 \\
0.0184 & 0 & -0.0111
\end{bmatrix} + \begin{bmatrix}
0 & 0 & -0.0184 \\
-0.0184 & 0 & 0 \\
0 & -0.0184 & 0.0184
\end{bmatrix} + \begin{bmatrix}
0.0184 & 0 & -0.1516 \\
0 & 0.0184 & 0 \\
-0.1516 & 0 & 0.0184
\end{bmatrix} + \begin{bmatrix}
0 & 0 & 0.0184 \\
0 & 0 & 0.1516 \\
-0.1516 & 0 & 0
\end{bmatrix}}
\]

Hence:

\[
\sqrt{\text{trace}(W_{c}^\infty(0, \infty))^2} = \sqrt{0.0241} = 0.1554
\]

Further, according to Corollary (6.2) we know that the controllability Gramian of the system in controller companion form can be obtained by the last column of \((A_{c} \oplus A_{c})^{-1}\) as:

\[
W_{c}^\infty(0, \infty) = \begin{bmatrix}
0.0110 & -0.0000 & -0.0184 \\
-0.0000 & 0.0184 & 0.0000 \\
-0.0184 & 0.0000 & 0.1516
\end{bmatrix}
\]

The eigenvalues of the controllability Gramian are \(\lambda_1 = 0.1540, \lambda_2 = 0.0184, \lambda_3 = 0.0087\) and they verify the upper bound for the maximum eigenvalue of \(W_{c}^\infty(0, \infty)\), i.e.

\[
\lambda_1 = 0.1540 \leq \sqrt{\text{trace}(W_{c}^\infty(0, \infty))^2} = 0.1554
\]

In the sequel, a simple method for calculating the controllability Gramian of the stable LTI system in companion form is proposed. This method involves the solution of a linear system of equations based on the system's characteristic polynomial.
**Definition (6.1):** (Xiao, et al., 1992) A square $n \times n$ matrix $P = [p_{ij}]$ is said to be a Xiao matrix, if it has only $n$ independent elements, and has the property:

$$p_{ij} = \begin{cases} 0, & i + j \neq 2k, k = 1, \ldots, n \\ (-1)^{\frac{i-j}{2}}, & i + j = 2k, k = 1, \ldots, n \end{cases}$$

(6.22)

**Proposition (6.2):** (Xiao, et al., 1992) Let the stable LTI system (6.1) be in controller companion form. The controllability Gramian $W_c^\infty(0, \infty)$ is a Xiao matrix and can be obtained as the solution of the quasi Hankel form linear system of equations below:

If the dimension of the system is even, i.e. $A \in \mathbb{R}^{n \times n}, n = 2k, k = 1, 2, 3, \ldots$

$$\begin{bmatrix} -\beta_0 & \beta_2 & -\beta_4 & \beta_6 & \cdots & \cdots & 0 \\ 0 & -\beta_1 & \beta_3 & -\beta_5 & \cdots & \cdots & 0 \\ 0 & \beta_0 & -\beta_2 & \beta_4 & \cdots & \cdots & 0 \\ 0 & 0 & \beta_1 & -\beta_3 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & -2\beta_1 & \cdots & -2\beta_{n-3} & -2\beta_{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ -1 \end{bmatrix}$$

(6.23)

If the dimension of the system is odd, i.e. $A \in \mathbb{R}^{n \times n}, n = 2k - 1, k = 1, 2, 3, \ldots$

$$\begin{bmatrix} -\beta_0 & \beta_2 & -\beta_4 & \beta_6 & \cdots & \cdots & 0 \\ 0 & -\beta_1 & \beta_3 & -\beta_5 & \cdots & \cdots & 0 \\ 0 & \beta_0 & -\beta_2 & \beta_4 & \cdots & \cdots & 0 \\ 0 & 0 & \beta_1 & -\beta_3 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & -2\beta_0 & \cdots & -2\beta_{n-3} & -2\beta_{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ -1 \end{bmatrix}$$

(6.24)

Here $\beta_i, i = 0, \ldots, n$ are the coefficients of the characteristic polynomial, and the vector $x$ includes the elements of the Xiao matrix $W_c^\infty(0, \infty)$:

In an even dimensional system, i.e. $A \in \mathbb{R}^{n \times n}, n = 2k, k = 1, 2, 3, \ldots$
\[ W_c^r(0, \infty) = \begin{bmatrix} x_1 & 0 & -x_2 & 0 & x_3 & \cdots & 0 \\ 0 & x_2 & 0 & -x_3 & \ddots & \ddots & \vdots \\ -x_2 & 0 & x_3 & \ddots & \ddots & \ddots & x_{n-2} \\ 0 & -x_3 & \ddots & \ddots & \ddots & \ddots & 0 \\ x_3 & \ddots & \ddots & x_{n-2} & 0 & -x_{n-1} & \vdots \\ \vdots & \ddots & \ddots & 0 & x_{n-2} & 0 & x_{n-1} \\ 0 & \cdots & x_{n-2} & 0 & -x_{n-1} & 0 & x_n \end{bmatrix} \] (6.25)

In an odd dimensional system, i.e. \( A \in \mathbb{R}^{n \times n} , n = 2k - 1 , k = 1, 2, 3, \ldots \)

\[ W_c^r(0, \infty) = \begin{bmatrix} x_1 & 0 & -x_2 & 0 & x_3 & \cdots & \frac{x_{(n+1)}}{2} \\ 0 & x_2 & 0 & -x_3 & \ddots & \ddots & \vdots \\ -x_2 & 0 & x_3 & \ddots & \ddots & \ddots & x_{n-2} \\ 0 & -x_3 & \ddots & \ddots & \ddots & \ddots & 0 \\ x_3 & \ddots & \ddots & x_{n-2} & 0 & -x_{n-1} & \vdots \\ \vdots & \ddots & \ddots & 0 & x_{n-2} & 0 & x_{n-1} \\ \frac{x_{(n+1)}}{2} & \cdots & x_{n-2} & 0 & -x_{n-1} & 0 & x_n \end{bmatrix} \] (6.26)

The proof is given in (Xiao, et al., 1992).

In the above Proposition matrices (6.23), (6.24) are called quasi Hankel since by permuting rows and columns so that all the odd numbered rows and columns precede the even numbered ones, the matrix transforms into a 2-block diagonal matrix, each block being almost Hankel matrix.

Regarding Proposition (6.2) a simple method is proposed for the calculating the controllability Gramian in asymptotically stable systems, which involves evaluating only the elements of a Routh table and does not involve the solution of a linear system of equations or matrix inversion (Sreeram & Agathoklis, 1991).

**Proposition (6.3):** In n-dimensional stable LTI system in controller companion form, the zero-plaid structured controllability Gramian \( W_c^r(0, \infty) \) described in (6.25), (6.26) can be obtained as:

\[
\begin{cases}
  x_n = \frac{0.5}{R_{n1}} \\
  x_{n-k} = \sum_{i=1}^{m-1} (-1)^i \frac{R_{n-k,i+1} x_{n-k+i}}{R_{n-k,i}}, k = 1, \ldots, n-1
\end{cases}
\] (6.27)

where \( R_{ij} \) denotes the entry in the Routh table of the reciprocal system corresponding to the \( i \)th row and \( j \)th column and \( m \) denotes the number of elements in the \((n-k)\)th row of the Routh table.
The proof is given in (Sreeram & Agathoklis, 1991).

**Note:** Consider the characteristic polynomial of an LTI system \( p(s) = s^n + \beta_{n-1}s^{n-1} + \cdots + \beta_1s + \beta_0 \).
Then the Routh table for the reciprocal characteristic polynomial is defined as: \( \tilde{p}(s) = s^n p(s^{-1}) \).

**Example (6.4):** Consider the stable LTI system in the previous example. The reciprocal polynomial of the system is:
\[
\tilde{p}(s) = \beta_0 s^3 + \beta_1 s^2 + \beta_2 s + 1 = 6.8913 s^3 + 8.2417 s^2 + 4.1328 s + 1
\]
Then the Routh table of the reciprocal system is:
\[
\begin{array}{ccc}
  s^3 & 6.8913 & 4.1328 \\
  s^2 & 8.2417 & 1 \\
  s^1 & 3.2966 & 0 \\
  s^0 & 1 \\
\end{array}
\]
Then according to Proposition (6.3) we have:
\[
W_c(0, \infty) = \begin{bmatrix}
  x_1 & 0 & -x_2 \\
  0 & x_2 & 0 \\
- x_2 & 0 & x_3 \\
\end{bmatrix}
\]
where:
\[
x_3 = \frac{0.5}{3.2966} = 0.1517 \\
x_2 = \frac{1 \times 0.1517}{8.2417} = 0.0184 \\
x_1 = \frac{4.1328 \times 0.0184}{6.8913} = 0.0110
\]

**Remark (6.5):** The trace of the controllability Gramian in the controller companion form of the stable LTI system (6.1), is equal to:
\[
\text{trace}(W_c(0, \infty)) = \sum_{k=1}^{n} x_k \tag{6.28}
\]
where:
\[
\begin{aligned}
\begin{cases}
  x_n = \frac{0.5}{R_{n1}} \\
  x_i = \frac{-\sum_{j=1}^{m-1} (-1)^j R_{i,j+1} x_{i+j}}{R_{i,1}}, \quad i = 1, \ldots, n-1
\end{cases}
\end{aligned}
\] (6.29)

Here \(R_{ij}\) denotes the entry in the Routh table corresponding to the \(i^{th}\) row and \(j^{th}\) column and \(m\) denotes the number of elements in the \(j^{th}\) row of the Routh table.

\[\square\]

**Remark (6.6):** An upper bound on the maximum eigenvalue of the controllability Gramian of the stable LTI system (6.1) in controller form is equal to:

\[
\lambda_{\max} (W_c^c(0, \infty)) \leq \sqrt{\sum_{i=1}^{n} q_i x_i^2}
\] (6.30)

where:

\[
\begin{aligned}
\begin{cases}
  x_n = \frac{0.5}{R_{n1}} \\
  x_i = \frac{-\sum_{j=1}^{m-1} (-1)^j R_{i,j+1} x_{i+j}}{R_{i,1}}, \quad i = 1, \ldots, n-1
\end{cases}
\end{aligned}
\] (6.31)

and:

\[
\begin{aligned}
\begin{cases}
  q_i = 2i - 1, \quad i \leq \text{int}\{ \frac{n}{2} \} \\
  q_i = 2n - (2i - 1), \quad i > \text{int}\{ \frac{n}{2} \}
\end{cases}
\end{aligned}
\] (6.32)

Here \(R_{ij}\) denotes the entry in the Routh table corresponding to the \(i^{th}\) row and \(j^{th}\) column and \(m\) denotes the number of elements in the \(j^{th}\) row of the Routh table. Further, \(\text{int}\{ \alpha \}\) is defined \(\frac{\alpha}{2}\) if \(\alpha = 2k, k = 1, 2, \ldots\) and \(\frac{\alpha + 1}{2}\) if \(\alpha = 2k - 1, k = 1, 2, \ldots\).

\[\square\]

The above Remark states that the eigenvalues of the controllability Gramian of the stable LTI system in controller form is bounded from above by a function of \(x_i\) s defined in Proposition (6.3).

**Example (6.5):** Consider the LTI system in example (6.2). The Routh table of the reciprocal system is:
According to Remark (6.5) we have:

\[
\begin{aligned}
\text{trace}(W_c^r(0, \infty)) &= \frac{0.5}{1.9263} + \frac{0.2596}{185.0919} + \frac{5.8585 \times 0.0014}{727.8173} = 0.2596 + 0.0014 + 1.1290 \times 10^{-5} = 0.26101129 \\
\end{aligned}
\]

Remark (6.6) determines also an upper bound of the maximum eigenvalues of the controllability Gramian as:

\[
\lambda_{\text{max}}(W_c^r(0, \infty)) = 0.2596 \leq \sqrt{x_1^2 + 3x_2^2 + x_3^2} = \sqrt{0.0674} = 0.2596
\]

\[\square\]

### 6.3. Inverse of Controllability Gramian and Energy Calculation in Controller Canonical Form

In this section, we introduce a simple algorithm for finding the inverse controllability Gramian of a SISO stable system based on the Routh table. This helps in feedback design when a constraint on the input control energy is imposed. By relating the minimum input energy to the location of the eigenvalues of the system.

Note that in the previous section we introduced some methods for finding the controllability Gramian \(W_c^r(0, \infty)\) in controller companion form, which has close resemblance to our method for finding the inverse of controllability Gramian of a SISO stable LTI system in phase variable form (Controller companion form) \(L = (W_c^r(0, \infty))^{-1}\). Our method is important since it does not involve solution of a linear system of equations or matrix inversion and hence is a computationally efficient way for calculating the control input energy in various problems of feedback control design or pole placement with input energy constraints, compared to other techniques. This result is illustrated by a number of numerical examples.

**Theorem (6.1):** Consider the SISO stable LTI system (6.1) in controller companion form. The inverse of the controllability Gramian \(L = (W_c^r(0, \infty))^{-1}\) can be derived as the zero-plaid structured matrix in terms of the characteristic polynomial’s coefficients:

\[
|sI - A| = s^n + \beta_{n-1}s^{n-1} + \cdots + \beta_0
\]

(6.33)

If the dimension of the system is odd, i.e. \(A \in \mathbb{R}^{n \times n}, n = 2k + 1, k = 0, 1, 2, \ldots\):
\[
\begin{pmatrix}
    s^n & 1 & \beta_{n-2} & \ldots & \beta_1 \\
    s^{n-1} & \beta_{n-1} & \beta_{n-3} & \ldots & \beta_0 \\
    \vdots & \vdots & \ddots & \ddots & \vdots \\
    s^1 & \beta_0 & \ldots & \beta_2 & 0 \\
\end{pmatrix} \quad (6.34)
\]

and:

\[
L = L^T = \begin{pmatrix}
    L_1 & 0 & L_2 & 0 & \ldots & 0 & L_{k+1} \\
    L_{k+2} & 0 & L_{k+3} & \ldots & L_{(k+1)+k+1} & 0 \\
    L_{2k+2} & 0 & \ldots & 0 & L_{((2k+1)-(k+2)+1)+(2k+1)} & 0 \\
    L_{3k+2} & 0 & \ldots & L_{(4k-(2k+2)+1)+4k} & \ddots & \ddots \\
    \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\
    L_{2k(k+1)-n} & \ddots & \ddots & \ddots & \ddots & L_{2\beta_{n-1}} \\
\end{pmatrix} \quad (6.35)
\]

or:

\[
L = L^T = \begin{pmatrix}
    2\beta_0 \beta_1 & 0 & 2\beta_0 \beta_3 & 0 & \ldots & 0 & 2\beta_0 \\
    0 & 2\beta_0 \beta_2 - 2\beta_0 \beta_3 & 0 & 2\beta_0 \beta_4 - 2\beta_0 \beta_5 & \ldots & 2\beta_0 \beta_{n-1} - 2\beta_0 & 0 \\
    0 & 2\beta_0 \beta_2 - (2\beta_0 \beta_1 - 2\beta_0 \beta_3) & 0 & 2\beta_0 \beta_4 - \beta_0 \beta_5 & \ldots & 2\beta_0 \beta_{n-1} - 2\beta_0 & 0 \\
    \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
    \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
    0 & 2\beta_0 \beta_1 & \ldots & \beta_2 & 0 \\
\end{pmatrix} \quad (6.36)
\]

If the dimension is even, i.e. \( A \in \mathbb{R}^{2n \times n}, n = 2k \), \( k = 0, 1, 2, \ldots \):

\[
\begin{pmatrix}
    s^n & 1 & \beta_{n-2} & \ldots & \beta_0 \\
    s^{n-1} & \beta_{n-1} & \beta_{n-3} & \ldots & 0 \\
    \vdots & \vdots & \ddots & \ddots & \vdots \\
    s^1 & \beta_0 & \ldots & \beta_2 & 0 \\
\end{pmatrix} \quad (6.37)
\]

Then:
\[
L = L^T = \begin{bmatrix}
L_1 & 0 & L_2 & 0 & \cdots & L_k & 0 \\
L_{k+1} & 0 & L_{k+2} & 0 & \cdots & 0 & L_{(k-1)+1}+k \\
L_{2k+1} & 0 & \cdots & L_{(2k-(k+1)+2k} & 0 \\
L_{3k} & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \vdots \\
L_{n+1} & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \ddots \\
L_{n(n+2)/4} & 0 & \cdots & 0 \\
\end{bmatrix}
\]  
(6.38)

or:

\[
L = L^T = \begin{bmatrix}
2\beta_0\beta_1 & 0 & 2\beta_0\beta_3 & 0 & \cdots & 2\beta_0\beta_{n-1} & 0 \\
0 & 2\beta_0\beta_1 - 2\beta_0\beta_3 & 0 & 2\beta_0\beta_1 - 2\beta_0\beta_3 & \cdots & 0 & 2\beta_1 \\
0 & 2\beta_0\beta_3 - (2\beta_0\beta_1 - 2\beta_0\beta_3) & 0 & 2\beta_0\beta_1 - 2\beta_0\beta_3 & \cdots & 0 & 2\beta_3 \\
0 & 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\
2\beta_{n-1} & 0 & \cdots & 0 \\
\end{bmatrix}
\]  
(6.39)

**Proof:**

Assuming that \(A\) is Hurwitz and using the system’s controllability Lyapunov equation we get:

\[
AW_c + W_c A^T = -bb^T \Rightarrow LA + A^T L = -Lbb^T L
\]  
(6.40)

The system is in controller form then if \(A \in \mathbb{R}^{n \times n}, n = 2k + 1, k = 0, 1, 2, \ldots\) according to above Lyapunov equation we have:

\[
H = \begin{bmatrix}
L_1 & L_2 & L_3 & \cdots & L_n \\
L_{n+1} & L_{n+2} & L_{n+3} & \cdots & L_{2n-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
L_{n(n+1)/2} & L_{n(n+1)/2-1} & L_{n(n+1)/2-2} & \cdots & L_{n(n+1)/2-n} \\
\end{bmatrix} \begin{bmatrix}
0 & I_{(n-1)\times(n-1)} \\
-\beta_0 & -\beta_1 & \cdots & -\beta_{n-2} & -\beta_{n-1} \\
\end{bmatrix} + H_{n\times n}^T = -L \begin{bmatrix}
0 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 1 \\
\end{bmatrix}_{n\times n}
\]  
(6.41)

Furthermore, note that since the controllability Gramian in controller companion form is symmetric and zero plaid (Seeram & Agathoklis, 1991), (Xiao, et al., 1992), as it is stated in the previous section, the same must hold true for \(L\). Thus (6.35) reduces to:
By calculating the right hand side we obtain:

\[
H + H_n^{T} = -L \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}_{n \times n}
\]

(6.42)

Hence equation (6.42) defines the elements of symmetric matrix \( L \) as:
Thus (6.36) is verified.

Now consider the case that \( A \in \mathbb{R}^{m \times n}, n = 2k, k = 0, 1, 2, \ldots \). Regarding the zero-plaid structure of matrix \( L \) we can write the Lyapunov equation (6.40) as:

\[
\begin{bmatrix}
  L_1 & 0 & L_2 & 0 & \cdots & L_k & 0 \\
  0 & L_{k+1} & 0 & L_{k+2} & \cdots & 0 & L_{(k-1)+k+1} \\
  L_2 & 0 & L_{2k+1} & 0 & \cdots & L_{2k-(k+1)+2k} & 0 \\
  0 & L_{k+2} & 0 & L_{3k} & 0 & \cdots & L_{(3k-1)-(2k+1)+3(2k-1)} \\
  \vdots & \vdots & \vdots & 0 & L_{4k-1} & \cdots & 0 \\
  \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  0 & L_{2k} & 0 & L_{4k-2} & 0 & \cdots & L_{(m(n+2)-4)} \\
\end{bmatrix}
\begin{bmatrix}
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
\end{bmatrix}
= \begin{bmatrix}
  L_0 & \beta_0 \\
  L_0 & \beta_0 \\
  L_0 & \beta_0 \\
  L_0 & \beta_0 \\
  L_0 & \beta_0 \\
  L_0 & \beta_0 \\
  L_0 & \beta_0 \\
  L_0 & \beta_0 \\
\end{bmatrix}
\begin{bmatrix}
  0 \\
  0 \\
  0 \\
  0 \\
  0 \\
  0 \\
  0 \\
  0 \\
\end{bmatrix}
\]

which implies that:

\[
(6.45)
\]
Therefore, the nonzero elements of the symmetric matrix $L$ are readily obtained by Gaussian elimination with partial pivoting to equation (6.46) as:

$$
\begin{align*}
-2\beta_0 L_{k+1} &= -L_{k+1}^2 \\
L_1 - \beta_1 L_{k+1} &= 0 \\
-\beta_2 L_{k+1} - \beta_0 L_{2k+1} &= -L_{k+1} L_{3k+1} \\
L_2 - \beta_1 L_{k+1} &= 0 \\
-\beta_1 L_{k+1} - \beta_0 L_{5k+1} &= -L_{k+1} L_{5k+1} \\
\vdots
\end{align*}
\Rightarrow
\begin{align*}
L_{k+1} &= 2\beta_0 \\
L_1 &= 2\beta_1 \\
L_{2k+1} &= 2\beta_2 \\
L_{3k+1} &= 2\beta_3 \\
\vdots
\end{align*}
$$

Thus (6.39) is proved.

$$
\begin{align*}
L_{k+2} + L_2 - \beta_1 L_{5k+1} &= 0 \\
L_{k+3} + L_3 - \beta_1 L_{5k+1} &= 0 \\
L_{2k+2} + L_{k+3} - \beta_1 L_{9k+1} &= 0 \\
\vdots
\end{align*}
\Rightarrow
\begin{align*}
L_{k+2} &= 2\beta_2 \beta_2 - 2\beta_1 \beta_0 \\
L_{k+3} &= 2\beta_3 \beta_4 - 2\beta_1 \beta_0 \\
L_{2k+2} &= 2\beta_3 \beta_2 - 2\beta_1 \beta_4 + 2\beta_3 \beta_0 \\
\vdots
\end{align*}
$$

Considering the above Theorem, it is of interest to mention that the minimum energy and the inverse of the controllability Gramian of SISO stable LTI system can be directly obtained from the Routh–Hurwitz table through a simple algorithm.
1) At the beginning, we define the system in controller companion form. Then:

2) Define the first two rows of Routh–Hurwitz for the SISO stable LTI system

\[
A_1 \in \mathbb{R}^{n \times n}, n = 2k + 1, k = 0, 1, 2, \ldots
\]
\[
A_0 \in \mathbb{R}^{n \times n}, n = 2k, k = 0, 1, 2, \ldots
\]

\[
\begin{array}{cccc}
\ddots & \ddots & \cdots & \ddots \\
\ddots & \ddots & \ddots & \ddots \\
\ddots & \ddots & \ddots & \ddots \\
\ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & 0 \\
\end{array}
\]

\[
\begin{array}{cccc}
\ddots & \ddots & \cdots & \ddots \\
\ddots & \ddots & \ddots & \ddots \\
\ddots & \ddots & \ddots & \ddots \\
\ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & 0 \\
\end{array}
\]

3) The last column of \( L \) is obtained by multiplying the second row of coefficients in the Routh–Hurwitz table.

\[
A_1 \in \mathbb{R}^{n \times n}, n = 2k + 1, k = 0, 1, 2, \ldots
\]
\[
A_0 \in \mathbb{R}^{n \times n}, n = 2k, k = 0, 1, 2, \ldots
\]

\[
\begin{array}{cccc}
\times & 0 & \cdots & \times \\
0 & \times & \cdots & 0 \\
\times & 0 & \times & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \times & \times & 0 \\
2\beta_0 & 0 & \cdots & \times \\
\end{array}
\]

\[
\begin{array}{cccc}
\times & 0 & \cdots & \times \\
0 & \times & \cdots & 0 \\
\times & 0 & \times & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \times & \times & 0 \\
2\beta_0 & 0 & \cdots & \times \\
\end{array}
\]

4) The first row of \( L \) can be readily determined based on the last column elements and the first
row of coefficients in the Routh–Hurwitz table as:

\[
A_1 \in \mathbb{R}^{n \times n}, n = 2k + 1, k = 0, 1, 2, \ldots
\]

\[
2\beta_0 \beta_1 \quad 0 \quad 2\beta_0 \beta_3 \quad \cdots \quad 2\beta_0 \beta_{n-4} \quad 0 \quad 2\beta_0 \beta_{n-2} \quad 0 \quad 2\beta_0
\]

\[
A_1 \in \mathbb{R}^{n \times n}, n = 2k, k = 0, 1, 2, \ldots
\]

\[
2\beta_0 \beta_1 \quad 0 \quad 2\beta_0 \beta_3 \quad \cdots \quad 0 \quad 2\beta_0 \beta_{n-3} \quad 0 \quad 2\beta_0 \beta_{n-1} \quad 0
\]

5) The rest of the elements are determined by the opposite sign of their upper right hand side element plus a proper coefficient of the last column's arrays. These coefficients are selected from the first row of the corresponding Routh–Hurwitz table:
In the last step of the algorithm, we just need to use the non-singular similarity matrix:

\[
\begin{bmatrix}
\beta_1 & \cdots & \beta_{n-4} & 0 & \beta_{n-2} & 0 & 0 \\
0 & \cdots & 0 & \beta_1 \ast - \beta_{n-2} & 0 & \beta_1 \times - 0 & 0 \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\beta_{n-6} \ast & 0 & \beta_{n-2} \ast + \beta_{n-6} \ast - \beta_{n-2} \Delta & 0 & \beta_{n-2} \ast - (\beta_{n-6} \times - \Delta) & 0 & \beta_{n-4} \ast - (\beta_{n-6} \times - \Delta) \\
0 & \beta_{n-2} \ast + \beta_{n-6} \ast - \beta_{n-2} \Delta & 0 & \beta_{n-2} \ast - (\beta_{n-6} \times - \Delta) & 0 & \beta_{n-4} \ast - (\beta_{n-6} \times - \Delta) & 0 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
\beta_1 & \cdots & \beta_{n-4} & 0 & \beta_{n-2} & 0 & 0 \\
0 & \cdots & 0 & \beta_1 \ast - (\beta_{n-2} \times - \beta_0) & 0 & \beta_2 \times - 0 & 0 \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\beta_{n-6} \ast & 0 & \beta_{n-2} \ast + \beta_{n-6} \ast - \beta_{n-2} \Delta & 0 & \beta_{n-2} \ast - (\beta_{n-6} \times - \Delta) & 0 & \beta_{n-4} \ast - (\beta_{n-6} \times - \Delta) \\
0 & \beta_{n-2} \ast + \beta_{n-6} \ast - \beta_{n-2} \Delta & 0 & \beta_{n-2} \ast - (\beta_{n-6} \times - \Delta) & 0 & \beta_{n-4} \ast - (\beta_{n-6} \times - \Delta) & 0 \\
\end{bmatrix}
\]

6) In the last step of the algorithm, we just need to use the non-singular similarity matrix:

\[
T = \begin{bmatrix}
b & Ab & A^2 b & \cdots & A^{n-1} b \\
\end{bmatrix}
\begin{bmatrix}
\beta_1 & \beta_2 & \cdots & \beta_{n-1} & 1 \\
\beta_2 & \beta_3 & \cdots & 1 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
1 & 0 & \cdots & 0 & 0 \\
\end{bmatrix}
\]

to find the inverse of the controllability Gramian \( W_c^{-1} = T^{-T} L T^{-1} \), then the minimum input energy is readily obtained as \( E_{\text{min}} = x_f^T W_c^{-1} x_f \).
The importance of this method lies in the fact that given a stable linear time-invariant continuous-time system in the frequency domain (i.e. transfer function \( G(s) = \frac{b(s)}{a(s)} \)) or the time domain (i.e. state space), we are able to calculate the minimum input energy directly in terms of the coefficients of the transfer function (i.e. \( a(s), b(s) \)) or equivalently based on the Routh-Hurwitz table.

Let us to review the details of the proposed method by some examples:

**Example (6.6):** As in the first example, consider the stable 4-dimensional LTI system below:

\[
A = \begin{bmatrix}
-1.1577 & 0.0161 & -0.5121 & -0.1164 \\
0.0161 & -0.9540 & -0.1244 & 0.1954 \\
-0.5121 & -0.1244 & -1.7388 & 0.0544 \\
-0.1164 & 0.1954 & 0.0544 & -2.0123
\end{bmatrix}, \quad B = \begin{bmatrix}
-1.0690 \\
-0.8095 \\
-2.9440 \\
0
\end{bmatrix}
\]

We are interested to calculate the minimum energy to transfer the states of the system from the origin to the final state \( x_f = [1 0 0 0]^T \).

To be able to use our algorithm we consider the controller model of the system as:

\[
A_c = \begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
-3.2289 & -10.5455 & -12.1920 & -5.8628
\end{bmatrix}, \quad B_c = \begin{bmatrix}
0 \\
0 \\
0 \\
1
\end{bmatrix}
\]

Then regarding the dimension of the system (\( n = 4 \Rightarrow n = 2k, k = 2 \)), based on Theorem (6.1) we can define the symmetric matrix \( L \) as:

\[
L = \begin{bmatrix}
L_1 & 0 & L_2 & 0 \\
\times & L_3 & 0 & L_4 \\
\times & \times & L_5 & 0 \\
\times & \times & \times & L_6
\end{bmatrix}
\]

Then in the first step of the algorithm, we write the Routh–Hurwitz table for the system:

\[
\begin{array}{cccc}
 s^4 & 1 & 12.1920 & 3.2289 \\
 s^3 & 5.8628 & 10.5455 & 0 \\
 s^2 & & & \\
 s^1 & & & \\
 s^0 & & & 
\end{array}
\]

In the second step, the last column of matrix \( L \) is determined as the second row of the Routh–Hurwitz coefficients multiplied by two:
According to the next step of the algorithm we determine the nonzero elements in the first row of matrix $L$:

\[
L_0 = 2 \times 5.8628 = 11.7256 \\
L_4 = 2 \times 10.5455 = 21.0910
\]

In the fourth step, we calculate the rest nonzero elements of the symmetric matrix $L$ as:

\[
L = \begin{bmatrix}
68.1007 & 0 & 37.8608 & 0 \\
0 & 219.2807 & 0 & 21.0910 \\
37.8608 & 0 & 121.8675 & 0 \\
0 & 21.0910 & 0 & 11.7256
\end{bmatrix}
\]

Hence, we get the inverse of the controllability Gramian in the controller form as:

\[
L = \begin{bmatrix}
68.1007 & 0 & 37.8608 & 0 \\
0 & 219.2807 & 0 & 21.0910 \\
37.8608 & 0 & 121.8675 & 0 \\
0 & 21.0910 & 0 & 11.7256
\end{bmatrix}
\]

Now in the last step, we can easily use the similarity transformation $T$ in Lemma (6.2) to obtain the inverse of the controllability Gramian $W_c^{-1}$ and to calculate the minimum input energy of the system:

\[
T = \begin{bmatrix}
-1.0690 & 2.7322 & -6.0760 & 12.8639 \\
-0.8095 & 1.1215 & -1.7816 & 3.1611 \\
-2.9440 & 5.7671 & -11.5769 & 23.4956 \\
0 & -0.1940 & 0.6054 & -1.4895
\end{bmatrix}
\]

or:

\[
T = \begin{bmatrix}
-0.7199 & -3.0907 & -3.5351 & -1.0690 \\
-2.1476 & -5.0760 & -3.6244 & -0.8095 \\
-0.3051 & -0.5319 & -0.1940 & 0
\end{bmatrix}
\]

Hence:

\[
W_c^{-1} = T^{-T}LT^{-1} = 1.0e+06 \times \begin{bmatrix}
0.0168 & -0.0327 & 0.0028 & 0.1423 \\
-0.0327 & 0.5428 & -0.1369 & -1.4594 \\
0.0028 & -0.1369 & 0.0366 & 0.3482 \\
0.1423 & -1.4594 & 0.3482 & 4.1250
\end{bmatrix}
\]
and finally:

\[ E_{\text{min}} = x_f^T W_c^{-1} x_f = 1.6759 \times 10^4 \]

**Example (6.7):** In the second example, we calculate the minimum input energy for the 5-dimensional stable system below:

\[
A = \begin{bmatrix}
-1.0860 & 1.2673 & 0.2865 & -2.1291 & -0.7724 \\
-0.9140 & -0.4367 & -0.4485 & 0.5639 & 0.6074 \\
-0.5004 & 0.0405 & 0.6074 & -0.4485 & 0.5639 \\
0.2028 & 0.3004 & -0.9072 & -0.4105 & 0.5639 \\
1.2718 & -0.3384 & 0.1786 & 0.3489 & -0.3977
\end{bmatrix}, \quad \begin{bmatrix}
B \\
-0.2725 \\
1.0984 \\
-0.2779 \\
0
\end{bmatrix}
\]

Assume the terminal state is defined as: \( x_f = [1 \ 0 \ 0 \ 0 \ 0]^T \).

In the beginning, we need to define the controller form of the system:

\[
A_c = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
-0.1409 & -2.0769 & -9.2391 & -11.6151 & -3.3676
\end{bmatrix}, \quad \begin{bmatrix}
B_c \\
0 \\
0 \\
0 \\
1
\end{bmatrix}
\]

Noting that the dimension of the system is odd \( (n = 5 \Rightarrow n = 2k + 1, k = 2) \), we define the structure of the symmetric matrix \( L \) using Theorem (6.1) as:

\[
L = \begin{bmatrix}
L_1 & 0 & L_2 & 0 & L_3 \\
\times & L_4 & 0 & L_5 & 0 \\
\times & \times & L_6 & 0 & L_7 \\
\times & \times & \times & L_8 & 0 \\
\times & \times & \times & \times & L_9
\end{bmatrix}
\]

In the first step, we write the first two rows of Routh–Hurwitz table for the system:

\[
\begin{align*}
s^5 & 11.6151 & 2.0769 \\
s^4 & 3.3676 & 9.2391 & 0.1409 \\
s^3 & 3.3676 & 9.2391 & 0.1409 \\
s^2 & 3.3676 & 9.2391 & 0.1409 \\
s^1 & 3.3676 & 9.2391 & 0.1409 \\
s^0 & 3.3676 & 9.2391 & 0.1409
\end{align*}
\]
Then the nonzero elements in the last column of $L$ are determined as the second row's elements of the Routh–Hurwitz multiplied by two:

$$L_y = 2 \times 3.3676 = 6.7352$$
$$L_z = 2 \times 9.2391 = 18.4782$$
$$L_3 = 2 \times 0.1409 = 0.2818$$

In the next step of the algorithm, the first row of matrix $L$ is described as:

$$L_x = 11.6151 \times 0.2817 = 3.2720$$
$$L_4 = 2.0769 \times 0.2817 = 0.5851$$

Then in the third step of the algorithm, we readily obtain the rest of nonzero elements of the symmetric matrix $L$ as:

$$L = \begin{bmatrix}
L_1 & 0 & L_2 & 0 & L_3 \\
\times & \times & 2.0769L_2 - L_2 & 0 & 2.0769L_y - L_3 & 0 \\
\times & \times & 11.6151L_4 - (2.0769L_y - L_3) & 0 & L_4 \\
\times & \times & \times & 11.6151L_4 - L_7 & 0 \\
\times & \times & \times & \times & L_y 
\end{bmatrix}$$

Thus:

$$L = \begin{bmatrix}
0.5851 & 0 & 3.2720 & 0 & 0.2818 \\
0 & 35.1051 & 0 & 13.7066 & 0 \\
3.2720 & 0 & 200.9196 & 0 & 18.4782 \\
0 & 13.7066 & 0 & 59.7518 & 0 \\
0.2818 & 0 & 18.4782 & 0 & 6.7352 
\end{bmatrix}$$

Now in the last stage, we can readily calculate the minimum input energy as:

$$T = \begin{bmatrix}
-0.2725 & 1.6084 & 0.0997 & -14.9854 & 35.1977 \\
1.0984 & -0.1060 & -2.3810 & 5.2822 & 9.3239 \\
-0.2779 & 0.3309 & -0.7859 & -1.4947 & 11.2815 \\
0 & -0.6072 & 4.6541 & -6.4990 & -26.2780 \\
0 & -0.7679 & 2.2339 & 1.5276 & -23.9877 
\end{bmatrix}$$

or:
\[
T = \begin{bmatrix}
0.1836 & 1.5140 & 2.3513 & 0.6908 & -0.2725 \\
0.7597 & 6.1815 & 10.0204 & 3.5931 & 1.0984 \\
-0.3997 & -2.8647 & -2.8989 & -0.6048 & -0.2779 \\
0.2832 & 2.1218 & 2.6093 & -0.6072 & 0 \\
0.0098 & 0.1321 & -0.3519 & -0.7679 & 0 \\
\end{bmatrix}
\]

Hence:

\[
W_c^{-1} = T^{-T}LT^{-1} = 1.0e+03 \times \begin{bmatrix}
0.4138 & 0.0699 & -0.0493 & -0.5970 & 1.1412 \\
0.0699 & 0.0246 & 0.0176 & -0.1033 & 0.2358 \\
-0.0493 & 0.0176 & 0.0825 & 0.0978 & -0.0763 \\
-0.5970 & -0.1033 & 0.0978 & 0.9071 & -1.6991 \\
1.1412 & 0.2358 & -0.0763 & -1.6991 & 3.4126 \\
\end{bmatrix}
\]

and then:

\[
E_{\min} = x_f^TW_c^{-1}x_f = 413.7565
\]

### 6.4. Controllability Gramian Computation in Diagonal Canonical Form

In this section, we investigate the controllability Gramian of the stable linear continuous time-invariant system (6.1) as the solution of the Lyapunov equation (6.2), in the case that matrix \(A\) is diagonalizable.

Let \(T\) be a non-singular transformation matrix such that:

\[
T^{-1}AT = \Sigma
\]

where \(\Sigma\) is diagonal matrix.

Note that \(T\) is the matrix of the eigenvectors and \(\Sigma\) is the diagonal matrix of the eigenvalues of \(A\).

Define a new state vector \(\tilde{x} = T^{-1}x\), then the linear system (6.1) transforms to:

\[
\begin{cases}
\dot{\tilde{x}} = \Sigma \tilde{x} + T^{-1}Bu \\
y = CT\tilde{x} + Du
\end{cases} \quad (6.48)
\]

which represents \(n\) decoupled first order equations.

Denoting the solution of the Lyapunov equation associated with above system as \(W_c^{\text{diag}}\), the controllability Gramian of the original system (6.1) will be equal to:

\[
W_c = TW_c^{\text{diag}}T^T \quad (6.49)
\]

The Proposition below presents the explicit solution of the Lyapunov equation in diagonal canonical form.
**Proposition (6.4):** Consider the stable LTI system (6.48). The controllability Gramian $W_{c}^{\text{diag}}(0,t_f)$ is obtained as:

$$W_{c}^{\text{diag}}(0,t_f)_{i,j} = (T^{-1}B)_{i}(T^{-1}B)_{j}^{T} \left( e^{(\lambda_i + \lambda_j)\tau} - 1 \right)$$

where $(T^{-1}B)_{i}, (T^{-1}B)_{j}^{T}$ represent $i$th and $j$th rows of $T^{-1}B$ and $\lambda_i$ denotes $i$th eigenvalue of $A$ and $T$ is the eigenvectors matrix of $A$.

**Proof:**

According to the definition of the controllability Gramian, we have:

$$W_{c}^{\text{diag}}(0,t_f) = \int_{0}^{t_f} e^{\sum_{i} e^{-\tau}BB^{T}T^{-T}e^{\sum_{i} e^{\lambda_i\tau}} d\tau}$$

Since $\Sigma$ is a diagonal matrix, exponentiation can be performed simply by exponentiating each of the diagonal elements:

$$W_{c}^{\text{diag}}(0,t_f) = \int_{0}^{t_f} \begin{bmatrix} e^{\lambda_1\tau} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2\tau} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n\tau} \end{bmatrix} \begin{bmatrix} e^{\lambda_1\tau} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2\tau} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n\tau} \end{bmatrix} d\tau$$

Using the Kronecker product we obtain:

$$\text{vec}(W_{c}^{\text{diag}}(0,t_f)) = \int_{0}^{t_f} \left( \begin{bmatrix} e^{\lambda_1\tau} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2\tau} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n\tau} \end{bmatrix} \otimes \begin{bmatrix} e^{\lambda_1\tau} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2\tau} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n\tau} \end{bmatrix} \right) d\tau \text{vec}(H)$$

where:

$$\text{vec}(W_{c}^{\text{diag}}(0,t_f)) = \int_{0}^{t_f} e^{\Sigma \otimes \Sigma \tau} d\tau \text{vec}(H)$$

or:

$$\text{vec}(W_{c}^{\text{diag}}(0,t_f)) = \int_{0}^{t_f} \begin{bmatrix} e^{2\lambda_1\tau} & 0 & \cdots & 0 \\ 0 & e^{\lambda_1 + \lambda_2\tau} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{2\lambda_n\tau} \end{bmatrix} d\tau \text{vec}(H)$$

Equation (6.55) yields:
\[
\begin{align*}
W_{e}^{\text{diag}}(0,t_{f})_{i,j} &= \left( \int_{0}^{t} e^{2\lambda \tau} d\tau \right) \left( T^{-1}B \right)_{i}^{T} \\
W_{e}^{\text{diag}}(0,t_{f})_{2,i} &= \left( \int_{0}^{t} e^{(\lambda_{i}+\lambda_{j}) \tau} d\tau \right) \left( T^{-1}B \right)_{2}^{T} \\
&\vdots \\
W_{e}^{\text{diag}}(0,t_{f})_{i,j} &= \left( \int_{0}^{t} e^{(\lambda_{i}+\lambda_{j}) \tau} d\tau \right) \left( T^{-1}B \right)_{i}^{T} \\
&\vdots \\
W_{e}^{\text{diag}}(0,t_{f})_{n,n} &= \left( \int_{0}^{t} e^{(\lambda_{n}+\lambda_{n}) \tau} d\tau \right) \left( T^{-1}B \right)_{n}^{T}
\end{align*}
\]

(6.56)

and the Proposition is proved.

\[\square\]

**Corollary (6.3):** The solution of the Lyapunov equation of the stable LTI system (6.41) is:

\[
W_{e}^{\text{diag}}(0,\infty)_{i,j} = -\left( T^{-1}B \right)_{i}^{T} \left( T^{-1}B \right)_{j} \left( \lambda_{i} + \lambda_{j} \right)^{-1}
\]

(6.57)

where \( (T^{-1}B)_{i} \) and \( (T^{-1}B)_{j}^{T} \) represent the \( i \)th and \( j \)th rows of \( T^{-1}B \), \( \lambda_{i} \) denotes the \( i \)th eigenvalue of \( A \) and \( T \) is the eigenvectors matrix of \( A \).

\[\square\]

**Corollary (6.4):** Consider the stable LTI system (6.48) in diagonal form. Then the trace of the controllability Gramian when \( t_{f} \to \infty \) is obtained as:

\[
\text{trace}(W_{e}^{\text{diag}}(0,\infty)) = -0.5 \sum_{i=1}^{n} \frac{(T^{-1}BB^{T}T^{-T})_{ii}}{\lambda_{i}}
\]

(6.58)

Again \( (T^{-1}B)_{i} \) and \( (T^{-1}B)_{j}^{T} \) denote the \( i \)th and \( j \)th rows of \( T^{-1}B \), \( \lambda_{i} \) is the \( i \)th eigenvalue of \( A \) and \( T \) is the eigenvector matrix of \( A \).

\[\square\]
6.5. Summary

Two different canonical form structures of the system have been discussed in this chapter:

- Phase-variable canonical form
- Diagonal canonical form

We considered the stable LTI system in the above canonical form structures and introduced the interesting links that the entries of the controllability Gramian of the system have to the entries of the Routh Hurwitz table. Moreover, an expression for the minimum input energy is developed as a simple function of the coefficients of the characteristic polynomial from the fact that it is connected to the inverse of the controllability Gramian.

Furthermore, we devised an approach to determine the value of the trace and an upper-bound for the maximum eigenvalue of the controllability Gramian based directly on the coefficients of the characteristic polynomial of the system.

In the next chapter, the problem that will be considered is the variability of the input structure. This will lead to the investigation of the selection of the optimal input structure under the minimum average energy requirement. A variety of such problem will be examined next.
Chapter 7: Optimal Selection of Input Structure, Minimizing the Average of Input Energy

7.1. Introduction

In this Chapter, we propose a novel approach for selection of input (and by duality output) structure based on the energy type criteria. The main aim of this chapter is to propose a strategy for the selection of the proper input matrix $B$, based on the average of the minimum input energy. The overall aim is to use these results to define the proper number and location of the sensors and actuators in a control system.

In the following parts of the chapter we discuss the input structure selection problem in different cases. First, we consider the case that the input matrix $B$ is a single input vector. We look for the best input structure to minimise the energy requirements and we propose a solution for a general system and a finite terminal time $t_f$. Then, in the next part we assume the system to be stable and that the terminal time tends to infinity, using appropriate Lyapunov equations. This is followed by the input structure selection for a normal system where $AA^* = A^*A$. Finally, in the case of multi-input systems, we define the proper set of
input vectors that minimize the average of the minimum input energy, and we investigate the problem subject to different possible conditions.

The work here provides a new approach to those in existing literature, which consider the problem of input structure selection over a binary set, where the input matrix $B$ can only take the values $\{0,1\}$ (V. Tzoumas, 2015), (Kumar & Narayanan, 2008), (Shaker & Tahavori, 2013) or can be chosen among the given sets (Summers & Lygeros, 2014), (Cortesi, et al., 2015). In contrast, the current approach developed in this chapter doesn’t have such limitations.

7.2. Optimal Selection of Input to Minimize the Average of Minimum Input Energy

In previous chapters, the minimum input energy was defined as a function of the inverse of the controllability Gramian and the effect of input structure on the value of minimum energy requirements was examined. In this chapter, we are looking for the best selection of the input matrix $B$, such that the average of the minimum energy is minimized. The mathematical formulation of this optimization problem is presented in (7.1).

\[
\min_{\beta} \{ \text{ave} \ E_{\min[0,t_f]} \} = \min_{\beta} \{ \text{ave} \ x^Tf W_c^{-1}(0,t_f) x_f \} \tag{7.1}
\]

where $\beta$ is the set of all possible inputs and $E_{\min[0,t_f]}$ is the minimum input energy evaluated over the interval $[0,t_f]$. $x_f$ defines the final state and $W_c^{-1}(0,t_f)$ is the inverse of the controllability Gramian. $\| \|$ denotes the standard Euclidean 2-norm.

**Theorem (7.1):** The average value of the minimum input energy $E_{\min}$ over the unit hypersphere is equal to $n^{-1} \text{trace}(W_c^{-1}(0,t_f))$ where $n$ denotes the dimension of the system.

**Proof:**

Based on the results of chapter 4, $E_{\min}$, the minimum energy required for the state trajectory to reach the final state $x_f$ from the origin in a terminal time $t_f$, was found as:

\[
E_{\min[0,t_f]} = x_f^T W_c^{-1} x_f \tag{7.2}
\]

The average value of $E_{\min[0,t_f]}$ over the unit hypersphere is obtained by integrating (7.2), $\|x_f\| = 1$, (Summers & Lygeros, 2014):

133
\[
aveE_{\min(0,t_f)} = \frac{\int_{|y| \leq 1} x^T W_c^{-1}(0,t_f)x \, dx}{\int_{|y| \leq 1} \, dx} = \frac{\text{trace}(W_c^{-1}(0,t_f))}{n}
\]  

(7.3)

and hence the Theorem is proved.

Note that the trace of \( W_c(0,t_f) \) is inversely related to the trace of \( W_c^{-1}(0,t_f) \), and thus maximizing \( \text{trace}(W_c(0,t_f)) \) effectively minimizes the average of \( E_{\min} \) required to move around the state space in all directions (Le, et al., 2015), (Cortesi, et al., 2015).

For an asymptotically stable strictly-proper system its \( H_2 \) norm is an important design parameter and is defined as (Antoulas, 2005):

\[
\|H\|_2 = (\text{trace} \frac{1}{2\pi} \int_{-\infty}^{\infty} H(j\omega)H^*(j\omega) d\omega)^{\frac{1}{2}}
\]  

(7.4)

This can be interpreted as the RMS\(^3\) response of the system when it is driven by a white noise input.

**Proposition (7.1):** (Summers & Lygeros, 2014) Let \( H(s) \) be an asymptotically stable strictly proper matrix function with realization \( (A, B, C) \). Then:

\[
W_c(0, \infty) = \int_{0}^{\infty} e^{A\tau} B B^* e^{A^T\tau} \, d\tau
\]

is the controllability Gramian of the system, from which the system \( H_2 \) norm is obtained as:

\[
\text{trace}(C W_c(0, \infty) C^*) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \text{trace}(H(jw)H^*(jw)) \, dw = \|H(s)\|_2^2
\]  

(7.5)\(^4\)

**Proof:**

The controllability Gramian corresponding to a stable LTI system can be expressed in the frequency domain using Parseval’s Theorem as:

\[
W_c(0, \infty) = \int_{0}^{\infty} e^{A\tau} B B^* e^{A^T\tau} \, d\tau = \frac{1}{2\pi} \int_{-\infty}^{\infty} (jwI - A)^{-1} B B^* (jwI - A^T)^{-1} \, dw
\]  

(7.6)

Now if we multiply both sides by output matrix \( C \), we get:

---

\(^3\) RMS norm is not a norm, but only a semi-norm, since we can have nonzero signals with zero RMS value

\(^4\) For simplicity of notation, quantities in the time and frequency domains will be denoted by the same symbol.
\[
CW_\epsilon(0, \infty)C^* = \frac{1}{2\pi} \int_{-\infty}^{\infty} C(jwI - A)^{-1} BB^* (-jwI - A^*)^{-1} C^* dw = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(jw)G^*(jw) dw
\]

Then it can be easily inferred from the above equation:

\[
\text{trace}\{CW_\epsilon(0, \infty)C^*\} = \frac{1}{2\pi} \text{trace}\left[\int_{-\infty}^{\infty} G(jw)G^*(jw) dw\right] \tag{7.7}
\]

which proves the result.

The Proposition above shows that the $H_2$ norm of a stable strictly proper system is equal to the $L_2$ norm of its (causal) impulse response, i.e. the energy of the system response to a unit impulse input, and can be represented as a weighted trace of the controllability Gramian $W_c(0, \infty)$.

Thus, maximizing the $H_2$ norm, minimizes a weighted average energy required to move around the state space, with certain directions weighted differently, which can be encoded into the output matrix.

In the following, we discuss the proper selection of input matrix $B$, such that the average energy of the trajectory between the origin and a terminal state uniformly distributed over the unit hypersphere, which is formulated in (7.1) is minimized.

Thus, the main aim in this chapter is the optimization of the average of the minimum energy, over the set of all possible input matrices $B$.

According to what was stated above, the optimization problem can be equivalently expressed as the maximization of the controllability Gramian's trace over the set of all inputs defined over the interval $[0, t_f]$ and satisfying the condition $\|B\| = 1$:

\[
\max_{\|B\|=1} \{\text{trace}(W_c(0, t_f))\} \tag{7.8}
\]

Note again that $\beta$ is the set of all possible inputs.

In this chapter, considering the dimension of input matrix $B$, the solution to the presented optimization problem (7.8) is investigated in two main parts:

- **Single input:**
  Where matrix $B$ is an input vector as $b_{n \times 1}$

- **Multiple inputs:**
  Where input matrix $B$ is a matrix of dimension $n \times m$

### 7.2.1. Single Input Matrix Structure Selection Problem

Consider the case of a single input LTI system. The aim is to select the proper input vector $b_{n \times 1}$, such that the trace of the controllability Gramian $W_c(0, t_f)$ is maximized. This guarantees the least value of the average of the minimum input energy $E_{min}$. 
According to the equation (7.8), we describe the optimization problem in this section as:

$$\min_{b \in \mathbb{B}} \{ \text{ave} \left( E_{\min[0,t_f]} \right) \} = \max_{b \in \mathbb{B}} \{ \text{trace}(W_c(0,t_f)) \}$$

(7.9)

where $\mathbb{B}$ is the set of all possible inputs and $E_{\min[0,t_f]}$ is the minimum input energy evaluated over the interval $[0,t_f]$.

In the following, this problem is investigated by considering various separate cases, such as the finite and infinite terminal times, and different properties for the system matrix $A$.

### 7.2.1.a. Optimal Input Vector Selection Considering the Finite Terminal Time

Here, we investigate the solution to the optimization problem (7.9) for the general single input LTI system, while the terminal time $t_f$ is finite. No special assumptions on system matrix $A$ are imposed. Note that the system may be stable, or unstable.

**Theorem (7.2):** In a LTI single input system, the eigenvector corresponding to the maximum eigenvalue of matrix $Q$ defined below describes the optimal input vector $b$, which minimizes the average of the minimum input energy, i.e.

$$\min_{b \in \mathbb{B}} \{ \text{ave} \left( E_{\min[0,t_f]} \right) \} = \lambda_{\max} (Q)$$

$$Q_{\text{nec}} = \int_0^{t_f} e^{A't}e^{At} dt$$

(7.10)

$$\arg \min_{b \in \mathbb{B}} \{ \text{ave} \left( E_{\min[0,t_f]} \right) \} = b_{\text{nec}} = V_{\max} (Q)$$

$\mathbb{B}$ is the set of all possible inputs. $\lambda_{\max} (Q)$, $V_{\max} (Q)$ denote the largest eigenvalue of $Q$ and the corresponding eigenvector respectively.

**Proof:**

Considering the definition of the controllability Gramian:
\[ W_\varepsilon (0, t_f) = \int_0^{t_f} e^{At} b b^* e^{A^* t} dt \]  
(7.11)

we get:

\[ \text{trace}(W_\varepsilon (0, t_f)) = \text{trace} \left( \int_0^{t_f} e^{At} b b^* e^{A^* t} dt \right) \]  
(7.12)

Using the linearity property of trace, the equation above can be stated as:

\[ \text{trace}(W_\varepsilon (0, t_f)) = \int_0^{t_f} \text{trace}(e^{At} b b^* e^{A^* t}) dt \]  
(7.13)

Using the commutative property of trace, this can be rewritten as:

\[ \text{trace}(W_\varepsilon (0, t_f)) = \int_0^{t_f} \text{trace}(b^* e^{A^* t} e^{At} b) dt \]  
(7.14)

Since \( b^* e^{A^* t} e^{At} b \) is scalar, the trace function can be ignored, and we get:

\[ \text{trace}(W_\varepsilon (0, t_f)) = \int_0^{t_f} b^* e^{A^* t} e^{At} b dt \]  
(7.15)

and since \( b \) and \( b^* \) are constant, we obtain:

\[ \text{trace}(W_\varepsilon (0, t_f)) = b^* \left( \int_0^{t_f} e^{A^* t} e^{At} dt \right) b \]  
(7.16)

By defining the n-dimensional matrix \( Q \):

\[ Q_{n \times n} = \int_0^{t_f} e^{A^* t} e^{At} dt \]  
(7.17)

we get:

\[ \min_{b_{1 \times n}} \{ \text{ave} E_{\min[0, t_f]} \} = \max_{b_{1 \times n}} \{ \text{trace}(W_\varepsilon (0, t_f)) \} = \max_{b_{1 \times \beta}} \{ b^* Q b \} = \lambda_{\max} (Q) \]  
(7.18)

and:

\[ \arg \max_{b_{1 \times \beta}} \{ \text{trace}(W_\varepsilon (0, t_f)) \} = b_{n \times 1}^{\text{opt}} = V_{\max} (Q) \]  
(7.19)

Here \( \lambda_{\max} (Q) \) is the largest eigenvalue of \( Q \) and \( V_{\max} (Q) \) is the eigenvector corresponding to it.

The equation above shows that in a LTI single input system, the eigenvector corresponding to the maximum eigenvalue of \( Q_{n \times n} = \int_0^{t_f} e^{A^* t} e^{At} dt \), describes the input matrix \( b \), which optimizes the average of the minimum input energy.
**Example (7.1):** (Rubin, 2016) Consider the LTI system below, which may have a real-world example such as a rocket in vertical motion:

\[
\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + b_{2\omega} u
\]

Now the problem can be defined as the selection of input vector \( b \), which maximizes the trace of controllability Gramian, i.e. the minimum average of input energy.

Based on the Theorem above, the eigenvector corresponding to the maximum eigenvalue of matrix \( Q \) in (7.17) corresponds to the optimal input vector \( b \), which leads to the maximum trace of the controllability Gramian. In addition, as one can easily see, the system is unstable since the eigenvalues of the system are located in the origin, \( \lambda_1 = \lambda_2 = 0 \). Let \( t_f = 1 \) be the (finite) terminal time. Then based on (7.17) we get:

\[
Q = \int_0^1 e^{A t_f} e^{A^T t_f} dt = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.3333 \end{bmatrix}
\]

The characteristic polynomial of \( Q \) is:

\[
|\lambda I - A| = \begin{bmatrix} \lambda - 1 & -0.5 \\ -0.5 & \lambda - 1.3333 \end{bmatrix} \implies \lambda^2 - 2.3333\lambda + 1.0833 = 0
\]

and hence the two eigenvalues and corresponding eigenvectors are:

\[
\lambda_1 = 1.6937, \lambda_2 = 0.6396
\]

and

\[
V_1 = \begin{bmatrix} 0.5847 \\ 0.8112 \end{bmatrix}, V_2 = \begin{bmatrix} -0.8112 \\ 0.5847 \end{bmatrix}
\]

respectively.

Then based on Theorem (7.2) choosing \( b = V_1 \begin{bmatrix} 0.5847 \\ 0.8112 \end{bmatrix} \), we achieve the maximum trace of the controllability Gramian, that is equal to 1.6937. This is illustrated in Figure (7.1) for vector \( b \) described uniformly in the interval \([-1, 1]\) to verify the validity of this result.
The importance of Theorem (7.2) is that it can be effectively used to find the optimal input vector for any LTI-SISO system, stable or unstable, with the minimum possible complexity in computation. After matrix $Q$ is calculated, the eigenvector corresponding to its largest eigenvalue is the solution of the optimization problem.

### 7.2.1.b. Optimal Input Vector Selection for the Case of Stable System

**Proposition (7.2):** For single input LTI system, optimal selection of the input vector $b$, which maximizes the trace of the controllability Gramian $W_c(0, \infty)$, i.e. minimizes the average of the minimum input energy, is given by the normalized eigenvector corresponding to the maximum eigenvalue of the unique solution to the Lyapunov equation below:

For a single input LTI system, we now assume that the system is stable (i.e., all eigenvalues of $A$ are in the open left half plane). We also consider the limit of the terminal time $t_f$ tends to infinity, i.e. $t_f \to \infty$.

**Problem:**
Given a stable LTI SISO system, find

$$\max_{b \in \beta} \{\text{trace}(W_c(0, \infty))\},$$

where $\beta$ is the set of all possible input vector $b$ with $\|b\|_1 = 1$.

**Proposition (7.2):** For single input LTI system, optimal selection of the input vector $b$, which maximizes the trace of the controllability Gramian $W_c(0, \infty)$, i.e. minimizes the average of the minimum input energy, is given by the normalized eigenvector corresponding to the maximum eigenvalue of the unique solution to the Lyapunov equation below:
Further, the value of the maximum trace is equal to the maximum eigenvalue of $Q$, i.e.

$$\min_{b_{u_e} \in \beta} \{\text{ave} E_{\min(0,\infty)}\} = \lambda_{\text{max}}(Q) \quad (7.21)$$

$$\arg \min_{b_{u_e} \in \beta} \{\text{ave} E_{\min(0,\infty)}\} = b_{\text{opt}}^{\text{eq}} = V_{\text{max}}(Q) \quad (7.22)$$

$\lambda_{\text{max}}(Q)$, $V_{\text{max}}(Q)$ denote the largest eigenvalue of $Q$ and the corresponding eigenvector respectively, $\beta$ denotes the set of all possible input vector.

\[\square\]

**Proof:**

For simplicity, the largest eigenvalue of $Q$ is assumed to be distinct. The result is still true if this condition is violated but in this case the solution is not uniquely defined. The controllability Gramian has been defined as the integral (7.23):

$$W_c(0,\infty) = \int_0^\infty e^{-At}bb^*e^{At} dt \quad (7.23)$$

If the LTI system is stable, i.e. $A$ is Hurwitz, the controllability Gramian $W_c(0,\infty)$ is the unique solution for the Lyapunov equation (Raczyński & Stanisławski, 2012):

$$A^*W_c(0,\infty) + W_c(0,\infty)A + bb^* = 0 \quad (7.24)$$

Similar to section (7.2.1.a), we consider the trace of the controllability Gramian as:

$$\text{trace}(W_c(0,\infty)) = b^*(\int_0^\infty e^{-At}e^{At} dt)b \quad (7.25)$$

and then by introducing the constant matrix $Q$ as in (7.26) below, we derive the optimum input vector which minimizes the average of the minimum input energy:

$$\text{trace}(W_c(0,\infty)) = b^*Qb$$

$$Q = \int_0^\infty e^{-At}e^{At} dt$$

Clearly $Q$ is the unique solution of the Lyapunov equation:

$$A^*Q + QA + I = 0 \quad (7.27)$$

and the solution of trace optimization problem (7.9) is obtained as:

$$\min_{b_{u_e} \in \beta} \{\text{ave} E_{\min(0,\infty)}\} = \max_{b_{u_e} \in \beta} \{\text{trace}(W_c(0,\infty))\} = \max_{b_{u_e} \in \beta} \{b^*Qb\} = \lambda_{\text{max}}(Q) \quad (7.28)$$
and

\[
\arg \max_{b \in \beta} \{ \text{trace}(W_c(0, \infty)) \} = b_{n, 1}^{\text{opt}} = V_{\max}(Q) \tag{7.29}
\]

Again \(\beta\) is the set of all possible inputs.

\[\Box\]

### 7.2.1.c. Optimal Input Vector in The Case That “A” Is a Normal Matrix

**Problem:**
Consider the standard LTI SISO system and assume that \(A\) is normal. Find

\[
\max_{b \in \beta} \{ \text{trace}(W_c(0, t_f)) \}, \text{where } \beta \text{ is the set of all possible input vectors}
\]

**Definition (7.1):** In LTI system, if \(A\) is normal, then it satisfies the property in (7.30):

\[
AA^* = A^*A
\tag{7.30}
\]

Note that \(A + A^*\) is always a normal matrix and it is diagonalizable by unitary matrices.

\[\Box\]

**Lemma (7.1):** Consider the LTI system, in which \(A\) is normal. Then the optimal input vector \(b\), which minimizes the average input energy within finite time \(t_f > 0\), is given by the eigenvector, i.e. \(V_{\max}(A + A^*)\), corresponding to the maximum eigenvalue of \(A + A^*\), i.e. \(\lambda_{\max}(A + A^*)\).

Furthermore, maximum value of the trace of the controllability Gramian is equal to:

\[
\frac{e^{\lambda_{\max}t_f} - 1}{\lambda_{\max}} \tag{7.31}
\]

and

\[
\arg \min_{b \in \beta} \{ \text{ave}E_{\min}(0, t_f) \} = b_{n, 1}^{\text{opt}} = V_{\max}(A + A^*) \tag{7.32}
\]

where \(\beta\) represents the set of all possible inputs.

\[\Box\]
**Proof:**

Similar to part (7.2.1.a) by using the linearity property of the trace we define the optimization problem as the maximization of (7.33):

\[
\text{trace}(W_c(0,t_f)) = b^* \left( \int_0^{t_f} e^{A^t} e^{A t} dt \right) b
\]  

(7.33)

Considering the normal assumption for the finite terminal time \( t_f \), we have:

\[
Q = \int_0^{t_f} e^{A^t} e^{A t} dt = \int_0^{t_f} e^{(A^t + A^* t) dt}
\]

(7.34)

and hence:

\[
Q = \int_0^{t_f} P \begin{pmatrix}
\frac{e^{\lambda t}}{\lambda_1} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \frac{e^{\lambda_n t}}{\lambda_n}
\end{pmatrix} P^{-1} dt
\]

(7.35)

Where \( \lambda_i \) defines the \( i \)th eigenvalue of \( A + A^* \), and \( P, P^{-1} \) are orthogonal eigenvector matrices.

Now by calculating the integral we get:

\[
Q = P \begin{pmatrix}
\frac{e^{\lambda t} - 1}{\lambda_1} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \frac{e^{\lambda_n t} - 1}{\lambda_n}
\end{pmatrix} P^{-1}
\]

(7.36)

Then using matrix \( Q \) above, the solution of the optimization problem is obtained as:

\[
\min_{h_{n+1} e^\beta} \{ \text{ave}E_{\min[0,t_f]} \} = \max_{h_{n+1} e^\beta} \{ \text{trace}(W_c(0,t_f)) \} = \max_{h_{n+1} e^\beta} \{ b^* Q b \} = \frac{e^{\lambda t} - 1}{\lambda_1}
\]

(7.37)

and

\[
\arg \max_{h_{n+1} e^\beta} \{ \text{trace}(W_c(0,t_f)) \} = h_{n+1}^{e^\beta} = V_{\max}(Q) = V_{\max}(A + A^*)
\]

(7.38)

where \( \beta \) denotes the set of all possible inputs, \( \lambda_1 \) is the maximum eigenvalue of matrix \( A + A^* \), and \( V_{\max} \) is its corresponding eigenvector.
7.2.1.d. Optimal Input Vector in The Case That “A” Is Stable and Normal

Lemma (7.2): (Karcanias, et al., 2010) if $A \in \mathbb{C}^{n \times n}$ is asymptotically stable and normal, then it is also strongly stable, which is defined by the condition:

$$\lambda_i(A + A^*) < 0, i = 1, \ldots, n.$$  

Proof:

Using the real Schur form of matrix $A$ (Bai & Demmelt, 1992):

$$U^* A^* U = \text{block} - \text{diag} \left\{ \lambda_1, \ldots, \lambda_k, \ldots \begin{bmatrix} \sigma_i & -\omega_i \\ \omega_i & \sigma_i \end{bmatrix} \ldots \right\}$$  (7.39)

where $\lambda_i$ are the real eigenvalues and $\sigma_i, \pm j\omega_i$ the complex eigenvalues of $A$.

Thus:

$$U^* (A + A^*) U = \text{block} - \text{diag} \left\{ 2\lambda_1, \ldots, 2\lambda_k, \ldots \begin{bmatrix} 2\sigma_i & 0 \\ 0 & 2\sigma_i \end{bmatrix} \ldots \right\}$$

Then, by introducing $\hat{A}$:

$$\frac{A + A^*}{2} = \hat{A}$$

we have:

$$U^* \hat{A} U = \text{block} - \text{diag} \left\{ \lambda_1, \ldots, \lambda_k, \ldots, \sigma_i, \sigma_i, \ldots \right\}$$  (7.40)

and from asymptotic stability $\lambda_i < 0$, $\sigma_i < 0$ and this establishes the negative-definiteness of $\hat{A}$ and the Lemma is proved.
The above result establishes a sufficient condition for strong stability in terms of the property of normality.

According to (Ricardo, 2009), the spectral Theorem for normal matrices states:

**Theorem (7.3):** For \( A \in \mathbb{C}^{n \times n} \), there exists a unitary \( n \times n \) matrix \( P \in \mathbb{R}^{n \times n} \) such that \( P^*AP = D \) (\( D \) is a diagonal matrix) if and only if \( A \) is normal.

**Proof:**

By the Schur decomposition, we can write any matrix as \( A = P^*TP \), where \( P \) is unitary and \( T \) is upper-triangular. If \( A \) is normal clearly \( T \) is normal as well, since:

\[
AA^* = A^*A \Rightarrow P^*TT^*P = P^*T^*TP \Rightarrow TT^* = T^*T
\]

which implies that \( T \) must be diagonal. The converse is obvious.

Now, by combining the results of the last two sections, we can examine the problem of optimum input selection in the case that the LTI system is both normal and stable. We have the following results:

**Proposition (7.3):** Consider an LTI single input system with \( A \) is a normal and Hurwitz. Then, the optimal input vector \( b \), which maximizes the trace of the controllability Gramian, i.e. minimizes the average of the input energy, is the eigenvector corresponding to the largest eigenvalue of \( A + A^* \) and the maximum value of the trace is \( -\lambda_1^{-1} \), where \( \lambda_1 (< 0) \) is the maximum eigenvalue of \( A + A^* \). Equivalently:

\[
\min_{\|H\|_\infty} \{\text{ave}E_{\text{min}(0,\infty)}\} = \frac{-1}{\lambda_1} \quad (7.42)
\]

and

\[
\arg \min_{\|H\|_\infty} \{\text{ave}E_{\text{min}(0,\infty)}\} = b_{\min}^* V_{\max}(A + A^*) \quad (7.43)
\]

where \( \beta \) denotes the set of all possible input vectors and \( V_{\max} \) (i.e. optimal input \( b \)) is the normalised eigenvector associated with \( \lambda_1 \).

**Proof:**

Due to the assumption of stability for the system, we can increase the terminal time to infinity \( t_f \to \infty \), and similar to the last part we define matrix \( Q \) as:
\[ Q = \int_0^\infty e^{At} e^{At^*} dt = \int_0^\infty e^{(A^* + A)t^*} dt \]  \hspace{1cm} (7.44)

Note: the identity \( e^{At} e^{At^*} = e^{(A^* + A)t^*} \) follows directly from the assumed normal property of \( A \).

Then, regarding Theorem (7.3), \( Q \) in (7.44) can be written as (7.45):

\[
Q = P \begin{pmatrix} -1 & \cdots & 0 \\ \lambda_1 & \ddots & \vdots \\ 0 & \cdots & -1 \end{pmatrix} P^* \hspace{1cm} (7.45)
\]

Thus, by combination the properties of stability and normality for the LTI system, the optimal input vector minimizing the average of input energy is achieved by:

\[
\min_{b \in \beta} \{ \text{ave} E_{\min[0, \infty]} \} = \max_{b \in \beta} \{ \text{trace}(W_c(0, \infty)) \} = \max_{b \in \beta} \{ b^* Q b \} = \lambda_{\max} (Q) = -\frac{1}{\lambda_1} \hspace{1cm} (7.46)
\]

and

\[
\arg \max_{b \in \beta} \{ \text{trace}(W_c(0, \infty)) \} = b^{opt}_{n+1} = V_{\max} (Q) = V_{\max} (A + A^*) \hspace{1cm} (7.47)
\]

where \( \beta \) is the set of all possible inputs, \( \lambda_1 (\leq 0) \) is the largest eigenvalue of \( A + A^* \) and \( V_{\max} \) is its corresponding eigenvector.

7.2.2 Multi-Input Matrix Structure Selection

In this part, we assume that input \( B \in \mathbb{R}^{n \times m} \), is a \( n \times m \) matrix:

\[
\begin{cases} 
\dot{x} = A_{n \times n} x + B_{n \times m} u \\
y = C_{1 \times n} x + D_{1 \times m} u 
\end{cases} \hspace{1cm} (7.48)
\]

The aim is to select the proper input matrix \( B = [b_1, b_2, \ldots, b_m], b_i \in \mathbb{R}^{n \times 1} \) such that the trace of controllability Gramian matrix \( W_c(0, t_f) \) is maximized. Again, this is equivalent to the minimization of the average of minimum input energy \( E_{\min} \), i.e.

\[
\min_{b \in \beta} \{ \text{ave} E_{\min[0, t_f]} \} = \max_{b \in \beta} \{ \text{trace}(W_c(0, t_f)) \} \hspace{1cm} (7.49)
\]

Now \( \beta \) is the set of all possible input matrices of appropriate dimensions.

In this part, we also investigate the problem subject to various assumptions defined on input matrix \( B \). In each case, we suggest the selection of the proper set, over which the average energy minimization (i.e. maximization of the trace of the controllability Gramian) is carried out.
7.2.2.a. Input Matrix Selection Within a Finite Terminal Time Part

In this part, we find the optimal input structure for the general case, where \( B = [b_1, b_2, \ldots, b_m] \), \( b_i \in \mathbb{R}^{n \times 1} \) is the input matrix such that \( \|B\|_F \leq 1 \), and matrix \( Q = \int_0^T e^{A^T} e^{A} dt \) has \( n \) eigenvalues ordered as \( \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > 0 \). It is assumed for simplicity that \( \lambda_i \) is distinct.

Here we are looking for the input matrix \( B \) that maximizes the trace of the controllability Gramian, \( W_c(0, t_f) \), i.e. minimizes the average input energy.

**Theorem (7.4):** Consider the system in (7.48) and assume that \( Q = \int_0^T e^{A^T} e^{A} dt \) has a maximum eigenvalue of multiplicity one. Then, all optimal input \( B = [b_1, b_2, \ldots, b_m] \), \( b_i \in \mathbb{R}^{n \times 1} \), such that \( \|B\|_F \leq 1 \), which minimize the average input energy, i.e. maximize the trace of the controllability Gramian \( W_c(0, t_f) \), can be defined as:

\[
\begin{align*}
    b_1 &= \alpha_1 V_1 \\
    b_2 &= \alpha_2 V_1 \\
    \vdots \\
    b_m &= \alpha_m V_1
\end{align*}
\]  

(7.50)

where \( V_1 \) is the normalized eigenvector corresponding to the largest eigenvalue of \( Q \) and the \( \alpha_i \)'s \( i = 1, \ldots, m \) are scalars satisfying the condition \( \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_m^2 = 1 \).

\[\square\]

**Proof:**

We know that:

\[
\text{trace}\{W_c(0, t_f)\} = \text{trace}\{\int_0^T e^{A^T} B B^* e^{A} dt\} = \text{trace}\{B^* \int_0^T e^{A^T} e^{A} dt B\}
\]

(7.51)

Defined \( Q \) as:

\[
Q = \int_0^T e^{A^T} e^{A} dt
\]

(7.52)
Then by substituting (7.52) into (7.51) we obtain:

$$\text{trace}\{W_c(0, t_f)\} = \text{trace}\{B^*QB\}$$  \hspace{1cm} (7.53)

Equivalently:

$$\text{trace}\{W_c(0, t_f)\} = \text{trace}\left\{Q \begin{pmatrix} b_1^* & b_2^* & \cdots & b_m^* \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \cdots & b_m \end{pmatrix} \right\}$$  \hspace{1cm} (7.54)

or:

$$\text{trace}\{W_c(0, t_f)\} = \text{trace}\left\{Q \begin{pmatrix} b_{11}^* & b_{12}^* & \cdots & b_{1m}^* \\ \vdots & \ddots & \vdots & \vdots \\ b_{m1}^* & b_{m2}^* & \cdots & b_{mm}^* \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mm} \end{pmatrix} \right\}$$  \hspace{1cm} (7.55)

and hence:

$$\text{trace}\{W_c(0, t_f)\} = \sum_{i=1}^{m} b_i^*Qb_i$$  \hspace{1cm} (7.56)

Then according to the Rayleigh quotient eigenvalue problem (Golub & Loan, 1983), (Watkins, 1982), (Uper, July 2002), the solution for the maximization of the trace of controllability Gramian is achieved by selecting input vectors $b_i, i = 1, \ldots, m$, along the eigenvector corresponding to the largest eigenvalue $\lambda$ of $Q$:

$$\max_{\|B\|_{\text{F}} \leq 1} \{\text{trace}(W_c(0, t_f))\} = \max_{\|B\|_{\text{F}} \leq 1} \{\text{trace}(B^*QB)\} = \max_{\|B\|_{\text{F}} \leq 1} \sum_{i=1}^{m} b_i^*Qb_i$$  \hspace{1cm} (7.57)

$$= \max_{\|B\|_{\text{F}} \leq 1} \left\{ b_1^*b_1 + b_2^*b_2 + \cdots + b_m^*b_m \right\}$$

so that:

$$b_1 = \alpha_1 V_1, b_2 = \alpha_2 V_1, \ldots, b_m = \alpha_m V_1$$

$$\|B\|_{\text{F}} = 1 \Rightarrow \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_m^2 = 1$$  \hspace{1cm} (7.58)

where $\beta$ is the set of all possible input matrices, $V_1$ is the eigenvector corresponding to $\lambda_1$, the maximum eigenvalue of $Q$.

Note that optimality over the set $\|B\|_{\text{F}} \leq 1$ occurs on the boundary, i.e. $\|B\|_{\text{F}} = 1$. 

$\square$
Then all columns of the optimal solution set are aligned along the direction of $V_i$ while we can distribute their lengths arbitrarily subject to the condition $\sum_{i=1}^{m} \|b_i\|^2 = 1$. This generalizes the solution of the single-input problem. In the next subsection, we investigate the optimal input structure $B \in \mathbb{R}^{n \times m}$ where the largest eigenvalue of $Q$ has multiplicity $r \leq n$.

### 7.2.2.a. Input Matrix Selection Within a Finite Terminal Time Part II

**Problem:**

Consider the LTI MIMO system (7.48) and suppose that $Q = \int_0^T e^{A_i} e^{A_i^T} dt$ has maximum eigenvalue $\lambda_i$ of multiplicity $r \leq n$. Find

$$\max_{B \in \beta} \{ \text{trace}(W_c(0, t_j)) \} ,$$

where $\beta$ is the set of all possible input matrices $B \in \mathbb{R}^{n \times m}$ such that $\|B\|_F \leq 1$, and matrix $Q = \int_0^T e^{A_i} e^{A_i^T} dt$ has $n$ eigenvalues ordered as

$$\lambda_1 = \lambda_2 = \cdots \lambda_i > \lambda_{r+1} \geq \cdots \geq \lambda_n > 0 ,$$

where $\lambda_i$ is the maximum eigenvalue of multiplicity $r \leq n$.

**Lemma (7.3):** For the LTI MIMO system (7.48), the columns of the optimal input matrix $B = [b_1, b_2, \ldots, b_m] \in \mathbb{R}^{n \times 1}$, which maximizes the trace of controllability Gramian $W_c(0, t_j)$, are equal to:

$$b_i = \sum_{k=1}^{r} \alpha_{ik} V_{ki}, i = 1, \ldots, m$$

(7.59)

where $V_{ki}, k = 1, \ldots, r$ are independent (orthonormal) eigenvectors corresponding to the maximum eigenvalue of $Q = \int_0^T e^{A_i} e^{A_i^T} dt$ and the $\alpha_i$'s $i = 1, \ldots, m$ are the constant coefficients that satisfy the condition $\sum_{i=1}^{m} \sum_{k=1}^{r} \alpha_{ik}^2 = 1$. In this case $\text{rank}(B) = \min (r, m)$, and the maximum value for the trace of $W_c(0, t_j)$ is equal to the largest eigenvalue of $Q$.

The dimension of optimal input matrix depends on the maximum eigenvalue of $Q$, assumed of multiplicity $r$. If $m \leq r$, then $B$ is full rank and the optimal solution would be $m$. 

148
dimensional input matrix. Otherwise, \( m - r \) of input vectors \( b_i, i = 1, \ldots, m \), would be dependent.

\[ \square \]

**Proof:**

Since the proof of this Lemma is quite similar to the proof of Theorem (7.4), then it is neglected.

\[ \square \]

**Remark (7.1):** Suppose that \( m \leq r \leq n \). Thus, among all optimal solutions with rank \( B = m \) the best (in the sense that the condition number of \( B \) is equal to one) is:

\[
B = [b_1, \ldots, b_m] = \frac{1}{\sqrt{m}} [V_1, V_2, \ldots, V_m]
\]

(7.60)

where \( V_1, V_2, \ldots, V_m \) are \( m \) orthonormal eigenvectors corresponding to the maximum eigenvalue of \( Q \).

\[ \square \]

### 7.2.2.a. Input Matrix Selection Within a Finite Terminal Time Part III

**Problem:**

Consider the LTI MIMO system (7.48) and suppose that \( Q = \int_0^t e^{A^T} e^{A} dt \) has maximum eigenvalue \( \lambda_i \) of multiplicity \( r < m \leq n \). Find

\[
\max_{B \in \beta, \text{rank}(B) = m} \{\text{trace}(W_c(0, t_{f}))\}, \quad \text{where } \beta \text{ is the set of all possible input matrices}
\]

Here we complete the solution of finding the optimal matrix \( B \) over a finite terminal time \( t_{f} \), by imposing the constraint on the rank of input matrix \( B \) to guarantee a full rank multiple input solution.

In other words, in this section we are looking for a full rank input matrix \( B = [b_1, b_2, \ldots, b_m] \), \( b_i \in \mathbb{R}^{n \times 1}, \|B\|_p \leq 1 \), that maximizes the trace of the controllability Gramian \( W_c(0, t_{f}) \), i.e. minimizes the average of input energy, where \( r \) is the order of the maximum eigenvalue of \( Q = \int_0^t e^{A^T} e^{A} dt \) is less than rank \( (B) = m \), i.e. \( r < m \leq n \).
Definition (7.2): Inner product of two vectors $a, b$ is defined as:

$$\langle a, b \rangle = \|a\| \|b\| \cos \theta$$  \hspace{1cm} (7.61)

From the Cauchy–Schwarz inequality it follows immediately that:

$$0 \leq \|\langle a, b \rangle\| \leq \|a\| \|b\|$$  \hspace{1cm} (7.62)

Next, we seek to find an optimal input matrix $B_{n \times m} = [b_1, b_2, \ldots, b_m]$ of full rank, which maximizes the trace of the controllability Gramian $W_t(0, t_f)$ subject to constraint $\|B\|_F \leq 1$.

For simplicity, we assume that $r < rank(b) = m$, where $r$ is the multiplicity of the largest eigenvalue of $Q = \int_0^t e^{A^T} e^{At} dt$. To define a meaningful problem, we need to redefine the constraint region as described below.

In effect, this replaces the full rank constraint on $B$ by a numerical-rank constraint of a certain tolerance. Specifically, the optimal solution requires that the first $r$ input vectors of matrix $B$ are aligned in the directions of the $r$ independent eigenvectors corresponding to the maximum eigenvalue of $Q$. Let $E$ be the spectral subspace generated by these eigenvectors. A numerical rank constraint on $B$ can be imposed by constraining the input vectors $b_i, i = r+1, \ldots, m$ to lie in directions whose angles with any vector in the spectral subspace $E$ exceed some minimum value $0 < \varepsilon$.

Thus, the problem becomes:

$$\max_{\alpha_j, j = 1, \ldots, j = r} \sum_{i = 1}^{r(m-r)} \beta_i \lambda_j$$  \hspace{1cm} (7.63)

where:

$$\beta_i \lambda_j$$
\[ \beta_i = \sum_{j=1}^{m} \alpha_i^j, i = 1, \ldots, 1 + (m - r) \]

\[ \sum_{j=1}^{r} (\alpha_i^j)^2 + \sum_{j=r+1}^{m} (\alpha_i^j)^2 = 1 \quad (7.64) \]

\[ 0 < \varepsilon \leq \cos \theta_i = \left( \frac{b_i \cdot b_j}{\|b_i\| \|b_j\|} \right), i = 1, \ldots, (m-r)((m-r)+1)/2, k, l = r, r+1, \ldots, m \]

Then, the optimal solution is obtained by selecting:

\[ b_1 = \alpha_1^1 V_1 \]
\[ \vdots \]
\[ b_r = \alpha_r^r V_r \]
\[ b_{r+1} = \sum_{j=1}^{r} \alpha_{r+1}^{r+1} V_j + \alpha_{r+1}^{r+1} V_{r+1} \quad (7.65) \]
\[ b_{r+2} = \sum_{j=1}^{r} \alpha_{r+2}^{r+2} V_j + \alpha_{r+2}^{r+2} V_{r+2} \]
\[ \vdots \]
\[ b_m = \sum_{j=1}^{r} \alpha_m^m V_j + \alpha_m^m V_{r+1} + \alpha_m^m V_{r+2} + \cdots + \alpha_m^m V_{(m-r)+1} \]

Here \( V_i, i = 1, \ldots, 1 + (m-r) \), is the normalized eigenvector corresponding to the \( i \)th largest eigenvalue of \( Q \), \( \lambda_i \) denotes the \( i \)th eigenvalue of \( Q \), \( i = 1, \ldots, 1 + (m-r) \), i.e. \( \lambda_1 = \cdots = \lambda_r > \lambda_{r+1} \geq \cdots \geq \lambda_m > 0 \).

Note that \( \alpha_i^j, i = 1, \ldots, 1 + (m-r), j = 1, \ldots, m \) are constant coefficients, which satisfy the condition \( \sum_{j=1}^{r} (\alpha_i^j)^2 + \sum_{j=r+1}^{m} (\alpha_i^j)^2 = 1 \). For notational simplicity coefficients \( \beta_i = \sum_{j=1}^{m} \alpha_i^j, i = 1, \ldots, 1 + (m-r) \) have also been defined.

Note further that in the special case where \( r = 1 \), (i.e. \( Q \) has a non-repeated maximum eigenvalue), the optimal input matrix is defined as:

\[ B = (\alpha_1^1 V_1, \alpha_2^2 V_2, \alpha_1^3 V_1 + \alpha_2^3 V_2, \alpha_1^4 V_1 + \alpha_2^4 V_2 + \alpha_3^4 V_3, \ldots, \alpha_1^m V_1 + \alpha_2^m V_2 + \alpha_3^m V_3 + \cdots + \alpha_m^m V_m) \quad (7.66) \]

where \( V_1, V_2, \ldots, V_m \) are the normalized eigenvectors corresponding to the \( m \) largest eigenvalues of \( Q = \int_0^t e^{A^T} e^{A} dt \) respectively. We can now state and prove the following Theorem.
**Theorem (7.5):** Under the constraint of matrix $B \in \beta$ given in equation (7.65) above, the trace of the controllability Gramian, $W_c(0,t_f)$, is maximized as:

$$\max_{b_i \in \beta} \{ \text{trace}(W_c(0,t_f)) \} = \sum_{i=1}^{\min(m-r)} b_i \lambda_i$$  \hspace{1cm} (7.67)

**Proof:**

It is assumed without loss of generality that the eigenvalues of $Q$ are distinct. Otherwise, the $b_i$'s could be selected as orthonormal bases of the corresponding spectral subspace spanned by the clusters of the repeated eigenvalues.

Similar to single input section, first we define the trace of controllability Gramian as:

$$\text{trace}\{W_c(0,t_f)\} = \text{trace}\{ \int_0^t e^{At}BB^*e^{A^T}dt \} = \text{trace}\left\{ B^T \int_0^t e^{A^T}e^{At}dB \right\}$$  \hspace{1cm} (7.68)

Again define:

$$Q = \int_0^t e^{A^T}e^{At}dt$$  \hspace{1cm} (7.69)

Then:

$$\text{trace}\{W_c(0,t_f)\} = \text{trace}\{B^TQB\} = \text{trace}(\begin{pmatrix} b_1^* \\
_2^* \\
_m^* \end{pmatrix} Q(b_1 \ b_2 \ \cdots \ b_m))$$  \hspace{1cm} (7.70)

or equivalently:

$$\text{trace}\{W_c(0,t_f)\} = \text{trace}(\begin{pmatrix} b_1^* Q_{11} \ b_1^* Q_{12} \ \cdots \ \times \ \cdots \\
_{m1}^* Q_{m1} \ b_{m1} \ \cdots \ \times \\
_{m1}^* Q_{m2} \ b_{m2} \ \cdots \ \times \ \cdots \\
_{m1}^* Q_{mn} \ b_{mn} \end{pmatrix})$$  \hspace{1cm} (7.71)

which implies that:

$$\text{trace}\{W_c(0,t_f)\} = \sum_{i=1}^m b_i^* Q b_i$$  \hspace{1cm} (7.72)

Then considering the Rayleigh quotient eigenvalue problem (Golub & Loan, 1983), (Watkins, 1982), (Uper, July 2002), the trace of the controllability Gramian is maximized as follows:
\[
\max_{\|b\| \leq 1, \|b\| = 1} \left\{ \text{trace}(W_t(0,t_J)) \right\} = \max_{\|b\| = 1} \left\{ b_1^* b_1 + b_2^* b_2 + \cdots + b_m^* b_m \right\} = \max_{\|b\| = 1} \sum_{i=1}^{m} b_i^* Q b_i \\
(7.73)
\]

It follows using Theorem (7.4) that one optimal solution can be obtained by choosing \( b_1 \) in the direction of the normalized eigenvector corresponding to the maximum eigenvalue of \( Q \), i.e. \( V_1, b_2 = \alpha_1^2 b_1 + \alpha_2^2 V_2, \) where \( V_2 \) is the normalized eigenvector corresponding to the second largest eigenvalue of \( Q \).

Using a similar procedure, we can define \( b_3 = \alpha_1^3 b_1 + \alpha_2^3 b_2 + \alpha_3^3 V_3, \) where \( V_3 \) is the normalized eigenvector corresponding to the third largest eigenvalue of \( Q \). In general:

\[
\begin{align*}
  b_1 &= \alpha_1^2 V_1 \\
  b_2 &= \alpha_1^2 V_1 + \alpha_2^2 V_2 \\
  &\vdots \\
  b_m &= \alpha_1^m V_1 + \alpha_2^m V_2 + \cdots + \alpha_m^m V_m \\
\end{align*}
(7.74)
\]

Nothing that:

\[
\|B\|_F = 1 \Rightarrow \sum_{i=1, \ldots, j} (\alpha_i')^2 = 1 \\
(7.75)
\]

and the Cauchy–Schwarz inequality:

\[
0 \leq \langle b_1, b_2 \rangle \leq \|b_1\| \|b_2\| \\
(7.76)
\]

Show the optimality of (7.65).

\[\square\]

A geometric interpretation of Theorem (7.5) is given below.

Let \( b_1 \) be defined in the direction of the eigenvector \( V_1 \). To satisfy the numerical rank constraint on input matrix \( B \) and minimize the angle between \( b_1 \) and \( b_2 \), \( b_2 \) must be located on the boundary of the cone shown in Figure (7.3) below. Constraining \( b_2 \) to lie outside this cone ensures that the angle between \( b_1 \) and \( b_2 \) is always greater than the required tolerance \( \epsilon \).
Similarly, with $b_1$ and $b_2$ fixed, $b_i$ is selected to lie on the intersection of the two cones centered at vectors $b_1$, $b_2$ respectively as shown in Figure (7.4).

This procedure could be continued to construct all $b_i, i = 1, \ldots, m$.

An alternative approach for defining the sub-optimal solution is based on the following Lemma:
**Lemma (7.4):** For any matrix $A_{m \times n}$, $m \leq n$, $\text{rank}(A) = m$, if and only if the minimum singular value of $A$ is greater than zero, i.e. $\sigma_m > 0$.

The proof is given in the Appendix. Using the above Lemma, we have:

$$\text{rank}(B) = m \Rightarrow \sigma_m(B) \geq \varepsilon > 0 \quad (7.77)$$

Then the requirement that the numerical rank of $B$ is $m$ can be expressed as $\sigma_m(B) \geq \varepsilon$ for a sufficiently small number $\varepsilon > 0$ then the corresponding optimization problem takes the form:

$$\max_{\alpha_j^i, \beta_i, j=1,...,m} \left\{ \sum_{j=1}^{1+(m-r)} \beta_i \lambda_i \right\} \quad (7.78)$$

Subject to:

$$0 < \varepsilon \leq \sigma_m(B)$$

Here $\alpha_j^i, i = 1, \ldots, 1 + (m - r), j = 1, \ldots, m$ are constant coefficients, which satisfy condition

$$\sum_{j=1}^{r} (\alpha_j^i)^2 + \sum_{j=1}^{m} (\alpha_j^i)^2 = 1$$

and coefficients $\beta_i$ are defined as:

$$\beta_i = \sum_{j=1}^{m} \alpha_j^i, i = 1, \ldots, 1 + (m - r)$$

Then, the optimal solution is obtained by selecting:

$$b_1 = \alpha_1^i V_i$$
$$\vdots$$
$$b_r = \alpha_r^i V_r$$

$$b_{r+1} = \sum_{j=1}^{r} \alpha_j^{r+1} V_j + \alpha_{r+1}^{r+1} V_{r+1} \quad (7.79)$$

$$b_{r+2} = \sum_{j=1}^{r} \alpha_j^{r+2} V_j + \alpha_{r+1}^{r+2} V_{r+1} + \alpha_{r+2}^{r+2} V_{r+2}$$
$$\vdots$$

$$b_m = \sum_{j=1}^{r} \alpha_j^{m} V_j + \alpha_{r+1}^{m} V_{r+1} + \alpha_{r+2}^{m} V_{r+2} + \cdots + \alpha_{(m-r)+1}^{m} V_{(m-r)+1}$$

where $V_i, i = 1, \ldots, 1 + (m - r)$, is the normalized eigenvector corresponding to $\lambda_i$, the $i$th largest eigenvalue of $Q$, i.e. $\lambda_1 = \cdots = \lambda_r > \lambda_{r+1} \geq \cdots \geq \lambda_m > 0$.

An alternative formulation of the optimization problem relies on the determinant of the Gram matrix of $B$. 

155
**Definition (7.3):** If \( A \in \mathbb{C}^{m \times n} \), then the square matrix \( G = A^* A \in \mathbb{C}^{n \times n} \) is known as the Gram matrix of \( A \) and \( |G| \) as the Gram determinant of \( A \) (Mirsky, 2012).

**Theorem (7.6):** (Mirsky, 2012) If \( G \) is the Gram matrix of \( A \), then \( \text{rank}(G) = \text{rank}(A) \).

The proof is given in the Appendix.

**Proposition (7.4):** (Peter Lancaster, Miron Tismenetsky, 1985) Let \( \mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m], \mathbf{b}_i \in \mathbb{R}^{n \times 1}, \ m \leq n \). Then, \( \text{rank}(\mathbf{B}) = m \) if and only if the Gram matrix of \( \mathbf{B} \) is non-singular, i.e. \( |G| \neq 0 \).

The proof of the above Proposition, which is given by (Peter Lancaster, Miron Tismenetsky, 1985) can be found in the Appendix, however, an easier proof might be using Theorem (7.6), it states that \( \text{rank}(G) = \text{rank}(B) \) then if \( \text{rank}(B) = m \), it implies that \( \text{rank}(G) = m \), and \( G \) is a symmetric positive definite matrix, which means Gram determinant of \( B \) is nonzero, i.e. \( \prod_{i=1}^{m} \lambda_i \neq 0 \) where \( \lambda_i \)’s are the eigenvalues of \( G \). Conversely, if \( G \) is non-singular, with regard to Definition (7.3) \( G \) is a symmetric positive definite matrix, then \( \text{rank}(G) = m \) thus using the Theorem above \( \text{rank}(B) = m \), and the proof of Proposition is completed.

Note that if \( \mathbf{B} \in \mathbb{R}^{n \times m} \), then \( G_{ij} = \mathbf{b}_i^T \mathbf{b}_j \). Thus, if \( \theta_{ij} \) is the angle between vectors \( \mathbf{b}_i \) and \( \mathbf{b}_j \) and \( \alpha, \beta \in \mathbb{R} \) such that \( \alpha \leq \theta_{ij} \leq \beta \), then it follows from the monotonicity of function \( \cos^{-1}(\cdot) \) that:

\[
\mathbf{b}_i \mathbf{b}_j \cos \beta \leq G_{ij} \leq \mathbf{b}_i \mathbf{b}_j \cos \alpha
\]

Which is a pair of linear inequalities on the elements of \( G \). Thus, we can always maximize or minimize angle \( \theta_{ij} \) by maximizing or minimizing \( G_{ij} \) (Stephen P. Boyd, Lieven Vandenberghe, 2004).

Using \( |G| \) as a measure f the numerical rank of \( B \), the optimization problem can be formulated as:

\[
\max_{\sum_{j=1}^{m} \beta_j \lambda_j} \sum_{i=1}^{1+(m-r)} \alpha_i
\]

Such that:
\[
\beta_i = \sum_{j=1}^{m} \alpha_{ij}, \quad i = 1, \ldots, 1 + (m - r)
\]
\[
\sum_{j=1}^{r} (\alpha_{ij})^2 + \sum_{j=r+1}^{m} (\alpha_{ij})^2 = 1
\]
\[
0 < \varepsilon \leq |G|
\]  

(7.82)

where:

\[
b_1 = \alpha_1^i V_i \\
\vdots \\
b_r = \alpha_r^i V_r \\
b_{r+1} = \sum_{i=1}^{r} \alpha_{i}^{r+1} V_i + \alpha_{r+1}^{r+1} V_{r+1} \\
b_{r+2} = \sum_{i=1}^{r} \alpha_{i}^{r+2} V_i + \alpha_{r+1}^{r+2} V_{r+1} + \alpha_{r+2}^{r+2} V_{r+2} \\
\vdots \\
b_m = \sum_{i=1}^{r} \alpha_{i}^{m} V_i + \alpha_{r+1}^{m} V_{r+1} + \alpha_{r+2}^{m} V_{r+2} + \cdots + \alpha_{(m-r)+1}^{m} V_{(m-r)+1}
\]  

(7.83)

To summarize the results of this sub-section, the optimization of the trace of the controllability Gramian \(W_c(0, t_f)\) under the constraint that \(\|B\|_F \leq 1\) and the input matrix \(B = [b_1, b_2, \ldots, b_m], b_i \in \mathbb{R}^{n \times 1}\) has full (numerical) rank, i.e. \(\text{rank}(B) = m\), is the solution of one of the suboptimal optimization problems proposed in (7.65), (7.79) or (7.83).

**Example (7.2):** Consider the problem of multiple input structure selection of a heating furnace (Yadykin, 2011), which can be represented by the following equations:

\[
\dot{x} = \begin{bmatrix} -0.5 & 0 \\ 0 & -1 \end{bmatrix} x + Bu
\]
\[
y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x
\]

subject to the constraints: \(\text{rank}(B) = 2\) and \(B = [b_1, b_2], b_i \in \mathbb{R}^{2 \times 1}, i = 1, 2,\)
\[
\|B\|_F^2 = \|b_1\|^2 + \|b_2\|^2 = 1.\]

Assume the terminal time \(t_f = 1\).

According to the results stated above the optimal solution can be obtained by minimizing the positive angle between the two input vectors, while choosing the first input vector along the eigenvector corresponding to the maximum eigenvalue of \(Q = \int_0^{t_f} e^{A(t)} e^{A^T} dt.\)

In this case:
\[ Q = \int_0^t e^{A't} e^{A} dt = \begin{bmatrix} 0.6321 & 0 \\ 0 & 0.4323 \end{bmatrix} \]

and

\[ \lambda_1 = 0.6321 \Rightarrow V_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \]

\[ \lambda_2 = 0.4323 \Rightarrow V_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \]

Thus, the optimization problem can be stated as:

\[
\max_{\alpha_1, \alpha_2, \alpha_3} \{0.6321(\alpha_1^2 + (\alpha_2^2) + 0.4323(\alpha_3^2) \}
\]

Such that:

\[
b_1 = \alpha_1^1 \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \alpha_1^1 \neq 0
\]

\[
b_2 = \alpha_2^2 b_1 + \alpha_2^3 b_2, \alpha_2^1, \alpha_2^2 \neq 0
\]

\[
0 < \varepsilon \leq \cos^{-1} \left( \frac{\alpha_2^3}{\sqrt{\alpha_2^2 + \alpha_3^2}} \right)
\]

\[
(\alpha_1^1)^2 + (\alpha_2^1)^2 + (\alpha_2^2)^2 = 1
\]

(7.84)

Taking the minimum angle \(\varepsilon\) between \(b_1\) and \(b_2\) as \(\varepsilon = 0.0367 \pi\), then by solving optimization problem in (7.84), the suboptimal solution for \(B\) is obtained as:

\[
\varepsilon = \cos^{-1} (0.9993) = 0.0367 \pi \Rightarrow \alpha_1^1 = 0.3134, \alpha_2^1 = -0.9490, \alpha_2^2 = 0.0349
\]

and hence:

\[
B = \begin{bmatrix} 0.3134 & -0.9490 \\ 0 & 0.0349 \end{bmatrix}
\]

In this case the maximum trace of controllability Gramian \(W_c(0,1)\) is:

\[
\max \{\text{trace}(W_c(0,1))\} = 0.6321((\alpha_1^1)^2 + (\alpha_2^1)^2) + 0.4323(\alpha_2^2)^2
\]

\[
= 0.6321(0.3134^2 + 0.9490^2) + 0.4323(0.0349^2) = 0.6319
\]

Similarly, if we solve the optimization problem (7.84) for a minimum angle \(\varepsilon = 0.0736 \pi\), the input matrix will be achieved as:

\[
B = \begin{bmatrix} 0.1432 & 0.9870 \\ 0 & -0.0728 \end{bmatrix}
\]

for which the value of trace of \(W_c(0,1)\) is 0.6311.
In Figures (7.5) the validity of the optimization results is investigated using a brute force method. The values of $b_1, b_2$ are continuously changed between -1 and 1 (with a step of 0.2), in this case, the maximum value of trace of controllability Gramian $W_c(0,1)$ is obtained as 0.6229, where the suboptimal input matrix is $B = \begin{bmatrix} 0.8 & 0.56 \\ 0.08 & 0.2 \end{bmatrix}$ and the angle between $b_1, b_2$ is $\varepsilon = 0.2435 \pi$. The corresponding results obtained by reducing the step to 0.1 are shown in Figure (7.6). In this case, the maximum value of trace of controllability Gramian $W_c(0,1)$ is obtained as 0.6229 and the suboptimal solution is $B = \begin{bmatrix} -0.64 & -0.76 \\ 0.08 & 0.08 \end{bmatrix}$, here the angle between $b_1, b_2$ is $\varepsilon = 0.0195 \pi$. This confirms that as the accuracy increases $b_1$ tends to align itself with the eigenvector corresponding to the maximum eigenvalue of $Q$, $V_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, and the angle between $b_1, b_2$, tends to zero.

Figure (7.7) illustrates changes in the value of trace $\{ W_c(0,1) \}$, when $b_1$ is selected along the eigenvector corresponding to the maximum eigenvalue of $Q$ and $b_2$ is formed as a linear combination of first and second eigenvectors of $Q$, i.e. $b_1 = \alpha_1^1 V_1, b_2 = \alpha_1^2 V_1 + \alpha_2^2 V_2$. In the first subplot, the values of trace are demonstrated as a function of $\cos(\varepsilon)$, while the second subplot displays the values of trace for different values of $\alpha_1^2, \alpha_2^2$. This figure also confirms that we can reach the suboptimal solution by decreasing the angle between input vectors $b_1, b_2$, i.e. increasing $\cos(\varepsilon)$. The maximum value of trace in Figure (7.7) is equal to 0.6320 and it is reached where input matrix is $B = \begin{bmatrix} 0.7680 & -0.64 \\ 0 & 0.024 \end{bmatrix}$ and $\varepsilon = 0.0374 \pi$, comparing this results with the results in plot (7.6) it can be easily inferred that though the optimal angle between input vectors is greater in Figure (7.7), however since in Figure (7.7) $b_1$ is selected align the eigenvector corresponding to the maximum eigenvalue of $Q$ the optimal value of trace (0.6320) is greater than optimal result (0.6296), shown in Figure (7.6).
Figure (7.5): $\text{trace}\{W_c(0,1)\}$ according to the variable input matrix $B$ in the interval $[-1:0.2:1]^2$

Figure (7.6): $\text{trace}\{W_c(0,1)\}$ respect to the variable input matrix $B$ in the interval $[-1:0.1:1]^2$
Figure (7.7): \( \text{trace}[W_c(0,1)] \) based on the angle \( \varepsilon \) between \( b_1, b_2 \) (top), and \( \text{trace}[W_c(0,1)] \) as a function of \( \alpha_1^2, \alpha_2^2 \) (bottom).

Next, consider optimization method (7.79), in which numerical rank is imposed via a lower bound on the minimum singular value of input matrix \( B \):

\[
\max_{\alpha_1^2, \alpha_2^2} \{ 0.6321((\alpha_1^2)^2 + (\alpha_2^2)^2) + 0.4323(\alpha_2^2)^2 \}
\]

where:

\[
b_1 = \alpha_1^1 \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \alpha_1^1 \neq 0
\]

\[
b_2 = \alpha_1^2 b_1 + \alpha_2^2 b_2, \alpha_1^2, \alpha_2^2 \neq 0
\]

\[
0 < \varepsilon \leq \sigma_m(B), B = [b_1, b_2]
\]

\[
(\alpha_1^1)^2 + (\alpha_1^2)^2 + (\alpha_2^2)^2 = 1
\]

The optimal input matrix for \( \varepsilon = 0.017 \) is given as \( B = \begin{bmatrix} 0.1776 & 0.9794 \\ 0 & 0.0958 \end{bmatrix} \), and the corresponding value of \( \text{trace}[W_c(0,1)] \) is 0.6303. In this case, the positive angle between \( b_1, b_2 \) is equal to \( \cos^{-1}(0.99952) = 0.098\pi \). By decreasing \( \varepsilon \) to 0.0011, i.e. \( \varepsilon = 0.0011 \), changes the suboptimal solution to \( B = \begin{bmatrix} 0.0434 & -0.9988 \\ 0 & 0.0243 \end{bmatrix} \), while the optimal value of \( \text{trace}[W_c(0,1)] \) becomes 0.6320. Further the positive angle between input vectors \( b_1, b_2 \) reduces to \( \cos^{-1}(0.9997) = 0.0245\pi \).

Repeating the procedure with the Gram determinant constraint (optimization (7.83)) the problem becomes:
max \{0.6321((\alpha_1^2)^2 + (\alpha_2^2)^2) + 0.4323(\alpha_2^2)^2\}

Such that:

\begin{align*}
b_1 &= \alpha_1^1 \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \alpha_1^1 \neq 0 \\
b_2 &= \alpha_2^2 b_1 + \alpha_2^2 b_2, \alpha_1^2, \alpha_2^2 \neq 0 \\
0 < \varepsilon &\leq |G(B)|, B = [b_1, b_2] \\
(\alpha_1^1)^2 + (\alpha_1^2)^2 + (\alpha_2^2)^2 &= 1
\end{align*}

The solution of this problem for \(\varepsilon = 0.0022\) is:

\[ B = \begin{bmatrix} -0.8753 & 0.4805 \\
0 & 0.0542 \end{bmatrix} \]

\[ \text{trace}(W_c(0,1)) = 0.6315 \]

The positive angle between input vectors \(b_1, b_2\) is \(\cos^{-1}(0.9937) = 0.1123\pi\) and the minimum singular value of \(B\) is 0.0475.

Setting \(\varepsilon = 1.2771 \times 10^{-4}\) gives:

\[ B = \begin{bmatrix} -0.3279 & -0.9441 \\
0 & -0.0345 \end{bmatrix} \]

\[ \text{trace}(W_c(0,1)) = 0.6319 \]

The positive angle between input vectors \(b_1, b_2\) reduces to \(\cos^{-1}(0.9993) = 0.0374\pi\) and the minimum singular value of \(B\) decreases to 0.0113.

In conclusion, the example confirms the validity of the results stated in this section, and shows that the optimization problems (7.65), (7.79), (7.83) produce equivalent suboptimal solutions.

**7.2.2.b. Optimal Multi-Input Selection in a Stable System**

Problem:

For stable LTI MIMO system (7.48) find

\[ \max_{b_{\text{in} \in \beta}, \n \in L_{\infty}} \{\text{trace}(W_c(0,1))\}, \text{ where } \beta \text{ is the set of all possible input matrices} \]
In this sub-section, we assume the multi input LTI system (7.48) is stable and \( B = [b_1, \ldots, b_m] \), \( b_i \in \mathbb{R}^{n \times 1} \) with \( \|B\|_F \leq 1 \) and \( \text{rank}(B) = m \). In this case, \( Q \) can be calculated as the unique solution to the Lyapunov equation:

\[
A^*Q + QA + I = 0 \tag{7.85}
\]

Using the results described in the previous sub-section the average of minimum input energy can be obtained by solving one of sub-optimization problems stated in (7.65), (7.79), (7.83).

**Lemma (7.5):** If \( A \) is stable, then \( \text{trace}(W_{c}(0, \infty)) = \left\| (sI - A)^{-1} B \right\|_2 \), where \( \| \cdot \|_2 \) denotes the \( H_2 \) norm of the system.

**Proof:**

Let \( h(t) = e^{At}B \) be the impulse response of the system:

\[
\begin{align*}
\dot{x} &= Ax + Bu \\
y &= x
\end{align*} \tag{7.86}
\]

Then:

\[
\|h\|_2^2 = \text{trace}\left\{ \int_0^\infty h(t)h^*(t)dt \right\}, \quad h(t) = e^{At}B
\]

\[
= \text{trace}\left\{ \int_0^\infty e^{At}BB^*e^{A^t}dt \right\}
\]

\[
= \text{trace}\{B^*\int_0^\infty e^{A^t}e^{At}dB\}
\]

Hence:

\[
\|H(s)\|_2^2 = \left\| (sI - A)^{-1} B \right\|_2^2 \tag{7.87}
\]

So, if we assume all states are observed in output, i.e. \( C = I \) then we can write the optimization problem of minimizing the average of input energy as the maximization problem of \( H_2 \) norm of the stable LTI system.

**7.2.2.c. Optimal Input Matrix Selection Subject to the Condition On 2-Norm of the Input Matrix B**

**Problem:**

For the LTI MIMO system (7.48), find

\[
\delta_0 = \max_{\|B\|_F \leq 1, \text{rank}(B) = m, \beta} \{ \text{trace}(W_c(0,t_f)) \}, \text{ where } \beta \text{ is the set of all possible input matrices}
\]
In this part, we look for an optimal solution to the problem of multi input structure selection corresponding to the constraint \( \|B_{n \times m}\| \leq 1 \), \( \|\| \) denotes the spectral norm (i.e. maximum singular value) and \( \text{rank}(B) = m \) note that the condition \( \|B_{n \times m}\| \leq 1 \) can be expressed equivalently as \( B^* B \leq I \).

**Theorem (7.7):** The optimal matrix \( B \), which maximizes the trace of controllability Gramian subject to \( \text{rank}(B) = m \) and \( B^* B \leq I \) is \( B = [V_1(Q), V_2(Q), ..., V_m(Q)] \) where \( V_i, i = 1, ..., m \)

is the eigenvector corresponding to the \( i^{th} \) largest eigenvalue of \( Q = \int_0^t e^{A^T}e^{At} dt \) The optimal value of \( \delta_0 \), in this case, is \( \sum_{i=1}^m \lambda_i \), where \( \lambda_i \) denotes the \( i^{th} \) largest eigenvalue of \( Q \).

**Proof:**

The constraint \( \|B\| \leq 1 \) can be written as:

\[
B^* B \leq I \Rightarrow \begin{bmatrix}
    b_1^* & b_2^* & \ldots & b_m^*
    \\
    b_1 & b_2 & \ldots & b_m
\end{bmatrix} \leq I_{m \times m}
\]

Then by considering the trace of the controllability Gramian, and defining \( Q = \int_0^\infty e^{A^T}e^{At} dt \) we have:

\[
\text{trace}(W(c,0,t)) = \text{trace}(B^*QB) = \text{trace} \left( \begin{bmatrix}
    b_1^* & b_2^* & \ldots & b_m^*
    \\
    b_1 & b_2 & \ldots & b_m
\end{bmatrix} \right)
\]

\[
= \begin{bmatrix}
    b_1^*Q_{11}b_1 & \ldots & \times
    \\
    \vdots & \ddots & \vdots
    \\
    \times & \ldots & b_m^*Q_{m1}b_m
\end{bmatrix}
= \sum_{i=1}^m b_i^* Q b_i
\]

Thus, the optimization problem is equivalent to:

\[
\min_{\beta, \min(0,t_i)} \{ \text{ave}E_{\min(0,t_i)} \} = \max_{\beta, \text{rank}(B)=m} \{ \text{trace}(W(c,0,t_i)) \} = \max_{\beta, \text{rank}(B)=m} \{ \text{trace}(B^*QB) \} = \max_{\beta, \text{rank}(B)=m} \sum_{i=1}^m b_i^* Q b_i
\]

(7.91)
This is equivalent to the Rayleigh quotient form of \( \sum_{i=1}^{m} \lambda_i(Q) \).

The results follow by expanding \( B_{n \times m} = X_{n \times n} Y_{n \times m} \) where \( X \) is the matrix whose columns are the eigenvectors of \( Q^5 \), and \( Y \in \mathbb{R}^{n \times m} \). Then:

\[
trace(B^*QB) = trace(Y^*\Lambda Y) = trace(\Lambda_{n \times n} Y_{n \times m} Y^*_{m \times n})
\]

(7.92)

where \( \Lambda \) is the diagonal matrix of the eigenvalues. The condition \( B^*B = I \), implies \( Y^*Y = I_{m \times m} \) and so \( trace \ Y^*Y = trace \ YY^* = trace \ I = m \).

Hence \( trace(\Lambda YY^*) \) is maximized when \( Y = \begin{bmatrix} I \\ 0 \end{bmatrix} \), which concludes the proof.

\( Q \) is symmetric matrix then it is diagonalizable
7.3. Summary:

In this chapter, we proposed a novel solution to the problem of optimum input structure selection, considering the average of minimum input energy, and we solved the problem for two main cases, single input where input matrix $B$ is a vector, and the case of multi inputs, where input matrix $B$ has the dimension of $n \times m$.

We also considered various interesting cases, in which matrix $A$ has special properties or the input matrix $B$ satisfies different constraints.

The results of this chapter are summarized in tables (7.1) and (7.2).

These results can be used as the solution to the problem of actuator localization, where we are looking for the optimum location of actuators, considering the minimum achievable input energy averaged over all possible directions.

<table>
<thead>
<tr>
<th>$b \in \mathbb{R}^{n \times 1}$</th>
<th>$\max \text{trace}{W_c(0,t_f)}$</th>
<th>$\arg \max \text{trace}{W_c(0,t_f)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1) $b \in \mathbb{R}^{n \times 1}$</strong></td>
<td>$\lambda_{\max}(Q)$</td>
<td>$V_{\max}(Q)$</td>
</tr>
<tr>
<td>$t_f$ finite</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_f \to \infty$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A stable</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **2) $b \in \mathbb{R}^{n \times 1}$** | $\lambda_{\max}(Q)$ | $V_{\max}(Q)$ |
| $t_f \to \infty$ |                                |                                |
| A stable       | $A^*Q + QA + I = 0$ |                                |

| **3) $b \in \mathbb{R}^{n \times 1}$** | $\lambda_{\max}(Q) = \frac{e^{\lambda_{1}t_f} - 1}{\lambda_{1}}$ | $V_{\max}(A + A^*)$ |
| $t_f$ finite |                                |                                |
| A normal      | $\lambda_1 = \lambda_{\max}(A + A^*)$ |                                |

| **4) $b \in \mathbb{R}^{n \times 1}$** | $\lambda_{\max}(Q) = -\lambda_{1}^{-1}$ | $V_{\max}(A + A^*)$ |
| $t_f \to \infty$ |                                |                                |
| A stable       | $\lambda_1 = \lambda_{\max}(A + A^*) < 0$ |                                |
| A normal       |                                |                                |

Table (7.1): Optimum input vector in single input cases, which minimizes the average of $E_{\min}$. 

166
<table>
<thead>
<tr>
<th>$B \in \mathbb{R}^{n \times m}$</th>
<th>$\max \text{trace} { W_c(0, t_f) }$</th>
<th>$\arg \max \text{trace} { W_c(0, t_f) }$</th>
</tr>
</thead>
</table>
| 1) $B \in \mathbb{R}^{n \times m}$  
$\|B\|_F \leq 1$  
$\text{rank}(B) = m$  
$t_f$ finite | $\sum_{i=1}^{m} \beta_i \lambda_i$  
A weighted function of $m$ first largest eigenvalues of $Q$ | $b_i$ is the linear function of $\{V_{\max 1}(Q), \ldots, V_{\max m}(Q)\}$  
$V_{\max i}, i = 1, \ldots, m$  
Eigenvectors corresponding to $\lambda_i^*, i$th largest eigenvalue of $Q$ |
| 2) $B \in \mathbb{R}^{n \times m}$  
$\|B\|_F \leq 1$  
$\text{rank}(B) = m$  
A stable  
$t_f \to \infty$ | $\text{If: } C = I$  
$\text{trace} \{ W_c(0, \infty) \}$  
$= \|h(t)\|_2^2$  
$= \|(sI - A)^{-1}B\|_2^2$  
$A^*Q + QA + I = 0$ | $b_i$ is the linear function of $\{V_{\max 1}(Q), \ldots, V_{\max m}(Q)\}$  
$V_{\max i}, i = 1, \ldots, m$  
Eigenvectors corresponding to $\lambda_i^*, i$th largest eigenvalue of $Q$ |
| 3) $B \in \mathbb{R}^{n \times m}$  
$\|B_{n \times m}\| \leq 1$  
$\text{rank}(B) = m$  
$t_f$ finite | $\sum_{i=1}^{m} \lambda_i^*(Q)$  
$\lambda_i^*$ are the $m$ largest eigenvalues of $Q$ | $B = [V_{\max 1}(Q), \ldots, V_{\max m}(Q)]$  
$V_{\max i}, i = 1, \ldots, m$  
Eigenvectors corresponding to $\lambda_i^*, i$th largest eigenvalue of $Q$ |

Table (7.2): Optimum input matrix in multi-input cases, which minimizes the average of $E_{\min}$
Chapter 8: Kamineh Algorithm Optimization

8.1. Introduction

In this chapter, a novel metaheuristics optimization algorithm namely Kamineh Algorithm (KA) is proposed. The KA starts from a random initial population and requires them to move towards the best solution using Logistic equation like functions.

A metaheuristic is formally defined as an iterative generation process which guides a subordinate heuristic by combining intelligently different concepts for exploring and exploiting the search space, learning strategies are used to structure information in order to find efficiently near-optimal solutions (Osman & Laporte, 1996).

Metaheuristics have admirable ability to avoid trapping in local optima, this important property is due to the stochastic nature of metaheuristic algorithms, which make them suitable for optimizing challenging real problems where the search space is usually unknown and very complex with a massive number of local optima.

Different metaheuristic algorithms exist in literature, some more popular ones are Genetic algorithms (GAs), Simulated annealing (SA), Differential evolution (DE), Ant colony optimization (ACO), Bat algorithms (BA), Particle swarm optimization (PSO), Tabu search (TS), Gravitational search algorithm (GSA), and Firefly algorithms (FA). Some algorithms have also...
been proposed integrating or improving the current techniques, to make them more efficient and applicable.

Two main components of any metaheuristic algorithm are intensification and diversification, or exploitation and exploration. Exploration is the ability of the algorithm to search for new individuals far from the current solution in the search space and to explore the search space on a global scale. Exploitation is to search the surrounding search area nearby the current solution, something like local search. Exploitation ensures that the solutions will converge to the optimality, whereas the exploration via randomization avoids the solutions being trapped at local optima and, at the same time, increases the diversity of the solutions.

Finding an algorithm that can handle both (exploitation and exploration) is still challenging because they are two different contrast objectives.

The No Free Lunch (NFL) Theorem of (Wolpert & Macready, 1997) states that “There is no strategy or algorithm that generally behaves better than another for the entire set of possible problems”. (Ho & Pepyne, 2002) also expressed that: “Universal optimizers are impossible”. This motivated us to develop a new effective metaheuristic algorithm called KA (Kamineh Algorithm), using the mathematical logistic equation like functions, which shows superior performance in terms of exploration, exploitation, and convergence comparable with other state-of-the-art methods. It makes KA a powerful optimization algorithm in solving complex multi-parameter optimization problems, such as the actuators (sensors) placement. In this thesis, the optimization criteria for the selection of actuators are based on the eigenvalues of controllability Gramian matrix.

The structure of this chapter is as follow:

Section 8.2 outlines the proposed KA optimization method. The results and discussion of benchmark functions and some case studies are presented in Sections 8.3. Finally, Section 8.4 concludes the work.

### 8.2. Kamineh Optimization Algorithm (KA)

In this section, the inspiration of the proposed method is first discussed. Then, the mathematical model is provided.

#### 8.2.1. Inspiration

The logistic equation (also called Verhulst model or logistic growth curve) is a model of population growth first published by (Verhulst, 1845), (Verhulst, 1847). The model is continuous and when is translated into mathematics, results to the differential equation below:

$$\frac{dP(t)}{dt} = rP(t) \left[ 1 - \frac{P(t)}{K} \right], P(0) = P_0$$

(8.1)
where \( t \) denotes time, \( P_0 \) is the initial population and \( r, K \) are constants associated with the growth rate and the carrying capacity of the population.

However, in many applications, when modelling a realistic problem, one may decide to describe the continuous-time logistic equations in terms of discrete-time model. Thus, the differential equation (8.1) can be formulated instead, as an initial value problem of a difference equation below:

\[
P_{n+1} = rP_n \left(1 - P_n \right) \tag{8.2}
\]

where \( r \) is the Malthusian parameter (rate of maximum population growth). The above statement is non-linear, since it involves a term \( P_n^2 \). Because of this non-linearity, the equation has remarkable non-trivial properties, but cannot be solved analytically. The discrete version of the logistic equation has the fascinating properties and can be used in many real-world examples such as:

- The concentration of oxygen in the lungs after the \( i^{th} \) breath
- The concentration in the blood of a drug after the \( i^{th} \) dose
- The size of a population of mosquitoes in year \( n \)
- The number of cells in a bacterial culture on day \( i \)

The steady state \( P_{ss} \) of difference equation (8.2) is defined by:

\[
P_{n+1} = P_n = P_{ss} \tag{8.3}
\]

which returns:

\[
P_{ss} = rP_{ss} \left(1 - P_{ss} \right) \Rightarrow \begin{cases} P_{ss1} = 0 \\ P_{ss2} = 1 - \frac{1}{r} \end{cases} \tag{8.4}
\]

Now consider a small perturbation \( \varepsilon \) that moves the system out of its steady state:

\[
P_n = P_{ss} + \varepsilon_n \Rightarrow \\
P_{n+1} = P_{ss} + \varepsilon_{n+1} \Rightarrow \\
\varepsilon_{n+1} = f \left(P_{ss} + \varepsilon_n\right) - P_{ss} \tag{8.5}
\]

Since \( \varepsilon_n \ll P_{ss} \) we can use the Taylor expansion around \( P_{ss} \):

\[
f \left(P_{ss} + \varepsilon_n\right) = f \left(P_{ss}\right) + \left(\frac{df}{dP}\right)_{P=P_{ss}} P_n + O\left(P_n^2\right) \tag{8.6}
\]

Then equation (8.5) can be approximated as:

\[
\varepsilon_{n+1} \approx \left(\frac{df}{dP}\right)_{P=P_{ss}} P_n \tag{8.7}
\]
Clearly, if \( \left. \frac{df}{dP} \right|_{P=P_0} < 1 \), the steady state is stable and the perturbation tends to zero as \( n \) increases, otherwise the steady state is unstable. In the case of the logistic equation (8.2), for the first steady state, i.e. \( P_{s1} = 0 \) we have:

\[
\left. \frac{df}{dP} \right|_{P=0} = (r - 2rP)_{P=0} = r \tag{8.8}
\]

Thus, the steady state \( P_{s1} = 0 \) is stable when \( 0 < r < 1 \).

Similarly, for the second steady state \( P_{s2} = 1 - \frac{1}{r} \), we have:

\[
\left. \frac{df}{dP} \right|_{P=1} = (r - 2rP)_{P=1} = 2 - r \tag{8.9}
\]

Thus, the steady state \( P_{s2} = 1 - \frac{1}{r} \) is stable when \( 1 < r < 3 \).

**Figure (8.1): Steady states of logistic equation (8.2)**

In Figure (8.2), several time sequences, corresponding to different values of \( 0 < r \) are shown. When \( r \) increases, the steepness of the parabola increases as well, which makes the slope of this tangent steeper, so that eventually the stability condition is violated. The steady state then becomes unstable and the oscillations appear. (May, 1976) proves that as \( r \) increases just beyond \( r = 3 \), stable oscillations of period 2 (sometimes called two-point cycles) appear. A stable oscillation means a periodic behavior that is maintained despite small perturbations
and period 2 implies that successive generations alternate between two fixed values of $P$. When $r$ further increases, the periodic solution becomes unstable and then higher period oscillations appear. When all the cycles become unstable, chaos is observed.

The schematic representation shown in Figure (8.3) illustrates that as $r$ increases, the system (8.2) undergoes successively cycles of period 2, 4, 8, ..., and ultimately leads to a chaotic behavior.
Figure (8.2): Graphical resolution for various values of $r$
In KA optimization method, an agent represents a diversity of species with different populations. Over the iterations, the species reach new populations different from the previous ones. The change of each species' population is correlated with the fitness of the agent and can affect all its other species' populations as well. The species on the same agent share their individual information with others to reach their destination population, which leads to the optimum fitness value.

A function of the distance between the current agent's fitness and the best existing objective value updates the rate of maximum population growth of the species in each iteration. Then the closer the objective value is to the current optimum fitness, the smaller the Malthusian parameter $r$ is and the species are closer to the stable steady state region. Moreover, as $r$ increases the chaotic growth of the species guarantees that KA algorithm intrinsically benefits
from high exploration and local optima avoidance. However, in the exploitation phase when \( r \) is small there are gradual changes due to the stable behavior and the cyclic oscillations of the logistic equations and random variations are considerably less than those in the exploration phase. This enhances the convergence capability of KA method towards the optimal solution.

In the next subsection, KA optimization method is proposed based on the mathematical model.

**8.2.2. Mathematical Model and Algorithm**

As shown in Figure (8.4), KA algorithm starts with randomly generated populations of species for each candidate/agent called individual and the fitness function is evaluated for each candidate.

We represent the set of current populations in a matrix as:

\[
X^r = \begin{bmatrix}
X_{11} & X_{12} & \cdots & X_{1n} \\
X_{21} & X_{22} & \cdots & X_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
X_{m1} & X_{m2} & \cdots & X_{mn}
\end{bmatrix}
\]

(8.10)

where \( n \) is the number of agents and \( m \) is the number of species (i.e. dimension). For all agents \( X_i, i = 1, \ldots, n \), we also assume a vector corresponding to the fitness values as:

\[
f_X = \begin{bmatrix} f_{X_1} & f_{X_2} & \cdots & f_{X_n} \end{bmatrix}
\]

(8.11)

Each agent passes through the fitness function and the obtained output updates the corresponding fitness value of the associated agent in the matrix \( f_X \).

We also assume that there is an array for storing the optimum populations for the best agent \( X_{best} = X_i, i = 1, \ldots, n \) and the corresponding fitness value \( f_{X_i} \).

![Figure (8.4): Schematic representation of the KA optimization method](image)
The KA algorithm approximates the global solution of the optimization problem based on the triplet:

\[(\text{Initialize, Update, Terminate}) = (I, U, T)\]

\(I\) is a function that generates some random populations of the species for individuals and corresponding fitness values:

\[I : \emptyset \rightarrow \{X, f_X\}\]  

(8.12)

Any random distribution can be used in \(I\) function. What we used in KA algorithm is as follows:

\[X(:, j) = \text{rand}(n, 1) \times (ub_j - lb_j) + lb_j, \quad j = 1, \ldots, m\]

\[f_X = \text{Obj}(X)\]  

(8.13)

where \(ub_j, lb_j, j = 1, \ldots, m\) denote the upper-bound and lower-bound of the \(j^{th}\) species' population (i.e. the \(j^{th}\) decision variable) respectively. \(\text{Obj}\) defines the objective function, which is used for the optimization problem. \(X(:, j)\) is a \(n\)-dimensional vector, which consists the \(j^{th}\) species' population of all individuals and \(\text{rand}(n, 1)\) is a \(n\)-dimensional random vector.

The \(U\) function, which is the main function, moves the agents around the search space. This function receives the populations \(x_{ij}\) where \(i = 1, \ldots, n, j = 1, \ldots, m\) and returns their updated ones eventually:

\[U : x_{ij}(t) \rightarrow x_{ij}(t + 1)\]  

(8.14)

The \(T\) function returns \(\text{Stop}\) if the termination criterion is satisfied and \(\text{Continue}\) if the termination criterion is not satisfied. The KA algorithm terminates the optimization process when the iteration counter goes higher than the maximum number of iterations by default. However, any other termination condition can also be considered easily wherever it is needed, e.g. maximum number of function evaluation or the accuracy of the global optimum obtained.

\[T : X \rightarrow \{\text{Stop, Continue}\}\]  

(8.15)

The \(U\) function iteratively run until the \(T\) function returns \(\text{Stop}\). With defined three-tuple \((I, U, T)\) the general framework of the KA algorithm is defined as:
\[ X = I(\ ) \]

While \( T = \text{Continue} \)

for \( i = 1, \ldots, n \)

Update \( \overline{X}_{best}, f_{\overline{X}_{best}} \)

Update species' populations of \( i \)-th individual \( x_{ij} = U(x_{ij}, f_{X_j}), \quad j = 1, \ldots, m \)

end

end

(8.16)

As mentioned in the previous subsection the inspiration of this algorithm is the fascinating properties of the discrete version of the logistic equation. In order to mathematically model this behavior, we update each species with respect to its current value of population via the equation below:

\[
x_{ij} (t+1) = \begin{cases} 
    rx_{ij} (t)(1-x_{ij} (t)) & O(t) = 1 \\
    \text{rand} \times x_{best,j} (t) & O(t) = 0 
\end{cases}
\]

(8.17)

where \( x_{ij} \) indicates the \( j \)-th species' population of the individual \( i \) and \( x_{best,j} \) denotes the \( j \)-th species' population of the current optimal solution. \( \text{rand} \) denotes a random number generated with uniform distribution over the interval of \([0,1]\). \( O(t) \) is a stochastic function defined as:

\[
O(t) = \begin{cases} 
    1 & \text{if } \text{rand} \geq 0.5 \\
    0 & \text{if } \text{rand} < 0.5 
\end{cases}
\]

(8.18)

In equation (8.17), \( r \) is the rate of maximum population growth and is defined as:

\[
r = \begin{cases} 
    1 < \beta_1 \leq 2 & S(f_{\overline{X}}, f_{\overline{X}_{best}}) \leq \alpha_1 \\
    2 < \beta_2 \leq 3.4 & \alpha_1 < S(f_{\overline{X}}, f_{\overline{X}_{best}}) \leq \alpha_2 \\
    3.4 < \beta_3 \leq 4 & \alpha_2 < S(f_{\overline{X}}, f_{\overline{X}_{best}}) 
\end{cases}
\]

(8.19)

where \( f_{\overline{X}} \) is the objective value of the associated agent and \( f_{\overline{X}_{best}} \) denotes the fitness value of the current optimal solution. \( S(f_{\overline{X}}, f_{\overline{X}_{best}}) \) dictates a logarithmic function of the difference between \( f_{\overline{X}} \) and \( f_{\overline{X}_{best}} \):

\[
S(f_{\overline{X}}, f_{\overline{X}_{best}}) = \log \left( 1 + \left| f_{\overline{X}} - f_{\overline{X}_{best}} \right| \right) \times \text{rand}
\]

(8.20)

The \( \text{rand} \) parameter in (8.20) brings a random weight for the logarithmic function, i.e. \( \log \left( 1 + \left| f_{\overline{X}} - f_{\overline{X}_{best}} \right| \right) \) in order to stochastically determine the effect of the difference between the current solution and the best one in defining the behavior of the logistic equations, which are used to update the current populations via (8.17).
In KA method, we employ two user-controlled parameters to guide the searching behavior:

- \( \beta_i\), \( i = 1, 2, 3 \): This parameter defines the rate of maximum population growth and determines the population's behavior.
- \( 0 < \alpha_1 < \alpha_2 < 1 \): defines random values in (0,1). This parameter controls the probability of the species' population changing rate and assists KA method to show a more random behavior throughout optimization through a balanced exploration and exploitation.

Choosing proper parameters of KA for numerical examples and real-world optimization problems can be done through different schemes:

- The trial-and-error scheme: May reveal the best possible performance over the parameter space at the expense of high computational cost. In real-world optimization problems, evaluating the fitness function may take a long time, much longer than evaluating our benchmark functions, then it might be impractical to use this scheme.
- Fixed parameter schemes: As the optimization benchmark set is generally divided into three categories (i.e., unimodal, high-dimensional multimodal, and low-dimensional multimodal), we try to determine a parameter value combination generally suitable for each category. This combination remains constant throughout the whole search (Yao, et al., 1999), (Lam, et al., 2012), (Price, et al., 2005).
- Deterministic parameter schemes: Change the parameter values throughout the search using some pre-defined rules (Lam, et al., 2012), (Chen, et al., 2012).
- Adaptive parameter schemes: Change the parameter values by adaptively learning the impact of changing parameters on the searching performance throughout the search (Qi, et al., 2009). Some schemes encode the parameters into the solution and evolve the parameters together with the population (Vrugt, et al., 2009).

Here, we use the deterministic scheme to choose the parameters, which can consistently lead to satisfactory results on a wide range of different kinds of optimization problems. According to (Bergh & Engelbrecht, 2006), there should be abrupt changes in the movement of search agents over the initial steps of optimization. This assists a meta-heuristic to explore the search space extensively. Therefore, in the KA algorithm, the first step is to combine the random solutions in the set of solutions abruptly with a high rate of randomness to find the promising regions of the search space. Hence, at the beginning of the optimization, the appropriate values of the parameters are selected, which lead to a higher rate of maximum population growth, i.e. \( r \). However, according to (Bergh & Engelbrecht, 2006) the changes should be reduced to emphasize exploitation at the end of optimization. Therefore, after finding the promising regions of the search space, the KA algorithm converges to the global optimum by gradual changes in the random solutions. The random variations are considerably less than those in the exploration phase. Then, the parameters change adaptively such that \( r \), i.e. the rate of maximum population growth, decreases over the course of iterations. Figure (8.5) shows that how the parameters of the KA algorithm change adaptively through the iterations. The following formulas are utilized in this regard:
\[
\beta_1 = 2 \left( 1 - \frac{t}{T} \right), \quad \beta_2 = 3.4 \left( 1 - \frac{t}{T} \right), \quad \beta_3 = 4 \left( 1 - \frac{t}{T} \right)
\]
\[
\alpha_1 = \min \left( \left( 1 - \frac{t}{T} \right), \alpha_2 - 0.3 \right), \quad \alpha_2 = \min \left( \left( 1 - \frac{t}{T} \right) + 0.3, 0.8 \right)
\] (8.20)

where \( t \) is the current iteration and \( T \) is the maximum number of iterations.

Figure (8.5): Change of parameters over the course of iterations.

Figure (8.6) illustrates how the parameters in equation (8.20) change the behavior of the logistic equation and decrease the range of the movement of the solutions over the iterations. It may be inferred from figure (8.6) that the KA algorithm explores different regions of the search space with a higher degree of randomness via increasing the range of the chaotic behavior of the logistic equation in (8.17), when the optimization starts, so this can assist resolving local optima stagnations. Then the algorithm gradually increases the probability of
the steady state behavior of the logistic equation through the iterations to emphasize more exploitation around the most promising regions of the search space. Then, using the adaptive parameters in the proposed method leads to a smooth transition from exploration to exploitation during optimization process, which guarantees the improved accuracy of local search over the iterations and the convergence of the proposed algorithm. In Figure (8.6), the red point shows a search agent, the blue star is the best solution found so far and the three colored circles around the red point describe the search spaces associated with 3 different logistic equation regions:

- Steady state
- Cyclic
- Chaos

Figure (8.6): The movement of a solution around the destination via change of the parameters over the iterations

After all, the pseudo code of the KA algorithm in (8.16) can be presented as follows:
Step 1: Initialization
Step 2: Iteration
   Do
   for each individual repeat
   Update \( \bar{X}_{best} \), \( f_{X_{best}} \)
   Update the parameters \( \beta_k, \alpha_l, \quad k = 1, 2, 3, \quad l = 1, 2 \) \( (8.21) \)
   Update the rate of maximum population growth, i.e. \( r \)
   Update the population of the species \( x_j = U(x_j, f_x); \quad j = 1, \ldots, m \)
   end
While stopping criteria
Step 3: Return the best solution obtained

The general steps defined in \( (8.21) \) shows that the KA algorithm starts the optimization process with creating a set of random solutions. The algorithm then saves the best solutions obtained so far, assigns it as the destination point, and updates other solutions with respect to it. Meanwhile, the parameters and the rate of maximum population growth, i.e. \( r \), are updated to emphasize more exploration of the search space at the beginning and higher exploitation of the search space as the iteration increases. The KA algorithm terminates the optimization process when the termination criterion is satisfied and the best solution obtained so far is returned as the global optimum.

The next section employs a wide range of test problems to investigate, analyze and confirm the effectiveness of the KA algorithm.

### 8.3. Experimental Results and Discussion

For any new optimization algorithm, it is essential to validate its performance and compare it with other existing algorithms over a good set of test functions.

In this section, the proposed KA algorithm is tested against different sets of benchmark functions and compared with five current popular metaheuristic methods introduced over a long range of time starting from 1975 onwards: Genetic Algorithms (GA) (Holland, 1975), Particle Swarm Optimization (PSO) (Kennedy & Eberhart, 1995), Firefly Algorithm (FA) (Yang, 2008), Gravitational Search Algorithm (GSA) (Rashedi & Nezamabadi-pour, 2009), and Collective Animal Behavior Algorithm (CAB) (Cuevas, et al., 2012).

In this work, the maximum number of iterations and the size of population for the KA algorithm are set to 100 and 80 respectively. The results are averaged over 30 independent runs and the mean value and standard deviation of the best solutions are obtained.

The performance of optimization algorithms were compared considering three criteria: 1) the final results (i.e. average value of cost functions and the value of standard deviations), 2) the convergence rate and 3) the results produced by non-parametric statistical Wilcoxon rank sum
test (Wilcoxon, 1945), (García, et al., 2009). We can also use the following efficiency measure to evaluate the performance of the algorithm (de-los-Cobos-Silva, et al., 2015):

\[
\text{efficiency} = \begin{cases} 
1 - \frac{f_{X_{\text{best}}}}{f_{\text{min}}} & \text{if } f_{\text{min}} \neq 0 \\
1 - \frac{f_{X_{\text{best}}}}{f_{\text{min}}} & \text{if } f_{\text{min}} = 0
\end{cases}
\]

where \( f_{X_{\text{best}}} \) is the best value of fitness found by the algorithm and \( f_{\text{min}} \) is the optimum value of the cost function. (Note: considering the accuracy of Matlab, when the efficiency is greater than 0.999999, it is rounded to 1.)

### 8.3.1. Benchmark Functions

Test functions are important to evaluate reliability, efficiency and validation of optimization algorithms. There have been many test or benchmark functions reported in the literature; however, there is no standard list or set of benchmark functions. The number of test functions in most papers varied from a few to about two-dozen, and ideally, the test functions used should be diverse and unbiased to cover a wide variety of problems, such as unimodal, multimodal, separable, non-separable and multi-dimensional problems to confidently make sure that the achieved results of our algorithm are not by chance. (Jamil & Yang, 2013)

A complete detailed description of optimization test functions can be found in (Jamil & Yang, 2013), (Ali, et al., 2005), (Adorio & Dilman, 2005), (Botev, et al., 2004), (Yang, 2010).

In this work, we are using a comprehensive set of 23 functions, which are used in most literature (Cuevas, et al., 2012), (Yao & Liu, 1996), (Gaviano, et al., 2003), (Ali, et al., 2005) to test the performance of the proposed approach. Such functions are classified into three different categories: unimodal test functions (F1-F7), which have only one local optimum and allow to evaluate the exploitation capability of the investigated optimization algorithm, multimodal test functions with variable dimensions (F8-F13) and multimodal fixed-low-dimensional test functions (F14-F23), which both have more than one local optimum. These functions are used to test the capability of the KA algorithm to escape from local minima, and to investigate the exploration ability of the algorithm. The test functions are given in Tables (8.1), (8.2) and (8.3), where V-no is the dimension of the function, Range is the boundary of the function’s search space and \( f_{\text{min}} \) shows the optimum value of the function. Figures (8.7), (8.8) and (8.9) represent two-dimensional form of some benchmark functions out of each category.
### Table (8.1): Unimodal test functions

<table>
<thead>
<tr>
<th>Function</th>
<th>( V_{\text{no}} )</th>
<th>Range</th>
<th>( f_{\text{min}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_1(x) = \sum_{i=1}^{n} x_i^2 )</td>
<td>30</td>
<td>([-100,100])</td>
<td>0</td>
</tr>
<tr>
<td>( f_2(x) = \sum_{i=1}^{n}</td>
<td>x_i</td>
<td>+ \prod_{i=1}^{n}</td>
<td>x_i</td>
</tr>
<tr>
<td>( f_3(x) = \sum_{i=1}^{n} (\sum_{j=1}^{i} x_j)^2 )</td>
<td>30</td>
<td>([-100,100])</td>
<td>0</td>
</tr>
<tr>
<td>( f_4(x) = \max_i {</td>
<td>x_i</td>
<td>, \ 1 \leq i \leq n} )</td>
<td>30</td>
</tr>
<tr>
<td>( f_5(x) = \sum_{i=1}^{n-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2] )</td>
<td>30</td>
<td>([-30,30])</td>
<td>0</td>
</tr>
<tr>
<td>( f_6(x) = \sum_{i=1}^{n} (\frac{x_i + 0.5}{2})^2 )</td>
<td>30</td>
<td>([-100,100])</td>
<td>0</td>
</tr>
<tr>
<td>( f_7(x) = \sum_{i=1}^{n} x_i^2 + \text{random}[0,1] )</td>
<td>30</td>
<td>([-1.28,1.28])</td>
<td>0</td>
</tr>
</tbody>
</table>

### Figure (8.7): Typical two-dimensional form of unimodal benchmark functions

### Table (8.2): Multimodal test functions with variable dimensions

<table>
<thead>
<tr>
<th>Function</th>
<th>( V_{\text{no}} )</th>
<th>Range</th>
<th>( f_{\text{min}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_8(x) = \sum_{i=1}^{n} -x_i \sin(\sqrt{</td>
<td>x_i</td>
<td>}) )</td>
<td>30</td>
</tr>
<tr>
<td>( f_9(x) = \sum_{i=1}^{n} x_i^2 - 10 \cos(2\pi x_i) + 10 )</td>
<td>30</td>
<td>([-5.12,5.12]^n)</td>
<td>0</td>
</tr>
<tr>
<td>( f_{10}(x) = -20 \exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}\right) - \exp\left(\frac{1}{n}\sum_{i=1}^{n} \cos(2\pi x_i)\right) + 20 + \varepsilon )</td>
<td>30</td>
<td>([-32,32]^n)</td>
<td>0</td>
</tr>
<tr>
<td>( f_{11}(x) = \frac{1}{3000} \sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos(x_i) + 1 )</td>
<td>30</td>
<td>([-600,600]^n)</td>
<td>0</td>
</tr>
<tr>
<td>( f_{12}(x) = \frac{n}{2} \left[10 \sin(\pi y_1) + \sum_{i=1}^{n-1} (y_i - 1)^2 [1 + 10 \sin^2(\pi y_{i+1})] + (y_n - 1)^2 + \sum_{i=1}^{n} u(x_i, 10, 100, 4)\right] )</td>
<td>30</td>
<td>([-50,50]^n)</td>
<td>0</td>
</tr>
<tr>
<td>( y_i = 1 + \frac{x_i + 1}{2} u(x_i, a, k, m) )</td>
<td>30</td>
<td>([-50,50]^n)</td>
<td>0</td>
</tr>
<tr>
<td>( k(x_i - a)^m ) \quad x_i &gt; a</td>
<td>0 \quad a &lt; x_i &lt; a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( k(-x_i - a)^m ) \quad x_i &lt; -a</td>
<td>0 \quad a &lt; x_i &lt; a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( f_{13}(x) = 0.1 [\sin^2(3\pi x_1) + \sum_{i=1}^{n} (x_i - 1)^2 [1 + \sin^2(3\pi x_i + 1)] + (x_n - 1)^2 [1 + \sin^2(2\pi x_n)] + \sum_{i=1}^{n} u(x_i, 5, 100, 4))</td>
<td>30</td>
<td>([-50,50]^n)</td>
<td>0</td>
</tr>
</tbody>
</table>
8.3.1.1. Unimodal Test Functions

Functions F1-F7 in Table (8.1) are unimodal functions. Unimodal functions are those functions which have only single local minima and these functions are suitable for testing the convergence rate and exploitation of algorithms. Minimization results of unimodal benchmark functions are given in Table (8.4) (due to accuracy of Matlab software, the values below $10^{-16}$ are assumed to be 0).
As Table (8.4) shows, the KA algorithm outperforms the other algorithms on 6 out of 7 unimodal test functions. The results are followed by the CAB, GSA, PSO, FA and GA algorithms. Only in the case of function F6 the CAB algorithm provides the better mean value over 30 runs, however, in this case, the standard deviation (i.e. SD) of the CAB algorithm is much higher than the SD value in the KA method. F6 poses a difficulty for the KA algorithm, since the flatness of the function does not give the algorithm any information to direct the search process towards the minima. Furthermore, the KA algorithm does provide very competitive results compared to the CAB on F5 and F7. The benchmark F7 is a noisy function, which contains a random term. The best results are those of KA. As Table (8.4) shows all algorithms are able to optimize F1 with a more or less acceptable error. Unlike the other algorithms, KA and GSA are able to reach the global optimum without any error. Although none of the six tested algorithms could optimize F5 which is a non-separable function the KA was able to provide a better result. Note that in general, inseparable functions are relatively difficult to solve, when compared with their separable counterpart, where each variable is independent of the other variables. If all the variables are independent, then a sequence of \( n \) independent optimization processes can be performed and then each design variable or parameter can be optimized independently (Salomon, 1996). F3 is also a non-separable function, for which the CAB algorithm comes close to the global optimum (which is equal to zero) with an error of about \( 10^{-9} \), while the best result obtained by the KA algorithms without any error. Some algorithms such as GSA and GA are not able to find good solutions for this benchmark. Moreover, to significantly compare the algorithms, we conducted a pairwise comparison to detect significant performance between all the algorithms with the non-parametric Wilcoxon rank test (Wilcoxon, 1945), (García, et al., 2009). The analysis is performed considering a 5% significance level over the mean value data. Table (8.5) reports the p-values produced by Wilcoxon’s test for the pairwise comparison of the five groups. Such groups are formed by KA versus GA, KA versus PSO, KA versus FA, KA versus GSA, and CAB versus CAB. As a null hypothesis, it is assumed that there is no significant difference between mean values of the two algorithms. The alternative hypothesis considers a significant difference between the mean values of both approaches. The most of p-values reported in Table (8.5) are less than 0.05 (5% significance level), which is a strong evidence against the null hypothesis, indicating that the KA results are statistically significant and that it has not occurred by coincidence i.e., due to the normal noise contained in the process.

In total, the results of Table (8.4) show that the KA algorithm is able to provide very competitive results on the unimodal benchmarks. The p-values in Table (8.5) also prove that the superiority is significant in the majority of the cases. This testifies that the proposed algorithm has a high exploitation ability, which is due to the integrated adaptive growth rate parameter that assists the KA optimization method to provide very good exploitation. High exploitation assists the KA algorithm to rapidly converge towards the optimum and exploit it accurately as can also be inferred from Figures (8.10)-(8.16).
### Table (8.4): Results of unimodal test functions

<table>
<thead>
<tr>
<th>Function</th>
<th>KA</th>
<th>GA</th>
<th>PSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>F1</td>
<td>0</td>
<td>0</td>
<td>0.8078</td>
</tr>
<tr>
<td>F2</td>
<td>0</td>
<td>0</td>
<td>0.269483</td>
</tr>
<tr>
<td>F3</td>
<td>0</td>
<td>0</td>
<td>0.13902</td>
</tr>
<tr>
<td>F4</td>
<td>0</td>
<td>0</td>
<td>1.59375</td>
</tr>
<tr>
<td>F5</td>
<td>6.2295</td>
<td>0.0813</td>
<td>369.7545</td>
</tr>
<tr>
<td>F6</td>
<td>3.6042e-05</td>
<td>9.8070e-06</td>
<td>6.984222</td>
</tr>
<tr>
<td>F7</td>
<td>9.4939e-05</td>
<td>2.9781e-05</td>
<td>0.047174</td>
</tr>
</tbody>
</table>

### Table (8.5): P-values produced by Wilcoxon’s test comparing KA versus GA, PSO, FA, GSA and CAB over all runs for unimodal test functions

<table>
<thead>
<tr>
<th>Function</th>
<th>KA</th>
<th>GA</th>
<th>PSO</th>
<th>FA</th>
<th>GSA</th>
<th>CAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>F1</td>
<td>8.0065e-09</td>
<td>5.4955e-05</td>
<td>8.0065e-09</td>
<td>N/A</td>
<td>3.9926e-06</td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>2.1025e-07</td>
<td>5.4987e-04</td>
<td>8.0065e-09</td>
<td>5.4987e-04</td>
<td>3.9926e-06</td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td>3.9926e-06</td>
<td>0.4643</td>
<td>2.1025e-07</td>
<td>8.0065e-09</td>
<td>0.2534</td>
<td></td>
</tr>
<tr>
<td>F4</td>
<td>8.0065e-09</td>
<td>8.0065e-09</td>
<td>8.0065e-09</td>
<td>8.0065e-09</td>
<td>0.2534</td>
<td></td>
</tr>
<tr>
<td>F5</td>
<td>1.6098e-04</td>
<td>6.7956e-08</td>
<td>6.7956e-08</td>
<td>0.0012</td>
<td>1.2009e-06</td>
<td></td>
</tr>
<tr>
<td>F6</td>
<td>3.5557e-08</td>
<td>2.4844e-04</td>
<td>6.7956e-08</td>
<td>6.7956e-08</td>
<td>0.2853</td>
<td></td>
</tr>
<tr>
<td>F7</td>
<td>1.5997e-05</td>
<td>6.7956e-08</td>
<td>8.4848e-09</td>
<td>6.7956e-08</td>
<td>6.7956e-08</td>
<td></td>
</tr>
</tbody>
</table>

Table (8.5): P-values produced by Wilcoxon’s test comparing KA versus GA, PSO, FA, GSA and CAB over all runs for unimodal test functions.
Figure (8.10): Benchmark function F1-The convergence rate, average fitness of all individuals, trajectory and the fitness history.

Figure (8.11): Benchmark function F2-The convergence rate, average fitness of all individuals, trajectory and the fitness history.
Figure (8.12): Benchmark function F3-The convergence rate, average fitness of all individuals, trajectory and the fitness history.

Figure (8.13): Benchmark function F4-The convergence rate, average fitness of all individuals, trajectory and the fitness history.
Figure (8.14): Benchmark function F5-The convergence rate, average fitness of all individuals, trajectory and the fitness history.

Figure (8.15): Benchmark function F6-The convergence rate, average fitness of all individuals, trajectory and the fitness history.
8.3.1.2. Multimodal Test Functions

Functions F8-F23 given in Table (8.2), (8.3) are multimodal functions. Multimodal functions are those functions which have more than one local optimum. These functions are used to test the ability of an algorithm to escape from any local minimum. If the exploration process of an algorithm is poorly designed, then it cannot search the function landscape effectively. This, in turn, leads to an algorithm getting stuck at a local minimum. Multimodal functions with many local minima are among the most difficult class of problems for optimization algorithms. The results reported in Table (8.6), (8.7) show that KA algorithm has a very good exploration capability, since it is the most efficient or the second-best algorithm in the majority of test problems. This is due to integrated mechanisms of exploration in the proposed algorithm that leads this algorithm towards the global optimum.

F9 is one of the multimodal functions that is difficult to optimize. Thus, only the KA algorithm was able to reach the global optimum. The only function that the KA algorithm cannot optimize is F8, which is a complex multimodal function with many local minima and very difficult to optimize. However, as the results show, the KA optimizer outperformed the other algorithms in the optimization of this benchmark function. The results of 30 runs with KA for the multimodal functions are often better than the results found by concurrent algorithms, with the exception of PSO and CAB which show a performance almost equal to that of the F16-F20 KA results. In benchmark functions F10 and F14 the area that contains the global minima are very small, when compared to the whole search space. Optimizing this kind of functions is really challenging. In addition, KA is the only one who can optimize F9-F11 and the
results found by the KA algorithm for F12, F13, F15, F20, F23 are the best. Only KA and CAB with theirs two variants reached the solutions very close to the global minimum of the function F21-F23. For the benchmark F15 the global minimum is located very close to the local minima. This function is notoriously difficult. Furthermore, in optimization, a problem that algorithms may suffer is the scaling problem with many orders of magnitude differences between the domain and the function hyper-surface (Junior, et al., 2004), it can be seen in benchmark F18. Only KA could optimize it without any error. The results obtained with KA in Tables (8.6), (8.7) evidence that even the worst outcomes of KA outperformed those of the other algorithms in the majority of the test. The p-values reported in Tables (8.8) and (8.9) also show that the KA algorithm has significantly better results. Therefore, it can be concluded that the KA algorithm is efficient algorithm with a high level of exploration, which assists it to explore the promising regions of the search space and to avoid all of the local optima and approach the global optimum.

<table>
<thead>
<tr>
<th></th>
<th>KA Mean</th>
<th>SD</th>
<th>GA Mean</th>
<th>SD</th>
<th>PSO Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>F8</td>
<td>-3004.4716</td>
<td>400.9120</td>
<td>-2091.64</td>
<td>2.47235</td>
<td>-1367.01</td>
<td>146.4089</td>
</tr>
<tr>
<td>F9</td>
<td>0</td>
<td>0</td>
<td>0.659271</td>
<td>0.815751</td>
<td>0.278588</td>
<td>0.218991</td>
</tr>
<tr>
<td>F10</td>
<td>0</td>
<td>0</td>
<td>0.956111</td>
<td>0.807701</td>
<td>1.11e-09</td>
<td>2.39e-11</td>
</tr>
<tr>
<td>F11</td>
<td>0</td>
<td>0</td>
<td>0.487809</td>
<td>0.217782</td>
<td>0.009215</td>
<td>0.007724</td>
</tr>
<tr>
<td>F12</td>
<td>1.2623e-06</td>
<td>2.993e-06</td>
<td>0.110769</td>
<td>0.002152</td>
<td>0.009617</td>
<td>0.026301</td>
</tr>
<tr>
<td>F13</td>
<td>5.7287e-06</td>
<td>9.5067e-04</td>
<td>0.129</td>
<td>0.068851</td>
<td>0.006675</td>
<td>0.008907</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>FA Mean</th>
<th>SD</th>
<th>GSA Mean</th>
<th>SD</th>
<th>CAB Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>F8</td>
<td>-1245.59</td>
<td>353.2667</td>
<td>-1108.01</td>
<td>574.7</td>
<td>-1200</td>
<td>50.42824</td>
</tr>
<tr>
<td>F9</td>
<td>0.263458</td>
<td>0.182824</td>
<td>25.96841</td>
<td>7.470068</td>
<td>0.001</td>
<td>8.45e-06</td>
</tr>
<tr>
<td>F10</td>
<td>0.168306</td>
<td>0.050796</td>
<td>0.062087</td>
<td>0.23628</td>
<td>0.0031</td>
<td>0.003096</td>
</tr>
<tr>
<td>F11</td>
<td>0.099815</td>
<td>0.024466</td>
<td>27.70154</td>
<td>5.040343</td>
<td>0.122678</td>
<td>0.049673</td>
</tr>
<tr>
<td>F12</td>
<td>0.126076</td>
<td>0.263201</td>
<td>1.799617</td>
<td>0.95114</td>
<td>5.60e-06</td>
<td>1.58e-8</td>
</tr>
<tr>
<td>F13</td>
<td>0.00213</td>
<td>0.001238</td>
<td>8.899084</td>
<td>7.126241</td>
<td>9.5067e-06</td>
<td>6.09e-07</td>
</tr>
</tbody>
</table>

Table (8.6): Results of high-dimensional multimodal test functions
<table>
<thead>
<tr>
<th>FA</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>F14</td>
<td>1.22</td>
<td>0.56</td>
<td>5.859838</td>
<td>3.831299</td>
</tr>
<tr>
<td>F15</td>
<td>0.00058</td>
<td>0.000324</td>
<td>0.003673</td>
<td>0.001647</td>
</tr>
<tr>
<td>F16</td>
<td>-1.03163</td>
<td>4.2e-07</td>
<td>-1.03163</td>
<td>4.88e-16</td>
</tr>
<tr>
<td>F17</td>
<td>0.397914</td>
<td>2.7e-05</td>
<td>0.397887</td>
<td>0</td>
</tr>
<tr>
<td>F18</td>
<td>3</td>
<td>4.22e-15</td>
<td>3</td>
<td>4.17e-15</td>
</tr>
<tr>
<td>F19</td>
<td>-3.85616</td>
<td>0.002706</td>
<td>-3.86278</td>
<td>2.29e-15</td>
</tr>
<tr>
<td>F20</td>
<td>-3.27</td>
<td>0.059</td>
<td>-3.31778</td>
<td>0.023081</td>
</tr>
<tr>
<td>F21</td>
<td>-6.04918</td>
<td>3.629551</td>
<td>-5.95512</td>
<td>3.737079</td>
</tr>
<tr>
<td>F22</td>
<td>-8.18178</td>
<td>3.829202</td>
<td>-9.68447</td>
<td>2.014088</td>
</tr>
<tr>
<td>F23</td>
<td>-9.34238</td>
<td>2.414737</td>
<td>-10.5364</td>
<td>2.6e-15</td>
</tr>
</tbody>
</table>

Table (8.7): Results of fixed-dimensional multimodal test functions
<table>
<thead>
<tr>
<th>KA versus</th>
<th>GA</th>
<th>PSO</th>
<th>FA</th>
<th>GSA</th>
<th>CAB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F8</strong></td>
<td>6.7956e-08</td>
<td>6.7956e-08</td>
<td>6.7956e-08</td>
<td>9.1728e-08</td>
<td>6.7956e-08</td>
</tr>
<tr>
<td><strong>F9</strong></td>
<td>5.4987e-04</td>
<td>5.4955e-05</td>
<td>5.4955e-05</td>
<td>8.0065e-09</td>
<td>8.0065e-09</td>
</tr>
<tr>
<td><strong>F10</strong></td>
<td>5.4987e-04</td>
<td>8.0065e-09</td>
<td>8.0065e-09</td>
<td>0.0215</td>
<td>5.4955e-05</td>
</tr>
<tr>
<td><strong>F11</strong></td>
<td>8.0065e-09</td>
<td>5.4955e-05</td>
<td>8.0065e-09</td>
<td>8.0065e-09</td>
<td>8.0065e-09</td>
</tr>
<tr>
<td><strong>F12</strong></td>
<td>6.7956e-08</td>
<td>0.5979</td>
<td>0.0315</td>
<td>6.7956e-08</td>
<td>1.2009e-06</td>
</tr>
<tr>
<td><strong>F13</strong></td>
<td>6.7956e-08</td>
<td>1.0373e-04</td>
<td>1.4309e-07</td>
<td>1.5997e-05</td>
<td>0.1075</td>
</tr>
</tbody>
</table>

Table (8.8): P-values produced by Wilcoxon’s test comparing KA versus GA, PSO, FA, GSA and CAB over all runs for high-dimensional multimodal test functions

<table>
<thead>
<tr>
<th>KA versus</th>
<th>GA</th>
<th>PSO</th>
<th>FA</th>
<th>GSA</th>
<th>CAB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F14</strong></td>
<td>0.0388</td>
<td>6.6725e-05</td>
<td>0.0058</td>
<td>2.1882e-15</td>
<td>6.4918e-18</td>
</tr>
<tr>
<td><strong>F15</strong></td>
<td>5.8913e-04</td>
<td>3.7961e-13</td>
<td>3.2874e-08</td>
<td>5.5727e-10</td>
<td>0.3790</td>
</tr>
<tr>
<td><strong>F16</strong></td>
<td>1.2118e-12</td>
<td>1.1275e-12</td>
<td>1.2118e-12</td>
<td>1.1092e-12</td>
<td>0.6411</td>
</tr>
<tr>
<td><strong>F17</strong></td>
<td>1.2118e-12</td>
<td>1.6853e-14</td>
<td>2.2081e-06</td>
<td>1.6853e-14</td>
<td>1.2118e-12</td>
</tr>
<tr>
<td><strong>F18</strong></td>
<td>0.6411</td>
<td>0.6293</td>
<td>0.0960</td>
<td>0.6409</td>
<td>0.3393</td>
</tr>
<tr>
<td><strong>F19</strong></td>
<td>1.2118e-12</td>
<td>1.1661e-12</td>
<td>1.2118e-12</td>
<td>1.1882e-12</td>
<td>0.2706</td>
</tr>
<tr>
<td><strong>F20</strong></td>
<td>2.6816e-07</td>
<td>0.1568</td>
<td>0.9246</td>
<td>0.3794</td>
<td>0.0771</td>
</tr>
<tr>
<td><strong>F21</strong></td>
<td>3.0199e-11</td>
<td>1.1077e-06</td>
<td>1.1077e-06</td>
<td>9.5139e-06</td>
<td>3.0199e-11</td>
</tr>
<tr>
<td><strong>F22</strong></td>
<td>3.0199e-11</td>
<td>9.5139e-06</td>
<td>1.1077e-06</td>
<td>0.0315</td>
<td>1.5997e-05</td>
</tr>
<tr>
<td><strong>F23</strong></td>
<td>1.1077e-06</td>
<td>9.5139e-06</td>
<td>0.0271</td>
<td>0.0263</td>
<td>0.1537</td>
</tr>
</tbody>
</table>

Table (8.9): P-values produced by Wilcoxon’s test comparing KA versus GA, PSO, FA, GSA and CAB over all runs for fixed-dimensional multimodal test functions
Figure (8.17): Benchmark function F8-The convergence rate, average fitness of all individuals, trajectory and the fitness history.

Figure (8.18): Benchmark function F9-The convergence rate, average fitness of all individuals, trajectory and the fitness history.
Figure (8.19): Benchmark function F10-The convergence rate, average fitness of all individuals, trajectory and the fitness history.

Figure (8.20): Benchmark function F11-The convergence rate, average fitness of all individuals, trajectory and the fitness history.
Figure (8.21): Benchmark function F12-The convergence rate, average fitness of all individuals, trajectory and the fitness history.

Figure (8.22): Benchmark function F13-The convergence rate, average fitness of all individuals, trajectory and the fitness history.
Figure (8.23): Benchmark function F14-The convergence rate, average fitness of all individuals, trajectory and the fitness history.

Figure (8.24): Benchmark function F15-The convergence rate, average fitness of all individuals, trajectory and the fitness history.
Figure (8.25): Benchmark function F16-The convergence rate, average fitness of all individuals, trajectory and the fitness history.

Figure (8.26): Benchmark function F17-The convergence rate, average fitness of all individuals, trajectory and the fitness history.
Figure (8.27): Benchmark function F18-The convergence rate, average fitness of all individuals, trajectory and the fitness history.

Figure (8.28): Benchmark function F19-The convergence rate, average fitness of all individuals, trajectory and the fitness history.
Figure (8.29): Benchmark function F20-The convergence rate, average fitness of all individuals, trajectory and the fitness history.

Figure (8.30): Benchmark function F21-The convergence rate, average fitness of all individuals, trajectory and the fitness history.
Figure (8.31): Benchmark function F22-The convergence rate, average fitness of all individuals, trajectory and the fitness history.

Figure (8.32): Benchmark function F23-The convergence rate, average fitness of all individuals, trajectory and the fitness history.
8.3.2. Analysis of KA Optimizer

This subsection aims for identifying and confirming the convergence and potential behavior of the KA algorithm when solving real problems. Although the results discussed in the previous section prove the high performance of the KA algorithm, there are several quantitative metrics, which can be employed to confidently confirm the performance of this algorithm in solving real problems such as actuator selection. In other words, the change of populations of the species and the obtained fitness values during optimization should be monitored to observe:

- how the average fitness of search agents changes from the first to the last iteration
- how the individuals converge towards the global optima in the search space
- how the algorithm improves the initial random solutions, and their fitness values over the iterations
- if there are abrupt changes in the populations of species in the initial stages of optimization to explore the search space and gradual changes in the final steps of iteration to exploit the search space

The second plots of the first column in Figures (8.10)-(8.32) show the fluctuations of the variables in the search agents. It can be seen that the populations of the species face abrupt fluctuations in the early steps of optimization. However, the sudden changes are decreased gradually over the iterations. This confirms that the search candidates first explore the search space to find the promising regions and then converge around the best solution obtained in the exploration phase. There is a question here as how to make sure that all of the search agents are improved during optimization despite the rapid and steady changes in the aforementioned plots. In order to confirm the improvement of all solutions, the average fitness of all search agents during optimization is illustrated in the first plots of the second column in Figures (8.10)-(8.32). It can be seen that the average fitness of all individuals tends to be decreased over the course of iterations. The interesting pattern that can be observed in this plots is the high fluctuation of the average fitness in the exploration phase and low changes in the average fitness in the last iterations, when the exploitation phase begins. Deterioration of the fitness of some of the search agents is inevitable in the exploration. However, the general pattern in the average plots illustrates that the fitness of candidates has a descending behavior over the iterations. This proves that the proposed KA algorithm is able to eventually improve the fitness of initial random solutions for a given optimization problem. In the previous section, it was claimed that the search agents of the KA algorithm tend explore the promising regions of the search space and exploit the best one eventually. However, the convergence behavior of the algorithm was not observed and verified. Although this can be inferred indirectly from the trajectory and average fitness, the convergence curve of KA is depicted in the first plots of Figures (8.10)-(8.32). These plots illustrate the best solutions obtained so far by KA during optimization. The descending trend is quite evident in the convergence curve of KA on all of the test functions investigated. This strongly evidences the ability of the proposed algorithm in obtaining a better approximation of the global optimum over the course of iterations. All the results and discussions of this section prove that the KA algorithm is suitable for solving real-world optimization problems. This is due to the integrated adaptive parameter selection scheme in the algorithm and employed mechanism of updating the best feasible solutions obtained during the change of populations of individuals, in which
the species are guided towards promising feasible regions of the search space. It leads to boosting the exploration of the feasible areas of a search space and therefore, KA can solve the real-world challenging optimization problems efficiently with low computational cost.

8.3.3. Optimization Problems Using the KA Algorithm

As the proposed algorithm considers optimization problem as black-boxes, it is readily incorporable to problems in different fields subject to proper problem formulation. In this section, first, some well-studied engineering design problems that have been solved by various optimization methods in the literature are used to examine the efficiency of KA algorithm. Then we used the KA algorithm to solve the problem of optimal actuator placement in some real-world examples.

**Example (8.1):** (Kaveh & Talatahari, 2010) Consider the well-known optimization problem of tension/compression spring design as depicted in Figure (8.34). The objective of this optimization problem is to minimize the weight of tension/compression spring. The optimization problem can be formulated as:

\[
\begin{align*}
\min J(x) &= (x_3 + 2)x_1^2 x_2 \\
h_1(x) &= 1 - \frac{x_2^3 x_3}{71785 x_1^4} \leq 0 \\
h_2(x) &= \frac{4x_2^2 - x_2 x_3}{12566(x_2 x_1^2 - x_1^2)} - \frac{1}{5108 x_1^2} \leq 0 \\
h_3(x) &= 1 - \frac{140.45 x_1}{x_2^2 x_3} \leq 0 \\
h_4(x) &= \frac{x_1 + x_2}{1.5} - 1 \leq 0 \\
0.05 \leq x_1 \leq 2, \quad 0.25 \leq x_2 \leq 1.3, \quad 2 \leq x_3 \leq 15
\end{align*}
\]

Here, the decision variables are the mean coil diameter \(D = x_1\), the wire diameter \(d = x_2\), and the number active coils \(N = x_3\). The results of KA optimization method are illustrated in Table (8.10) and compared to some existing results of the other optimization methods.
Example (8.2): Consider a pressure vessel design problem (Sandgren, 1988), (Kaveh, 2016), where the objective is to minimize the cost of fabricating a pressure vessel which is clapped at both ends by hemispherical heads as depicted in Figure (8.35). The design variables are the
thickness of the shell $T_s(=x_1)$, the thickness of the head $T_h(=x_2)$, the inner radius $R(=x_3)$ and the length of cylindrical section of the vessel $L(=x_4)$. The optimization problem can be defined as:

$$\begin{align*}
\min \ J(x) &= 0.622x_1x_3x_4 + 1.7781x_1^2x_4^2 + 3.1661x_1^4 + 19.84x_1^2x_3 \\
h_1(x) &= -x_1 + 0.0193x_3 \leq 0 \\
h_2(x) &= -x_2 + 0.0095x_3 \leq 0 \\
h_3(x) &= -\pi x_3^2x_4 - \frac{4}{3}\pi x_3^3 + 1296000 \leq 0 \\
h_4(x) &= x_4 - 240 \leq 0 \\
0 \leq x_1 \leq 99, & \quad 0 \leq x_2 \leq 99, & \quad 10 \leq x_3 \leq 200, & \quad 10 \leq x_4 \leq 200
\end{align*}$$

Figure (8.35): A pressure vessel, and its design variables

The results of KA optimization method are illustrated in Table (8.11) and compared to some existing results of the other optimization methods.

![Convergence curve of KA in Example (8.2)](image-url)

Figure (8.36): Convergence curve of KA in Example (8.2)
Table (8.11): The numerical results of Example (8.2)

<table>
<thead>
<tr>
<th>Method</th>
<th>$T_s$</th>
<th>$T_h$</th>
<th>$R$</th>
<th>$L$</th>
<th>$J_{opt}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KA Optimization</td>
<td>0.804799</td>
<td>0.4026198</td>
<td>41.68858</td>
<td>181.8004</td>
<td>5949.0906</td>
</tr>
<tr>
<td>PSO (He &amp; Wang, 2007)</td>
<td>0.8125</td>
<td>0.4375</td>
<td>42.091266</td>
<td>176.746500</td>
<td>6061.0777</td>
</tr>
<tr>
<td>GA (Coello, 2000)</td>
<td>0.8125</td>
<td>0.4345</td>
<td>40.323900</td>
<td>200.000000</td>
<td>6288.7445</td>
</tr>
<tr>
<td>GA (Coello &amp; Montes, 2002)</td>
<td>0.8125</td>
<td>0.4375</td>
<td>42.097398</td>
<td>176.654050</td>
<td>6059.9463</td>
</tr>
<tr>
<td>GA (Deb, 1997)</td>
<td>0.9375</td>
<td>0.5000</td>
<td>48.329000</td>
<td>112.679000</td>
<td>6410.3811</td>
</tr>
<tr>
<td>ES (Montes &amp; Coello, 2008)</td>
<td>0.8125</td>
<td>0.4375</td>
<td>42.098087</td>
<td>176.640518</td>
<td>6059.7456</td>
</tr>
<tr>
<td>DE (Li, et al., 2007)</td>
<td>0.8125</td>
<td>0.4375</td>
<td>42.098411</td>
<td>176.637690</td>
<td>6059.7340</td>
</tr>
<tr>
<td>ACO (Kaveh &amp; Talatahari, 2010)</td>
<td>0.8125</td>
<td>0.4375</td>
<td>42.103624</td>
<td>176.572656</td>
<td>6059.0888</td>
</tr>
<tr>
<td>Lagrangian Multiplier</td>
<td>1.1250</td>
<td>0.6250</td>
<td>58.291000</td>
<td>43.6900000</td>
<td>7198.0428</td>
</tr>
<tr>
<td>(Kannan &amp; Kramer, 1994)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Branch-bound (Sandgren, 1988)</td>
<td>1.1250</td>
<td>0.6250</td>
<td>47.700000</td>
<td>117.701000</td>
<td>8129.1036</td>
</tr>
</tbody>
</table>

In the next example, we apply KA optimization method to solve the problem of designing a robust PID controller. The results obtained are compared to those of ALPSO (Bellman, 2003). The problem is mathematically formulated as follows:

$$\min J(x) = \arg \max_{\lambda_i(x)} \{ \Re(\lambda_i(x)), \forall i \}$$

$$\left\{ \begin{array}{l}
h_i(x) = \sup_{a>0} \left[ W_s(s) \right]_{\text{Re}^+} \left[ S(s,x) \right]_{\text{Re}^+} - 1 \leq 0 \\
\text{s.t.} \ h_2(x) = \sup_{a>0} \left[ W_r(s) \right]_{\text{Re}^+} \left[ T(s,x) \right]_{\text{Re}^+} - 1 \leq 0 \\
\text{lb}_i \leq x_i \leq \text{ub}_i, \ i = 1, \ldots, 4
\end{array} \right.$$  

(8.23)

where $\lambda_i(x)$ denotes the $i$th pole of the closed-loop system. $W_s(s)$ and $W_r(s)$ are weighting matrices.
\[ x = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}^T \] is the vector of variables, \( l_b, u_b \) are the lower and upper bounds respectively. \( s \) is the Laplace variable, \( \omega \) is the frequency (rad/s). \( S(s, x) \) denotes the sensitivity function defined as:

\[ S(s, x) = \frac{1}{1 + L(s, x)}. \]

\( T(s, x) \) is the closed-loop system, i.e.:

\[ T(s, x) = \frac{L(s, x)}{1 + L(s, x)} \]

and \( L(s, x) \) is the open-loop transfer function:

\[ L(s, x) = P(s) K(s, x) \]

where \( P(s) \) is the transfer function of the system and \( K(s, x) \) the transfer function of the PID controller:

\[ K(s, x) = 10^{\frac{1}{5}} \left( 1 + 10^{-s} s^{-1} + \frac{10^{\frac{1}{5}} s}{1 + 0^{(s-x_1)}} s \right) \]

This optimization problem has been solved for a magnetic levitation system in the next example. For more details see (Bellman, 2003) and references therein.

**Example (8.3):** The transfer function of a magnetic levitation system is defined as:

\[ P(s) = \frac{7.147}{(s - 22.55)(s + 20.9)(s + 13.99)} \]

The objective is to design the robust PID controller:

\[ K(s, x) = 10^{\frac{1}{5}} \left( 1 + 10^{-s} s^{-1} + \frac{10^{\frac{1}{5}} s}{1 + 0^{(s-x_1)}} s \right) \]

Let us define the weighting matrices in (8.23) as:

\[ W_s(s) = \frac{5}{s + 0.1}, \quad W_f(s) = \frac{43.867(s + 0.066)(s + 88)(s + 31.4)}{(s + 10^5)^2} \]

Consider the search space as follows:

\[ 2 \leq x_1 \leq 4, \quad -1 \leq x_2 \leq 1, \quad -1 \leq x_3 \leq 1, \quad 1 \leq x_4 \leq 3 \]

The results are shown in Table (8.12), from which we can observe that the better value of the objective function obtained with ALPSO is due to the violation of the constraint \( h_1(x) \); this is not the case at all in our solution for which all constraints are satisfied. Therefore, the best solution found by KA method is significantly better than the solution found by ALPSO. Figure (8.37) shows the convergence curve of the KA optimization for this example.

<table>
<thead>
<tr>
<th></th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>( h_1(x) )</th>
<th>( h_2(x) )</th>
<th>( J_{opt} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>KA optimization</td>
<td>3.2614</td>
<td>-0.8085</td>
<td>-0.7433</td>
<td>2.3026</td>
<td>-9.1778e-04</td>
<td>-0.0045</td>
<td>-1.7124</td>
</tr>
<tr>
<td>ALPSO</td>
<td>3.2548</td>
<td>-0.8424</td>
<td>-0.7501</td>
<td>2.3137</td>
<td>6.1 e-03</td>
<td>-4 e-04</td>
<td>-1.7197</td>
</tr>
</tbody>
</table>

**Table (8.12):** The numerical results of Example (8.3)
Example (8.4): Consider the control of the mass-damper-spring system shown in the figure below:

By applying the fundamental principle of dynamics, we get the following equation of motion 
\[ m\ddot{x} + c\dot{x} + kx = u \]
where \( x \) is the position of the mass from its equilibrium position, and \( u \) is the control variable which represents the force applied to the mass. The state space representation of the system is then given by:

\[
\begin{bmatrix}
\dot{x} \\
y
\end{bmatrix} =
\begin{bmatrix}
0 & 1 \\
-k/m & -c/m
\end{bmatrix}
\begin{bmatrix}
x \\
u
\end{bmatrix}
\]

where \( x_1 \) is the position and \( x_2 \) the velocity of the mass. The parameters \( m, c \) and \( k \) are not known exactly, but it is assumed that their values are within the following known intervals: 
\( m \in [m_{\text{min}}, m_{\text{max}}] \), \( c \in [c_{\text{min}}, c_{\text{max}}] \) and \( k \in [k_{\text{min}}, k_{\text{max}}] \). These parameters can be also written as:
The block diagram of the system is shown below:

\[ m = \frac{m_{\text{max}} + m_{\text{min}}}{2}, \quad \| \delta_m \| \leq 1 \]
\[ c = \frac{c_{\text{max}} + c_{\text{min}}}{2}, \quad \| \delta_c \| \leq 1 \]
\[ k = \frac{k_{\text{max}} + k_{\text{min}}}{2}, \quad \| \delta_k \| \leq 1 \]

Then, the corresponding uncertain system can be described as follows:

\[
\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2 \\
y
\end{bmatrix} = \hat{G} \begin{bmatrix} x_1 \\
x_2 \\
u \end{bmatrix}
\]

\[
\hat{G} \in \Gamma = \left\{ G(\Delta) = F_x^\text{sys} (\Delta) : \Delta = \begin{bmatrix} \delta_m & 0 & 0 \\
0 & \delta_c & 0 \\
0 & 0 & \delta_k \end{bmatrix}, \| \Delta \| \leq 1 \right\}
\]

where the matrices \( \text{sys} \) and \( \Delta \) are:

\[
\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2 \\
q_m \\
q_c \\
q_k \\
y
\end{bmatrix} = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
-\frac{k_v}{m_o} & -\frac{c_v}{m_o} & -\frac{m_d}{m_o} & -\frac{c_d}{c_m} & -\frac{k_d}{k_m} & 1/m_o & 0 \\
-\frac{k_v}{m_o} & -\frac{c_v}{m_o} & -\frac{m_d}{m_o} & -\frac{c_d}{c_m} & -\frac{k_d}{k_m} & 1/m_o & 0 \\
0 & c_0 & 0 & 0 & 0 & 0 & 0 \\
k_v & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix} \begin{bmatrix}
x_1 \\
x_2 \\
v_m \\
v_c \\
v_k \\
u \end{bmatrix}
\]

where:
Hence, the extended or augmented system $G_{aug}$, which is a perfectly known LTI system is defined as:

$$
\begin{bmatrix}
q_m \\
q_c \\
q_k \\
v_m \\
v_c \\
v_k
\end{bmatrix} = \Delta
\begin{bmatrix}
v_m \\
v_c \\
v_k
\end{bmatrix}
$$

Therefore, the multi-objective robust structured control design problem can be stated as follows. For the uncertain system $G \in \Gamma$, defined above, find a robust PID controller $K$, which minimizes the decay rate while ensuring the robust stability of the closed-loop system. The optimization problem in KA can be formulated based on the minimization of the maximum decay rate as:
or based on the average decay rate minimization as follows:

\[
\begin{align*}
\min_k \alpha \\
\Phi\left(\left[\tilde{G}^{(i)}, K\right]\right) &= \max_{1 \leq j \leq n} \Re\left[\lambda_j(A_i(K, \Delta^{(i)}))\right] \\
\max_{1 \leq q \leq Q} \left\{\Phi\left(\left[\tilde{G}^{(q)}, K\right]\right)\right\} &\leq \alpha < 0 \\
\tilde{G}^{(i)} &\in \left[\tilde{G}^{(1)}, \tilde{G}^{(2)}, \ldots, \tilde{G}^{(q)}\right] \\
K &\in K_G \\
0 &\leq K_p \leq 150, 0 \leq K_i \leq 150, 0 \leq K_D \leq 150, 0.0001 \leq \tau \leq 0.1
\end{align*}
\]

\(A_{ij}\) is the state matrix of the nominal closed-loop system and \(\lambda_j, i = 1, \ldots, n\) is the \(i^{th}\) eigenvalue of \(A_{ij}\). \(K_G\) is the set of structured controllers that stabilize the uncertain system \(\tilde{G}\). \(\eta\) denotes the number of samples, which can be determined using the inequality below (Toscano, 2013):

\[
\eta \geq \ln(1 - \rho)/\ln(1 - e)
\]

where \(\rho \in (0, 1), e \in (0, 1)\) are two given numbers, which state with a confidence \(\rho\), the obtained solution is a probable global minimum of the objective function to level \(e\). \(\tilde{G}^{(i)}\) is the sample system with the uncertainty \(\Delta^{(i)}\). Table (8.13) shows the results of the worst-case decay rate optimization and the results of the average decay rate optimization are also given in Table (8.14). The obtained results evidence that KA algorithm outperforms the other methods.

<table>
<thead>
<tr>
<th></th>
<th>(K_p)</th>
<th>(K_i)</th>
<th>(K_D)</th>
<th>(\tau)</th>
<th>(J_{opt})</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KA Optimization</strong></td>
<td>150</td>
<td>58.12941</td>
<td>131.5966</td>
<td>0.0007594358</td>
<td>-0.71043</td>
</tr>
<tr>
<td><strong>HKA</strong></td>
<td>150</td>
<td>59.14118</td>
<td>132.7368</td>
<td>0.0001095862</td>
<td>-0.62616</td>
</tr>
<tr>
<td><strong>GA</strong></td>
<td>148.1950</td>
<td>48.5724</td>
<td>148.6651</td>
<td>0.0064</td>
<td>-0.58759</td>
</tr>
</tbody>
</table>

Table (8.13): Numerical results for worst-case optimization case of Example (8.4)
The next example discusses the problem of optimal distribution of actuators in a three-story building. Let us assume the actuators are the viscous dampers. Viscous dampers greatly reduce the earthquake excitation of a structure and permit it to remain linearly elastic during a seismic event. They work by adding damping to a structure, which significantly lowers its resonant response to an earthquake and reduces stress and deflection because the force from the damping is completely out of phase with stresses. This provides a significant decrease in earthquake excitation (Lee, 2016). By installing as many dampers as necessary, damping can be selectively inserted into the structure at optimal positions. Several dangerous modes are effectively reduced, and resonance effects are eliminated. Therefore, damper locations can be regarded as the problem of maximizing the efficiency taking into account the energy dissipated in dampers as well as minimizing the number of devices, which has great influence on costs in case of retrofitting. (Wu & Ou, 1997), (Takewaki, et al., 1999). Clearly, minimum vibration magnitudes are criteria for the effectiveness (Takewaki, 2000).

**Example (8.5):** Consider the three-story building presented in (Gluck, et al., 1996), (Takewaki, 1997):

\[
M \ddot{x}(t) + C \dot{x}(t) + Kx(t) = -M \dddot{x}_g + \Gamma u(t)
\]

\[
M = \begin{bmatrix}
200.4 & 0 & 0 \\
0 & 200.4 & 0 \\
0 & 0 & 178
\end{bmatrix} (Kg) , \quad K = 10^7 \times \begin{bmatrix}
238.932 & -119.466 & 0 \\
-119.466 & 238.932 & -119.466 \\
0 & -119.466 & 119.466
\end{bmatrix} (N/m)
\]

\[
C = \begin{bmatrix}
264.99 & -78.09 & -16.08 \\
-78.09 & 246.89 & -92.15 \\
-16.08 & -92.15 & 162.02
\end{bmatrix}
\]
where \( \mathbf{x} \) is the \( n \)-dimensional displacement vector relative to the ground. \( \mathbf{M}, \mathbf{C}, \mathbf{K} \) are the mass, damping and stiffness matrices respectively. \( \mathbf{\Gamma} \) is the \( n \times m \) distribution matrix for the damper force \( \mathbf{u}(t) \) and \( \mathbf{l} \) is an \( n \)-dimensional distribution vector for the ground acceleration \( \ddot{\mathbf{x}}_g \), with all elements being unity. \( m \) denotes the number of actuators.

An augmented system state equation can be formulated as:

\[
\dot{\mathbf{z}} = \mathbf{A}\mathbf{z} + \mathbf{B}\mathbf{u} + \mathbf{W}
\]

\[
\mathbf{A} = \begin{bmatrix} 0 & I \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{C} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ \mathbf{M}^{-1}\mathbf{\Gamma} \end{bmatrix}
\]

Here, \( \mathbf{z} = \begin{bmatrix} \mathbf{x}^T & \ddot{\mathbf{x}}_g^T \end{bmatrix}^T \) is the state vector. \( \mathbf{W} = [0 \quad -\mathbf{1}] \ddot{\mathbf{x}}_g \) denotes the earthquake induced disturbance.

The response of the building structure, under the effect of random earthquake excitation, largely depends on the system matrix \( \mathbf{A} \) and the control effort \( \mathbf{u} \). In fact, installation of dampers will affect the control gain matrix \( \mathbf{B} \). A zero entry of \( \mathbf{\Gamma} \) is related to the absence of a damper installed at the corresponding story.

As (Hac \& Liu, 1993) states the controllability Gramian \( \mathbf{W}_c(0,t_f) \) of the above system tends to a diagonal form when the terminal time, \( t_f \), is large enough. Therefore, we have:

\[
(W_c)_{ii} = (W_c^u)_{ii} = \sum_{j=1}^{m} \frac{b_{ij}^2}{4\xi_j \omega_j} = \frac{1}{4\xi_j \omega_j} \sum_{j=1}^{m} b_{ij}^2 \quad i = 1, \ldots, N
\]

and

\[
(W_c^{r})_{ii} = (W_c^r)_{ii} = \sum_{j=1}^{m} \frac{(b_{ij}^{r})^2}{4\xi_j^{r} \omega_j^{r}} = \frac{1}{4\xi_j^{r} \omega_j^{r}} \sum_{j=1}^{m} (b_{ij}^{r})^2 \quad i = 1, \ldots, N^{r}
\]
where $(W^T)_{ii}$, $(W^r)_{ii}$ equal to the energy transmitted from the actuators to the structure for the $i^{th}$ used and residual eigenmode. $N$ denotes the number of the eigenmodes, which are considered, and $N'$ is the number of residual eigenmodes.

Hence, according to the previous chapters one approach to determine the actuator placement is to minimize the control energy required by maximizing a measure of the controllability Gramian $W_c$.

However, one must consider that if one eigenvalue of $(W^r)_{ii}$ corresponding to a residual mode is high, the induced spill-over effect can be important. As our objective is to control the modes considered, without exciting residual modes, by transmitting a maximum control force with a minimum energy, and to achieve optimum suppression of quake-induced vibrations, the control problem is cast as a multi-objective optimization process. To manage this in our algorithm, we defined the objective function as a single fitness by the use of importance weights. Thus, the optimization problem can be defined as follows (Arbel, 1981), (Hac & Liu, 1993), (Collet, 2001):

$$
\min_b J = \alpha_1 \max_{i=1, \ldots, N'} (W^r)_{ii} - \alpha_2 \min_{i=1, \ldots, N} (W^u)_{ii} + \alpha_3 \int_0^{t_r} z^T z \, dt
$$

Here $\alpha_i, \ i = 1, 2, 3$ is a weighting constant, $z$ is the state vector of displacements and velocities and $\int_0^{t_r} z^T z \, dt$ reflects upon the desire to minimize the structural response. This optimization problem has $m$ decision variables (the locations of the actuators) and $N + N'$ optimization functions. This criterion ensures global controllability of the system for the $N$ first eigenmodes and try to minimize the global excitation of the $N'$ residual modes.

In this example, we considered El Centro earthquake, on soil, with a peak ground acceleration of 0.30g, and the North-East (21º) component recorded during the 1952 Kern County earthquake, Taft, on rock, with a peak ground acceleration of 0.16g.

According to (Bruant & Proslier, 2005), as the components of $W_c$ have not the same range then the above optimization problem becomes, by using the homogeneous components:

$$
\min_b J = \alpha_1 \max_{i=1, \ldots, N'} (W^r)_{ii} - \alpha_2 \min_{i=1, \ldots, N} (W^u)_{ii} + \alpha_3 \int_0^{t_r} z^T z \, dt
$$

Hence, the objective function in the KA algorithm can be formulated as:

$$
\min_b J = \alpha_1 \max_{i=1, \ldots, N'} \frac{\sum_{j=1}^{m} (b^r_j)^2}{\max_{i=1, \ldots, N} \sum_{j=1}^{m} (b^u_j)^2} - \alpha_2 \min_{i=1, \ldots, N} \frac{\sum_{j=1}^{m} b^2_j}{\max_{i=1, \ldots, N} \sum_{j=1}^{m} b^2_j} + \alpha_3 \int_0^{t_r} z^T z \, dt
$$

or (regarding the type of the selected seismic, Elcentro):
\[
\min_b J = \alpha_1 \max_{j=1, \ldots, \text{number of floors}} \frac{\sum_{j=1}^{m} (b_{ij})^2}{\max_{\text{locations}} \sum_{j=1}^{m} (b_{ij})^2} - \alpha_2 \min_{j=1, \ldots, \text{number of floors}} \frac{\sum_{j=1}^{m} b_{ij}^2}{\max_{\text{locations}} \sum_{j=1}^{m} b_{ij}^2} + \alpha_3 \sum_{k=1}^{\text{number of floors}} \max_i (z_k)
\]

In this example, for simplicity, we consider that we are given a fixed number of the actuators. We are interested in finding the optimal distribution for the available actuators. The results are shown in Figures (8.43)-(8.52). Figure (8.42) demonstrates the fluctuation in the convergence curve of the optimization due to the multi-objective intrinsic of the problem. The results of this particular example show that when only one damper is placed this should be located at the first story in order to obtain the best overall performance. When more than one damper is available, the best damper placement is one damper per story; if the number of dampers is less than the number of stories, one damper per story beginning at the lowest story is the best choice.

Figure (8.42): The best objective values of Example (8.5) using KA optimization
Figure (8.43): Top-floor displacement response of the 3-storey building

Figure (8.44): Maximum displacement of the floors (One damper)
Figure (8.45): Inter-story displacement response of the floors for one damper in different locations

Figure (8.46): Acceleration of the floors using two dampers
Figure (8.47): Floors’ maximum displacement (Two dampers)

Figure (8.48): Inter-story displacement response of the floors for two dampers in different locations
Figure (8.49): Acceleration of the floors using two dampers

Figure (8.50): Floors’ maximum displacement (Three dampers)
Figure (8.51): Inter-story displacement response of the floors for three dampers in different locations

Figure (8.52): Acceleration of the floors using three dampers
In the next example, we consider the actuator placement problem for uncertain linear systems, where the aim is to minimize the control input energy to drive the system from an arbitrary initial condition on the unit sphere to the origin in the worst case, i.e., $\min_{B \in \beta} E_{\text{max}}$.

$E_{\text{max}}$ depicts the maximum value of the minimum input energy and $\beta$ is the set of all possible input matrices. According to the previous chapters, the above optimization problem can be viewed as the optimization problem of minimizing the inverse of the smallest eigenvalue of the associated controllability Gramian via the placement of an actuator.

In spite of its simple and natural formulation, this problem is difficult to solve.

**Example (8.6):** (Sojoudi, et al., 2009) (Sojoudi, et al., 2009 (b)) Consider the uncertain fourth-order LTI system below:

$$
A(\alpha_1, \alpha_2) =
\begin{bmatrix}
-0.65 & -0.6 & -1 & 1.5 \\
-0.05 & -0.95 & -1.85 & -0.4 \\
-0.7 & 1.6 & -3.05 & 0.2 \\
-1.5 & -0.1 & -2.85 & -0.35
\end{bmatrix}
+ \begin{bmatrix}
\alpha_1 & & & \\
\alpha_2 & & & \\
\alpha_1 & & & \\
\alpha_2 & & &
\end{bmatrix}
\begin{bmatrix}
-0.75 & 1.4 & 0 & 2.3 \\
-1.35 & -1.75 & -0.65 & -1.1 \\
0.8 & 2.2 & -3.25 & -2.5 \\
-0.5 & 1.7 & 0.15 & 0.45
\end{bmatrix}
$$

$$
B(\alpha_1, \alpha_2) =
\begin{bmatrix}
-1.2 & -0.3 & -0.85 & 1.55 \\
0.35 & -1.05 & -1.3 & -1.8 \\
0.25 & 1.2 & -2.5 & -0.8 \\
-0.3 & 1.45 & -1 & -0.3
\end{bmatrix}
+ \begin{bmatrix}
\alpha_1 & & & \\
\alpha_2 & & & \\
\alpha_1 & & & \\
\alpha_2 & & &
\end{bmatrix}
\begin{bmatrix}
0.45 & 1.7 & 0.85 & 0.75 \\
-1.7 & -0.7 & 0.65 & 0.7 \\
0.55 & 1 & -0.75 & -1.7 \\
-0.2 & 0.25 & 1.15 & 0.75
\end{bmatrix}
$$

where $\alpha_1$ and $\alpha_2$ are the uncertain parameters of the system, which belong to the polytope:

$$S = \{(\alpha_1, \alpha_2) | \alpha_1 + \alpha_2 = 1, \alpha_1, \alpha_2 \geq 0\}$$

Regard $A(\alpha_1, \alpha_2)$ is robustly Hurwitz over the region $S$. Assume that four actuators are available, the aim is to find the minimum set of actuators, which play a vital role in controlling the system. To this end, for the given set of actuator $s_{\text{actuator}} \subset \{1, 2, 3, 4\}$, we consider four cases as follows:

- $\{1, 2, 3, 4\}$
- $\{1, 2, 3\}$
- $\{2, 3, 4\}$
- $\{1, 4\}$

The optimization problem can be formulated using SDP as:

$$
\min_{\alpha_1, \alpha_2} x
\begin{cases}
W_c(\alpha_1, \alpha_2) - xI \geq 0 \\
x > 0 \\
\alpha_1 + \alpha_2 = 1, \alpha_1, \alpha_2 \geq 0
\end{cases}
$$
Here, $W_c(\alpha_1, \alpha_2)$ denotes the controllability Gramian, which is dependent to the values of the parameters $(\alpha_1, \alpha_2) \in \mathcal{S}$.

The optimal results obtained by applying KA algorithm are given in Table (8.15).

<table>
<thead>
<tr>
<th>Actuator Set</th>
<th>$(\alpha_1, \alpha_2)$</th>
<th>$\text{obj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${1, 2, 3, 4}$</td>
<td>$0.23665, 0.25416$</td>
<td>$0.25311$</td>
</tr>
<tr>
<td>${1, 2, 3}$</td>
<td>$0.22288, 0.15434$</td>
<td>$0.19853$</td>
</tr>
<tr>
<td>${2, 3, 4}$</td>
<td>$0.23629, 0.4307$</td>
<td>$0.074542$</td>
</tr>
<tr>
<td>${1, 4}$</td>
<td>$0.82059, 0.172$</td>
<td>$0.0074269$</td>
</tr>
</tbody>
</table>

Table (8.15): Numerical results for Example (8.6)

The results show that the last actuator can be neglected due to its weak contribution to the value of the minimum input energy in the worst case. In contrast, the first actuator is not neglectable as removing it increases the maximum value of the minimum input energy drastically. Moreover, in the case that the second and third actuators are removed concurrently, although the system remains robustly controllable by the remaining inputs, the required control energy would be huge.

The resulting input and state of the system for the interval $t \in [-10, 0]$ are plotted in Figures (8.53)-(8.56), which confirm the results obtained in Table (8.15). One can observe that if the complete set of actuators, i.e. $s_{\text{actuator}} = \{1, 2, 3, 4\}$ is selected the worst-case optimal input of the system has the minimum overshoot occurring at $t = 0$ (about 2 in magnitude) while if both actuators 2 and 3 are removed, i.e. $s_{\text{actuator}} = \{1, 4\}$, the worst-case optimal input of the system has a very large overshoot (about 10 in magnitude).

Figure (8.53): The input and state of the system with the actuator set $s_{\text{actuator}} = \{1, 2, 3, 4\}$.
Example (8.7): The purpose of this example is to accomplish the pole placement, finding the optimal location of actuators in a mechanical system to control the vibration response with the minimum control effort.

Consider the open-loop system shown in figure below:
Define the mass damping and stiffness matrix for this system as:

\[
M = \begin{bmatrix}
1 & 0 & 0 \\
0 & 2 & 0 \\
0 & 0 & 1
\end{bmatrix}, \quad C = \begin{bmatrix}
0.2 & -0.2 & 0 \\
-0.2 & 0.2 & 0 \\
0 & 0 & 0
\end{bmatrix}, \quad K = \begin{bmatrix}
15 & -10 & 0 \\
-10 & 15 & -5 \\
0 & -5 & 5
\end{bmatrix}
\]

We wish to assign the eigenvalues of the system to the set:

\[\Lambda = \{-1, -3, -0.5 \pm 2i, -0.75 \pm 5i\}\]

by using optimal actuation with the minimum control effort. Therefore, the optimization problem can be defined as:

\[
\begin{aligned}
\min_b & \quad \|F\| \\
\text{s.t.} & \quad |A_{cl} - \lambda I| = 0, \quad \forall \lambda \in \Lambda \\
& \quad \|b\| \leq 1
\end{aligned}
\]

where \(A_{cl}\) denotes the closed-loop system and \(F\) is the associated feedback gain. Note that the force selection vector could theoretically be made very large to allow the position and velocity gain vectors to be very small, with the same control effect. However, physically this would not minimize the actuation needed. To consider this, we defined the last constraint \(\|b\| \leq 1\). Using the KA optimization algorithm, we obtain:

\[
b_{opt} = \begin{bmatrix}
0.5828 \\
-0.3996 \\
0.7076
\end{bmatrix}, \quad F_{opt} = \begin{bmatrix}
-10.8490 \\
14.0029 \\
-2.7915
\end{bmatrix} \begin{bmatrix}
2.3196 \\
-7.4201 \\
-11.4613
\end{bmatrix}
\]
Figure (8.59) shows the change of cost function when the entries of the input matrix $b$ change over the interval $[-1,1]$. The graph of cost function in the neighborhood of the input matrix $b$ associated with the global optimization is also shown in Figure (8.60).

The validity of the optimization results is investigated using brute force method. The values of $b_1$, $b_2$ are continuously changed between -1 and 1 with a step of 0.01 and $b_3 = \sqrt{b_1^2 + b_2^2}$, in
this case, the minimum value of the feedback gain associated with the desired eigenvalues is obtained when the input matrix $b$ is:

$$b = \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix}^T = \begin{bmatrix} 0.5800 & -0.4000 & 0.7096 \end{bmatrix}^T$$

It yields the feedback gain vector as follows:

$$F = \begin{bmatrix} -10.9032 & 14.0101 & -7.3940 & 2.2987 \end{bmatrix} - \begin{bmatrix} 2.8320 & -11.4136 \end{bmatrix}$$

Furthermore, given an initial position and velocity of:

$$x_0 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \quad \dot{x}_0 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

using the optimal input $b$ and the associated gain vector, the control input necessary over time is shown in Figure (8.61). To compare the results, Figure (8.62) demonstrates the control input for some other random selected input matrices:

$$b^1 = \begin{bmatrix} 0.6948 & 0.3171 & 0.9502 \end{bmatrix}, \quad b^2 = \begin{bmatrix} 0.0344 & 0.4387 & 0.3816 \end{bmatrix}$$

$$b^3 = \begin{bmatrix} 0.7655 & 0.7952 & 0.1869 \end{bmatrix}, \quad b^4 = \begin{bmatrix} 0.4898 & 0.4456 & 0.6463 \end{bmatrix}$$
As it is discussed in chapter 3, many system properties are related to the controllability Gramian. Moreover, the eigenvalues and their corresponding eigenvectors of the controllability Gramian determine the speed and direction of energy dissipation, therefore, it is worth investigating the optimal location of the actuators, which use the minimum feedback gain, assigning a specific controllability Gramian to the LTI system. This assignability problem defines the best possible set of input matrices to steer the LTI system to a target state, under energy limitations imposed by a given controllability Gramian.

**Example (8.8):** Consider the mass-spring-damper system in the previous example. We apply the KA optimization to find the optimal actuation, which assigns the below controllability Gramian to the system when the control terminal time is very large $t_f \to \infty$:

$$
W_c^o = 
\begin{bmatrix}
0.1372 & 0.1809 & 0.2113 & 0.0000 & -0.0020 & 0.0019 \\
0.1809 & 0.2567 & 0.3014 & 0.0020 & 0.0000 & 0.0032 \\
0.2113 & 0.3014 & 0.3926 & -0.0019 & -0.0032 & -0.0000 \\
0.0000 & 0.0020 & -0.0019 & 0.2489 & 0.1467 & 0.1558 \\
-0.0020 & 0.0000 & -0.0032 & 0.1467 & 0.2737 & 0.2292 \\
0.0019 & 0.0032 & -0.0000 & 0.1558 & 0.2292 & 0.4616 \\
\end{bmatrix}
$$

Let $A_{cl}$, $A_{ol}$ denote the state matrices of the closed-loop and the open loop system respectively. Using the Lyapunov equation, we obtain:

$$
A_{cl}W_c^o + W_c^o A_{cl}^T = -BB^T
$$

$$(A_{ol} - BF)W_c^o + W_c^o (A_{ol} - BF)^T = -BB^T
$$

Then, we have:

$$
A_{cl}W_c^o + W_c^o A_{cl}^T + BB^T = BF W_c^o + W_c^o F^T B^T
$$

Using the Kronecker sum, we get:

$$
vec(A_{cl}W_c^o + W_c^o A_{cl}^T + BB^T) = (W_c^o \otimes B) + (B \otimes W_c^o) vec(F)
$$

$B$ is the input matrix, $F$ is the feedback gain matrix, vec(.) is the vectorization operator and $\otimes$ denotes the Kronecker product. Then $F$ can be readily obtained by solving a linear system of equations.

Hence, the assignability problem can be formulated as:

$$
\begin{align*}
\min_{B} & \quad \|W_c - W_c^o\|_F + \|F\| \\
W_c &= -vec^{-1}((A_{cl} \oplus A_{cl})^{-1} vec(BB^T)) \\
\|B\| &\le 1 \\
A_{cl} &\in H
\end{align*}
$$
\[ \| \cdot \|_F \text{ denotes the Frobenius norm, } A_t \text{ is the state matrix of the closed-loop system and } H \text{ is the set of Hurwitz matrices with the appropriate dimension. } \text{vec}^{-1}(\cdot) \text{ is de-vectorization operator.} \]

Using the KA optimizer, the optimal actuation is obtained as:

\[ B = \begin{bmatrix} 0.5030 & 0.4993 & 0.7055 \end{bmatrix}^T \]

which results in the feedback matrix:

\[ F = \begin{bmatrix} 0.0007 & -0.0133 & 0.0050 & -0.4041 & 0.2113 & -0.3628 \end{bmatrix} \]

and produces the controllability Gramian:

\[
W_c = \begin{bmatrix}
0.1434 & 0.1881 & 0.2208 & -0.0000 & -0.0026 & 0.0002 \\
0.1881 & 0.2680 & 0.3148 & 0.0026 & 0.0000 & 0.0023 \\
0.2208 & 0.3148 & 0.4089 & -0.0002 & -0.0023 & -0.0000 \\
-0.0000 & 0.0026 & -0.0002 & 0.2714 & 0.1419 & 0.1647 \\
-0.0026 & 0.0000 & -0.0023 & 0.1419 & 0.2842 & 0.2362 \\
0.0002 & 0.0023 & -0.0000 & 0.1647 & 0.2362 & 0.4718
\end{bmatrix}
\]

The convergence curve is shown in Figure (8.63). To verify the accuracy of the results, Table (8.16) compares the eigenvalues of the desired controllability Gramian with the eigenvalues of the controllability Gramian produced by the KA optimizer:

<table>
<thead>
<tr>
<th>( \lambda_1 )</th>
<th>( \lambda_2 )</th>
<th>( \lambda_3 )</th>
<th>( \lambda_4 )</th>
<th>( \lambda_5 )</th>
<th>( \lambda_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W''_c )</td>
<td>0.0055</td>
<td>0.0228</td>
<td>0.1002</td>
<td>0.1720</td>
<td>0.7121</td>
</tr>
<tr>
<td>( W'_c )</td>
<td>0.0066</td>
<td>0.0227</td>
<td>0.1146</td>
<td>0.1795</td>
<td>0.7333</td>
</tr>
</tbody>
</table>

Table (8.16): The eigenvalues of the controllability Gramians in Example (8.8)
Figure (8.63): The convergence curve of the KA algorithm in Example (8.8)
8.4. Summary:

In this chapter, a novel optimization algorithm was proposed as an alternative for solving optimal actuator selection problems among the current techniques in the literature. In the proposed algorithm, the logistic equation like functions fluctuated the solutions outwards or towards the current optimum point via switching amongst the steady state, cyclic and chaotic behavior to guarantee exploration and exploitation of the search space, respectively. A combination of random and adaptive variables also facilitated divergence and convergence of the search agents in the algorithm. To benchmark the performance of the proposed algorithm, a set of well-known test cases including unimodal, fixed- dimensional multi-modal, and varying-dimensional multi-modal functions were employed to test exploration, exploitation, local optima avoidance, and convergence of the KA algorithm for the various convex, non-convex, separable, non-separable cases. To quantitatively and qualitatively verify the performance of the KA algorithm several performance metrics were employed:

- The average and standard deviation of the results, which define the convergence curve
- The average fitness of solutions over the iterations
- The trend of the change of the solutions throughout the iterations
- Non-parametric statistical Wilcoxon rank sum test results

The results of the metrics proved that KA algorithm required the search individuals to change abruptly in the initial stage of optimization and gradually in the final steps. The results showed that this behavior caused exploration of the search space extensively and exploitation of the most promising regions. The convergence curves and the average fitness of solutions also confirmed the improvement of initial random populations and the best current solution obtained during the optimization by the KA algorithm.

Finally, a wide range of challenging case studies is employed to demonstrate the performance of the proposed algorithm in solving real-world problems with constrains. The results demonstrate the merits of KA in solving real problems.

Amongst the case studies, the KA algorithm is used to solve the problem of the optimal actuation in a typical 3-storey building to decrease seismic effects. The problem of input selection in optimal feedback design is also investigated in two different cases:

- Considering the pole placement
- Assigning the controllability Gramian

The KA algorithm also demonstrates the ability to solve the problem of optimal actuation for the robust DOC in the case of uncertainty.

It can be concluded that the KA algorithm not only is a very suitable alternative compared the current selection methods in the literature for solving optimal actuator problems, but also it can be applied in many different optimization problems. Therefore, the KA algorithm is offered to researchers in different fields.

This chapter opens up several research directions for future studies. The binary version of this algorithm can be proposed to solve problems with binary objectives. The KA algorithm can
also be hybridized with other methods and algorithms in the field of optimization to improve its performance. Finally, investigation of the application of KA in different fields would be a valuable contribution.
Chapter 9: Conclusions & Future Research

9.1. Summary and Contribution of Thesis

This thesis has contributed to the development of methodologies for the selection of the input and output structures, that is selection of the matrix $B$ and the matrix $C$ in a continuous-time LTI system described by the triple $S(A, B, C)$. This has been achieved by considering the implications of such selections as far as the required energy to achieve control and state reconstruction. The main criterion for this selection has been the minimization of the required energy for control and state reconstruction and the notions of relative degree of controllability and observability have been instrumental for the development of the overall methodology. The methodologies developed may be used as design tools where apart from energy requirements other design criteria may be also incorporated for the selection of inputs and outputs. The research undertaken in this thesis is an integral part of the effort to develop an integrated approach to Control and Global Process Instrumentation. The developed methodology has been supported by computational tools for the relevant energy indicators (Gramians) and the development of an optimization based approach that can be used for the design process.

A general research agenda to determine the quality of input and output structure based on fundamental criteria such as performance, stability and so on has been presented, which highlights the significance of the input-output structure selection to meet a certain control objective. The developed research agenda is novel and has introduced new aspects to the design of desirable input and output structures by exploring the state input, state output criteria based on energy and linking them to the
relative degree of controllability and observability. In the case of controllability, the transfer a state from the origin to some point, i.e. to a terminal state location on the boundary of a disc of radius $R$ from the origin has been the key issue, whereas, in the case of observability, it has been the evaluation of the cost of energy to reconstruct the initial state from the distance $R$ from the origin. To address the above challenges properly, it has been required to provide solutions to some specific problems such as:

I. The computation of the energy required for each of the above cases for a given time of transition and a fixed value of $R$ as a function of terminal time and $R$.

II. Expressing the functions derived in the first problem as a function of the input matrix, output matrix, and then evaluate the quality of the given input-output structure selection, it should be as a function of the input-output matrix.

The main contributions of the thesis have been in the following areas:

- Provided a literature review for link of energy to controllability and observability and a review of methodologies for selection of inputs and outputs. The review on relative measure of controllability and observability properties covers measures such as Kalman: controllability and observability matrices rank tests, gramians: integral definitions (finite time) as well as the algebraic expression (Lyapunov and Sylvester for infinite time). These reviews also include Gramians (controllability, observability and cross gramians) for unstable systems with no poles on the imaginary axis. Additionally, eigenvectors measures, and input and output restriction pencil, which are measures of controllability and observability independent of state feedback.
- Proposed an overview of Controllability/ Observability Gramians, and methodologies for their computation. The proposed approach is based on the integral definitions (finite time) of the controllability Gramian hence it does not imply any restriction on the stability of the system. Another advantage of this method is its simplicity to be used in perturbation analysis and robust control design. The controllability Gramian assignment problem is also discussed.
- Provided a solution of minimum energy, minimum time problems for controllability and observability which has led to the definition of the constant energy surfaces linked to controllability and observability. This has been linked to clarifying how to transfer states from the origin to some point at defined time on the boundary of a disc of radius $R$ along with the transfer energy requirement may be achieved.
- Has produced computational tools for the Energy for unstable systems and measures of degree of controllability/observability. The proposed measures have been linked to the minimum required control input energy to change the state from an arbitrary initial condition to an arbitrary final condition. The improved measure has been introduced by generalizing the Kalman controllability and observability matrices rank tests. A new quantitative measure of the degree of disturbance rejection (DODR) has also been proposed representing how
controllable the system is under the presence of persistent disturbance. The measures are general in the sense that they cover both stable and unstable systems.

- Has developed some new methods for the computation of controllability Gramian in the stable system, which has particular canonical form structures. The value of minimum energy has been evaluated directly via the coefficients of the characteristic polynomial of the system. The trace and some upper bounds for the largest eigenvalue of the controllability Gramian has been derived based on the coefficients of the Routh Hurwitz's table.

- Has introduced a number of partial problems linked to the overall selection of system of inputs, and thus of the input matrix $B$, to minimize the average required energy for controllability respectively observability. The new proposed approach provided an extension to the existing literature, in which the problem of input structure selection is investigated over the binary set $\{0,1\}$.

- Has developed an optimization method for the optimal selection of input/output system structure for large-scale systems. The proposed optimization algorithm exploring and exploiting the search space in order to find efficiently near-optimal solution of the problem of IO selection. We focused on the problem of cardinality-constrained actuator placement for minimum control effort, where the optimal actuator set is selected so that an input energy metric is minimized. While this is an NP-hard combinatorial problem, our proposed algorithm has proved its efficiency in converging for optimal solution.

9.2. Future Directions

The research agenda developed for the selection of the input and output structures has a number of problems which have not yet been handled in this thesis and remain open. A number of topics that require further research are:

- Energy methodology for the selection of the input, output structure is based on energy requirements. A related issue is the gain requirements for feedback design which leads to an alternative problem formulation. The link between energy and gain requirements is not well understood and this may also form the basis for further research. Of particular interest is exploring the link between energy and gain restrictions.

- Linked to the above investigation is exploring the relations between relative degrees of controllability and observability and the restrictions on the gains for state feedback design, state reconstruction respectively.

- The computational tools based on optimization need further development. The associated optimization problem is non-convex and computing approximate solutions in an efficient way is still challenging.

- The selection of the input and output structures has been considered so far as unconstrained problems. In many design cases, certain variables have to be used for control, measurement respectively and thus constrained versions emerge which are linked to constrained optimization problems.

- Two challenging theoretical problems that require attention are linked to introducing restricted notions of controllability and observability under energy, or gain constraints.
• The basic ideas developed for linear systems can be transferred with minor modification to the case of nonlinear systems. This is achieved by generating an efficient problem relaxation based on the steady state. The results of this thesis can also be extended to some other types of the system, e.g. discrete-time systems, time variant systems.

• When the parameters of the system are subject to perturbation in order to find a dominant subset of inputs and outputs for any given large-scale system, it is needed to determine the effectiveness of each input and output in the overall operation of the control system. Thus, the investigating the robust controllability (observability) degree of the system is worth pursuing.
References


243


Appendix chapter 3:

**Proposition**: (Parrilo, 2003), (Parrilo, 2000) A multivariate polynomial \( p(x) \in \mathbb{R} \) in \( n \) variables and of degree 2\( d \) is a sum of squares if and only if there exists a real symmetric positive semidefinite matrix \( Q \) such that:

\[
p(x) = z^T Qz
\]

where \( z \) is the vector of monomials of degree up to \( d \), we call matrix \( Q \) the Gram matrix.

\[
Z = [1, x_1, x_2, \ldots, x_n, x_1 x_2, \ldots, x_n^d]^T
\]

Given Gram matrix \( Q \), of rank \( t \), we can construct polynomial \( h_1, h_2, \ldots, h_t \), such that:

\[
p(x) = \sum_{i=1}^{t} (h_i)^2
\]

**Proof:**

If \( p(x) = \sum_{i=1}^{t} (h_i)^2 \) is true and is SOS, then as above we take \( Q = AA^T \), where \( A \) is the matrix whose columns are the coefficients of the \( h_i \)s.

Now suppose there exist a real symmetric psd matrix \( Q \), such that \( p(x) = z^T Qz \) and rank(\( Q \))=t, since \( Q \) is real symmetric of rank \( t \), there exists a real matrix \( V \) and a diagonal \( D = \text{diag}(d_1, \ldots, d_t, 0, \ldots 0) \), \( d_i \neq 0 \) such that \( Q = V^T D V \), since \( Q \) is psd we have \( d_i > 0 \), then:

\[
p(x) = z^T V^T D V z
\]

Now considering \( V = v_{i,j} , i = 1, \ldots, t \), we have:

\[
h_i = \sqrt{d_i} \sum_{j=1}^{k} v_{i,j} z^j \in \mathbb{R}
\]

It follows that \( p(x) = \sum_{i=1}^{t} (h_i)^2 \).

\[\Box\]
Appendix chapter 5:

**Proposition:** In an LTI system with no undamped modes, the controllability Gramian in the frequency domain is defined as:

\[
W_c = \frac{1}{2\pi} \int_{-\infty}^{\infty} (j\omega I - A)^{-1} BB^T (-j\omega I - A^T)^{-1} d\omega
\]

And it can be expressed as the block diagonal form:

\[
W_c = V \begin{bmatrix} W_s & 0 \\ 0 & W_u \end{bmatrix} V^T
\]

**Proof:**

Consider the steady state definition of controllability Gramian in the time domain:

\[
W_c(0, \infty) = \int_0^\infty e^{At} BB^T e^{A^Tt} dt
\]

Then based on the Parseval’s Theorem we can convert the definition above to the statement below:

\[
W_c = \frac{1}{2\pi} \int_{-\infty}^{\infty} (j\omega I - A)^{-1} BB^T (-j\omega I - A^T)^{-1} d\omega
\]

The statement above is clear for the stable systems, but let’s investigate what happens if the system has some eigenvalues on the right half plane.

Through non-singular similarity transformation \( V \) (the eigenvectors’ matrix) the LTI system is decomposed into stable and anti-stable diagonal blocks as it is explained before:

\[
\Lambda = \begin{bmatrix} A_s & 0 \\ 0 & A_u \end{bmatrix}, \Lambda = V^{-1}AV
\]

By imposing the similarity transformation \( V \) we will obtain:

\[
W_{\text{new}}(0, \infty) = \int_0^\infty e^{VAV^{-1}t}VB_{\text{new}}B_{\text{new}}^TV e^{V^{-1}AV^T} dt \\
\Rightarrow W_{\text{new}}(0, \infty) = V \int_0^\infty e^{At} B_{\text{new}}B_{\text{new}}^T e^{At} dt V^T \Rightarrow VW_{\text{new}} V^T = W_c
\]

Using the Parseval’s Theorem in the frequency domain we have:

\[
W_c = V \frac{1}{2\pi} \int_{-\infty}^{\infty} (j\omega I - \Lambda)^{-1} B_{\text{new}}B_{\text{new}}^T (-j\omega I - \Lambda^T)^{-1} d\omega V^T
\]

---

\(^6\) For the simplicity we use the same symbol for the controllability Gramian in the frequency domain and the time domain.
Considering the clockwise contour $\Gamma$ that encircles the right-half plane, we will have:

$$W_c = V \frac{1}{2\pi j} \oint_{\Gamma} \left[ (sI - A_y)^{-1} B_y B_y^T (sI - A_y) (sI - A_y)^{-1} \right] ds \mathcal{V}^T$$

Clearly off-diagonal blocks both are analytic in the right half plane (RHP) then their integral would be zero.

And the diagonal blocks both are stable. The stability of the first block is clear regarding this fact that it represents the stable modes of the system. For the second diagonal block that is related to the anti-stable subsystem, one can readily prove that:

$$\oint_{\Gamma} (sI - A_y)^{-1} B_y B_y^T (-j\omega - A_y)^{-1} d\omega = \oint_{\Gamma} ((-sI + A_y)^{-1} B_y B_y^T (sI - A_y))^{-1} ds = \oint_{\Gamma} (sI - (-A_y)^{-1} B_y B_y^T (-sI - (-A_y)^{T})^{-1} ds$$

$$\cdots$$

Example (5.4):

$$A_{ij3s3} = (a_{i1} + a_{i2} + a_{i3})$$

$$a_{i1} = \begin{bmatrix}
-0.0000 + 1.2900i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & -0.0000 - 1.2900i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & -0.0001 + 0.4826i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & -0.0000 - 0.4835i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
\end{bmatrix}$$
\[
\begin{bmatrix}
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
-0.0001 - 0.4835i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
\end{bmatrix}
\]

\[
a_{2} = 
\begin{bmatrix}
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
-0.0152 - 0.0336i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
\end{bmatrix}
\]

\[
a_{3} = 
\begin{bmatrix}
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
-0.0112 - 0.0228i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\
\end{bmatrix}
\]

\[
A_{5x5} = 
\begin{bmatrix}
462.7000 & 0 & 0 & 0 & 0 \\
0 & 191.4000 & 0 & 0 & 0 \\
0 & 0 & 243.4000 & 0 & 0 \\
0 & 0 & 0 & 272.4000 & 0 \\
0 & 0 & 0 & 0 & 284.5000
\end{bmatrix}
\]
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.0172</td>
<td>0.0124</td>
<td>-0.0166</td>
<td>-0.0136</td>
<td></td>
</tr>
<tr>
<td>-0.0172</td>
<td>0.0124</td>
<td>-0.0166</td>
<td>-0.0136</td>
<td></td>
</tr>
<tr>
<td>-1.0416</td>
<td>0.1889</td>
<td>1.2417</td>
<td>0.2170</td>
<td></td>
</tr>
<tr>
<td>-1.0416</td>
<td>0.1889</td>
<td>1.2417</td>
<td>0.2170</td>
<td></td>
</tr>
<tr>
<td>-0.1211</td>
<td>-1.0076</td>
<td>0.1456</td>
<td>-1.2523</td>
<td></td>
</tr>
<tr>
<td>-0.1211</td>
<td>-1.0076</td>
<td>0.1456</td>
<td>-1.2523</td>
<td></td>
</tr>
</tbody>
</table>

\[ B_{13x1} = \begin{bmatrix}  
-3.7286 & 3.0158 & -3.4040 & -2.4204 \\
-3.7286 & 3.0158 & -3.4040 & -2.4204 \\
-1.4743 & -0.9178 & -0.2739 & 2.9960 \\
-1.4743 & -0.9178 & -0.2739 & 2.9960 \\
0.7778 & 0.6500 & 1.2099 & 1.0762 \\
0.7778 & 0.6500 & 1.2099 & 1.0762 \\
-4.6704 & 2.8426 & -4.9413 & -4.7850 
\end{bmatrix} \]

\[ B_{u5x5} = \begin{bmatrix}  
2.6764 & -0.5699 & 6.9452 & -0.3500 \\
8.7810 & -21.0600 & 0.5281 & 33.6300 \\
-44.7700 & -16.00 & -53.4300 & 1.0640 \\
-50.9400 & -8.020 & 65.9800 & 13.8900 \\
34.1700 & -55.910 & -50.3000 & -44.6800 
\end{bmatrix} \]
Appendix chapter 7:

**Lemma:** For any matrix $A_{n \times m}$, $m \leq n$, $\text{rank}(A) = m$, if and only if the minimum singular value of $A$ is greater than zero, i.e. $\sigma_m > 0$.

**Proof:**

Based on singular value decomposition (Gene H. Golub, Charles F. Van Loan, 1996), (Horn, Roger A and Johnson, Charles R, 2012), (Marcus, M. and Minc, H, 1988), for any matrix $A_{n \times m}$ has a factorization of the form $U \Sigma V^*$, where $U$ is a $n \times n$ unitary matrix, $\Sigma$ is a $n \times m$ rectangular diagonal matrix with non-negative real numbers on the diagonal, and $V$ is a $m \times m$ unitary matrix. The diagonal entries $\sigma_i$'s of $\Sigma$ are the singular values of $A$.

Now assume that $\text{rank}(A) = m$ then clearly $\text{rank}(\Sigma) = m$, which means that all singular values of $A$ must be greater than zero, (i.e. $\sigma_m > 0$). Conversely, if minimum singular value of $A$ is greater than zero, then the number of nonzero columns of $\Sigma$, which shows the rank of $A$ would be equal to $m$, i.e. $\text{rank}(\Sigma) = \text{rank}(A) = m$, and the Lemma is proved.

$\square$

**Theorem:** If $G$ is the Gram matrix of $A$, then $\text{rank}(G) = \text{rank}(A)$.

**Proof:** (Mirsky, 2012)

Let $x$ be a vector such that $A^* Ax = 0$. Then $x^* A^* Ax = 0$ and so we have $(A^* x)^* A x = 0$, i.e. $|Ax|^2 = 0$. Therefore, we have $Ax = 0$. Conversely, if $x$ satisfies the relation $A x = 0$, then clearly $A^* Ax = 0$. It follows that the homogeneous system $A x = 0$ and $G x = 0$ are equivalent. Hence, $n - \text{rank}(A) = n - \text{rank}(G)$ where $n$ is the number of columns of $A$. The proof of the Theorem is completed.

$\square$

**Proposition:** Let $B = [b_1, b_2, \ldots, b_m]$, $b_j \in \mathbb{R}^{n \times 1}$, $m \leq n$. Then, $\text{rank}(B) = m$ if and only if the Gram matrix of $B$ is non-singular, i.e. $|G| \neq 0$.

**Proof:** (Peter Lancaster, Miron Tismenetsky, 1985)

Assume $\text{rank}(B) < m$, i.e. $[b_1, b_2, \ldots, b_m]$ are linearly dependent, let $a = [a_1, a_2, \ldots, a_m]$ be a nonzero vector such that $\sum_{i=1}^{m} a_i b_i = 0$ then for all $j$, $\sum_{i=1}^{m} a_i b_j^* b_i = 0$ so $a^T$ is in the kernel of the Gram matrix. Conversely, if the Gram matrix is singular, then there exists a nonzero vector $a = [a_1, a_2, \ldots, a_m]$, such that $\sum_{i=1}^{m} a_i b_j^* b_i = 0$, let $w = \sum_{i=1}^{m} a_i b_i = 0$ Then $w$ is orthogonal to every
and therefore orthogonal to itself. That is, \( w \cdot w = 0 \) then \( w = 0 \), thus \( [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n] \) are linearly dependent, and \( \text{rank}(B) < m \).

\[ \square \]