Springer Optimization and Its Applications 61

Vladimir L. Boginski
Clayton W. Commander
Panos M. Pardalos
Yinyu Ye *Editors*

Sensors: Theory, Algorithms, and Applications

Springer

Springer Optimization and Its Applications

VOLUME 61

Aims and Scope

Optimization has been expanding in all directions at an astonishing rate during the last few decades. New algorithmic and theoretical techniques have been developed, the diffusion into other disciplines has proceeded at a rapid pace, and our knowledge of all aspects of the field has grown even more profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in all areas of applied mathematics, engineering, medicine, economics, and other sciences.

The series *Springer Optimization and Its Applications* publishes undergraduate and graduate textbooks, monographs and state-of-the-art expository work that focus on algorithms for solving optimization problems and also study applications involving such problems. Some of the topics covered include nonlinear optimization (convex and nonconvex), network flow problems, stochastic optimization, optimal control, discrete optimization, multi-objective programming, description of software packages, approximation techniques and heuristic approaches.

For further volumes:
http://www.springer.com/series/7393

Vladimir L. Boginski • Clayton W. Commander
Panos M. Pardalos • Yinyu Ye
Editors

Sensors: Theory, Algorithms, and Applications

Springer

Editors
Vladimir L. Boginski
Department of Industrial and Systems Engineering
University of Florida
303 Weil Hall
Gainesville, FL 32611
USA
boginski@reef.ufl.edu

Panos M. Pardalos
Department of Industrial and Systems Engineering
University of Florida
303 Weil Hall
Gainesville, FL 32611
USA
pardalos@ufl.edu

Clayton W. Commander
Air Force Research Laboratory
Munitions Directorate
Eglin Air Force Base
101 West Eglin Boulevard
Eglin AFB, FL 32542
USA
clayton.commander@eglin.af.mil

Yinyu Ye
Department of Management Science and Engineering
Huang Engineering Center 308
School of Engineering
Stanford University
475 Via Ortega
Stanford, CA 94305
USA
yinyu-ye@stanford.edu

ISSN 1931-6828
ISBN 978-0-387-88618-3 e-ISBN 978-0-387-88619-0
DOI 10.1007/978-0-387-88619-0
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011941384

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In recent years, technological advances have resulted in the rapid development of a new exciting research direction – the interdisciplinary use of sensors for data collection, systems analysis, and monitoring. Application areas include military surveillance, environmental screening, computational neuroscience, seismic detection, transportation, along with many other important fields.

Broadly speaking, a sensor is a device that responds to a physical stimulus (e.g., heat, light, sound, pressure, magnetism, or motion) and collects and measures data regarding some property of a phenomenon, object, or material. Typical types of sensors include cameras, scanners, radiometers, radio frequency receivers, radars, sonars, thermal devices, etc.

The amount of data collected by sensors is enormous; moreover, this data is heterogeneous by nature. The fundamental problems of utilizing the collected data for efficient system operation and decision making encompass multiple research areas, including applied mathematics, optimization, and signal/image processing, to name a few. Therefore, the task of crucial importance is not only developing the knowledge in each particular research field, but also bringing together the expertise from many diverse areas in order to unify the process of collecting, processing, and analyzing sensor data. This process includes theoretical, algorithmic, and application-related aspects, all of which constitute essential steps in advancing the interdisciplinary knowledge in this area.

Besides individual sensors, *interconnected systems of sensors*, referred to as *sensor networks*, are receiving increased attention nowadays. The importance of rigorous studies of sensor networks stems from the fact that these systems of multiple sensors not only acquire individual (possibly complimentary) pieces of information, but also effectively exchange the obtained information. Sensor networks may operate in static (the locations of individual sensor nodes are fixed) or dynamic (sensor nodes may be mobile) settings.

Due to the increasing significance of sensor networks in a variety of applications, a substantial part of this volume is devoted to theoretical and algorithmic aspects of problems arising in this area. In particular, the problems of information fusion are especially important in this context, for instance, in the situations when the data collected from multiple sensors is synthesized in order to ensure effective operation of the underlying systems (i.e., transportation, navigation systems, etc.). On the other hand, the reliability and efficiency of the sensor network itself (i.e., the ability of the network to withstand possible failures of nodes, optimal design of the network in terms of node placement, as well as the ability of sensor nodes to obtain location coordinates based on their relative locations – known as *sensor network localization* problems) constitutes another broad class of problems related to sensor networks. In recent years, these problems have been addressed from rigorous mathematical modeling and optimization perspective, and several chapters in this volume present new results in these areas.

From another theoretical viewpoint, an interesting related research direction deals with investigating information patterns (possibly limited or incomplete) that are obtained by sensor measurements. Rigorous mathematical approaches that encompass dynamical systems, control theory, game theory, and statistical techniques, have been proposed in this diverse field.

Finally, in addition to theoretical and algorithmic aspects, application-specific approaches are also of substantial importance in many areas. Although it is impossible to cover all sensor-related applications in one volume, we have included the chapters describing a few interesting application areas, such as navigation systems, transportation systems, and medicine.

This volume contains a collection of chapters that present recent developments and trends in the aforementioned areas. Although the list of topics is clearly not intended to be exhaustive, we attempted to compile contributions from different research fields, such as mathematics, electrical engineering, computer science, and operations research/optimization. We believe that the book will be of interest to both theoreticians and practitioners working in the fields related to sensor networks, mathematical modeling/optimization, and information theory; moreover, it can also be helpful to graduate students majoring in engineering and/or mathematics, who are looking for new research directions.

We would like to take the opportunity to thank the authors of the chapters for their valuable contributions, as well as Springer staff for their assistance in producing this book.

Gainesville, FL, USA

Vladimir L. Boginski
Clayton W. Commander
Panos M. Pardalos
Yinyu Ye

Contents

Part I Models and Algorithms for Ensuring Efficient Performance of Sensor Networks

Part II Theoretical Aspects of Analyzing Information Patterns

Part III Sensors in Real-World Applications

Contributors

Ashwin Arulselvan Technische Universität Berlin, Berlin, Germany, arulsel@math.tu-berlin.de

Pia E.K. Berg-Yuen Air Force Research Laboratory Munitions Directorate, Eglin AFB, FL 32542, USA, piagreg@cox.net

Animesh Chakravarthy Wichita State University, Wichita, KS, USA, animesh.chakravarthy@wichita.edu

Emily M. Craparo Naval Postgraduate School, Monterey, CA, USA, emcrapar@nps.edu

J. Willard Curtis Air Force Research Laboratory Munitions Directorate, Eglin AFB, FL 32542, USA, jess.curtis@eglin.af.mil

Laura Di Giacomo Dipartimento di Statistica, Sapienza Universita’ di Roma, Italy

Eric Feron Georgia Tech Atlanta, GA, USA, feron@gatech.edu

David Yang Gao University of Ballarat, Mt. Helen, VIC 3350, Australia, d.gao@ballarat.edu.au

William MacKunis Air Force Research Laboratory Munitions Directorate, Eglin AFB, FL 32542, USA, mackunis@gmail.com

Mikel M. Miller Air Force Research Laboratory Munitions Directorate, Eglin AFB, FL 32542, USA, mikel.miller@eglin.af.mil

Meir Pachter Air Force Institute of Technology, Wright-Patterson AFB, OH 45433, USA, meir.pachter@afit.edu

Panos M. Pardalos University of Florida, Gainesville, FL 32611, USA, pardalos@ufl.edu

Giacomo Patrizi Dipartimento di Statistica, Sapienza Universita’ di Roma, Italy, g.patrizi@caspur.it

Jaime Peraire Massachusetts Institute of Technology, Cambridge, MA, USA, peraire@mit.edu

Khanh D. Pham Air Force Research Laboratory, Space Vehicles Directorate, Kirtland AFB, NM 87117, USA, khanh.pham@kirtland.af.mil

Ning Ruan Curtin University of Technology, Perth, WA 6845, Australia, mimiopt@gmail.com

Andrey Soloviev Research and Engineering Education Facility, University of Florida, Shalimar, FL 32579, USA, soloviev@ufl.edu

Kyungyeol Song McKinsey Corporation Seoul, South Korea, drsky@alum.mit.edu

My T. Thai University of Florida, Gainesville, FL 32611, USA, mythai@cise.ufl.edu

Ravi Tiwari University of Florida, Gainesville, FL 32611, USA, rtiwari@cise.ufl.edu

Chrysafis Vogiatzis University of Florida, Gainesville, FL 32611, USA, chvogiat@ufl.edu

Petros Xanthopoulos University of Florida, Gainesville, FL 32611, USA, petrosx@ufl.edu

Part I
Models and Algorithms for Ensuring Efficient Performance of Sensor Networks

On Enhancing Fault Tolerance of Virtual Backbone in a Wireless Sensor Network with Unidirectional Links

Ravi Tiwari and My T. Thai

Abstract A wireless sensor network (WSN) is a collection of energy constrained sensor node forming a network which lacks infrastructure or any kind of centralized management. In such networks, virtual backbone has been proposed as the routing infrastructure which can alleviate the broadcasting storm problem occurring due to consistent flooding performed by the sensor node, to communicate their sensed information. As the virtual backbone nodes needs to carry other nodes' traffic, they are more subject to failure. Hence, it is desirable to construct a fault tolerant virtual backbone. Most of recent research has studied this problem in homogeneous networks. In this chapter, we propose solutions for efficient construction of a fault tolerant virtual backbone in a WSN where the sensor nodes have different transmission ranges. Such a network can be modeled as a disk graph (DG), where link between the two nodes is either unidirectional or bidirectional. We formulate the fault tolerant virtual backbone problem as a k-Strongly Connected m-Dominating and Absorbing Set (k, m) SCDAS problem. As the problem is NP-hard, we propose an approximation algorithm along with the theoretical analysis and conjectured its approximation ratio.

1 Introduction

A wireless sensor network (WSN) is a collection of power constrained sensors nodes with a base station. The sensors are supposed to sense some phenomena and collect information, which is required to be sent to the base station for further forwarding or processing. As the sensors are power constraint, their transmission

R. Tiwari • M.T. Thai (✉)
Computer Science and Engineering Department, University of Florida, Gainesville, FL, USA
e-mail: rtiwari@cise.ufl.edu; mythai@cise.ufl.edu

V.L. Boginski et al. (eds.), *Sensors: Theory, Algorithms, and Applications*,
Springer Optimization and Its Applications 61, DOI 10.1007/978-0-387-88619-0_1,

ranges are small. Hence, the sensed information may be relayed on multiple intermediate sensor nodes before reaching the base station. As there is no fixed or predefined infrastructure, in order to enable data transfer in such networks, all the sensor nodes frequently flood control messages, thus causing a lot of redundancy, contentions, and collisions [20]. As a result, a virtual backbone has been proposed as the routing infrastructure of such networks for designing efficient protocols for routing, broadcasting, and collision avoidance [1]. With virtual backbone, routing messages are only exchanged between the sensor nodes in the virtual backbone, instead of being flooded to all the sensor nodes. With the help of virtual backbone, routing is easier and can adapt quickly to network topology changes. It has been seen that the virtual backbones could dramatically reduce routing overhead [18]. Furthermore, using virtual backbone as relay nodes can efficiently reduce the energy consumption, which is one of the critical issues in WSNs to maximize the sensor network lifetime.

However, transmission range of all the sensor nodes in the WSN are not necessarily equal. As the transmission range depends upon the energy level of a sensor node which can be different for different sensor nodes, this may result in sensor nodes having different transmission range. The sensor nodes can also tune their transmission ranges depending upon their functionality, or they may perform some power control to alleviate collisions or to achieve some level of connectivity. In some topology controlled sensor networks, sensor nodes may adjust their transmission ranges differently to obtain certain optimization goals. All these scenarios result into the WSN with different transmission ranges. Such a network can be modeled as a Disk Graph (DG) G. Note that G is a *directed* graph, consisting both bidirectional and unidirectional links.

Since the virtual backbone nodes in the WSN need to relay other sensor node's traffic, so, due to heavy load often they are vulnerable to frequent node or link failure which is inherent in WSNs. Hence, it is very important to study the fault tolerance of the virtual backbone in wireless sensor networks. Therefore, constructing a fault tolerant virtual backbone that continues to function during node or link failure is an important research problem, which has been not studied sufficiently. In [6, 7], the authors considered this problem in Unit Disk Graph (UDG) [2], in which all nodes have the same transmission ranges. When a wireless network has nodes with same transmission ranges then it will only have bidirectional links. In such a case the virtual backbone is represented by the connected dominating set (CDS) of the graph representing the wireless network. Whereas, when the wireless network has nodes with different transmission range then it will have both unidirectional and bidirectional links. In this case the virtual backbone is represented by a strongly connected dominating and absorbing set (SCDAS) [12], here, a node not in virtual backbone has at least one virtual backbone node as its incoming and outgoing neighbor, respectively.

Although the virtual backbone problem has been extensively studied in general undirected graphs and UDGs [3, 13, 16, 17, 19, 21–23], the construction of virtual backbone in wireless networks with different transmission ranges is explored to a little extent. In [8] and [5], the authors extended their marking process to networks with unidirectional links to find a SCDAS. However, the paper does not present any approximation ratio. Recently, we proposed a constant approximation algorithms for SCDAS problem [4, 10, 12, 14]. The construction of fault tolerant virtual backbone in general undirected graphs is also one of the newly studied problems. Dai et al. addressed the problem of constructing k-connected k dominating set ((k, k) CDS) [6] in UDG. In Feng et al. [7] introduced the problem of constructing $(2, 1)$ CDS in UDGs and proposed a constant approximation ratio. Note that the solutions of these two papers are applicable only to undirected graphs. In addition, the authors just considered a special case of the general problem, where $k = m$ or $k = 2$ and $m = 1$. In [24] Wu et al. studied the construction of (k, m) CDS but they considered undirected graph. Recently, we have considered the fault tolerant virtual backbone in heterogeneous networks with only bidirectional links [15]. We proposed a constant approximation algorithm for any value of k and m. In summary, no work has studied the (k, m) SCDAS in heterogeneous networks with unidirectional and bidirectional links for any value of k and m.

In this chapter we study the enhancing of fault tolerance of virtual backbone in WSN represented by a directed disk graph (DG). The virtual backbone in this case is represented as a strongly connected dominating and absorbing set (SCDAS). The fault tolerance of a virtual backbone can be enhanced in two aspects. Firstly, by increasing the dominance and absorption of the virtual backbone nodes, i.e., by increasing the number of virtual backbone nodes in the incoming and outgoing neighborhood of a non-virtual backbone nodes. Secondly, by increasing the connectivity of the virtual backbone, by ensuring the nodes in virtual backbone has multiple paths to each other in the subgraph induced by them. In order to generate a fault tolerant virtual backbone, we formulate the (k, m) SCDAS problem. The (k, m) SCDAS problem is to find an SCDAS of a directed graph $G = (V, E)$ such that the graph induced by the (k, m) SCDAS nodes is k-strongly node connected and any node not in (k, m) SCDAS has at least m nodes in its incoming and outgoing neighborhood, respectively.

The rest of this chapter is organized as follows. Section 2 describes the preliminaries, network model, and problem definition. The enhancement of fault tolerance of virtual backbone in terms of dominance and absorption is studied in Sect. 3. In Sect. 4 we conclude the chapter with a brief summary.

2 Network Model and Problem Definition

2.1 Preliminaries

Let a directed graph $G = (V, E)$ represent a network where V consists of all nodes in a network and E represents all the communication links.

For any vertex $v \in V$, the **incoming neighborhood** of v is defined as $N^-(v) = \{u \in V \mid (u,v) \in E\}$, and the **outgoing neighborhood** of v is defined as $N^+(v) = \{u \in V \mid (v,u) \in E\}$.

Likewise, for any vertex $v \in V$, the **closed incoming neighborhood** of v is defined as $N^-[v] = N^-(v) \cup \{v\}$, and the **closed outgoing neighborhood** of v is defined as $N^+[v] = N^+(v) \cup \{v\}$.

A subset $S \subseteq V$ is called a **dominating set** (DS) of G iff $S \cup N^+(S) = V$ where $N^+(S) = \bigcup_{u \in S} N^+(u)$ and $\forall v \in N^+(S), N^-(v) \cap S \neq \emptyset$. If $|N^-(v) \cap S| \geq m$, then S is said to be a m dominating set.

A subset $A \subseteq V$ is called an **absorbing set** (AS) of G iff $A \cup N^-(S) = V$ where $N^-(S) = \bigcup_{u \in S} N^-(u)$ and $\forall v \in N^-(S), N^+(v) \cap S \neq \emptyset$. If $|N^+(v) \cap S| \geq m$, then A is said to be a m absorbing set.

A subset $S \subseteq V$ is called an **independent set** (IS) of G iff $S \cup N^+(S) = V$ and $S \cap N^+(S) = \emptyset$.

A subset SI $\subseteq V$ is called a **Semi-independent Set** (SI) of G iff $u, v \in$ SI, then, $\{(u,v),(v,u)\} \notin E$, or if $(u,v) \in E$ then $(v,u) \notin E$ and vice-versa. Nodes u and v are said to be **Semi-independent** to each other.

Given a subset $S \subseteq V$, an **induced subgraph of** S, denoted as $G[S]$, obtained by deleting all vertices in the set $V - S$ from G.

A directed graph G is said to be **strongly connected** if for every pair of nodes $u, v \in V$, there exists a directed node disjoint path. Likewise, a subset $S \subseteq V$ is called a **strongly connected set** if $G[S]$ is strongly connected. If for every pair of nodes $u, v \in V$, there exists at least k directed node disjoint paths then the graph G is said to be ***k* strongly connected**, similarly a subset $S \subseteq V$ is called a ***k* strongly connected set** if $G[S]$ is k strongly connected.

A subset $S \subseteq V$ is called a **Strongly Connected Dominating Set** (SCDS) if S is a DS and $G[S]$ is strongly connected. S is called a **Strongly Connected Dominating and Absorbing Set** (SCDAS) if S is an SCDS and for all nodes $u \notin S$, $N^+(u) \cap S \neq \emptyset$ and $N^-(u) \cap S \neq \emptyset$. S is a (k,m) SCDAS if it is k strongly connected and m dominating and m absorbing.

2.2 Network Model and Problem Definition

In this chapter, we study the fault tolerant virtual backbone in wireless sensor networks with different transmission ranges. In this case, the WSN can be modeled as a directed graph $G = (V, E)$. The sensor nodes in V are located in the two dimensional Euclidean plane and each sensor node $v_i \in V$ has a transmission range $r_i \in [r_{\min}, r_{\max}]$. A directed edge $(v_i, v_j) \in E$ if and only if $d(v_i, v_j) \leq r_i$, where $d(v_i, v_j)$ denotes the Euclidean distance between v_i and v_j. Such a directed graphs G is called *Disk Graphs* (DG). An edge (v_i, v_j) is bidirectional if both (v_i, v_j) and (v_j, v_i) are in E, i.e., $d(v_i, v_j) \leq \min\{r_i, r_j\}$. Otherwise, it is a unidirectional edge. Figure 1 shows a disk graph (DG), here the black dots represents the sensor nodes and the dotted circles around them represents their transmission disks. The directed

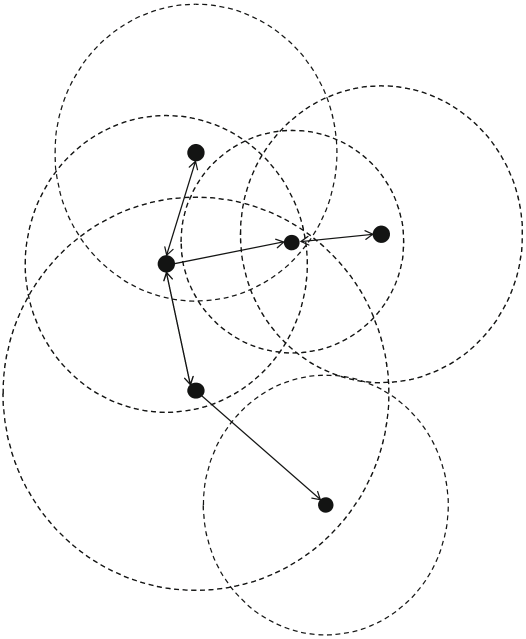


Fig. 1 A disk graph (DG) with unidirectional and bidirectional links

arrows represent the unidirectional links whereas the bidirected edge represents bidirectional links. In our network model, we consider both unidirectional and bidirectional edges.

The virtual backbone in a WSN that can be represented by the connected dominating set (CDS). In this chapter we studied the fault tolerance of the virtual backbone in the WSN modeled as a disk graph (DG). In this case the virtual backbone can be represented by a strongly connected dominating and absorbing set (SCDAS). There can be two kinds of faults occurring in WSN. A sensor node can become faulty or a link between two sensor nodes might go down. Hence, the fault tolerance of virtual backbone in WSN can be enhanced in two ways. Firstly, by enhancing the dominance and the absorption of the SCDAS representing the virtual backbone by ensuring more SCDAS nodes are there as incoming and outgoing neighbors to a non-SCDAS node. This ensures that a non-virtual backbone node has other options to forward its data, if one of its virtual backbone neighbor goes down due to some failure. Secondly, by ensuring there are multiple paths between the virtual backbone nodes in the subgraph generated by the virtual backbone nodes, so that if a link between two virtual backbone goes down it would not affect the connectivity of the virtual backbone. This can be achieved by increasing the connectivity of the SCDAS representing the virtual backbone.

Under such a model and requirements, we formulate the fault tolerant virtual backbone problem as follows:

k-Strongly Connected m Dominating and Absorbing Set problem ((k,m) SCDAS): Given a directed graph $G = (V, E)$ representing a sensor network and two positive integers k and m, find a subset $C \subseteq V$ with a minimum size and satisfying the following conditions:

- C is an SCDAS
- The subgraph $G(C)$ is k-connected
- Each node not in C is dominated and absorbed by at least m nodes in C

3 Enhancing Domination and Absorption of the Virtual Backbone

In this section we study enhancing fault tolerance of the virtual backbone in the WSN represented by a directed disk graph $G = (V, E)$. The fault tolerance of a virtual backbone needs to be enhanced in two aspects, firstly in terms of domination and absorbtion, secondly, in terms of connectivity of the subgraph induced by the virtual backbone nodes. As a fault tolerant virtual backbone in the WSN with unidirectional and bidirectional links can be represented by a (k, m) SCDAS, hence, we propose an approximation algorithm for constructing a (k, m) SCDAS for a directed disk graph $G = (V, E)$ representing the WSN for any value of k and m. We also provide the theoretical analysis of our algorithm and conjectured its approximation ratio. The (k, m) SCDAS of graph G represents the virtual backbone of the WSN such that any node v not in virtual backbone has at least m virtual backbone nodes in $N^+(v)$ and $N^-(v)$, respectively, and the graph induced by the virtual backbone nodes is k-strongly connected. This ensures that the virtual backbone can sustain $m - 1$ virtual backbone nodes failure without isolating any non-virtual backbone node from the virtual backbone and it can sustain $k - 1$ virtual backbone nodes failure without disconnecting the virtual backbone. In order to generate a (k, m) SCDAS we first generate an $(1, m)$ SCDAS, which is a special case of (k, m) SCDAS where $k = 1$. We then enhance the connectivity of the subgraph induced by the $(1, m)$ SCDAS nodes to make it k-node connected, which results in a (k, m) SCDAS.

3.1 An Approximation Algorithm for (k, m) SCDAS Problem

The algorithm for generating the (k, m) SCDAS is illustrated in Algorithm 1. In order to generate a (k, m) SCDAS we first generate a $(1, m)$ SCDAS, which is a special case of (k, m) SCDAS where $k = 1$. The algorithm for generating a $(1, m)$ SCDAS is illustrated in Algorithm 2.

The construction of $(1, m)$ SCDAS is divided into two phases. In the first phase a strongly connected dominating and absorbing set (SCDAS) is generated and then in the second phase extra nodes are iteratively added to make it m dominating. In the first phase a strongly connected dominating and absorbing set is generated by

Algorithm 1 Approximation Algorithm for (k, m)-Strongly Connected Dominating and Absorbing Set

```
INPUT: An $m$-connected directed graph $G = (V, E)$, here $m \geq k$
OUTPUT: A $(k, m)$ SCDAS $C$ of $G$
Run Algorithm 2 on $G$ to generated an $(1, m)$ SCDAS $C$.
for Every pair of black nodes $v_i, v_j \in C$ do
    $C = C \cup \mathrm{Find\,k\,Path}\,(G, i, j, k)$
end for
Return $C$ as the $k$-$m$-SCDAS
```

Algorithm 2 Algorithm for $(1, m)$ Strongly Connected Dominating and Absorbing Set

```
INPUT: An $m$ connected directed graph $G = (V, E)$
OUTPUT: A $(1, m)$ SCDAS $C$ of $G$
Generate a directed graph $G'$ by reversing the edges of graph $G$
Select a nodes $s$ as a root.
$C = \emptyset$
$C = C \cup$ Find DS1 (G,s);
$C = C \cup$ Find DS1 ($G'$,s);
for $i = 1; i \leq m - 1; i++$ do
    Color all the Gray nodes in G and $G'$ White
    $C = C \cup$ Find DS2 (G)
    $C = C \cup$ Find DS2 ($G'$)
end for
The set $C$ is the $(1, m)$ SCDAS
```

calling Algorithm 3 twice. When Algorithm 3 terminates there are three different color nodes in the graph; the black nodes, the blue nodes, and the gray nodes. In first call to Algorithm 3 the graph G and a node s are passed as the parameter and it returns a set of black and blue nodes forming a directed dominating tree for G rooted at s. The black nodes in the tree form the dominating set of G and they are semi-independent to each other, they dominate all the gray nodes in the graph. The blue nodes act as connectors: they connect the black nodes in a way to form a directed tree rooted at s, as shown in Fig. 2a. In the second call to Algorithm 3 the inverse graph G' and the node s are passed as parameters. Similarly it returns a set of blue and black nodes forming a directed dominating tree for G' rooted at s. As the graph G' is the inverse graph of G, hence, the set of blue and black nodes forming a directed dominating tree for G' equivalently forms a directed absorbing tree in G, as shown in Fig. 2b. For all the gray nodes in G' the corresponding nodes in G are absorbed by the nodes in G corresponding to all the black nodes in G'. The union of the set of blue and black nodes returned for G and the set of nodes in G corresponding to the set nodes returned for G' forms a strongly connected dominating and absorbing set for G.

In the second phase extra nodes are added to enhance the dominance and the absorption of the strongly connected dominated and absorbing set to m. In order to do this $m - 1$ iterations are performed and in each iteration the Algorithm 4

Algorithm 3 Find DS1 (G,s)

```
Set $S = \emptyset$
$S = S \cup s$
$BLACK = \emptyset$; $BLUE = \emptyset$
while There is a white node in $G$ do
    Select a White node $u \in S$ having maximum number of white nodes in $N^{+}(u)$
    Color $u$ black and remove it from $S$
    $BLACK = BLACK \cup u$
    if $u! = s$ then
        Color the $Parent$(u) Blue if it is Gray
        $BLUE = BLUE \cup Parent$(u)
    end if
    Color all the nodes $v \in N^{+}(u)$ Gray
    for All the White node $w \in N^{+}(v)$ do
        if $w \notin S$ then
            $S = S \cup w$
        end if
        Mark $v$ as the parent of $w$
    end for
end while
Return $BLACK \cup BLUE$
```

Algorithm 4 Find DS2(G)

```
$BLACK = \emptyset$
while There is a White node in $G$ do
    Select a White node $u$ having maximum number of white nodes in $N^{+}(u)$
    Color $u$ Black and all the White node $v \in N^{+}(u)$ Gray
    $BLACK = BLACK \cup u$
end while
Return $BLACK$
```

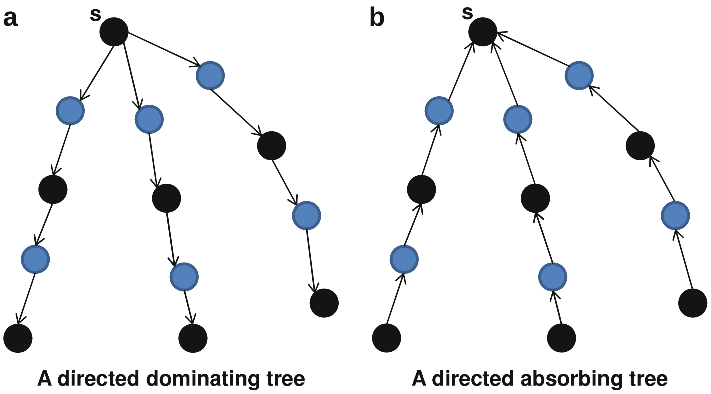


Fig. 2 A directed dominating and absorbing tree

Algorithm 5 Find k-Path(G, i, j, k)

```
Keeping vertex $v_i$ as source and $v_j$ as destination, construct a flow network $G_f$ of $G$ with $2 * |V-2| + 2$ vertices and $|E| + |V-2|$ edges
$S = \phi$
$G_r \leftarrow G_f$
$flow \leftarrow 0$
for $l = 1; l < k; l++$ do
    Find an augmented path from $v_i$ to $v_j$ in $G_r$ by increasing the flow by 1 unit.
    For all the saturated edges $(v_{in}, v_{out})$ on this augmented path color the corresponding nodes $v$ in $G$ blue if they are white and add them to $S$
    Update the residual network $G_r$
end for
Return $S$
```

is called twice. In the first call G is passed as a parameter, this results in the enhancement of the dominance of the SCDAS by one. In the second call G' is passed as a parameter, which results in the enhancement of the absorption of the SCDAS by one. After $m-1$ the SCDAS becomes $(1, m)$ SCDAS.

Once the $(1, m)$ SCDAS is formed, for each ordered pair of nodes in $(1, m)$ SCDAS, $k-1$ node disjoint paths are identified by running Find k-Path algorithm given in Algorithm 5. All white nodes on these paths are colored blue and are included in the virtual backbone. These nodes are called the connector nodes. Now, as the domination and the absorption of the virtual backbone is m and the connectivity of the subgraph generated by the virtual backbone nodes is k, hence, it forms a (k, m) SCDAS. One important thing to be noticed here is that to ensure that the subgraph $G((k, m)$ SCDAS) is k-connected and the graph G should be at least m-connected and $m \geq k$.

The Find k-Path Algorithm: The Find k-path Algorithm is illustrated in Algorithm 5. Given a k-connected directed graph $G = (V, E)$, and a pair of vertices $v_i, v_j \in V$, the algorithm finds the set of nodes on k node disjoint paths from v_i to v_j in graph G. The algorithm first generates a flow network G_f by partitioning each node $v \in V \setminus \{v_i, v_j\}$ into two nodes v_{in} and v_{out}. Then connecting v_{in} and v_{out} through a unidirectional edge $(v_{\text{in}}, v_{\text{out}})$ and assign this edge a capacity of 1 unit. All the incoming edges directed towards v in G are set as incoming edges to v_{in} in G_f whereas all the outgoing edges emanating from v in G are set as outgoing edges from v_{out} in G_f assign infinite capacity to these edges, this results in G_f having $2|V-2|+2$ nodes and $|E|+|V-2|$ edges. Once the flow network G_f of graph G is formed, we run k iterations and in each iteration an augmented path in G_f from v_i to v_j is determined by increasing the flow by 1 unit. We consider that a unit flow is indivisible. On each newly found augmented path, for any saturated edge $(v_{\text{in}}, v_{\text{out}})$ we select the corresponding vertex v in G as a node on the k disjoint path from v_i to v_j. Figure 3 show the iterations of finding augmented paths between v_i and v_j for $k = 2$.

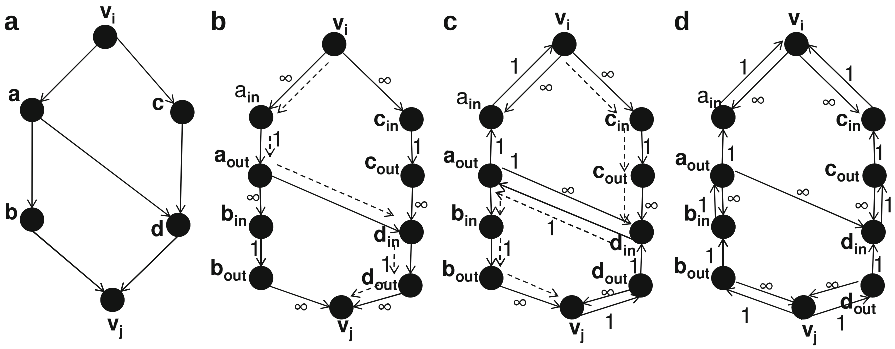


Fig. 3 Iterations for finding augmented paths for $k = 2$

3.1.1 Theoretical Analysis

Lemma 1. *The Algorithm 2 is correct and produces a virtual backbone which is* $(1, m)$ *SCDAS.*

Proof. In order to prove this lemma we need to show that the virtual backbone formed is strongly connected and m dominating and absorbing. The algorithm works in two phases, in the first phase it generates a dominating tree and an absorbing tree for the graph G. Let the set of blue and black nodes forming the dominating tree represented as D and the set of blue and black nodes forming the absorbing tree be represented as A. Now let the set of black nodes in D and A be represented as *Black*(D) and *Black*(A) respectively. The node s has a directed path using blue and black nodes to all the nodes in *Black*(D), and all the nodes in *Black*(A) has a directed path using blue and black nodes to s. Hence, the black nodes in *Black*(D)∩ *Black*(A) has a path using blue and black nodes to all the other blue and black node in $D \cup A$. The black nodes in *Black*(A) \ *Black*(D) have a directed blue–black path to s and as this node must be dominated by some black node in $Black(D)$ it must also have a directed blue–black path from s to it through its dominator. Similarly the nodes in *Black*(D) \ *Black*(A) will have a blue–black path from and to the root node s. As all the black nodes in *Black*(D)∪ *Black*(A) have a directed blue–black path from and to the root s, hence, all the nodes in $D \cup A$ are strongly connected and forms a SCDAS.

In the second phase extra nodes are added to enhance the domination and the absorption of the virtual backbone by $m - 1$. As all these extra nodes are dominated and absorbed by black nodes, hence, the extended virtual backbone will still be strongly connected. □

Lemma 2. *The number of Semi-Independent neighbors* K_{SI} *of any node u can be bounded by* $(2R + 1)^2$. *Here* $R = \frac{r_{max}}{r_{min}}$.

Proof. Let u be the node with transmission range $r_{\max}$. The number of semi-independent neighbors of u, i.e., $N^+(u) \cap SI$ can be bounded by K_{SI}. It can be noticed that the distance between any two nodes v and $w \in SI$, i.e., $d(v, w) > r_{\min}$. Hence, the size of $N^+(u) \cap SI$, i.e., K_{SI} is bounded by the number of disjoint disks with radius $r_{\min}/2$ packing in the disk centered at u with radius of $r_{\max} + r_{\min}/2$. So, we have:

$$\left|N^+(v) \cap S\right| \leq \frac{\pi(r_{\max} + r_{\min}/2)^2}{\pi(r_{\min}/2)^2} \leq (2R+1)^2. \tag{1}$$

□

Lemma 3. *Let $G = (V, E)$ be any directed graph (DG) with bounded transmission range ratio R, then the number of black nodes in a dominating set of G obtained on calling Algorithm 3 or the number of black nodes added on every call to Algorithm 4 in order to enhance the domination by 1 is bounded by: $|DS| \leq \left(\frac{K_{SI}}{m} + 1\right)|DS^*_m|$, here DS^*_m is the optimal solution for m dominating set of G.*

Proof. Let us consider DS and DS^*_m, there are two possible cases:

1. $\mathrm{DS} \subseteq \mathrm{DS}^*_m$
2. $\mathrm{DS} \subsetneq \mathrm{DS}^*_m$

Case (a): As $\mathrm{DS} \subseteq \mathrm{DS}^*_m$, we have $|\mathrm{DS}| \leq |\mathrm{DS}^*_m|$.
Case (b): $\forall u \in \mathrm{DS} \setminus \mathrm{DS}^*_m$, let $D_u = |\mathrm{DS}^*_m \cap N^-(u)|$. As DS^*_m, is an m dominating set of G, $D_u \geq m$ for each $u \in \mathrm{DS} \setminus \mathrm{DS}^*_m$ and we have:

$$\sum_{u \in \mathrm{DS} \setminus \mathrm{DS}^*_m} D_u \geq m\, |\mathrm{DS} \setminus \mathrm{DS}^*_m|\,. \tag{2}$$

For all $v \in \mathrm{DS}^*_m$, let

$$d_v = |(\mathrm{DS} \setminus \mathrm{DS}^*_m) \cap N^+(v)|. \tag{3}$$

As the black nodes in DS obtained on calling Algorithm 3 or obtained in every call to Algorithm 4 cannot have a bidirectional edge between each other, hence, they form a Semi-Independent set. From Lemma 2, we have $\forall v \in \mathrm{DS}^*_m$ there are at most K_{SI} Semi-independent nodes in its neighborhood, hence $d_v \leq K_{\mathrm{SI}}$. Therefore we have:

$$K_{SI}|\mathrm{DS}^*_m| \geq \sum_{v \in \mathrm{DS}^*_m} d_v. \tag{4}$$

However note that

$$\sum_{u \in \mathrm{DS} \setminus \mathrm{DS}^*_m} D_u = |\{(v, u) \in E | u \in \mathrm{DS} \setminus \mathrm{DS}^*_m, v \in \mathrm{DS}^*_m\}| = \sum_{u \in \mathrm{DS}^*_m} d_v. \tag{5}$$

From (2), (4) and (5) we have:

$$m|\mathrm{DS} \setminus \mathrm{DS}_m^*| \leq \sum_{u \in \mathrm{DS} \setminus \mathrm{DS}_m^*} D_u = \sum_{u \in \mathrm{DS}_m^*} d_v \leq K_{\mathrm{SI}}|\mathrm{DS}_m^*|. \tag{6}$$

Therefore,

$$m|\mathrm{DS} \setminus \mathrm{DS}_m^*| \leq K_{\mathrm{SI}}|\mathrm{DS}_m^*|. \tag{7}$$

Thus it follows that,

$$|\mathrm{DS}| \leq \left(\frac{K_{\mathrm{SI}}}{m} + 1\right)|\mathrm{DS}_m^*|. \tag{8}$$

Therefore from the two cases (a) and (b), we conclude that

$$|\mathrm{DS}| \leq \left(\frac{K_{\mathrm{SI}}}{m} + 1\right)|\mathrm{DS}_m^*|. \tag{9}$$

□

Lemma 4. *The number of nodes in m dominating set of G is at most $(K_{SI} + m)$ $|DS_{1,m}^*|$, here $DS_{1,m}^*$ is the optimal solution for $(1, m)$ SCDS.*

Proof. The number of nodes in m dominating set are $|\mathrm{DS}^1 \cup \mathrm{DS}^2 \cdots \cup \mathrm{DS}^n|$. Here DS^i is the set of nodes added in the ith iteration. Let $|\mathrm{DS}| = \max_i\{|\mathrm{DS}^i|\}$, then from Lemma 3, we have:

$$\left|\mathrm{DS}^1 \cup \mathrm{DS}^2 \ldots \cup \mathrm{DS}^n\right| \leq m|\mathrm{DS}| \leq m\left(\frac{K_{\mathrm{SI}}}{m} + 1\right)|\mathrm{DS}_m^*| \leq (K_{\mathrm{SI}} + m)|\mathrm{DS}_{1,m}^*|. \tag{10}$$

□

Theorem 1. *The Algorithm 2 produces a $(1, m)$ SCDAS with the size bounded by $2\left(K_{SI}\left(1 + \frac{1}{m}\right) + m + 1\right)|DAS_{1,m}^*|$, here $DAS_{1,m}^*$ is the optimal solution for $(1, m)$ SCDAS.*

Proof. Let C denotes our solution to the $(1, m)$ SCDS. Let $BLUE$ and $BLACK$ be the set of blue and black nodes in G and $BLUE'$ and $BLACK'$ be the set of blue and black nodes in G' respectively. Then we have:

$$|C| = |BLUE| + |BLACK| + \left|BLUE'\right| + \left|BLACK'\right|. \tag{11}$$

When the Algorithm 3 runs on G and G' it results in a dominating tree for each of them, respectively. For both G and G' the dominating tree is rooted at same node s. The dominating tree for G' is equivalent to the absorbing tree for G. On every branch of these trees black and blue nodes are placed in alternative sequence, starting from a black node. Hence, we have $|BLUE| = |BLACK| - 1$. According to Lemma 3:

The number of black nodes generated on calling Algorithm 3 is bounded by $\left(\frac{K_{\text{SI}}}{m} + 1\right) |\text{DS}^*_{1,m}|$.

Hence we have:

$$|BLUE| \leq \left(\frac{K_{\text{SI}}}{m} + 1\right) |\text{DS}^*_{1,m}| \leq \left(\frac{K_{\text{SI}}}{m} + 1\right) |\text{DAS}^*_{1,m}|, \tag{12}$$

$$|BLUE'| \leq \left(\frac{K_{\text{SI}}}{m} + 1\right) |\text{DS}^*_{1,m}| \leq \left(\frac{K_{\text{SI}}}{m} + 1\right) |\text{DAS}^*_{1,m}|. \tag{13}$$

According to Lemma 4 the total number of black nodes are bounded by $(K_{\text{SI}}+m) * |\text{DS}^*_{1,m}| \leq (K_{\text{SI}} + m) * |\text{DAS}^*_{1,m}|$. So we have:

$$|BLACK| \leq (K_{\text{SI}} + m)|\text{DAS}^*_{1,m}|, \tag{14}$$

$$|BLACK'| \leq (K_{\text{SI}} + m)|\text{DAS}^*_{1,m}|, \tag{15}$$

from (11) to (15) we have:

$$|C| \leq 2\left(K_{\text{SI}}\left(1 + \frac{1}{m}\right) + m + 1\right) |\text{DAS}^*_{1,m}|. \tag{16}$$

□

Lemma 5. *The Algorithm 5 for any two vertices $v_i, v_j \in G$ finds all the nodes on the k node disjoint directed paths from v_i to v_j.*

Proof. In order to find the nodes on k node disjoint directed paths from v_i to v_j, the Algorithm 5 run k iterations. In each iteration it finds a new augmented path from v_i to v_j by increasing the flow by 1 unit. As the graph G is k-connected, hence, there will be at least k augmented paths existing. As we consider that a unit flow is indivisible, hence, in each iteration exactly a single directed linear augmented path will be explored and determined. As the capacity of all edge $(v_{\text{in}}, v_{\text{out}})$ is 1 unit, hence, each of them can be used in a single augmented path. This ensures that any node v is selected only for a single path from v_i to v_j in G which ensures the node disjointness of the k paths explored. As the Algorithm 5 can find k-augmented paths ensuring any edge $(v_{\text{in}}, v_{\text{out}})$ can be used exactly in one augmented path, this will result into finding all the nodes on k node disjoint paths from v_i to v_j. □

Lemma 6. *The Algorithm 1 is correct and produces a virtual backbone which is m dominating and absorbing, and the subgraph generated by virtual backbone nodes is k-connected.*

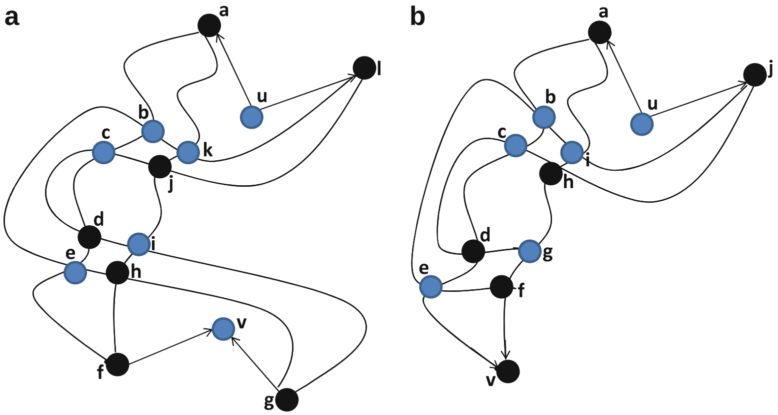


Fig. 4 Examples for showing the existence of k nodes disjoint path from a blue node to a blue or black node

Proof. Algorithm 1 first considers a virtual backbone as $(1, m)$ SACDAS generated by Algorithm 2 and extends it to a (k, m) SCDAS. As shown in Lemma 1 the $(1, m)$ SCDAS is m dominating and absorbing. Hence, the (k, m) SCDAS will also be m dominating and absorbing. Now we have to show that the subgraph $G((k, m)$ SCDAS) is generated by nodes in (k, m) SCDAS is k-connected. Using Algorithm 5 all the black nodes in $(1, m)$ SCDAS are connected to each other through k disjoint paths. This is proved in Lemma 1. Now as all the black nodes are connected to each other through k nodes disjoint paths we only need to show that all the blue and the black nodes as well as blue and blue nodes are connected to each other through k node disjoint paths. As any blue node v has at least m black nodes in $N^+(v)$ and $N^+(v)$ respectively and $m \geq k$. This will ensure that there will be at least k node disjoint paths from any blue node to another blue node through its absorbing black nodes. As shown in the example depicted in Fig. 4a, for $k = 2, m = 2$, their exists 2 node disjoint paths from a blue node u to another blue node v. Similarly there will be k node disjoint blue–black paths from any blue node to all the black nodes. Figure 4b shows that for $k = 2, m = 2$, here a blue node u is having 2 node disjoint directed blue–black paths to a black node v. □

Conjecture 1. In Algorithm 5 the number of blue connector nodes needed to make the black nodes in $(1, m)$ SCDAS connected to each other through k node disjoint paths are bounded by $2k * 2(K_{\mathrm{SI}} + m)|\mathrm{DAS}^*_{1,m}|$. Here, $2(K_{\mathrm{SI}} + m)|\mathrm{DAS}^*_{1,m}|$ is the bound on number of black nodes in $(1, m)$ SCDAS according to inequality in (14) and (15).

Conjecture 2. Following Lemma 1 and Conjecture 1 the Algorithm 1 provides an approximation ratio $2(2k + 1)(K_{\mathrm{SI}} + m)$ for the (k, m) SCDAS problem.

4 Conclusion

In this chapter we studied the fault tolerance of the virtual backbone in a sensor network having both unidirectional and bidirectional links. We modeled the sensor network as a directed disk graph and formulated the problem of finding a virtual backbone as a (k, m) SCDAS problem for any value of k and m. The (k, m) SCDAS of any directed graph $G = (V, E)$ represents a virtual backbone, such that, for every node not in the virtual backbone there are at least m virtual backbone nodes in its incoming and outgoing neighborhood, respectively and the subgraph induced by the virtual backbone nodes is strongly k-connected. As the problem is NP-hard, we proposed an approximation algorithm and provided a conjecture on its approximation ratio.

References

1. A. Ephremides, J. Wieselthier, and D. Baker, A Design Concept for Reliable Mobile Radio Networks with Frequency Hopping Signaling, *Proceedings of IEEE*, 75(1):56–73, 1987.
2. B. Clark, C. Colbourn, and D. Johnson, Unit Disk Graphs, *Discrete Mathematics*, vol. 86, pp. 165–177, 1990.
3. B. Das, R. Sivakumar, and V. Bharghavan, Routing in Ad Hoc Networks Using a Spine, *International Conferfence on Computers and Communication Networks*, 1997.
4. D.-Z. Du, M.T. Thai, Y. Li, D. Liu, S. Zhu, “Strongly Connected Dominating Sets in Wireless Sensor Networks with Unidirectional Links”, *In Proceedings of APWEB, LNCS*, 2006.
5. F. Dai and J. Wu, An Extended Localized Algorithms for Connected Dominating Set Formation in Ad Hoc Wireless Networks, *IEEE Transactions on Parallel and Distributed Systems*, vol. 15, no. 10, October 2004
6. F. Dai and J. Wu, On Construction *k*-Connected *k*-Dominating Set in Wireless Network, *Parallel and Distributed Processing Symposium, 2005. Proceedings.19th IEEE International*, 2005.
7. F. Wang, M.T. Thai, D.-Z. Du, “On the construction of 2-connected virtual backbone in wireless network”, *IEEE Transactions on Wireless Communications*, 2007.
8. J. Wu, Extended Dominating-Set-Based Routing in Ad Hoc Wireless Networks with Unidirectional Links, *IEEE Transactions on Parallel and Distributed Computing*, 22, 1–4, 2002, 327–340.
9. Kwang-Fu Li, Yueh-Hsai, Chia-Ching Li, Find-reassemble-path algorithm for finding node disjoint paths in telecommunications network with two technologies, *EUROCON 2001*.
10. M. Park, C. Wang, J. Willson, M.T. Thai, W. Wu, and A. Farago, A Dominating and Absorbent Set in Wireless Ad-hoc Networks with Different Transmission Range, *Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC)*, 2007.
11. M.R. Garey, D. S. Johnson, *Computers and Intractability. A guide to the Theory of NP-completeness*, Freeman, New York, 1979.
12. M.T. Thai, R. Tiwari, and D.-Z Du, On Construction of Virtual Backbone in Wireless Sensor Networks with Unidirectional Links *IEEE Transaction on Mobile Computing (TMC)*, vol. 7, no. 8, pp. 1–12, 2008.
13. M.T. Thai, D.-Z. Du, Connected Dominating Sets in Disk Graphs with Bidirectional Links, *IEEE Communications Letters*, vol. 10, no. 3, pp. 138–140, March 2006.

14. M.T. Thai, F. Wang, D. Liu, S. Zhu, and D. Z. Du, Connected Dominating Sets in Wireless Networks with Different Transmission Ranges, *IEEE Transactions on Mobile Computing,* vol. 3, no. 2, pp. 145–152, 2007.
15. M.T. Thai, N. Zhang, R. Tiwari, and X. Xu, On Approximation Algorithms of k-Connected m-Dominating Sets in Disk Graphs, *Theoretical Computer Science journal*, vol. 385, no. 1–3, pp. 49–59, 2007.
16. N. Zhang, I. Shin, F. Zou, W. Wu, and M.T. Thai, Trade-Off Scheme for Fault Tolerant Connected Dominating Sets on Size and Diameter, in *Proceedings of ACM International Workshop on Foudations of Wireless Ad Hoc and Sensor Networking and Computing (FOWANC)*, in conjunction with MobiHoc, 2008.
17. P.-J. Wan, K.M. Alzoubi, and O. Frieder, Distributed Construction on Connected Dominating Set in Wireless Ad Hoc Networks, *Proceedings of the Conference of the IEEE Communications Society (INFOCOM),* 2002.
18. P. Sinha, R. Sivakumar, and V. Bharghavan, Enhancing ad hoc routing with dynamic virtual infrastructures, in *Proceedings of Infocom*, 2001.
19. S. Guha and S. Khuller, Approximation Algorithms for Connected Dominating Sets, *Algorithmica*, vol. 20, pp. 374–387, 1998
20. S,-Y. Ni, Y.-C. Tseng, Y.-S. Chen, and J.-P. Sheu, The Broadcast Strom Problem in a Mobile ad hoc Network, *Proceedings of MOBICOM*, 1999.
21. X. Cheng, X. Huang, D. Li, W. Wu, and D.-Z. Du, Polynomial-Time Approximation Scheme for Minimum Connected Dominating Set in Ad Hoc Wireless Networks, *Networks,* vol. 42, no. 4, pp. 202–208, 2003.
22. Y. Li, S. Zhu, M.T. Thai, and D.-Z. Du, Localized Construction of Connected Dominating Set in Wireless Networks, *NSF International Workshop on Thoretical Aspects of Wireless Ad Hoc, Sensor and Peer-to-Peer Networks*, 2004.
23. Y. Li, M.T. Thai, F. Wang, C.-W. Yi, P.-J. Wang, and D.-Z. Du, On Greedy Construction of Connected Dominating Sets in Wireless Networks, *Special issue of Wireless Communications and Mobile Computing (WCMC),* vol. 5, no. 88, pp. 927–932, 2005.
24. Y. Wu, F. Wang, M.T. Thai, and Y. Li, Constructing k-Connected m-Dominating Sets in Wireless Sensor Networks, in *Proceedings of Military Communications Conference (MILCOM 2007)*, October, 2007.

Constrained Node Placement and Assignment in Mobile Backbone Networks

Emily M. Craparo

Abstract This chapter describes new algorithms for mobile backbone network optimization. In this hierarchical communication framework, *mobile backbone nodes* (MBNs) are deployed to provide communication support for *regular nodes* (RNs). While previous work has assumed that MBNs are unconstrained in position, this work models constraints in MBN location. This chapter develops an exact technique for maximizing the number of RNs that achieve a threshold throughput level, as well as a polynomial-time approximation algorithm for this problem. We show that the approximation algorithm carries a performance guarantee of $\frac{1}{2}$ and demonstrated that this guarantee is tight in some problem instances.

1 Introduction and Background

Data collected by distributed sensor networks often must be collected or aggregated in a central location. The mobile backbone network architecture has been proposed to alleviate scalability problems in ad hoc wireless networks [1, 2], which can hinder the deployment of large-scale distributed sensing platforms. Noting that most communication capacity in large-scale single-layer mobile networks is dedicated to packet-forwarding and routing overhead, Xu et al. propose a multi-layer hierarchical network architecture and demonstrate the improved scalability of a two-layer framework [2]. Srinivas et al. [3] define two types of nodes: regular nodes (RNs), which have restricted mobility and limited communication capability, and mobile backbone nodes (MBNs), which have superior communication capability and which can be deployed to provide communication support for the RNs.

E.M. Craparo (✉)
Department of Operations Research, Naval Postgraduate School, Monterey, CA, USA
e-mail: emcrapar@nps.edu

V.L. Boginski et al. (eds.), *Sensors: Theory, Algorithms, and Applications*,
Springer Optimization and Its Applications 61, DOI 10.1007/978-0-387-88619-0_2,

In addition to scaling well with network size, the mobile backbone network architecture naturally models a variety of real-world systems, such as airborne communication hubs that are deployed to provide communication support for ground platforms, or mobile agents that are positioned to collect data from stationary sensor nodes.

Srinivas et al. [4] and Craparo et al. [5] address problems involving simultaneous MBN placement and RN assignment. Both [4] and [5] seek to simultaneously *place* K MBNs, which can occupy any location in the plane, and *assign* N RNs to the MBNs, in order to optimize a various throughput characteristics of the network. Srinivas et al. describe an enumeration-based exact algorithm and several heuristics for maximizing the minimum throughput achieved by any RN [4]. Craparo et al. study the problem of maximizing the number of RNs that achieve a threshold throughput level $\tau_{\min}$; they propose an exact algorithm based on mixed-integer linear programming, as well as a polynomial-time approximation algorithm with a constant-factor performance guarantee [5].

A key feature of the formulations in [4] and [5] concerns the potential locations of the MBNs. Although the MBNs can feasibly occupy any locations in the plane, [4] and [5] demonstrate that the MBNs can be restricted to a relatively small set of locations ($O(N^3)$) without compromising the optimality of the overall solution. In particular, each MBN can be placed at the *1-center* of its assigned RNs. (A MBN is located at the *1-center* of a set of RNs if the maximum distance from the MBN to the any of the RNs in the set is minimized.) Additionally, each 1-center location l is associated with a unique radius of communication. This radius is the maximum possible distance between the MBN at location l and any of the RNs in subsets for which l is a 1-center [5]. Thus, the restriction of MBNs to 1-center locations not only dramatically reduces the size of the feasible set of MBN locations, but also removes the communication radius as an independent decision variable in the optimization problem.

In the formulations of [4] and [5], it is always possible to place MBNs in 1-center locations because the MBNs are assumed to be capable of occupying *any* location. In some applications, this assumption is valid. For instance, an airborne communication hub (e.g., a blimp) could easily be placed at the 1-center of its assigned RNs. In other applications, however, the potential locations of the MBNs may be limited. In hastily-formed networks operating in disaster areas, for instance, ground-based communication hubs are generally restricted to public spaces such as schools, hospitals, and police stations [6]. In this case, the mobile backbone network optimization problem is *constrained*, in the sense that the MBNs can occupy only a discrete set of locations, and these potential locations are given as input data. In this application, it is generally impossible to place each MBN at the 1-center of its assigned RNs. Although the restriction of MBNs to a finite set of locations can reduce the size of the solution space with respect to MBN placement, the maximum communication radius of each MBN is a separate decision variable in this case, and the formulations of [4] and [5] are inappropriate. This work describes a mobile backbone network optimization problem with MBN placement constraints and provides exact and approximation algorithms for solving this problem, along with full proofs of results as previously described in [7].

2 Problem Statement

We use the communication model of [4] and [5], in which the throughput τ that can be achieved between a RN n and a MBN k, is a *monotonically nonincreasing* function of two quantities: the *distance* between n and k, and the *number* of RNs that are assigned to k (and thus interfere with n's transmissions). We assume that each RNs are assigned to one MBN encounter, no interference from RNs assigned to other MBNs (for example, because each "cluster" consisting of an MBN and its assigned RNs operates on a dedicated frequency).

Under such a throughput model, we pose the *constrained placement and assignment* (CPA) problem as follows: given a set of N RNs distributed in a plane, *place* K MBNs in the plane while simultaneously *assigning* the RNs to the MBNs, such that the number of RNs that achieve throughput at least $\tau_{\min}$ is maximized. MBNs can occupy locations from the set $\mathrm{L} = \{1, \ldots, L\}$, $L \geq K$, and each RN can be assigned to at most one MBN.

We do not require the MBNs to be "connected" to one another; this model is appropriate for applications in which MBNs serve to provide a satellite uplink for RNs, such as in the hastily-formed networks as mentioned in Sect. 1. It is also appropriate for applications in which the MBNs are powerful enough to communicate effectively with one another over the entire problem domain. We also assume that the positions of RNs are known exactly, through the use of GPS, for example.

Problem CPA is similar to the message ferrying problem, in which RNs have a finite amount of data available to transmit, and MBNs must efficiently collect this data [8–11]. CPA differs in that as it does not assume that the RNs have a limited amount of data to transmit; rather, CPA seeks to provide throughput on a permanent basis. In this sense, CPA is similar to a facility location problem. However, whereas CPA seeks to efficiently utilize a limited resource (the MBNs), most facility location problems focus on servicing all customers at minimum cost. Additionally, the throughput model in this work does not correspond to a notion of "service" in any known facility location problem. CPA is also similar to cellular network optimization; however, most approaches to cellular network optimization involve decomposition of the problem. Some formulations take base station placement as an input and optimize over user assignment and transmission power, with the objective of minimizing total interference [12–15]. Others use a simple heuristic for the assignment of users to base stations and to focus on selection of base station locations [16, 17]. In contrast, CPA seeks to optimize the network *simultaneously* over MBN placement and RN assignment, without assuming that RNs have variable transmission power capabilities.

3 Network Design Formulation

A key insight concerning the structure of the throughput function facilitates solution of CPA. Consider a cluster of nodes consisting of an MBN and its assigned RNs. Note that if the RN that is farthest away from the MBN achieves throughput of at least $\tau_{\min}$, then all other RNs in the cluster also achieve throughput of at least $\tau_{\min}$. Thus, in order to guarantee that all regular nodes in a cluster achieve adequate throughput, we need only to ensure that the most distant RN in the cluster achieves throughput of at least $\tau_{\min}$ [5].

Leveraging this insight, we can obtain an optimal solution to the simultaneous MBN placement and RN assignment problem via a *network design* formulation. In network design problems, a given network can be augmented with additional arcs for a given cost, and the objective is to "purchase" a set of augmenting arcs, subject to a budget constraint, in order to optimize flow in some way [18]. The formulation of the network design problem used in this work is similar to that presented in [5], in that the geometry and throughput characteristics of the problem are captured in the structure of the network design graph. Relative to the formulation in [5], however, we must use additional constraints in the network design problem. These constraints account for the fact that the communication radius of each MBN is an independent decision variable, i.e., it is not uniquely determined by the selection of the MBN location.

Our network design problem is formulated on a graph $G = (\mathcal{N}, \mathcal{A})$ of the form as shown schematically in Fig. 1. The graph G is constructed as follows:

The nodes of G consist of a source s, a sink t, and two node sets, $\mathrm{N} = \{n_1, \ldots, n_N\}$ and $\mathrm{M} = \{m_1^1, \ldots, m_L^N\}$. N represents the RNs, while M represents possible combinations of MBN locations and communication radii; node m_l^n represents the MBN at location l and that communicates with RNs within radius r_l^n of l, where r_l^n is the distance from location l to RN n. The source s is connected to each of the nodes in N via an arc of unit capacity. For each RN i, candidate MBN location l, and communication radius r_l^n, n_i is connected to node m_l^n if and only if $r_l^i \leq r_l^n$. All of the arcs connecting nodes in N to nodes in M have unit capacity. Finally, each node in M is connected to the sink, t. The capacity of the arc connecting node m_l^n to t is the product of a binary variable y_l^n and a constant c_l^n. The binary variable y_l^n represents the decision of whether to place the MBN at location l with maximum communication radius r_l^n. The constant c_l^n is the maximum number of RNs that can be assigned to the MBN at location l such that an RN at a distance r_l^n from l achieves throughput of at least $\tau_{\min}$. This quantity can be computed by means of a given throughput function, τ, and a desired minimum throughput level, $\tau_{\min}$. For an invertible throughput function, one can take the inverse of the function with respect to cluster size, evaluate the inverse at the desired minimum throughput level $\tau_{\min}$, and take the floor of the result to obtain an integer value for c_l^n. If the throughput function cannot easily be inverted with respect to cluster size, one can perform a search for the largest cluster size $c_l^n \leq N$ such that $\tau(c_l^n, r_l^n) \geq \tau_{\min}$. A binary search for c_l^n would involve $O(\log(N))$ evaluations of the function τ for each radius.

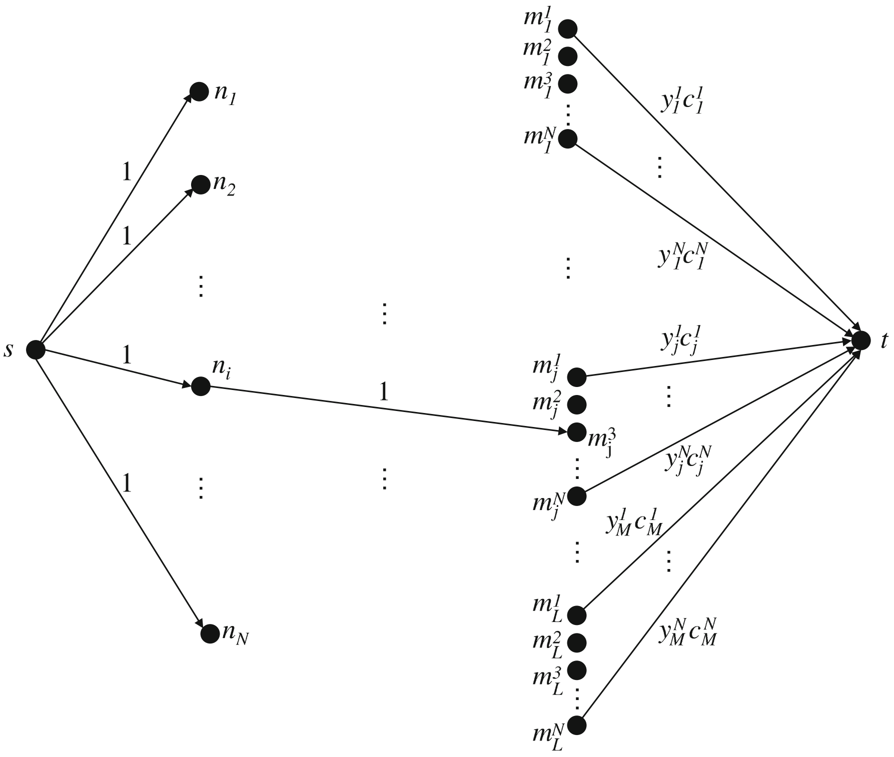


Fig. 1 Schematic representation of the graph on which an instance of the network design problem is posed

The objective of the network design problem is to “activate” a subset of the arcs entering t in such a way as to maximize the volume of flow that can travel from s to t. In addition to the capacity and flow conservation constraints typical of network models, the network design problem also includes cardinality and multiple-choice constraints. The cardinality constraint states that exactly K arcs are to be activated, reflecting the fact that K MBNs are available for placement. The multiple-choice constraints state that at most one arc with subscript l can be activated for each $l = 1, \ldots, L$. These constraints allow at most one MBN to be placed at each location; in other words, the locations $1, \ldots, L$ represent item classes, while the possible radii $r_l^1, \ldots, r_l^N$ represent items within each class, and the multiple-choice constraints state that at most one item can be selected from each class.

We denote the network design problem on G as the Multiple-Choice Network Design (MCND) problem. MCND can be solved via the following mixed-integer linear program (MILP):

$$\max_{\mathbf{x},\mathbf{y}} \sum_{i=1}^{N} x_{sn_i} \tag{1a}$$

$$\text{subject to} \quad \sum_{l=1}^{L}\sum_{n=1}^{N} y_l^n = K \tag{1b}$$

$$\sum_{n=1}^{N} y_l^n \leq 1 \qquad \forall\ l = 1,\ldots,L \tag{1c}$$

$$\sum_{i:(i,j)\in\mathcal{A}} x_{ij} = \sum_{k:(j,k)\in\mathcal{A}} x_{jk} \qquad j \in \mathcal{N} \setminus \{s,t\} \tag{1d}$$

$$x_{ij} \geq 0 \qquad \forall\ (i,j) \in \mathcal{A} \tag{1e}$$

$$x_{ij} \leq 1 \qquad \forall\ (i,j) \in \mathcal{A} : j \in \mathcal{N} \setminus \{t\} \tag{1f}$$

$$x_{m_l^n t} \leq y_l^n c_l^n \qquad \forall\ l,n \tag{1g}$$

$$x_{n_i m_l^n} \leq y_l^n \qquad \forall\ i,l,n \tag{1h}$$

$$y_l^n \in \{0,1\} \qquad \forall\ l,n. \tag{1i}$$

The objective of MCND is to maximize the flow of $\mathbf{x}$ that traverses G, which corresponds to the total number of RNs that can be assigned at throughput $\tau_{\min}$. Constraint (1b) states that K arcs (MBN locations) are to be selected, and constraint (1c) states that at most one MBN can be placed at each location. Constraints (1d)–(1g) are network flow constraints, stating that flow through all internal nodes must be conserved (1d) and that arc capacities must be observed (1e)–(1g). Constraint (1h) is a valid inequality that improves computational performance by reducing the size of the feasible set in the LP relaxation. Constraint (1i) ensures that y_l^n is binary for all l,n. Note that, for a given specification of the $\mathbf{y}$ vector, all flows of $\mathbf{x}$ are integer in all basic feasible solutions of the resulting linear network flow problem.

An optimal solution to a instance of MCND provides both the placement of MBNs and the assignment of RNs to MBNs. The MBN is placed at location l if $y_l^n = 1$ for some n. RN i is assigned to the MBN at location l if and only if the flow from node n_i to node m_l^j is equal to 1 for some j. The equivalence between MCND and the original problem CPA is more formally stated in Theorem 1.

Theorem 1 *Given an instance of CPA, the solution to the corresponding instance of MCND yields an optimal MBN placement and RN assignment.*

Proof. The proof of Theorem 1 appears in Appendix 1. □

3.1 *Hardness of Network Optimization*

Although an optimal solution to MCND provides an optimal solution to the corresponding instance of CPA, the MILP approach described above is not computationally tractable from a theoretical perspective. This fact motivates consideration of the fundamental tractability of CPA itself. If CPA is NP-hard, it may be difficult or impossible to find an exact algorithm that is significantly more efficient than the MILP approach. Unfortunately, CPA is indeed NP-hard.

Theorem 2 *Problem CPA is NP-hard.*

Proof. The proof of Theorem 2 appears in Appendix 2. □

4 Approximation Algorithm

The probable intractability of CPA motivates consideration of approximate techniques. This section describes the approximation algorithm for MCND that runs in polynomial time and has a constant-factor performance guarantee.

The approximation algorithm is based on the insight that the maximum number of RNs that can be assigned is a *submodular* function of the set of mobile MBN locations and communication radii that are selected. Given a finite ground set $D = \{1, \ldots, d\}$, a set function $f(S)$, $S \subseteq D$, is submodular if

$$f(S \cup \{i, j\}) - f(S \cup \{i\}) \leq f(S \cup \{j\}) - f(S) \tag{2}$$

for all $i, j \in D$, $i \neq j$ and $S \subset D \setminus \{i, j\}$ [19]. Theorem 3 describes the submodularity of the objective function in the context of problem MCND.

Theorem 3 *Given an instance of MCND on a graph G, the maximum flow that can be routed through G is a submodular function of the set of arcs incident to t that are selected.*

Proof. The proof of Theorem 3 is similar to that of Lemma 1 in [5] and will not be presented here. □

4.1 Submodular Maximization with Multiple-Choice and Cardinality Constraints

Submodular maximization has been studied in many contexts, and with a variety of constraints. Nemhauser et al. [20] showed that for maximization of a nondecreasing, nonnegative submodular function subject to a cardinality constraint, a greedy selection technique produces a solution whose objective value is within $1 - \frac{1}{e}$ of the optimal objective value, where e is the base of the natural logarithm [21]. Approximation algorithms have also been developed for submodular maximization subject to other constraints, for example, Sviridenko [23] described a polynomial-time algorithm for maximizing a nondecreasing, nonnegative submodular function subject to a knapsack constraint.

In MCND, we aim to maximize a nonnegative, nondecreasing submodular function subject to L multiple-choice constraints and one cardinality constraint.

Algorithm 1

```
$S \leftarrow \emptyset$
$maxflow \leftarrow 0$
$U \leftarrow \{1,\ldots,L\}$
for $k$=1 to $K$ do
    for $l \in U$ do
        for $n$=1 to $N$ do
            if $f(S \cup \{y_l^n\}) \geq maxflow$ then
                $maxflow \leftarrow f(S \cup \{y_l^n\})$
                $y^* \leftarrow y_l^n$
                $l^* \leftarrow l$
            end if
        end for
    end for
    $S \leftarrow S \cup \{y^*\}$
    $U \leftarrow U \setminus \{l^*\}$
end for
return $S$
```

This is a special case of the problem of submodular maximization under multiple linear constraints as described by Kulik et al. [24]. Kulik et al. described the approximation algorithm for this problem, however, their approximation algorithm runs in polynomial time only if the number of linear constraints is a fixed constant [24]. Because MCND has $O(L)$ linear constraints, this algorithm can be quite computationally intensive. Fortunately, a simple greedy approach provides a provably good solution to MCND.

Consider Algorithm 1. Algorithm 1 starts with an empty set of selected arcs, S, and iteratively adds the arc that produces the maximum increase in the objective value, f, while maintaining feasibility with respect to the multiple choice constraints. After K iterations, Algorithm 1 produces a solution that obeys both the multiple-choice and cardinality constraints of MCND. The running time of Algorithm 1 is polynomial in K, L, and N; it requires solution of $O(KLN)$ maximum flow problems on bipartite networks with at most $N + K + 2$ nodes each. Moreover, Algorithm 1 carries a theoretical performance guarantee, as stated in Theorem 4.

Theorem 4 *Algorithm 1 is an approximation algorithm for MCND with approximation guarantee* $\frac{1}{2}$.

Proof. The proof of Theorem 4 appears in Appendix 3. □

That is, if the optimal solution to an instance of MCND has objective value OPT, then Algorithm 1 produces a solution S such that $f(S) \geq \frac{1}{2}$OPT.

The performance guarantee of $\frac{1}{2}$ shown in Theorem 4 is indeed tight for some problem instances. For example, consider the instance of CPA shown in Fig. 2b, with $K = 2$, $\tau(c,r) = \frac{1}{cr^2}$, and $\tau_{\min} = 1$. The corresponding instance of MCND is shown in Fig. 2a. Note that on the first iteration of the greedy algorithm, nodes m_1^1, m_1^2, and m_2^1 are all optimal; each allows one unit of flow to traverse the graph.

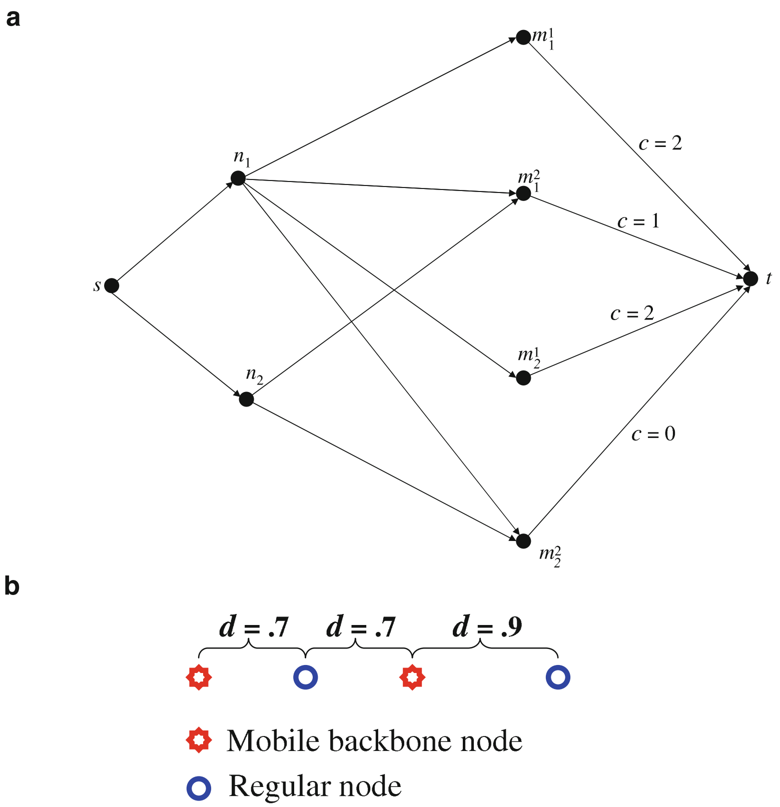


Fig. 2 Example of an instance of CPA for which the $\frac{1}{2}$ approximation guarantee of Algorithm 1 is tight. From left to right, the nodes shown are MBN 2, RN 1, MBN 1, and RN 2. **(a)** Example of an instance of MCND for which Algorithm 1 exactly achieves its performance guarantee. **(b)** A network optimization problem that yields the network design problem shown in Fig. 2a, for $\tau(c, r) = \frac{1}{cr^2}$ and $\tau_{\min} = 1$

Assume that the greedy algorithm selects node m_1^1. Then, on the greedy algorithm's second iteration, nodes m_2^1 and m_2^2 remain available for selection. However, neither of these nodes allows any additional flow to traverse the graph; thus, the total objective value obtained by the greedy algorithm is equal to 1, while an exact algorithm would have selected nodes m_1^2 and m_2^1 to obtain an objective value of 2.

While a theoretical performance guarantee is useful, the empirical performance of Algorithm 1 is also of interest. Figure 3 shows the average performance of Algorithm 1 relative to an exact (MILP) algorithm, for randomly-generated instances of CPA and their corresponding instances of MCND. Both RN locations and candidate MBN locations were generated according to a uniform distribution in a square area. As the figure indicates, Algorithm 1 tends to significantly outperform its performance guarantee, achieving average objective values up to 90% of those obtained by the exact algorithm, with a dramatic reduction in computation time. These results indicate that Algorithm 1 is a promising candidate for large-scale network design problems.

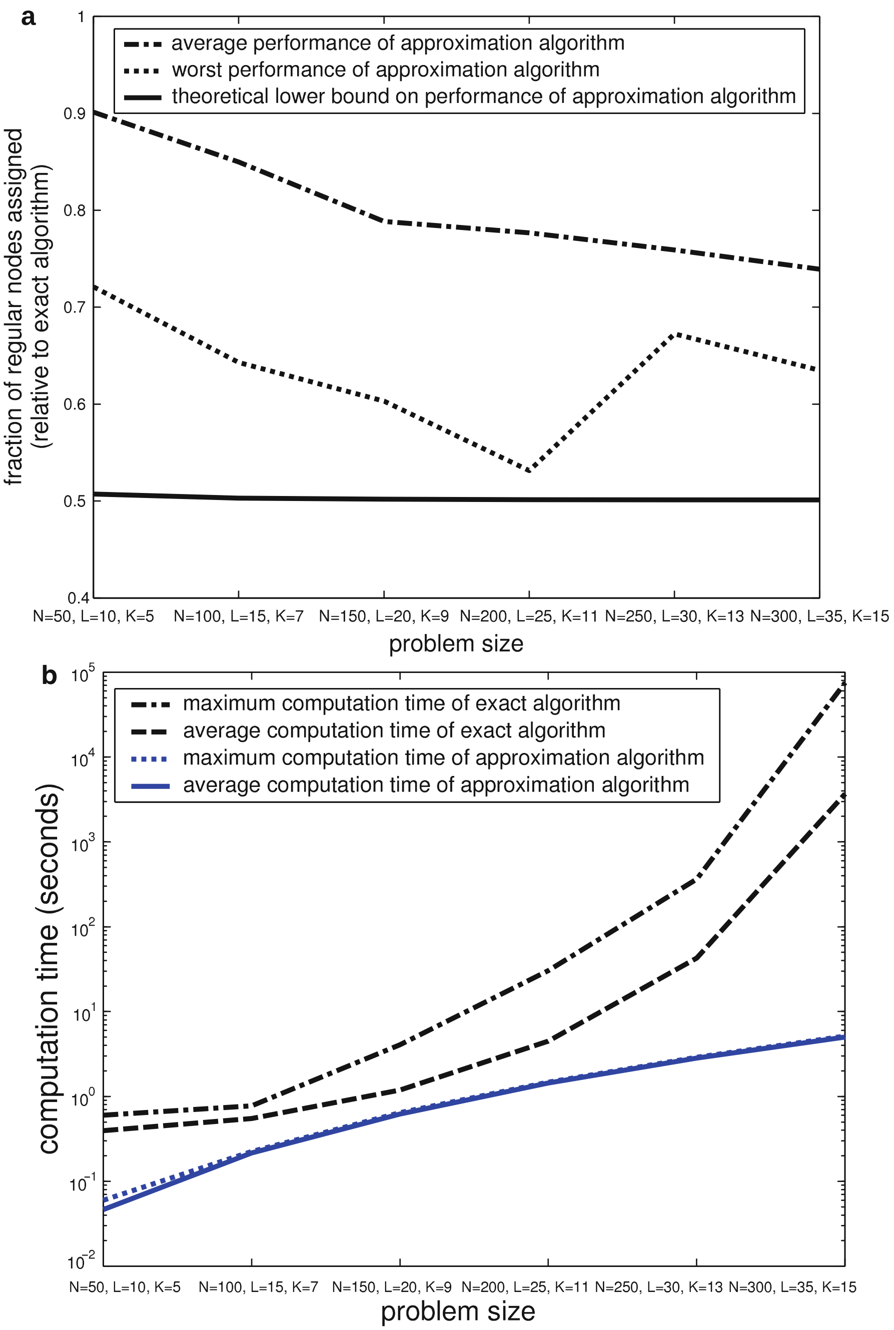


Fig. 3 Comparison of the exact and approximation algorithms developed in this work. **(a)** Performance of the approximation algorithm developed in this work, relative to an exact solution technique, in terms of number of RNs assigned at the given throughput level **(b)** Computation time of the approximation algorithm and the exact (MILP) algorithm for various problem sizes. Due to the large range of values represented, a logarithmic scale is used

5 Conclusion

This chapter has described algorithms for maximizing the number of RNs that achieve a threshold throughput level in a mobile backbone network. While previous work on this topic has assumed that MBNs are unconstrained in position, we model constraints in MBN location. Techniques described in this work include an exact algorithm based on mixed-integer linear programming (MILP) and polynomial-time approximation algorithm. Experimental results indicate that the approximation algorithm achieves good performance with a drastic reduction in computation time, making it suitable for a large-scale applications. We have also shown that the approximation algorithm carries a theoretical performance guarantee, and that this performance guarantee can indeed be tight in some instances, although the empirical performance of the approximation algorithm tends to exceed the performance guarantee.

Appendix 1. Proof of Theorem 1

Fix an instance of CPA, and consider a feasible solution to this instance, that is, a solution in which at most one MBN is placed in each location, each RN is assigned to at most one MBN, and each RN that is assigned to an MBN achieves throughput at least $\tau_{\min}$. Let A_k denote the set of RNs assigned to MBN k. Then, the objective value of this solution is $\sum_k |A_k|$. A solution to the corresponding instance of MCND with objective value $\sum_k |A_k|$ can be constructed as follows.

Consider MBN k occupying location l. Let f denote the most distant RN from k that is in set A_k, and let d_{fk} denote the distance from f to k. Then, set y_l^f equal to 1. For each RN $i \in A_k$, set x_{sn_i} and $x_{n_i m_l^f}$ equal to 1, and set $x_{n_i m}$ equal to zero for all other m. Note that the arc from node n_i to node $m_l^{n_f}$ is guaranteed to exist by construction. Flow conservation (constraint (1d)) is now satisfied at node n_i; repeating this process for each MBN $k = 1, \ldots, K$ results in constraint (1d) being satisfied for all $i \in \cup_k A_k$. For all $i \notin \cup_k A_k$, set x_{sn_i} and $x_{n_i m}$ equal to zero for all m. Flow conservation is now satisfied at all nodes $n_1, \ldots, n_N$, and capacity constraints (1e) and (1f) are now satisfied for all arcs except those entering the sink t. Setting the remaining binary variables to zero results in constraints (1b), (1c), and (1i) being satisfied. Finally, consider the arcs entering the sink t. If the MBN is not placed at location l with radius r_l^n, the arc connecting node m_l^n to t has capacity zero. Therefore, set $x_{m_l^n t}$ to zero, and note that because node m_l^n has no incoming flow, constraint (1d) is satisfied at node m_l^n. On the other hand, if the MBN is placed at location l and has RN n as its most distant assigned RN, then the arc connecting node m_l^n to t has capacity c_l^n. Node m_l^n has $|A_k|$ units of incoming flow, and by definition of c_l^n and our assumption that our original solution to CPA is feasible, we know that $|A_k| \leq c_l^n$. Therefore, we can set $x_{m_l^n t}$ equal to $|A_k|$, thus satisfying constraints (1d) and (1g) for all nodes and arcs. The objective value of this solution is $\sum_k |A_k|$.

We have shown that for every feasible solution to CPA, there is a corresponding feasible solution to MCND with the same objective value. It remains to be shown that for every optimal solution to MCND, there is a corresponding solution to CPA with the same objective value.

Consider an optimal solution to an instance of MCND, and assume that all of the flows in this solution are integer. (Due to total unimodularity of the network flow constraint matrix, all flows $\mathbf{x}$ are integer in all basic feasible solutions of the linear maximum flow problem induced by a specification of the $\mathbf{y}$ vector.) For each $l = 1, \ldots, L$, place in the MBN at location l if and only if $y_l^n = 1$ for some n. K MBNs have now been placed.

For all l and n, if $y_l^n = 1$, assign to the MBN at location l the set of RNs for which $x_{n_i m_l^n} = 1$ in the solution to MCND. Note that, by definition of c_l^n, all of the assigned RNs achieve throughput at least $\tau_{\min}$. Furthermore, because all arcs originating at s have unit capacity, and because all flows in the solution are integer, each RN can be assigned to at most one MBN. Thus, we have obtained a feasible solution to CPA. Furthermore, the value of this solution is equal to that obtained in CPA: each unit of flow represents exactly one RN that is successfully assigned at throughput at least $\tau_{\min}$. Thus, MCND yields an optimal solution to CPA.

Appendix 2. Proof of Theorem 2

The proof of Theorem 2 reduces an instance of the *Euclidean K-center problem on points* to CPA. In the *Euclidean K-center problem*, the input is a set of N points on the plane and a positive real number r, and the objective is to determine whether it is possible to place K discs of radius r in the plane such that every input point is within distance at most r from the center of at least one disc, i.e., every point is covered by at least one disc. The *Euclidean K-center problem on points* has the additional restriction that the center of each disc must coincide with one of the N input points. Both versions of the problem are known to be NP-complete [25].

Proof. Fix an instance of the Euclidean K-center problem on points. Denote the input points by $\mathrm{N} = \{1, \ldots, N\}$ and the radius by r. This instance can be reduced to an instance of CPA as follows: Define N RNs, and let their locations coincide with the input points. Next, define N candidate MBN locations also coinciding with the input points, and let K be the number of MBNs to be placed. Fix $\tau_{\min}$, and define the throughput function τ as follows:

$$\tau(A_k, d_{nk}) = \tau(d_{nk}) = \begin{cases} \tau_{\min} & \text{if } d_{nk} \le r, \\ 0 & \text{if } d_{nk} > r. \end{cases} \tag{3}$$

Note that τ fits the assumptions stated in Sect. 2; it is monotonically nonincreasing with d_{nk} and does not vary with A_k.

Denote an optimal solution to CPA by (A^*, B^*), where B^* denotes the placement of the MBNs (i.e., the subset of the candidate locations $1, \dots, N$ that are occupied by MBNs) and A^* denotes the optimal assignment of RNs to MBNs. Assume without loss of generality that the nodes are numbered such that $B^* = \{1, \dots, K\}$. Let A_k denote the set of RNs assigned to MBN k in solution (A^*, B^*).

If the optimal objective value of this instance of CPA is equal to N, then the answer to the original Euclidean K-center problem on points is YES. Given a solution to CPA (A^*, B^*) in which $\sum_k |A_k| = N$, a solution to the Euclidean K-center problem on points in which all points are covered can be constructed by placing discs at locations B^*. By our assumption that all RNs in the set A_k achieve throughput at least $\tau_{\min}$, it follows that all RNs in the set A_k are within radius r of the disc at location k and thus are covered by that disc. Furthermore, since each RN can be assigned to at most one MBN, the fact that $\sum_k |A_k| = N$ implies that *all* RNs achieve throughput at least $\tau_{\min}$. Therefore, all nodes in the original Euclidean K-center problem on points are covered by discs placed at locations B^*.

Likewise, if the answer to the original Euclidean K-center problem on points is YES, then the optimal objective value the corresponding instance of CPA must be equal to N. Let B^* denote a placement of discs such that each input point is covered by at least one disc, and again denote this placement by $B^* = \{1, \dots, K\}$. Let $C_n \in B^*$ denote the set of discs that cover point n. If point n is covered by the disc at location $k \in C_n$, then the RN at location n can be assigned to the MBN at location k and achieve throughput at least $\tau_{\min}$ in CPA. Since throughput is not a function of cluster size in (3), a feasible solution to CPA consists of a placement of MBNs at the locations in B^* and an assignment A in which each RN n is assigned to exactly one of the MBNs occupying locations in C_n.

Thus, the Euclidean K-center problem on points can be reduced to CPA. The time required to perform this reduction is polynomial in the number of input points; therefore, CPA is NP-hard. □

Appendix 3. Proof of Theorem 4

Proof. Consider the problem of maximizing a nonnegative, nondecreasing submodular set function $f(S)$ subject to multiple-choice and cardinality constraints. Items eligible for inclusion in S belong to a ground set D that is divided into C disjoint subsets called *classes*. A set S is feasible if $|S| \leq K$ and no two items in S belong to the same class. Note that because f is nondecreasing, there always exists an optimal solution such that $|S| = K$.

Let S^g denote the set of items selected by the greedy algorithm, and let S^* denote the set of items selected by an exact algorithm (i.e., the optimal solution). We wish to find a lower bound on the ratio of $f(S^g)$ to $f(S^*)$.

Consider first the special case in which $C = K$, i.e., the number of elements to be selected is equal to the number of item classes. In this case, exactly one item from each class is to be selected.

Assume without loss of generality that the item classes are numbered such that the item from class k was chosen by the greedy algorithm during its kth iteration, for $k = 1, \dots, K$. Denote the kth item selected by the greedy algorithm by i_k, and denote the item from class k selected by the exact algorithm by i_k^*. Furthermore, denote the set of items selected by the greedy algorithm up to iteration k by S_k^g, i.e., $S_k^g = \{i_1, \dots, i_k\}$. Finally, denote the marginal increase in the objective value obtained by adding item i_k to the set S_{k-1}^g by δ_k, i.e., $\delta_k = f(S_k^g) - f\left(S_{k-1}^g\right)$. Note that

$$f(S^g) = \sum_{k=1}^{K} \delta_k$$

and

$$\delta_k \leq \delta_{k-1}.$$

A set function g defined over a ground set U is submodular if and only if [21]

$$\begin{aligned} g(T) \leq g(S) + &\sum_{j \in T \setminus S} (g(S \cup \{j\}) - g(S)) \\ &- \sum_{j \in S \setminus T} (g(S \cup T) - g(S \cup T \setminus \{j\})) \forall S,\ T \subseteq U. \end{aligned}$$

Because f is a nonnegative, nondecreasing submodular function, the final term $\sum_{j \in S \setminus T} (f(S \cup T) - f(S \cup T \setminus \{j\}))$ is nonnegative, and therefore

$$f(T) \leq f(S) + \sum_{j \in T \setminus S} (f(S \cup \{j\}) - f(S)) \ \ \forall S, T \subseteq U. \tag{4}$$

In particular,

$$f(S^*) \leq f(S^g) + \sum_{j \in S^* \setminus S^g} (f(S^g \cup \{j\}) - f(S^g)). \tag{5}$$

Consider item $i_k^* \notin S^g$. Because i_k^* was not chosen by the greedy algorithm, it follows that

$$f(S_{k-1}^g \cup \{i_k^*\}) - f(S_{k-1}^g) \leq \delta_k.$$

By (2),

$$f(S^g \cup \{i_k^*\}) - f(S^g) \le f(S_{k-1}^g \cup \{i_k^*\}) - f(S_{k-1}^g), \tag{6}$$
$$\le \delta_k. \tag{7}$$

Substituting this into (5), we obtain:

$$\begin{aligned} f(S^*) &\le f(S^g) + \sum_{j \in S^* \setminus S^g} f(S^g \cup \{j\}) - f(S^g) \\ &\le f(S^g) + \sum_{k=1}^{K} \delta_k \\ &= 2f(S^g). \end{aligned}$$

Thus, we have obtained a bound on the ratio of $f(S^g)$ to $f(S^*)$ for the special case in which $C = K$:

$$\frac{f(S^g)}{f(S^*)} \ge \frac{1}{2}. \tag{8}$$

Now consider the case in which $C > K$. In this case, neither the greedy algorithm nor the exact algorithm can select an item from every class, and the two algorithms will not necessarily select items from the same classes. Let S_s^* denote the set of items in the optimal solution that belong to classes from which the greedy algorithm also selected an item, and let S_d^* denote the set of items in the optimal solution that belong to classes from which the greedy algorithm did not select an item. Note that $S^* = S_s^* \cup S_d^*$. Denote the item classes by $c = 1, \ldots, C$.

Consider item $i \in S_s^*$ belonging to class c, and denote the item selected from class c by the greedy algorithm as i_c. Assume that the greedy algorithm selected item i_c in iteration k. Then, by the same argument used in the case of $C = K$,

$$f(S^g \cup \{i\}) - f(S^g) \le f(S_{k-1}^g \cup \{i\}) - f(S_{k-1}^g), \tag{9}$$
$$\le \delta_k, \tag{10}$$

where S_{k-1}^g is again the set of items selected by the greedy algorithm at the beginning of iteration k, and δ_k is the marginal increase in the objective value obtained by adding item i_c to the set S_{k-1}^g, i.e., $\delta_k = f(S_{k-1}^g \cup \{i_c\}) - f(S_{k-1}^g)$.

Now consider item $i \in S_d^*$ belonging to class c'. The greedy algorithm did not select an item from class c'; therefore:

$$f(S^g \cup \{i\}) - f(S^g) \le f(S_{K-1}^g \cup \{i\}) - f(S_{K-1}^g),$$
$$\le \delta_K,$$

where δ_K is the marginal increase in the objective value obtained when the greedy algorithm adds the final element i_K to the set S^g_{K-1} to obtain $S^g = S^g_{K-1} \cup \{i_K\}$.

Substituting these inequalities into (5), we obtain:

$$\begin{aligned} f(S^*) \leq f(S^g) &+ \sum_{j \in S^*_s \setminus S^g} f(S^g \cup \{j\}) - f(S^g) \\ &+ \sum_{j \in S^*_d \setminus S^g} f(S^g \cup \{j\}) - f(S^g) \\ &\leq f(S^g) + \sum_{k: i_k \in U_c, U_c \cap S^* \neq \emptyset} \delta_k + |S^*_d| \delta_K \\ &\leq f(S^g) + \sum_{k=1}^{K} \delta_k \\ &= 2f(S^g), \end{aligned}$$

where we have used the fact that $\delta_K \leq \delta_k$ for $k \leq K$.

Thus, the approximation ratio for the case of $C > K$ is the same as in the case of $C = K$, i.e., $\frac{1}{2}$.

We note that an alternative proof of this performance guarantee can be obtained by demonstrating the matroid structure of the feasible set [22]. □

References

1. I. Rubin, A. Behzadm R. Zhang, H. Luo, and E. Caballero, TBONE: a Mobile-Backbone Protocol for Ad Hoc Wireless Networks, *Proc. IEEE Aerospace Conference*, 6, 2002.
2. K. Xu, X. Hong, and M. Gerla, Landmark Routing in Ad Hoc Networks with Mobile Backbones, *Journal of Parallel and Distributed Computing*, 63, 2, 2003, pp. 110–122.
3. A. Srinivas, G. Zussmanm and E. Modiano, Construction and Maintenance of Wireless Mobile Backbone Networks, *IEEE/ACM Trans. on Networking*, Vol. 17, No. 1, pp. 239–252, Feb. 2009.
4. A. Srinivas and E. Modiano, Joint node placement and assignment for throughput optimization in mobile backbone networks, *Proc. IEEE INFOCOM 08*, Apr. 2008.
5. E. M. Craparo, J. P. How and E. Modiano, Throughput Optimization in Mobile Backbone Networks, *IEEE Transactions on Mobile Computing*, Vol. 10, No. 4, pp. 560–572, Apr. 2011.
6. B. Steckler, B. Bradford, and S. Urrea, Hastily Formed Networks For Complex Humanitarian Disasters After Action Report and Lessons Learned from the Naval Postgraduate School's Response to Hurricane Katrina, http://faculty.nps.edu/dl/HFN/documents/NPS_Katrina_AAR-LL_04-MAY-06.pdf, Sept. 2005 (accessed Sept. 2009).
7. E. M. Craparo, Constrained Node Placement and Assignment in Mobile Backbone Networks, *Proc. IEEE Conference on Decision and Control*, Dec. 2010.
8. D. Jea, A. A. Somasundara, and M. B. Srivastava, Multiple controlled mobile elements (data mules) for data collection in sensor networks, In *Proc. IEEE/ACM DCOSS 05*, Jun. 2008.
9. R. Shah, S. Roy, S. Jain, and W. Brunette, Data MULEs: Modeling a three-tier architecture for sparse sensor networks, In *Proc. IEEE SNPA03*, May 2003.

10. M. M. Bin Tariq, M. Ammar, and E. Zegura, Message ferry route design for sparse ad hoc networks with mobile nodes, In *Proc. ACM MobiHoc 06*, May 2006.
11. W. Zhao, M. Ammar, and E. Zegura, A message ferrying approach for data delivery in sparse mobile ad hoc networks, In *Proc. ACM MobiHoc 04*, May 2004.
12. S. V. Hanly, An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity, *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 7, pp. 1332–1340, Sept. 1995.
13. G. J. Foschini and Z. Milzanic, A simple distributed autonomous power control algorithm and its convergence, *IEEE Transactions on Vehicular Technology*, Vol. 40, pp. 641–646, 1993.
14. R. Yates and C. Y. Huang, Integrated power control and base station assignment, *IEEE Transactions on Vehicular Technology*, Vol. 44, No. 3, pp.638–644, 1995.
15. R. Mathar and T. Niessen, Integrated power control and base station assignment, *Wireless Networks*, Vol. 6, No. 6, pp. 421–428, Dec. 2000.
16. C. Glaßer, S. Reith and H. Vollmer, The complexity of base station positioning in cellular networks, *Discrete Applied Mathematics*, Vol. 148, No. 1, pp. 1–12, 2005.
17. E. Amaldi, A. Capone, F. Malucelli, Radio planning and coverage optimization of 3G cellular networks, *Wireless Networks*, Vol. 14, pp. 435–447, 2008.
18. A. Ahuja, T. Magnanti and J. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, New Jersey, 1993.
19. D. Bertsimas and R. Weismantel, *Optimization over Integers*, Dynamic Ideas, 2005, p. 88.
20. G. Nemhauser and L. Wolsey, Maximizing submodular set functions: Formulations and Analysis of Algorithms, *Studies on Graphs and Discrete Programming*, P. Hansen, ed., 1981, pp. 279–301.
21. G. Nemhauser, L. Wolsey, and M. Fisher, An analysis of the approximations for maximizing submodular set functions - I, *Mathematical Programming*, vol. 14, 1978, pp. 265–294.
22. M. Fisher, G. Nemhauser, and L. Wolsey, An analysis of the approximations for maximizing submodular set functions - II, *Mathematical Programming Studies*, vol. 8, 1978, pp. 73–83.
23. M. Sviridenko, A note on maximizing a submodular set function subject to a knapsack constraint, *Operations Research Letters*, Vol. 32, No. 1, Jan. 2004, pp. 41–43.
24. A. Kulik, H. Shachnai and T. Tamir, Maximizing submodular set functions subject to multiple linear constraints, *Proc. of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms*, 2009.
25. S. Masuyama, T. Ibaraki and T. Hasegawa, The computational complexity of the m-center problems on the plane, *Trans. IECE of Japan*, E-64, pp. 57–64, 1981.

Canonical Dual Solutions to Sum of Fourth-Order Polynomials Minimization Problems with Applications to Sensor Network Localization

David Yang Gao, Ning Ruan, and Panos M. Pardalos

Abstract This chapter presents a canonical dual approach for solving a general sum of fourth-order polynomial minimization problem. This problem arises extensively in engineering and science, including database analysis, computational biology, sensor network communications, nonconvex mechanics, and ecology. We first show that this global optimization problem is actually equivalent to a discretized minimal potential variational problem in large deformation mechanics. Therefore, a general analytical solution is proposed by using the canonical duality theory developed by the first author. Both global and local extremality properties of this analytical solution are identified by a triality theory. Application to sensor network localization problem is illustrated. Our results show when the problem is not uniquely localizable, the "optimal solution" obtained by the SDP method is actually a local maximizer of the total *potential energy*. However, by using a perturbed canonical dual approach, a class of Euclidean distance problems can be converted to a unified concave maximization dual problem with zero duality gap, which can be solved by well-developed convex minimization methods. This chapter should bridge an existing gap between nonconvex mechanics and global optimization.

Keywords Sensor network localization • Nonconvex programming • Global optimization • Nonlinear algebraic equations • Canonical duality theory • Triality

D.Y. Gao (✉)
School of Science, Information Technology, and Engineering, University of Ballarat,
Mt. Helen, VIC 3350, Australia
e-mail: d.gao@ballarat.edu.au

N. Ruan
Curtin University of Technology, Perth, WA 6845, Australia
e-mail: mimiopt@gmail.com

P.M. Pardalos
Department of Industrial and System Engineering, University of Florida, Gainesville,
FL 32611, USA
e-mail: pardalos@ufl.edu

V.L. Boginski et al. (eds.), *Sensors: Theory, Algorithms, and Applications*,
Springer Optimization and Its Applications 61, DOI 10.1007/978-0-387-88619-0_3,

1 Primal Problems and Connections to Finite Deformation Mechanics

We are interested in solving the following general nonlinear programming problem:

$$(\mathcal{P}): \quad \min \left\{ P(\mathbf{x}) = \sum_{e=1}^{m} W_e(\mathbf{x}) + \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{x}^T\mathbf{f} \ \ : \ \ \mathbf{x} \in \mathbb{R}^n \right\}, \tag{1}$$

where

$$W_e(\mathbf{x}) = \frac{1}{2}\alpha_e \left(\frac{1}{2}\mathbf{x}^T A_e \mathbf{x} + \mathbf{b}_e^T \mathbf{x} + c_e \right)^2, \tag{2}$$

and $A_e = A_e^T$, $Q = Q^T \in \mathbb{R}^{n\times n}$ are indefinite symmetrical matrices, $\mathbf{f}$, $\mathbf{b}_e \in \mathbb{R}^n$ are given vectors, $c_e \in \mathbb{R}$ and α_e are given constants. Without loss generality, we assume that $\alpha_e > 0, \quad \forall e = 1, \ldots, m$. The criticality condition $\delta P(x) = 0$ leads to a nonlinear equilibrium equation:

$$\sum_{e=1}^{m} \alpha_e \left(\frac{1}{2}\mathbf{x}^T A_e \mathbf{x} + \mathbf{b}_e^T \mathbf{x} + c_e \right) (A_e\mathbf{x} + \mathbf{b}_e) + Q\mathbf{x} - \mathbf{f} = 0. \tag{3}$$

Direct methods for solving this coupled nonlinear algebraic system are very difficult. Also (3) is only a necessary condition for global minimizer of the problem $(\mathcal{P})$. A general sufficient condition for identifying the global minimizer is a fundamental task in global optimization.

The nonconvex minimization problem $(\mathcal{P})$ arises naturally in a wide range of applications, including chaotical dynamical systems [18], chemical database analysis [47], information theory, large deformation computational mechanics [10], location/allocation, network communication, and phase transitions of solids [34].

For example, the sensor network location problem is to solve the following system of nonlinear equations [3, 6, 38]:

$$\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 = d_{ij}^2, \ \ \forall (i,j) \in \mathcal{I}_p, \ \ \mathbf{u}_k = \mathbf{a}_k, \quad \forall k \in \mathcal{I}_b, \tag{4}$$

where the vectors $\mathbf{u}_i = \{u_i^\alpha\} \in \mathbb{R}^d$ $(i = 1, \ldots, p)$ represent the locations of the unknown sensors, $\mathcal{I}_p = \{(i,j) : \ i < j, \ d_{ij}$ is specified$\}$ and $\mathcal{I}_b = \{k : \ \mathbf{u}_k = \mathbf{a}_k$ is specified$\}$ are two given index sets, d_{ij} are given distances for $(i,j) \in \mathcal{I}_p$, the given vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_q \in \mathbb{R}^d$ are the so called anchors. By using the least squares method, the quadratic equations (29) of the sensor localization problem can be reformulated as an optimization problem:

$$\min \left\{ P(\mathbf{u}) = \sum_{(i,j)\in\mathcal{I}_p} \frac{1}{2} \left(\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 - d_{ij}^2 \right)^2 : \ \ \mathbf{u}_i \in \mathcal{U}_a \right\}, \tag{5}$$

where

$$\|\mathbf{u}_i - \mathbf{u}_j\|_2 = \sqrt{\sum_{\alpha=1}^{d}(u_i^\alpha - u_j^\alpha)^2}$$

denotes the Euclidian distance between $\mathbf{u}_i$ and $\mathbf{u}_j$, and $\mathcal{U}_a = \{\mathbf{u} \in \mathbb{R}^{d\times p}|\ \mathbf{u}_k = \mathbf{a}_k \forall k \in \mathcal{I}_b\}$ is a feasible space. Let $\mathbf{x} = \{\{u_1^1, \ldots, u_1^d\}, \ldots, \{u_p^1, \ldots, u_p^d\}\} \in \mathbb{R}^n$ $(n = d \times p)$ denote an extended vector. By using Lagrange multiplier method to relax the boundary conditions in $\mathcal{U}_a$, the least squares method for the sensor localization problem (5) can be written in the problem (1) for certain properly defined matrices $\{A_e\}$.

The sensor network localization problem can also be viewed as a variant of the Graph Realization problem, or a distance geometry problem [1, 44] which has been studied extensively in computational biology, Euclidean ball packing, molecular confirmation, and recently, wireless network communication. In general, the sensor network localization problem is considered to be NP-hard even for the simplest case $d = 1$ [38, 43]. Recent result of Aspnes et al. [3] shows that the problem of computing a realization of the sensors on the plane is NP-complete in general. So and Ye confirmed that this is true even when the instance has a unique solution on a given plane [44]. Therefore, many approximation methods have been proposed for solving this nonconvex, nonsmooth global optimization problem approximately. The semi-definite programming (SDP) and second-order cone programming (SOCP) relaxations are two of the popular methods studied recently [1, 2, 4–6, 44–46]. Similar to the linear primal–dual interior point methods, many numerical schemes for SDP or SOCP often return to the analytic center of the solution set. Therefore, one common feature of SDP and SOCP relaxations is that the computed sensor locations are very inaccurate when the solution of the localization problem is not unique. Very often, these numerical solutions could be even local maximizers of the least squares objective function (see the last section and [42]).

Mathematics and mechanics have been two complementary partners since the Newton times. Many fundamental ideas, concepts, and mathematical methods extensively used in calculus of variations and optimization are originally developed from mechanics. For examples, the Lagrange multiplier method was first proposed by Lagrange from the classical analytic mechanics; while the concepts of super-potential and sub-differential in modern convex analysis were introduced by Moreau from frictional mechanics [39, 40]. From the point view of computational large deformation mechanics, both the fourth-order polynomial minimization problem $(\mathcal{P})$ and the sensor localization problem (5) are actually two special cases of discretized finite deformation problems [10]. In continuum mechanics and differential geometry, the deformation $\mathbf{u}(x) : \Omega \to \mathbb{R}^d$ is a vector field over an open domain $\Omega \subset \mathbb{R}^d$, and the nonconvex function

$$\min\left\{P(\mathbf{u}) = \int_{\Omega}\left[W(\nabla\mathbf{u}) - \mathbf{u}^T\mathbf{f}\right] d\Omega \ :\ \ \mathbf{u} \in \mathcal{U}_a\right\}, \tag{6}$$

where $W(\mathbf{F})$ is the so-called *stored strain energy*, which is usually a nonconvex function of the deformation gradient $\mathbf{F}=\nabla\mathbf{u}$, the feasible set $\mathcal{U}_a$ in this nonconvex variational problem is called the *kinetically admissible space*, in which, certain boundary conditions are prescribed. According to the hyper-elasticity law [14, Chapter 6.1.2], the stored strain energy should be a function of the *right Cauchy–Green* strain tensor $\mathbf{C} = (\nabla\mathbf{u})^T(\nabla\mathbf{u})$, i.e., there exists an isotropic function $V(\mathbf{C})$ such that:

$$W(\nabla\mathbf{u}) = V(\mathbf{C}(\mathbf{u})). \tag{7}$$

Particularly, for the simplest St. Venant–Kirchhoff material, $V(\mathbf{C})$ is a quadratic function of $\mathbf{C}$, and the stored energy $W(\mathbf{F})$ is a fourth-order polynomial tensor function of $\mathbf{F}$. In terms of $\mathbf{u}$, we have [13, 14, 32]:

$$W(\nabla\mathbf{u}) = \frac{1}{2}\left[\left(\frac{1}{2}(\nabla\mathbf{u})^T(\nabla\mathbf{u}) - \boldsymbol{\epsilon}\right) : \mathbf{H} : \left(\frac{1}{2}(\nabla\mathbf{u})^T(\nabla\mathbf{u}) - \boldsymbol{\epsilon}\right)\right], \tag{8}$$

where $\mathbf{H} = \{H^{\alpha\beta\gamma\delta}\}$ is a fourth-order (Hooke elastic) tensor; $\boldsymbol{\epsilon} = \{\epsilon_{\alpha\beta}\}$ is a given *internal variable*, which could be either residual strain tensor, or dislocation [13]. In the case that $\boldsymbol{\epsilon} = \mathbf{I}$, an identity tensor, $E = \frac{1}{2}(\nabla\mathbf{u})^T(\nabla\mathbf{u}) - \mathbf{I}$ is the well-known *Green–St. Venant strain tensor*. The double dot product

$$\mathbf{H} : E = \sum_{\gamma=1}^{d}\sum_{\delta=1}^{d} H^{\alpha\beta\gamma\delta}E_{\gamma\delta},$$

is a standard notation in finite deformation theory. In the most simple one-dimensional problem, $W(u_x) = \frac{1}{2}\left(\frac{1}{2}u_x^2 - \epsilon\right)^2$ is the well-known *double-well* potential, first proposed by van der Waals in thermodynamics in 1895 [15, 18]. By using finite difference method (FDM), the deformation gradient $\nabla\mathbf{u}$ can be directly approximated by the difference $\mathbf{u}(x_i) - \mathbf{u}(x_j) = \mathbf{u}_i - \mathbf{u}_j$. While in finite element method (FEM), the domain $\Omega = \bigcup_e^m \Omega^e$ is discretized by a finite number of elements $\Omega^e \subset \Omega$ and in each element, the deformation filed $\mathbf{u}(x) = \sum_i \mathbf{N}_i(x)\mathbf{u}_i$ is numerically represented by the nodal vectors $\mathbf{u}_i$ via piecewise interpolation (polynomial) function $\mathbf{N}_i(x)$ [10]. Therefore, by either FDM or FEM, the minimal potential variational problem (6) can be eventually reduced to a very complicated large-scale fourth-order polynomial minimization problem with both the problems $(\mathcal{P})$ and (5) as its two special cases.

In order to solve the nonconvex variational problem (6), a unified *canonical duality theory* has been developed from nonlinear analysis of finite deformation theory [12, 14, 15, 31] and finite element analysis of large scale computational mechanics [10, 34]. A general analytical solution of the problem (6) has been obtained in [12,13]. It is now realized that this canonical duality theory is potentially powerful for solving a large class of challenging problems in global optimization [7,8,16,19,21,23,24,32]. The purpose of this chapter is to demonstrate the potential of the canonical duality theory by solving the proposed primal problem $(\mathcal{P})$. The rest of this chapter is arranged as follows. In the next section, we show how to use the canonical dual transformation to convert the nonconvex problem into a canonical dual problem. An analytical solution form is proposed. The extremality conditions of this analytical solution are specified in Sect. 3. Examples are given in Sect. 4 to illustrate the powerful canonical dual approach. Application to the sensor network localization shows that the canonical dual problem is a concave maximization over a convex set, which can be easily solved to obtain a unique solution. The concluding remarks are made in the last section.

2 Canonical Dual Problem

Following the standard procedure of the canonical dual transformation, we introduce a Gâteaux differentiable *geometrical operator*

$$\boldsymbol{\xi} = \Lambda(\mathbf{x}) = \left\{ \frac{1}{2}\mathbf{x}^T A_k \mathbf{x} + \mathbf{b}_k^T \mathbf{x} + c_k \right\} : \mathbb{R}^n \to \mathbb{R}^m, \tag{9}$$

which is a map from $\mathbb{R}^n$ into $\mathcal{V}_a \subset \mathbb{R}^m$. Thus, the nonconvex function $W(\mathbf{x})$ can be written in the canonical form:

$$W(\mathbf{x}) = V(\Lambda(\mathbf{x})), \tag{10}$$

where

$$V(\boldsymbol{\xi}) = \sum_{k=1}^{m} \frac{1}{2}\alpha_k \xi_k^2 = \frac{1}{2}\boldsymbol{\alpha}^T(\boldsymbol{\xi} \circ \boldsymbol{\xi})$$

is a quadratic function, where $\boldsymbol{\alpha} = \{\alpha_k\} \in \mathbb{R}^m$, and $\boldsymbol{\xi} \circ \boldsymbol{\xi} = \{\xi_k \xi_k\} \in \mathbb{R}^m$ represents the Hadamard product. Thus, the duality relation

$$\boldsymbol{\varsigma} = \delta V(\boldsymbol{\xi}) = \boldsymbol{\alpha} \circ \boldsymbol{\xi} \tag{11}$$

is invertible for any given $\boldsymbol{\xi} \in \mathcal{V}_a$.

Let $\mathcal{V}_a^*$ be the range of the duality mapping $\boldsymbol{\varsigma} = \delta V(\boldsymbol{\xi}) : \mathcal{V}_a \to \mathcal{V}_a^* \subset \mathbb{R}^m$. Thus, for any given $\boldsymbol{\varsigma} \in \mathcal{V}_a^*$, the Legendre conjugate V^* can be uniquely defined by:

$$V^*(\boldsymbol{\varsigma}) = \text{sta}\left\{\boldsymbol{\xi}^T\boldsymbol{\varsigma} - V(\boldsymbol{\xi})\right\} = \sum_{k=1}^{m} \frac{1}{2}\alpha_k^{-1}\varsigma_k^2,$$

where sta{} denotes finding stationary points of the statement in {}. So $(\boldsymbol{\xi}, \boldsymbol{\varsigma})$ forms a *canonical duality pair* on $\mathcal{V}_a \times \mathcal{V}_a^*$ [14] and the following canonical duality relations hold on $\mathcal{V}_a \times \mathcal{V}_a^*$:

$$\boldsymbol{\varsigma} = \delta V(\boldsymbol{\xi}) \quad \Leftrightarrow \quad \boldsymbol{\xi} = \delta V^*(\boldsymbol{\varsigma}) \quad \Leftrightarrow \quad \boldsymbol{\xi}^T\boldsymbol{\varsigma} = V(\boldsymbol{\xi}) + V^*(\boldsymbol{\varsigma}). \tag{12}$$

Replacing $W(\mathbf{x}) = V(\varLambda(\mathbf{x}))$ by $\varLambda(\mathbf{x})\boldsymbol{\varsigma} - V^*(\boldsymbol{\varsigma})$, the Gao–Strang *generalized complementary function* [14, 31] can be defined by:

$$
\begin{aligned}
\varXi(\mathbf{x}, \boldsymbol{\varsigma}) &= \varLambda(\mathbf{x})\boldsymbol{\varsigma} - V^*(\boldsymbol{\varsigma}) + \mathbf{x}^T Q\mathbf{x} - \mathbf{x}^T\mathbf{f} \\
&= \sum_{k=1}^{m} \left[\left(\frac{1}{2}\mathbf{x}^T A_k \mathbf{x} + \mathbf{b}_k^T \mathbf{x} + c_k \right) \varsigma_k - \frac{1}{2}\alpha_k^{-1}\varsigma_k^2 \right] \\
&\quad + \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{x}^T\mathbf{f}.
\end{aligned}
\tag{13}
$$

For a fixed $\boldsymbol{\varsigma} \in \mathcal{V}_a^*$, the criticality condition $\delta_x \varXi(\mathbf{x}, \boldsymbol{\varsigma}) = 0$ leads to the following *canonical equilibrium equation*:

$$
G(\boldsymbol{\varsigma})\mathbf{x} - F(\boldsymbol{\varsigma}) = 0,
\tag{14}
$$

where $F(\boldsymbol{\varsigma}) = \mathbf{f} - \sum_{k=1}^m \boldsymbol{\varsigma}_k \mathbf{b}_k$, $G(\boldsymbol{\varsigma}) = Q + \sum_{k=1}^m \boldsymbol{\varsigma}_k A_k$. Clearly, for any given $\boldsymbol{\varsigma} \in \mathcal{V}_a^*$, if the vector $F(\boldsymbol{\varsigma})$ is in the column space of $G(\boldsymbol{\varsigma})$, denoted by $\mathcal{C}_{ol}(G(\boldsymbol{\varsigma}))$, the canonical equilibrium equation has at least one solution $\bar{x} = G^+(\boldsymbol{\varsigma})F(\boldsymbol{\varsigma})$, where G^+ denotes the Moore–Penrose generalized inverse of G. Therefore, the dual feasible space defined by:

$$
\mathcal{S}_a = \{\boldsymbol{\varsigma} \in \mathbb{R}^m \mid F(\boldsymbol{\varsigma}) \in \mathcal{C}_{ol}(G(\boldsymbol{\varsigma}))\},
\tag{15}
$$

the canonical dual function can be formulated as:

$$
\begin{aligned}
P^d(\boldsymbol{\varsigma}) &= \text{sta}\{\varXi(\mathbf{x}, \boldsymbol{\varsigma}) : \mathbf{x} \in \mathcal{X}_a\} \\
&= \sum_{k=1}^{m} \left(c_k\varsigma_k - \frac{1}{2}\alpha_k^{-1}\varsigma_k^2 \right) - \frac{1}{2}F^T(\boldsymbol{\varsigma})G^+(\boldsymbol{\varsigma})F(\boldsymbol{\varsigma}),
\end{aligned}
\tag{16}
$$

which is piecewise smooth on $\mathcal{S}_a$. Thus, the canonical dual problem can be finally proposed as the following:

$$
(\mathcal{P}^d) : \text{sta}\left\{ P^d(\boldsymbol{\varsigma}) = \sum_{k=1}^{m} \left(c_k\varsigma_k - \frac{1}{2}\alpha_k^{-1}\varsigma_k^2 \right) - \frac{1}{2}F^T(\boldsymbol{\varsigma})G^+(\boldsymbol{\varsigma})F(\boldsymbol{\varsigma}) \ : \ \boldsymbol{\varsigma} \in \mathcal{S}_a \right\}.
\tag{17}
$$

Theorem 1 (Complementary-Dual Principle). *The problem $(\mathcal{P}^d)$ is canonically dual to the primal problem $(\mathcal{P})$ in the sense that if $\bar{\boldsymbol{\varsigma}}$ is a critical point of $(\mathcal{P}^d)$, then the vector*

$$
\bar{\mathbf{x}} = G^+(\bar{\boldsymbol{\varsigma}})F(\bar{\boldsymbol{\varsigma}})
\tag{18}
$$

is a critical point of $(\mathcal{P})$ and

$$
P(\bar{\mathbf{x}}) = P^d(\bar{\boldsymbol{\varsigma}}).
\tag{19}
$$

Proof. Suppose that $\bar{\boldsymbol{\varsigma}}$ is a critical point of $(\mathcal{P}^d)$, then we have:

$$\frac{\partial P^d(\bar{\boldsymbol{\varsigma}})}{\partial \boldsymbol{\varsigma}_k} = c_k - \alpha_k^{-1}\boldsymbol{\varsigma}_k + \mathbf{b}_k^T\bar{\mathbf{x}} + \frac{1}{2}\bar{\mathbf{x}}^T A_k \bar{\mathbf{x}} = 0, \;\; k = 1, \ldots, m, \tag{20}$$

where $\bar{\mathbf{x}} = G^+(\bar{\boldsymbol{\sigma}})F(\bar{\boldsymbol{\varsigma}})$. The criticality condition (20) is actually the canonical duality relation (constitutive equation) (11), i.e., $\boldsymbol{\varsigma}_k = \alpha_k\left(\frac{1}{2}\bar{\mathbf{x}}^T A_k\bar{\mathbf{x}} + \mathbf{b}_k^T\bar{\mathbf{x}} + c_k\right)$. Thus, we have:

$$\begin{aligned}
\bar{\mathbf{x}} &= G^+(\bar{\boldsymbol{\sigma}})F(\bar{\boldsymbol{\varsigma}})\\
&= \left[Q + \sum_{k=1}^m \alpha_k\left(\frac{1}{2}\bar{\mathbf{x}}^T A_k\bar{\mathbf{x}} + \mathbf{b}_k^T + c_k\right)A_k\right]^+\\
&\quad \times\left[\mathbf{f} - \sum_{k=1}^m \alpha_k\left(\frac{1}{2}\bar{\mathbf{x}}^T A_k\bar{\mathbf{x}} + \mathbf{b}_k^T\bar{\mathbf{x}} + c_k\right)\mathbf{b}_k\right].
\end{aligned}$$

This shows that $\bar{\mathbf{x}}$ is a critical point of the primal problem $(\mathcal{P})$. Moreover, in term of $\bar{\mathbf{x}} = G^+(\bar{\boldsymbol{\varsigma}})F(\bar{\boldsymbol{\varsigma}})$, we have:

$$\begin{aligned}
P^d(\bar{\boldsymbol{\varsigma}}) &= \sum_{k=1}^m\left(c_k\bar{\varsigma}_k - \frac{1}{2}\alpha_k^+\bar{\varsigma}_k^2\right) - \frac{1}{2}F^T(\bar{\boldsymbol{\varsigma}})G^+(\bar{\boldsymbol{\varsigma}})F(\bar{\boldsymbol{\varsigma}})\\
&= \sum_{k=1}^m\left(c_k\bar{\varsigma}_k - \frac{1}{2}\alpha_k^+\bar{\varsigma}_k^2\right) - \frac{1}{2}\left(\mathbf{f} - \sum_{k=1}^m\bar{\varsigma}_k b_k\right)^T\left(Q + \sum_{k=1}^m\bar{\varsigma}_k A_k\right)^+\\
&\quad \times\left(\mathbf{f} - \sum_{k=1}^m\bar{\varsigma}_k b_k\right)\\
&= \sum_{k=1}^m\left(c_k\bar{\varsigma}_k - \frac{1}{2}\alpha_k^+\bar{\varsigma}_k^2\right) + \frac{1}{2}\bar{\mathbf{x}}^T\left(Q + \sum_{k=1}^m\bar{\varsigma}_k A_k\right)\bar{\mathbf{x}} - \bar{\mathbf{x}}^T\left(\mathbf{f} - \sum_{k=1}^m\bar{\varsigma}_k b_k\right)\\
&= \sum_{k=1}^m\left[\left(\frac{1}{2}\bar{\mathbf{x}}^T A_k\bar{\mathbf{x}} + \mathbf{b}_k^T\bar{\mathbf{x}} + c_k\right)\bar{\varsigma}_k - \frac{1}{2}\alpha_k^+\bar{\varsigma}_k^2\right] + \frac{1}{2}\bar{\mathbf{x}}^T Q\bar{\mathbf{x}} - \bar{\mathbf{x}}^T\mathbf{f}\\
&= \sum_{k=1}^m\left[\left(\frac{1}{2}\bar{\mathbf{x}}^T A_k\bar{\mathbf{x}} + \mathbf{b}_k^T\bar{\mathbf{x}} + c_k\right)^2\alpha_k - \frac{1}{2}\alpha_k\left(\frac{1}{2}\bar{\mathbf{x}}^T A_k\bar{\mathbf{x}} + \mathbf{b}_k^T\bar{\mathbf{x}} + c_k\right)^2\right]\\
&\quad + \frac{1}{2}\bar{\mathbf{x}}^T Q\bar{\mathbf{x}} - \bar{\mathbf{x}}^T\mathbf{f}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{m} \frac{1}{2}\alpha_k \left(\frac{1}{2}\bar{\mathbf{x}}^T A_k \bar{\mathbf{x}} + \mathbf{b}_k^T \bar{\mathbf{x}} + c_k\right)^2 + \frac{1}{2}\bar{\mathbf{x}}^T Q \bar{\mathbf{x}} - \bar{\mathbf{x}}^T \mathbf{f} \\
&= P(\bar{\mathbf{x}}).
\end{aligned}
$$

This proves the theorem. □

Theorem 1 presents an analytic solution (18) for the critical point of the primal problem $(\mathcal{P})$. This solution is actually a special case of the general analytical solution form proposed in nonconvex variational problems [12, 13, 15]. In finite deformation elasticity, this theorem solves an open problem left by Hellinger (1914) and Reissner (1954) and is recognized as the Gao principle [37]. Applications of this complementary-dual principle have been given to a series of nonconvex minimization and integer/fractional programming problems in global optimization [7, 16, 19, 23, 24]. It is known that the criticality condition is only necessary for local minimizers. In the next section, we will study sufficient conditions for both global and local extrema.

3 Global and Local Optimality Criteria

In order to identify global and local extremality properties of the analytical solution (18), we need to introduce some useful feasible spaces:

$$\mathcal{S}_a^+ = \{\boldsymbol{\varsigma} \in \mathcal{S}_a \mid G(\boldsymbol{\varsigma}) \succeq 0\}, \tag{21}$$

$$\mathcal{S}_a^- = \{\boldsymbol{\varsigma} \in \mathcal{S}_a \mid G(\boldsymbol{\varsigma}) \prec 0\}. \tag{22}$$

By the canonical duality theory developed in [14], we have the following results.

Theorem 2 (Triality Theorem). *Suppose that the vector $\bar{\boldsymbol{\varsigma}}$ is a critical point of the canonical dual function $P^d(\bar{\boldsymbol{\varsigma}})$. Let $\bar{\mathbf{x}} = G^+(\bar{\boldsymbol{\varsigma}})F(\bar{\boldsymbol{\varsigma}})$.*

If $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_a^+$, then $\bar{\boldsymbol{\varsigma}}$ is a global maximizer of P^d on $\mathcal{S}_a^+$ if and only if the vector $\bar{\mathbf{x}}$ is a global minimizer of P on $\mathbb{R}^n$, i.e.,

$$P(\bar{\mathbf{x}}) = \min_{\mathbf{x}\in\mathbb{R}^n} P(\mathbf{x}) \;\Leftrightarrow\; \max_{\boldsymbol{\varsigma}\in\mathcal{S}_a^+} P^d(\boldsymbol{\varsigma}) = P^d(\bar{\boldsymbol{\varsigma}}). \tag{23}$$

If $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_a^-$, then on the neighborhood $\mathcal{X}_0 \times \mathcal{S}_0 \subset \mathbb{R}^n \times \mathcal{S}_a^-$ of $(\bar{\mathbf{x}}, \bar{\boldsymbol{\varsigma}})$, the vector $\bar{\mathbf{x}} \in \mathcal{X}_0$ is a local maximizer of $P(\mathbf{x})$ if and only if $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_0$ is a local maximizer of $P^d(\boldsymbol{\varsigma})$, i.e.,

$$P(\bar{\mathbf{x}}) = \max_{\mathbf{x}\in\mathcal{X}_0} P(\mathbf{x}) \;\Leftrightarrow\; \max_{\boldsymbol{\varsigma}\in\mathcal{S}_0} P^d(\boldsymbol{\varsigma}) = P^d(\bar{\boldsymbol{\varsigma}}). \tag{24}$$

If $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_a^-$ and $n = m$, then on the neighborhood $\mathcal{X}_0 \times \mathcal{S}_0 \subset \mathbb{R}^n \times \mathcal{S}_a^-$ of $(\bar{\mathbf{x}}, \bar{\boldsymbol{\varsigma}})$, the vector $\bar{\mathbf{x}} \in \mathcal{X}_0$ is a local minimizer of $P(\mathbf{x})$ if and only if $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_0$ is a local minimizer of $P^d(\boldsymbol{\varsigma})$, i.e.,

$$P(\bar{\mathbf{x}}) = \min_{\mathbf{x}\in\mathcal{X}_0} P(\mathbf{x}) \;\Leftrightarrow\; \min_{\varsigma\in\mathcal{S}_0} P^d(\varsigma) = P^d(\bar{\varsigma}). \tag{25}$$

If $n \neq m$, the double-min duality (25) holds conditionally.

Proof. By Theorem 1 and the canonical duality theory [14], we know that vector $\bar{\varsigma} \in \mathcal{S}_a$ is a critical point of the problem $(\mathcal{P}^d)$ if and only if $\bar{\mathbf{x}} = G^+(\bar{\varsigma})F(\bar{\varsigma})$ is a critical point of the problem $(\mathcal{P})$, and

$$P(\bar{\mathbf{x}}) = \Xi(\bar{\mathbf{x}}, \bar{\varsigma}) = P^d(\bar{\varsigma}).$$

By the fact that the canonical dual function $P^d(\varsigma)$ is concave on $\mathcal{S}_a^+$, the critical point $\bar{\varsigma} \in \mathcal{S}_a^+$ is a global maximizer of $P^d(\varsigma)$ over $\mathcal{S}_a^+$. Since $(\bar{\mathbf{x}}, \bar{\varsigma})$ is a saddle point of the total complementary function $\Xi(\mathbf{x}, \varsigma)$ on $\mathbb{R}^n \times \mathcal{S}_a^+$, i.e., Ξ is convex in $\mathbf{x} \in \mathbb{R}^n$ and concave in $\varsigma \in \mathcal{S}_a^+$, by the canonical min–max duality theory [14], we have:

$$\begin{aligned}
P^d(\bar{\varsigma}) &= \max_{\varsigma\in\mathcal{S}_a^+} P^d(\varsigma) = \max_{\varsigma\in\mathcal{S}_a^+}\min_{\mathbf{x}\in\mathbb{R}^n} \Xi(\mathbf{x},\varsigma) = \min_{\mathbf{x}\in\mathbb{R}^n}\max_{\varsigma\in\mathcal{S}_a^+} \Xi(\mathbf{x},\varsigma) \\
&= \min_{\mathbf{x}\in\mathbb{R}^n}\left\{\frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{f}^T\mathbf{x} + \sum_{k=1}^m \max_{\varsigma_k\in\mathcal{S}_a^+}\left\{\left(\frac{1}{2}\mathbf{x}^T A_k\mathbf{x} + \mathbf{b}_k^T\mathbf{x} + c_k\right)\varsigma_k \right.\right. \\
&\qquad \left.\left. - \frac{1}{2}\alpha_k^+\varsigma_k^2\right\}\right\} \\
&= \min_{\mathbf{x}\in\mathbb{R}^n}\left\{\frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{f}^T\mathbf{x} + \sum_{k=1}^m \frac{1}{2}\alpha_k\left(\frac{1}{2}\mathbf{x}^T A_k\mathbf{x} + \mathbf{b}_k^T\mathbf{x} + c_k\right)^2\right\} \\
&= \min_{\mathbf{x}\in\mathbb{R}^n} P(\mathbf{x}) = P(\bar{\mathbf{x}}).
\end{aligned}$$

This proves the statement (23).

If $\bar{\varsigma} \in \mathcal{S}_a^-$, the matrix $G(\bar{\varsigma})$ is a negative definite. In this case, the Gao–Strang complementary function $\Xi(\bar{\mathbf{x}}, \bar{\varsigma})$ is a so-called super-Lagrangian [14], i.e., it is locally concave in both $\mathbf{x} \in \mathcal{X}_0 \subset \mathcal{X}_a$ and $\varsigma \in \mathcal{S}_0 \subset \mathcal{S}_a^-$. By the fact that

$$\max_{\mathbf{x}\in\mathcal{X}_0}\max_{\varsigma\in\mathcal{S}_0} \Xi(\mathbf{x},\varsigma) = \max_{\varsigma\in\mathcal{S}_0}\max_{\mathbf{x}\in\mathbb{R}^n} \Xi(\mathbf{x},\varsigma) \tag{26}$$

holds on the neighborhood $\mathcal{X}_0 \times \mathcal{S}_0$ of $(\bar{\mathbf{x}}, \bar{\varsigma})$, we have the double-max duality statement (24). If $n = m$, we have [33]:

$$\min_{\mathbf{x}\in\mathcal{X}_0}\max_{\varsigma\in\mathcal{S}_0} \Xi(\mathbf{x},\varsigma) = \min_{\varsigma\in\mathcal{S}_0}\max_{\mathbf{x}\in\mathbb{R}^n} \Xi(\mathbf{x},\varsigma) \tag{27}$$

which leads to the double-min duality statement (25). This proves the theorem. □

Theorem 2 shows that the extremality condition of the analytical solution (18) is controlled by the critical point of the canonical dual problem, i.e., if $\bar{\varsigma} \in \mathcal{S}_a^+$, the solution $\bar{\mathbf{x}}(\bar{\varsigma})$ is a global minimizer of $(\mathcal{P})$; if $\bar{\varsigma} \in \mathcal{S}_a^-$, then $\bar{\mathbf{x}}(\bar{\varsigma})$ is a local maximizer (or minimizer when $n = m$) of $(\mathcal{P})$ if and only if the critical point $\bar{\varsigma}$ is a local maximizer (or minimizer when $n = m$) of P^d on $\mathcal{S}_a^-$. When $n \neq m$, the double-min duality holds conditionally, which was an open problem left in [18, 19]. This open problem is solved recently in [33]. Therefore, based on this triality theory, if $\mathcal{S}_a^+ \neq \emptyset$, the primal problem is canonical dual to

$$(\mathcal{P}^d_{\max}): \quad \max\{P^d(\varsigma): \quad \varsigma \in \mathcal{S}_a^+\}, \tag{28}$$

which is a concave maximization problem over a convex set and can be solved easily via well-developed convex minimization algorithms. Existence and uniqueness of the canonical dual solutions are discussed in [28, 29].

4 Applications

We now present examples to illustrate the applications of the theory proposed in this chapter.

Example 4.1. Unconstrained two-dimensional polynomial minimization.

$$\min\left\{P(x_1,x_2) = \sum_{k=1}^{2}\frac{1}{2}\alpha_k\left(\frac{1}{2}(a_{k1}x_1^2 + a_{k2}x_2^2) + c_k\right)^2 + \frac{1}{2}(q_1x_1^2 + q_2x_2^2)\right.$$
$$\left. - \sum_{i=1}^{2} f_i x_i \; : x \in \mathbb{R}^2\right\}.$$

This fourth-order polynomial minimization problem is actually a discretized form of a nonconvex variational problem in phase transitions studied recently in [26, 27]. On the dual feasible set

$$\mathcal{S}_a = \{\varsigma \in \mathbb{R}^2 \mid (q_1 + \varsigma_1 a_{11} + \varsigma_2 a_{21})(q_2 + \varsigma_1 a_{12} + \varsigma_2 a_{22}) \neq 0\},$$

the canonical dual function has the form of

$$P^d(\varsigma) = \sum_{k=1}^{2}\left(c_k\varsigma_k - \frac{1}{2\alpha_k}\varsigma_k^2\right) - \frac{1}{2}[f_1, f_2]$$
$$\times \begin{bmatrix} (q_1 + \varsigma_1 a_{11} + \varsigma_2 a_{21})^{-1} & 0 \\ 0 & (q_2 + \varsigma_1 a_{12} + \varsigma_2 a_{22})^{-1} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}.$$

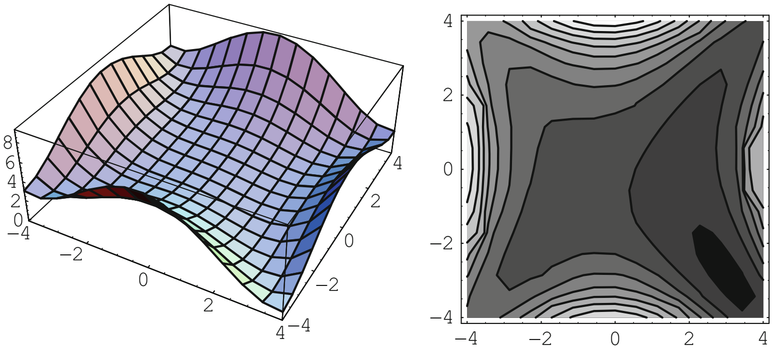


Fig. 1 Graph of $P(\mathbf{x})$ (*left*) and contour of $P(\mathbf{x})$ (*right*)

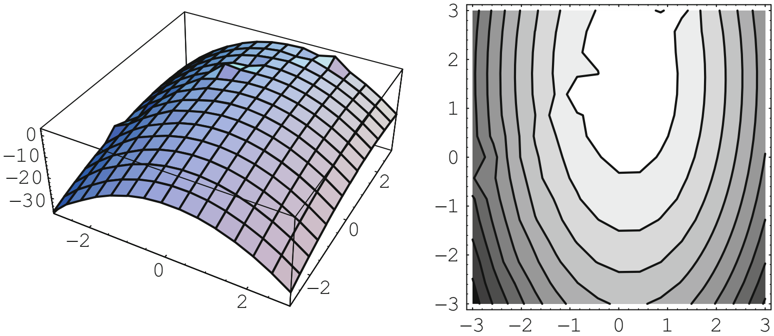


Fig. 2 Graph of $P^d(\varsigma)$ (*left*) and contour of $P^d(\varsigma)$ (*right*)

If we let $a_{11} = -0.4$, $a_{12} = 0.6$, $a_{21} = 0.5$, $a_{22} = -0.3$, $q_1 = -1$, $q_2 = 0.6$, $\mathbf{f} = \{0.3, -0.2\}$, $\mathbf{c} = \{1, 2\}$, $\boldsymbol{\alpha} = \{0.2, 0.8\}$, the graphs and contours of the primal and dual functions are illustrated in Figs. 1 and 2. In this case, the dual problem has a unique critical point $\bar{\varsigma} = \{0.3467, 2.4700\}$ in the space

$$\mathcal{S}_a^+ = \{\varsigma \in \mathbb{R}^2 \mid (q_1 + \varsigma_1 a_{11} + \varsigma_2 a_{21})(q_2 + \varsigma_1 a_{12} + \varsigma_2 a_{22}) > 0\}.$$

Therefore, by Theorem 2, we know that

$$\begin{aligned}\bar{\mathbf{x}} &= \{f_1/(q_1 + \bar{\varsigma}_1 a_{11} + \bar{\varsigma}_2 a_{21}), f_2/(q_2 + \bar{\varsigma}_1 a_{12} + \bar{\varsigma}_2 a_{22})\} \\ &= \{3.1146, -2.9842\}\end{aligned}$$

is a global minimization. It is easy to verify that

$$P(\bar{\mathbf{x}}) = 0.4075 = P^d(\bar{\varsigma}).$$

Example 4.2. Minimization problem of Colville Function.

$$\begin{aligned}\min P(\mathbf{x}) = {} & 100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 90(x_4 - x_3^2)^2 + (1 - x_3)^2 \\ & +10.1((x_2 - 1)^2 + (x_4 - 1)^2) + 19.8(x_2 - 1)(x_4 - 1) \\ \text{s.t.} \quad & -10 \le x_i \le 10, i = 1, 2, 3, 4.\end{aligned}$$

This is a well-known test problem for global optimization. On the dual feasible set

$$\mathcal{S}_a = \{\boldsymbol{\varsigma} \in \mathbb{R}^2 \mid (1 - \varsigma_1)(1 - \varsigma_2) \neq 0\},$$

the canonical dual function has the form of

$$\begin{aligned}P^d(\boldsymbol{\varsigma}) = {} & 42 - \frac{1}{400}\varsigma_1^2 - \frac{1}{360}\varsigma_2^2 \\ & -\frac{1}{2}\begin{bmatrix} 2 \\ 40 - \varsigma_1 \\ 2 \\ 40 - \varsigma_2 \end{bmatrix} \begin{bmatrix} 2 - 2\varsigma_1 & & & \\ & 20.2 & & 19.8 \\ & & 2 - 2\varsigma_2 & \\ & 19.8 & & 20.2 \end{bmatrix}^+ \begin{bmatrix} 2 \\ 40 - \varsigma_1 \\ 2 \\ 40 - \varsigma_2 \end{bmatrix} \\ \Xi(\mathbf{x}, \boldsymbol{\varsigma}) = {} & 42 + (x_2 - x_1^2)\boldsymbol{\varsigma}_1 + (x_4 - x_3^2)\boldsymbol{\varsigma}_2 - \frac{1}{400}\boldsymbol{\varsigma}_1^2 - \frac{1}{360}\boldsymbol{\varsigma}_2^2 \\ & + \left(x_1^2 + 10.1x_2^2 + x_3^2 + 10.1x_4^2 + 19.8x_2x_4\right) \\ & - \left(2x_1 + 40x_2 + 2x_3 + 40x_4\right).\end{aligned}$$

By solving the criticality condition $\nabla\Xi(\mathbf{x}, \boldsymbol{\varsigma}) = 0$, we get three critical points:

$$\begin{aligned}&\bar{x}^1 = (1, 1, 1, 1), \quad \bar{\boldsymbol{\varsigma}}^1 = (0, 0), \\ &\bar{x}^2 = (-0.967974, 0.947139, -0.969516, 0.951248), \quad \bar{\boldsymbol{\varsigma}}^2 = (2.03309, 2.03144), \\ &\bar{x}^3 = (-0.031251, 0.165971, -0.0312582, 0.184264), \quad \bar{\boldsymbol{\varsigma}}^3 = (32.999, 32.9916).\end{aligned}$$

By the fact that $\bar{\boldsymbol{\varsigma}}^1$ is a unique solution in $\mathcal{S}_a^+$, Theorem 2 tells that $\bar{x}^1$ is global minimizer of $P(\mathbf{x})$. Since $n = 4 > m = 2$ and $\bar{\boldsymbol{\varsigma}}^2, \bar{\boldsymbol{\varsigma}}^3 \notin \mathcal{S}_a^-$, we know that both $\bar{x}^2$ and $\bar{x}^3$ are stationary points. But, it is easy to check out that $P(\bar{x}^i) = \Xi(\bar{x}^i, \bar{\boldsymbol{\varsigma}}^i) = P^d(\bar{\boldsymbol{\varsigma}}^i) \;\; \forall i = 1, 2, 3$ and

$$P(\bar{x}^1) = 0 < P(\bar{x}^2) = 0.127006 < P(\bar{x}^3) = 34.84.$$

Example 4.3. Sensor network location problem in $\mathbb{R}^2$.

We consider the simplest sensor network problem with one sensor and two anchors, as shown in Fig. 3. The network is as follows:

$$\mathcal{I}_p = \{\, \{1, 2\}, \{1, 3\} : d_{12} \text{ and } d_{13} \text{ are specified}\},$$

$$\mathcal{I}_b = \{2, 3 : \mathbf{u}_2 = \mathbf{a}_2 \text{ and } \mathbf{u}_3 = \mathbf{a}_3 \text{ are specified}\}.$$

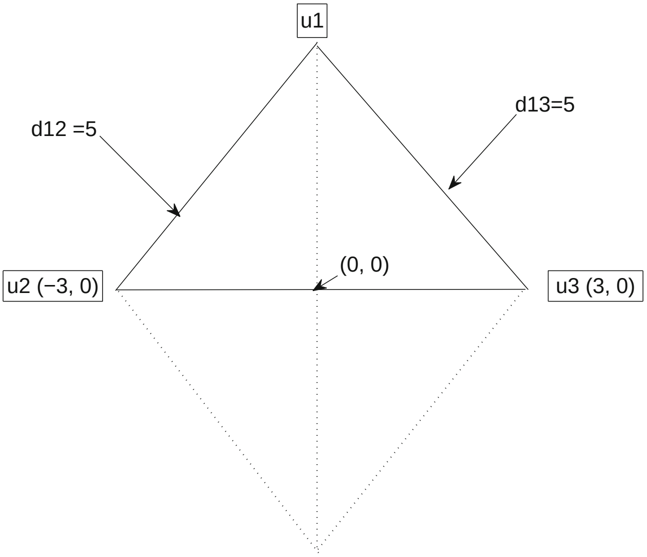


Fig. 3 Sensor network

Let $\mathbf{u}_1 = \{u_1^1, u_1^2\} \in \mathbb{R}^2$ denote the location of sensor and $\mathbf{u}_i = \mathbf{a}_i \ (i = 2, 3)$ denote the location of the two anchors, the sensor network location problem is to solve the following system of nonlinear equations:

$$\begin{aligned} (\mathcal{P}_0) \quad & \|\mathbf{u}_1 - \mathbf{u}_2\|^2 = d_{12}^2, \ \ \{1,2\} \in \mathcal{I}_p, \ \ 2 \in \mathcal{I}_b, \\ & \|\mathbf{u}_1 - \mathbf{u}_3\|^2 = d_{13}^2, \ \ \{1,3\} \in \mathcal{I}_p, \ \ 3 \in \mathcal{I}_b, \end{aligned} \tag{29}$$

where

$$d_{12} = d_{13} = 5, \ \ \mathbf{u}_2 = \mathbf{a}_2 = \{-3, 0\}, \ \ \mathbf{u}_3 = \mathbf{a}_3 = \{3, 0\}.$$

The polynomial defined by (5) for this problem is:

$$\begin{aligned} P(\mathbf{u}) = & \frac{1}{2}\left(\left(u_1^1 - u_2^1\right)^2 + \left(u_1^2 - u_2^2\right)^2 - (d_{12})^2\right)^2 + \frac{1}{2}\left(\left(u_1^1 - u_3^1\right)^2 \right. \\ & \left. + \left(u_1^2 - u_3^2\right)^2 - (d_{13})^2\right)^2. \end{aligned}$$

Let $\mathbf{x} = \{x_1, x_2\} = \mathbf{u}_1 \in \mathbb{R}^2$. On the feasible space

$$\mathcal{U}_a = \left\{\mathbf{u}_i = \left\{u_i^1, u_i^2\right\}, \ i = 1, 2, 3 | \ \ \mathbf{u}_2 = \mathbf{a}_2, \mathbf{u}_3 = \mathbf{a}_3\right\},$$

the fourth-order polynomial $P(\mathbf{x})$:

$$P(\mathbf{x}) = \frac{1}{2}\left((x_1 + 3)^2 + x_2^2 - 5^2\right)^2 + \frac{1}{2}\left((x_1 - 3)^2 + x_2^2 - 5^2\right)^2$$

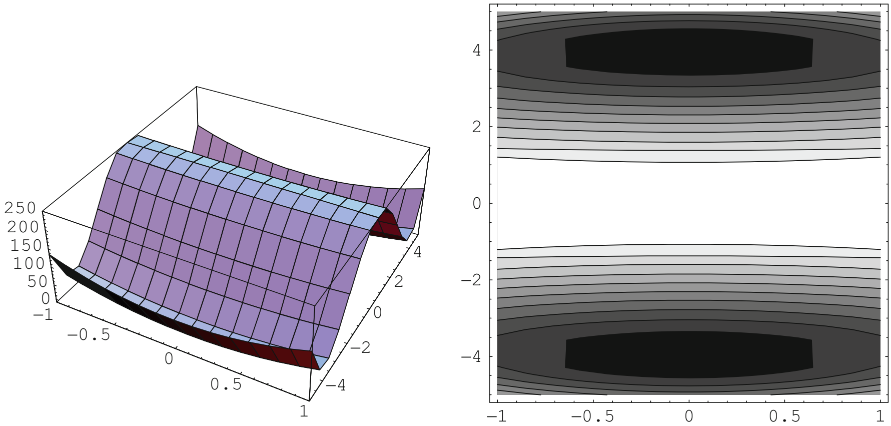


Fig. 4 Double-well function $P(\mathbf{x})$ (*left*) and its contours (*right*) for Example 4.3

is a double-well function as shown in Fig. 4, which has three critical points: two minimizers $\mathbf{x}_1 = \{0, 4\}$ and $\mathbf{x}_2 = \{0, -4\}$, and one local maximizer $\bar{\mathbf{x}}_3 = \{0, 0\}$. It is easy to check that the two minimizers $\bar{\mathbf{x}}_i, \ i = 1, 2$ are solutions to the problem $(\mathcal{P}_0)$. According to [44], this problem is not localizable. The "optimal solution" by the SDP approach is $\bar{\mathbf{x}}_3 = \{0, 0\}$, which is a local maximizer of the function $P(\mathbf{x})$, not a solution to the problem $(\mathcal{P}_0)$. Therefore, this simple example shows that the popular SDP method for solving nonconvex minimization problems usually leads to wrong results.

From the point view of nonconvex structural mechanics, this one sensor–two anchors network system is equivalent to a post-buckling analysis of a large deformed beam model as proposed in [11, 18]: each local minimizer represents a possible buckling (locally stable) state of the beam, while the local maximizer represents the un-bucked (unstable) state. Theoretically speaking, if the beam is made by perfect material, the buckling will never occur unless there is a distributed lateral load, or perturbation $\boldsymbol{\epsilon}$. Therefore, in order to solve the sensor network problem $(\mathcal{P})$ in a realistically way, we introduce the following perturbed problem:

$$(\mathcal{P}_\epsilon): \quad \min\left\{P_\epsilon(\mathbf{x}) = P(\mathbf{x}) - \boldsymbol{\epsilon}^T\mathbf{x} \ : \ \ \mathbf{x} \in \mathbb{R}^2\right\}, \tag{30}$$

where $\boldsymbol{\epsilon} = \{\epsilon^1, \epsilon^2\} \geq 0$ is a given perturbation vector. On the canonical dual feasible space $\mathcal{S}_a$ is defined by:

$$\mathcal{S}_a = \{(\varsigma_1, \varsigma_2) \mid \varsigma_1 + \varsigma_2 \neq 0\},$$

the perturbed canonical dual problem is

$$(\mathcal{P}^d_\epsilon)_{\max}: \quad \max\left\{P^d_\epsilon(\boldsymbol{\varsigma}) : \ \boldsymbol{\varsigma} \in \mathcal{S}^+_a\right\}, \tag{31}$$

where

$$\begin{aligned} P_\epsilon^d(\boldsymbol{\varsigma}) &= -\frac{1}{2}F_\epsilon(\boldsymbol{\varsigma})^T G^{-1}(\boldsymbol{\varsigma})F_\epsilon(\boldsymbol{\varsigma}) + \big(\mathbf{a}_2^T\mathbf{a}_2 - (d_{12})^2\big)\varsigma_1 \\ &\quad+\big(\mathbf{a}_3^T\mathbf{a}_3 - (d_{13})^2\big)\varsigma_2 - \frac{1}{2}\varsigma_1^2 - \frac{1}{2}\varsigma_2^2, \\ F_\epsilon(\boldsymbol{\varsigma}) &= \boldsymbol{\epsilon} + 2\varsigma_1\mathbf{a}_2 + 2\varsigma_2\mathbf{a}_3, \\ G(\boldsymbol{\varsigma}) &= 2\mathrm{diag}(\varsigma_1 + \varsigma_2), \end{aligned}$$

and the canonical dual feasible space $\mathcal{S}_a^+ = \{\boldsymbol{\varsigma} \in \mathcal{S}_a \mid \varsigma_1 + \varsigma_2 > 0\}$.

If we assume $\boldsymbol{\epsilon} = \{0.05, 0.05\}$, the canonical dual problem has a unique solution $\bar{\boldsymbol{\varsigma}}^1 = \{0.00729064, -0.00104125\} \in \mathcal{S}_a^+$. By the triality theory we know that

$$\bar{\mathbf{x}}_1 = G(\bar{\boldsymbol{\varsigma}}_1)F_\epsilon(\bar{\boldsymbol{\varsigma}}_1) = \{0.000694324, 4.000390438\}$$

is a global minimizer of $P_\epsilon(\mathbf{x})$.

On the dual feasible space $\mathcal{S}_a^- = \{\boldsymbol{\varsigma} \in \mathcal{S}_a \mid \varsigma_1 + \varsigma_2 < 0\}$, the canonical dual function $P_\epsilon^d(\boldsymbol{\varsigma})$ has two critical points (see Fig. 4):

$$\begin{aligned} \bar{\boldsymbol{\varsigma}}^2 &= \{0.00104208, -0.0072927\} \in \mathcal{S}_a^-, \\ \bar{\boldsymbol{\varsigma}}^3 &= \{-15.9625, -16.0375\} \in \mathcal{S}_a^-. \end{aligned}$$

By the triality theory we know that $\bar{\mathbf{x}}_2 = \{0.0006944565, -3.9960323936\}$ is a local minimizer, while $\bar{\mathbf{x}}_3 = \{0.00624986, -0.00078125\}$ is a local maximizer (see Fig. 4). It is easy to verify that

$$\begin{aligned} P_\epsilon(\bar{\mathbf{x}}_1) &= -0.200027 = P_\epsilon^d(\bar{\boldsymbol{\varsigma}}^1) < P_\epsilon(\bar{\mathbf{x}}_2) = 0.200791 \approx 0.199973 \\ &= P_\epsilon^d(\bar{\boldsymbol{\varsigma}}^2) < P_\epsilon(\bar{\mathbf{x}}_3) = 256 = P_\epsilon^d(\bar{\boldsymbol{\varsigma}}^3). \end{aligned}$$

By the fact that

$$P(\bar{\mathbf{x}}_1) = 0.0000271153 < P(\bar{\mathbf{x}}_2) = 0.00102382,$$

both the perturbed solutions $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ can be considered as the global minima to the original problem $(\mathcal{P}_0)$ and it is easy to verify that

$$\begin{aligned} \|\mathbf{x}_1 - \mathbf{a}_2\|^2 &= 25.0073, \quad \|\mathbf{x}_1 - \mathbf{a}_3\|^2 = 25.0130, \\ \|\mathbf{x}_2 - \mathbf{a}_2\|^2 &= 24.9724, \quad \|\mathbf{x}_2 - \mathbf{a}_3\|^2 = 24.9641. \end{aligned}$$

5 Conclusions

We have presented a detailed application of the canonical duality theory for solving general sum of fourth-order polynomial optimization problem. An analytical solution is obtained by the complementary-dual principle and its extremality property is classified by the triality theory. Results show that by using the canonical dual transformation, the nonconvex primal problem in $\mathbb{R}^n$ can be converted to a concave maximization dual problem $(\mathcal{P}^d_{\max})$ in $\mathbb{R}^m$, which can be solved by well-developed convex minimization techniques. Application to sensor network localization problem shows that this NP-hard problem in global optimization is actually a special case of discretized finite deformation problem. We demonstrated that the "optimal solution" by the SDP approach is actually a local maximizer of the total *strain potential*. A connection of this network problem with the buckling analysis of large deformed beam model is revealed. By using a perturbed problem, both global and local extrema are obtained. The idea and the method presented in this article can be used and generalized to solve many difficult problems in global optimization, network communication, and scientific computations. Comprehensive review of the canonical duality and its applications can be found in [14, 24, 25, 32].

Acknowledgements The work of David Gao and N. Ruan was supported by NSF grant CCF-0514768, US Air Force (AFOSR) grants FA9550-09-1-0285 and FA9550-10-1-0487. The first author is grateful to Drs. Jiapu Zhang and Changzhi Wu at the University of Ballarat for the helpful comments and suggestions.

References

1. Alfakih A.Y. (2000). Graph rigidity via Euclidean distance matrices, *Linear Algebra and Its Applications*, 310:149–165.
2. Alfakih, A.Y., Khandani, A., and Wolkowicz, H. (1999). Solving Euclidean distance matris completion problems via Semidefinite programming. *Comput. Opt. and Appl.*, 12:13–30.
3. Aspnes, J., Goldberg, D., and Yang, Y.R. (2004). On the computational complexity of sensor network localization. *Lecture Notes in Computer Science* (3121), Springer-Verlag, pp. 32–44.
4. Barvinok, A. (1995). Problems of distance geometry and convex properties of quadratic maps. *Disc. Comp. Geom.*, 13:189–202.
5. Biswas, P. and Ye, Y. (2004). Semidefinite programming for ad hoc wireless sensor network localization. *Proc. 3rd IPSN*, 46–54.
6. Biswas, P., Liang, T.C., Toh, K.C., Wang T.C., and Ye. Y. (2006) Semidefinite Programming Approaches for Sensor Network Localization with Noisy Distance Measurements. *IEEE Transactions on Automation Science and Engineering* 3(4), 360–371.
7. Fang, S.C., Gao D.Y., Sheu R.l. and Wu S.Y. (2008). Canonical dual approach for solving 0-1 quadratic programming problems, *J. Industrial and Management Optimization*, 4.
8. Fang, S.-C., Gao, D.Y., Sheu, R.L., and Xin, W.X. (2009). Global optimization for a class of fractional programming problems, *J. Global Optimization*, 45:337–353, DOI 10.1007/s10898-008-9378-7.
9. Floudas, C. A. and Visweswaran, V. (1995). Quadratic optimization, in *Handbook of Optimization*, R. Horst and P.M. Pardalos (eds). Kluwer Academic Publishers, Dordrecht/Boston/London, pp. 217–270.

10. Gao, D.Y. (1996). Complementary finite element method for finite deformation nonsmooth mechanics, *J. Eng. Math.*, 30, pp. 339–353.
11. Gao, D.Y. (1996). Nonlinear elastic beam theory with applications in contact problem and variational approaches, *Mech. Research Commun.*, 23 (1), 11–17.
12. Gao, D.Y.(1998). Duality, triality and complementary extremun principles in nonconvex parametric variational problems with applications, *IMA J. Appl. Math.*, 61, 199–235.
13. Gao, D.Y. (1999). General Analytic Solutions and Complementary Variational Principles for Large Deformation Nonsmooth Mechanics. *Meccanica* 34, 169–198.
14. Gao, D.Y.(2000). *Duality Principles in Nonconvex Systems: Theory, Methods and Applications*. Kluwer Academic Publishers, Dordrecht/Boston/London, xviii+454pp.
15. Gao, D.Y. (2000). Analytic solution and triality theory for nonconvex and nonsmooth vatiational problems with applicatons, *Nonlinear Analysis*, 42(7), 1161–1193.
16. Gao, D.Y. (2000). Canonical dual transformation method and generalized triality theory in nonsmooth global optimization, *J. Global Optimization*, 17 (1/4), pp. 127–160.
17. Gao, D.Y. (2001). Tri-duality in global optimization, in *Encyclopedia of Optimization*, C. A. Floudas and P.D. Pardalos (eds). Kluwer Academic Publishers, Dordrecht/Boston/London, Vol. 1, pp. 485–491.
18. Gao, D.Y. (2003). Nonconvex semi-linear problems and canonical dual solutions. *Advances in Mechanics and Mathematics*, Vol. II, Kluwer Academic Publishers. pp. 261–312.
19. Gao, D.Y. (2003). Perfect duality theory and complete solutions to a class of global optimization problems, *Optimisation*, 52 (4–5), pp. 467–493.
20. Gao, D.Y. (2004). Complete solutions to constrained quadratic optimization problems, *J. Global Optimization*, special issue on Duality. 29, 377–399.
21. Gao, D.Y.(2005). Sufficient conditions and perfect duality in nonconvex minimization with inequality constraints, *J. Industrial and Management Optimization*, 1, 59–69.
22. Gao, D.Y.(2006). Complete solutions and extremality criteria to polynomial optimization problems, *J. Global Optimization*, 35, 131–143.
23. Gao, D.Y.(2007). Solutions and optimality criteria to box constrained nonconvex minimization problem, *J. Industrial and Management Optimization*, 3(2), 293–304.
24. Gao, D.Y. (2008). Advances in canonical duality theory with applications to global optimization, in *Proceedings of the Fifth International Conference on Foundations of Computer-Aided Process Operations*, Cambridge, MA, M. Ierapetriou, M. Bassett and S. Pistikopoulos (eds.), Omni Press, pp.73–82.
25. Gao, D.Y. (2009). Unified canonical dual solutions to a class of problems in global optimization, *Computers & Chemical Engineering* 33, 1964–1972.
26. Gao, D.Y. and Ogden, R.W. (2008). Closed-form solutions, extremality and nonsmoothness criteria in a large deformation elasticity problem, *Zeitschrift fur angewandte Mathematik und Physik (ZAMP)*, 59 (3), 498–517.
27. Gao, D.Y. and Ogden, R.W. (2008) Multiple solutions to non-convex variational problems with implications for phase transitions and numerical computation, *Quarterly J. Mech. Appl. Math.* 61 (4), 497–522.
28. Gao, D.Y. and Ruan, N. (2010). Solutions to quadratic minimization problems with box and integer constraints, *J. Global Optimization*, 47:463–484.
29. Gao, D.Y., Ruan, N., and Sherali, H.D. (2009). Solutions and optimality criteria for nonconvex constrained global optimization problems, *J. Global Optimization*, 45(3):473–497.
30. Gao, D.Y., Ruan, N. and Sherali, H.D. (2010). Canonical dual solutions for fixed cost quadratic program, *Optimization and Optimal Control*, A. Chinchuluun et al. (eds.), Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6-7.
31. Gao, D.Y. and Strang, G. (1989). Geometric nonlinearity: Potential energy, complementary energy, and the gap function. *Quart. Appl. Math.*, 47(3), 487–504.
32. Gao, D.Y. and Sherali, H.D.(2008). Canonical Duality Theory: Connections between nonconvex mechanics and global optimization. In: Gao, D. Y. & Sherali, H. D. (Editors), *Advances in applied mathematics and global optimization.* Including papers from CDGO, the 1st

International Conference on Complementarity, Duality, and Global Optimization, Blacksburg, August 15–17, 2005, pp. 257–326. Springer, New York. ISBN: 978-0-387-75713-1 90-02

33. Gao, D.Y. and Wu, C-Z. (2010). On the Triality Theory in Global Optimization, to appear in *J. Global Optimization* (published online arXiv:1104.2970v1 at http://arxiv.org/abs/1104.2970)
34. Gao, D.Y. and Yu, H.F. (2008). Multi-scale modelling and canonical dual finite element method in phase transitions of solids, *Int. J. Solids Struct.* 45, 3660–3673
35. Hansen, P., Jaumard, B., Ruiz, M., and Xiong, J.(1991). *Global minimization of indefinite quadratic functions subjects to box constraints.* Technical report, Technical Report G-91-54, Gread, École Polytechnique, Université McGill, Montreal.
36. Horst, R., Pardalos, P.M., and Thoai, N.V.(2000). *Introduction to Global Optimization*. Kluwer Academic Publishers.
37. Li, S.F. and Gupta, A. (2006). On dual configuration forces, *J. of Elasticity*, 84:13–31.
38. Moré, J. and Wu, Z. (1997). Global continuation for distance geometry problems, *SIAM Journal on Optimization*, 7, 814–836.
39. Moreau, J.J. (1968). La notion de sur-potentiel et les liaisons unilatérales en élastostatique, *C.R. Acad. Sc. Paris*, 267 A, 954–957.
40. Moreau, J.J., Panagiotopoulos, P.D. and Strang, G. (1988). *Topics in nonsmooth mechanics.* Birkhuser Verlag, Basel-Boston, MA.
41. Murty, K.G. and Kabadi, S.N.(1988), Some NP-hard problems in quadratic and nonlinear programming, *Math. Programming*, 39, 117–129.
42. Ruan, N., Gao, D.Y., and Jiao, Y. (2010). Canonical dual least square method for solving general nonlinear systems of equations, *Computational Optimization with Applications*, 47:335–347. DOI: 10.1007/s10589-008-9222-5
43. Saxe, J. (1979). Embeddability of weighted graphs in k-space is strongly NP-hard, in Proc. 17th Allerton Conference in Communications, Control, and Computing, Monticello, IL, 1979, 480–489.
44. So, A.M. and Ye, Y. Y. (2006). A semidefinite programming approach to tensegrity theory and realizability of graphs, *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, 766–775.
45. Tseng, P. (2007). Second-order cone programming relaxation of sensor network localization, August, 2005, *SIAM Journal on Optimization*, **18**(1) 156–185 (2007).
46. Waki, H., Kim, S., Kojima, M. and M. Muramatsu. Sums of squares and semidefinite programming relaxations for polynomial optimization problems with structured sparsity. *SIAM Journal on Optimization*, 17 (1) 218–242.
47. Xie, D., Singh, S.B. , Fluder, E.M., and Schlick, T. (2003). Principal component analysis combined with truncated-Newton minimization for dimensionality reduction of chemical databases. *Math. Program.*, Ser B 95: 161–185.

Part II
Theoretical Aspects of Analyzing Information Patterns

Optimal Estimation of Multidimensional Data with Limited Measurements

William MacKunis, J. Willard Curtis, and Pia E.K. Berg-Yuen

Abstract Recent results indicate how to optimally schedule transmissions of a measurement to a remote estimator when there are limited uses of the communication channel available. The resulting optimal encoder and estimation policies solve an important problem in networked control systems when bandwidth is limited. Previous results were obtained only for scalar processes, and the previous work was unable to address questions regarding informational relevance. We extend the state-of-the art by treating the case where the source process and measurements are multidimensional. To this end, we develop a nontrivial re-working of the underlying proofs. Specifically, we develop optimal encoder policies for Gaussian and Gauss–Markov measurement processes by utilizing a measure of the informational value of the source data. Explicit expressions for optimal hyper-ellipsoidal regions are derived and utilized in these encoder policies. Interestingly, it is shown in this chapter that analytical expressions for the hyper-ellipsoids exist only when the state's dimension is even; in odd dimensions (as in the scalar case) the solution requires a numerical look up (e.g., use of the erf function). We have also extended the previous analyses by introducing a weighting matrix in the quadratic cumulative cost function, whose purpose is to allow the system designer to designate which states are more important or relevant to total system performance.

W. MacKunis (✉)
Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA
e-mail: mackunis@gmail.com

J.W. Curtis • P.E.K. Berg-Yuen
Air Force Research Laboratory Munitions Directorate, Eglin AFB, FL 32542, USA
e-mail: jess.curtis@eglin.af.mil; piagreg@cox.net

V.L. Boginski et al. (eds.), *Sensors: Theory, Algorithms, and Applications*,
Springer Optimization and Its Applications 61, DOI 10.1007/978-0-387-88619-0_4,

1 Introduction

The current technological advances in circuit miniaturization are driving down the cost of producing many sensors, especially cameras and other electro-optical imaging sensors. This makes it feasible to deploy numerous high-resolution sensors to provide feedback for some control system. One of the many emerging challenges faced by modern feedback system designers is that such sensors are often not collocated with the plant to be controlled. This leads to difficult questions regarding bandwidth utilization, time-delays, and other effects that have been somewhat addressed in the networked control systems (NCS) literature (e.g., [2, 4–7, 10–12, 14, 15, 17, 19–29]).

Research and design of NCS is of critical importance for applications including remote surgery, chemical processes, automated highway systems, refineries, power plants, and cooperative control of unmanned aerial vehicles and unmanned ground vehicles. While the design of NCS for engineering systems with very high communication rate (i.e., bandwidth) can be simplified, significant challenges can arise in NCS design for systems with severe bandwidth limitations. Examples of applications afflicted by severe communication rate constraints include mobile telephony, sensor networks, remotely controlled systems, and unmanned aerial vehicles. In these applications, the resolution of the transmitted data can be reduced to adhere to the communication rate constraints. This can result in large quantization errors, which can significantly hinder control performance.

Estimation and control of NCS in the presence of communication rate limitations has been widely addressed in literature [5, 14, 15, 19, 21, 25, 28, 29]. In [28], the problem of state estimation in the presence of bandwidth constraints is addressed. Eschewing the classical assumption that the observation process is continuous with additive noises, Wong and Brockett approached the problem by imposing the condition that the observations must be coded and transmitted over a digital communication channel with limited bandwidth. To this effect, they introduced the concept of a *coder-estimator sequence* and a *finitely recursive coder-estimator sequence* [28]. Building on the research in [28], Wong and Brockett investigated the feedback control of a system with bandwidth-limited communication constraints in [29]. It is shown in [29] that the delay in the feedback control prevents asymptotic stabilization for communication constrained systems with uncontrolled dynamics that are unstable. The weaker stability notion of *containability* is introduced in [29] to describe the stability properties of such systems, and connections are shown to exist between containability and the communication data rate, and the rate of change of the state. Nair and Evans addressed stabilization of a linear networked system in [19], where the feedback link has a fixed data rate. Specifically, they calculate the minimum data rate that is needed to asymptotically stabilize the output (in the mth moment) of a discrete-time, linear, time-varying, infinite-dimensional plant with no process or measurement noise. In [6], De Persis shows that a nonlinear system can be stabilized by feedback data communicated through a channel with a finite data rate. Moreover, it is shown in [6] that this technique (termed *encoded feedback*) is capable of achieving asymptotic stability for a nonlinear system that falls within the class of feedforward systems. The researches in [6, 19, 28, 29] provide multiple techniques to stabilize systems subject to feedback communication rate limitations. Additional difficulties exist, however, in designing NCS for systems with imperfect communication systems.

For applications utilizing wireless and Internet channels, data packet losses (drop-outs, erasures) can cause significant complications in NCS design. In [24], the problem of state estimation in the presence of lossy measurements is addressed. While the scheme (termed *finite loss history estimator* (FLHE)) proposed in [24] is suboptimal, it has the practical advantage of handling lossy measurements in a manner that is computationally less expensive than that of the time-varying Kalman estimator (TVKE). This advantage is due to the fact that while, like the TVKE, the FLHE scheme uses a predictor/estimator structure, the corrector gains are precomputed in the FLHE, as opposed to being recursively updated as in the TVKE. Feedback stabilization of discrete-time linear systems utilizing communication channels subject to stochastic drop-outs is investigated in [7]. The research in [7] proves that the maximum erasure probability that can be tolerated using a linear receiver is directly related to the unstable eigenvalues of the plant. The effect of data losses over unreliable network links is examined in [23]. By modeling the arrival of an observation as a random process, a discrete-time Kalman filter is utilized in [23] to estimate the state of a controlled system. This estimate is then used for feedback control. Whereas in [23] the convergence properties of the Kalman filter are examined assuming complete reception or loss of the observation measurements, the research in [17] provides an extension by allowing partial packet losses in the observation. In [26], a technique is presented for handling missing or delayed information due to packet drops in a networked system. To that end, optimal conditions for stability of a NCS in the presence of packet drops are determined using a stochastic framework. The work in [26] is applicable to both continuous-time and discrete-time systems, and it allows for both measurement of noise and process noise.

A central issue in designing NCS that use slow and/or lossy communication channels is that of deciding which data is valuable enough to transmit (and thus consume relatively scarce bandwidth resources) and which data can be safely discarded. The seminal work by Meier et al. [18] addressed a version of this problem, where they assumed that only one sensor among a set of multiple sensors could be used at any given time. Meier et al. also proved a separation property between the optimal plant control policy and the measurement control policy. The measurement control problem, which is the sensor scheduling problem, was cast as a nonlinear deterministic control problem and shown to be solvable by a tree-search in general. It was proven that if the decision to choose a particular sensor rests with the estimator, an open-loop selection strategy is optimal for a cost based on the estimate error covariance [1]. Forward dynamic programming (DP) and a gradient method were proposed for this purpose.

In this chapter, we address the issue of informational value in a network estimation and control context. In these fields there has been a traditional bias towards using the entropic formulation [8, 30] as a standard measure of information, though there has recently been increasing interest in an alternative informational value metrics for cooperative and networked estimation [9] and control. In [13], the authors examine the optimal communication policy of an observer who is observing a random process and who must decide whether to send observations across a communication channel to an estimator. They discover jointly optimal policies for the observer and the estimator so as to minimize the mean-square estimation error of the observer in the case where the observer is limited in the number of times that it can transmit. A method is presented in [13] to compute the optimal transmission policy off-line via DP. A very similar problem was treated in [16], where the optimal policy involves transmitting a measurement only if it lies outside some symmetric region centered around the mean value of the observed process. Both the papers treat only the scalar case, and propose solutions of an optimization problem where the objective function considers only estimation and communication errors.

The constrained uses of the communication channel necessitate information arbitrage, and the sensing agent must decide which measurement will be of most value to the estimator. Moving beyond a scalar problem is essential for exploring informational relevance issues, because it forces the observer to consider which elements of the observation vector may be most relevant to some decision-maker (controller) who is being fed information by the estimator. The development in this chapter illustrates how an observation and an estimation techniques can be designed for optimal estimation of multidimensional data over a communication channel with limited uses. This entails a nontrivial reworking of the proofs found in [13]. In particular, we assume a symmetric threshold policy is optimal, then we show how to compute the optimal transmission region via DP. In the multidimensional case we consider, these regions are hyper-ellipsoids. We discover that for even dimensional spaces there exist analytical expressions for the optimal cost-to-go and the corresponding hyper-ellipsoids. This is in contrast to odd dimensional spaces (e.g., the scalar case presented in [13]), where the optimal cost-to-go must be numerically computed via the error function.

An important aspect of our extension to multiple dimensions is that, in the present work, the objective function can entertain notions of informational relevance through a cost-weighting matrix. In the scalar case the objective is to minimize the expected cumulative squared estimation error; in multiple dimensions one is forced to grapple with the fact that some states might be much more valuable to know precisely than others, and this leads to interesting future questions regarding how control systems being fed by this estimator might influence the choice of weights in the observer/estimator policy. Numerical simulation results are provided, which illustrate the performance of the value of information (VOI)-based optimal estimation strategy. Moreover, the results clearly show that reduced estimation error can be achieved through the utilization of a cost-weighting matrix.

1.1 Problem Statement

The problem of optimal estimation of data based on limited measurements will be addressed for the case where the source data is multidimensional. The problem will

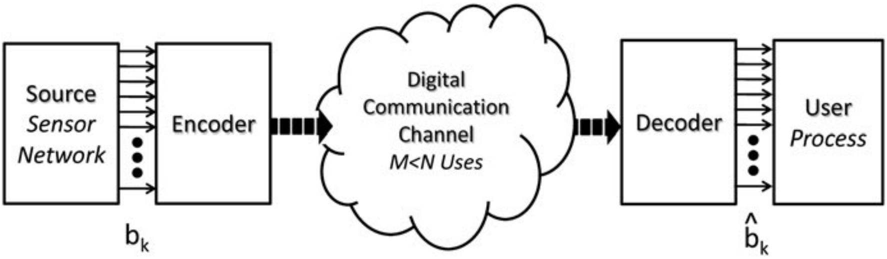


Fig. 1 Communication system with limited transmission channel uses

be framed in the context of communication of data across a channel with limited uses as depicted in Fig. 1. By extending a technique similar to that presented in [12], VOI-based encoder and decoder policies will be utilized for optimal sequential estimation of n dimensional data communicated over a channel with limited uses. Specifically, a source sequentially generates data $b_k \in \mathbb{R}^n$ over an N-step decision horizon $0 \leq k \leq N-1$, which must be transmitted over a noiseless channel. The data b_k are generated according to some a priori known as stochastic process (e.g., an independent identically distributed (I.I.D.) Gaussian random process or a correlated Gauss–Markov process). An encoder/observer is placed at the source output, and a decoder/estimator is placed at the channel output. Observer and estimator policies are utilized to optimize the accuracy of the communication system in the presence of limited channel uses.

The communication channel is restricted, such that, it can only be accessed for $M < N$ transmissions. The objective is to design observer and estimator policies that minimize the error between the source data b_k and its estimate $\hat{b}_k$ over the N-step decision horizon. At each time step k, the number of remaining time steps is denoted $1 \leq t_k \leq N$, and the remaining number of available transmissions is denoted $1 \leq s_k \leq M$.

2 Source Process is Gaussian

2.1 The Solution in the n-D Case

The total estimation error over the N-step horizon can be expressed as:

$$e^{\pi}_{(s,t)} = \sum_{k=0}^{N-1} \left\{ \left(b_k - \hat{b}_k \right)^T Q \left(b_k - \hat{b}_k \right) \right\}, \tag{1}$$

where $Q \in \mathbb{R}^{n \times n}$ denotes a user-defined weighting matrix. The estimate $\hat{b}_k \in \mathbb{R}^n$ of b_k is defined as the following conditional expectation:

$$\hat{b}_k = E \left\{ b_k \mid (s_k, t_k) ; x_k \right\},$$

where $x_k \in \mathbb{R}^n$ denotes the observer output. The observer policy μ_k can be expressed as

$$\mu_k = \begin{cases} b_k & \text{if } b_k \in \mathcal{J}_{(s_k,t_k)}, \\ \text{NT} & \text{if } b_k \in \mathcal{J}^c_{(s_k,t_k)}, \end{cases} \tag{2}$$

where $\mathcal{J}_{(s_k,t_k)}$ and $\mathcal{J}^c_{(s_k,t_k)}$ denote the optimal observation set and its complement, respectively. To simplify the notation in the subsequent analysis, these terms will be denoted as $\mathcal{J}$ and $\mathcal{J}^c$. It will be shown in the next section that $\mathcal{J}^c$ can be explicitly calculated as the n dimensional region that globally minimizes the *cost-to-go of the estimation error*. If the source data $b_k \in \mathcal{J}$, the observer transmits the data, and the estimator uses the transmitted data. If $b_k \in \mathcal{J}^c$, the observer does not transmit the data, but instead transmits a single bit datum indicating NT for no transmission. When the estimator receives the NT signal, it uses the expected value of the source data based on knowledge of the statistics of b_k. The observer policy utilizes the VOI of the source data to determine whether or not the data should be transmitted. Heuristically speaking, if the observer determines that the data falls within the region $\mathcal{J}^c$, the data is determined to have low informational value since it is close to the expected value of the source data. If the data falls within the set $\mathcal{J}$, it is far from the expected value, so it is determined to have high informational value. Data having low informational value is not transmitted by the observer; instead the NT datum is transmitted. Data having high informational value is transmitted, and the estimator simply uses the transmitted data. Since the estimator has knowledge of the statistics of the source data, it can minimize the overall error in the estimation of the source data by using the expected value of the data when the NT signal is received. In the following development, the procedure for calculating the optimal observation set $\mathcal{J}^*$ will be presented.

2.2 Optimal Observation Set

In this section, the cost-to-go equation in (3) will be utilized to calculate an optimal observation set $\mathcal{J}^*$ within which the source data possesses high VOI. To that end, the optimal region $\mathcal{J}^{c*}$ will be calculated from (3) as the range of b that globally minimizes $e^*_{(s,t)}$. The set $\mathcal{J}^*$ will be used to develop an observer policy that only transmits data having high VOI.

Based on (1), the DP equation can be used to express the optimal estimation error as [3]:

$$e^*_{(s,t)} = \min_{\mathcal{J}_{(s,t)}} \left\{ e^*_{(s-1,t-1)} - \int_{b\in\mathcal{J}^c} \left[\left(e^*_{(s-1,t-1)} - e^*_{(s,t-1)} \right) f(b) - b^T Q b f(b) \right] db \right\}, \tag{3}$$

given that b has zero mean, where the fact that

$$\int_{b \in \mathcal{J}} f\,(b)\,db = 1 - \int_{b \in \mathcal{J}^c} f\,(b)\,db$$

was utilized.

The Gaussian PDF $f\,(b)$ in (3) can be expressed as:

$$f\,(b) = \frac{1}{(2\pi)^{n/2}\,|P|^{1/2}} \exp\left\{-\frac{1}{2} b^T P^{-1} b\right\}, \tag{4}$$

where $P \in \mathbb{R}^{n \times n}$ denotes a positive definite, symmetric covariance matrix. To facilitate the following analysis, a linear transformation will be defined as:

$$x = P^{-1/2} b \qquad b = P^{1/2} x. \tag{5}$$

Using the Jacobian determinant, (5) can be used to express the integration differential db as:

$$db = \left|P^{1/2}\right| dx. \tag{6}$$

After using (5), the expression in (4) can be rewritten as:

$$f\,(x) = \frac{1}{(2\pi)^{n/2}\,|P|^{1/2}} \exp\left\{-\frac{1}{2} x^T x\right\}, \tag{7}$$

where the fact that

$$\left[P^{-1/2}\right]^T = P^{-1/2} \tag{8}$$

was utilized. The motivation behind the linear transformation in (5) is based on the desire to facilitate the subsequent evaluation of the integrals in (3). After substituting (5) in (3), the following is obtained:

$$\begin{aligned} e^*_{(s,t)} = \min_{\mathcal{J}_{(s,t)}} \Bigg\{ & e^*_{(s-1,t-1)} - \left(e^*_{(s-1,t-1)} - e^*_{(s,t-1)}\right) \int_{x \in \mathcal{J}_x^c} \left|P^{1/2}\right| f\,(x)\,dx \\ & + \int_{x \in \mathcal{J}_x^c} \left|P^{1/2}\right| x^T G x f\,(x)\,dx \Bigg\}, \end{aligned} \tag{9}$$

where the positive semi-definite, symmetric matrix G is defined as:

$$G \triangleq P^{1/2} Q P^{1/2}.$$

After transforming (9) into spherical coordinates, the integration region minimizing the estimation error can be calculated as:

$$r^{*2} = x^{*T} x^* = \frac{n\left(e^*_{(s-1,t-1)} - e^*_{(s,t-1)}\right)}{\mathrm{tr}\,(QP)}, \tag{10}$$

where $\mathrm{tr}\,(\cdot)$ denotes the trace of a matrix, x^* denotes the optimal value of x, and r^* denotes the corresponding optimal distance as expressed in spherical coordinates.

Hence, transforming (10) back into the b domain using (5), the optimal region is obtained as:

$$b^{*T}P^{-1}b^{*} = \frac{n\left(e^{*}_{(s-1,t-1)} - e^{*}_{(s,t-1)}\right)}{\operatorname{tr}(QP)}. \tag{11}$$

3 Estimation Error Recursion

In this section, an analytical formulation for the estimation error recursion formula is derived. To derive an explicit mathematical expression for the optimal estimation error, the two integrals in (9) must be evaluated. Since the integration variable in this case is a vector, the expression will be transformed into spherical coordinates to facilitate the derivation.

After transforming the Gaussian PDF given in (7) to spherical coordinates, the PDF can be expressed as:

$$f(r) = \left(\frac{1}{(2\pi)^{\frac{n}{2}}\,|P|^{1/2}}\right)\exp\left\{-\frac{1}{2}r^{2}\right\}, \tag{12}$$

where $r \in \mathbb{R}$ denotes the radial distance in spherical coordinates.

3.1 Evaluating the Recursion Formula

By transforming the integral expression in (9) to spherical coordinates and utilizing the region defined in (11) to define the limits of integration, the optimal estimation error (cost-to-go) can be calculated as:

$$\begin{aligned} e^{*}_{(s,t)} = e^{*}_{(s-1,t-1)} - &\left(e^{*}_{(s-1,t-1)} - e^{*}_{(s,t-1)}\right)\left|P^{1/2}\right| \\ &\times \int_{0}^{2\pi}\int_{0}^{\pi}\cdots\int_{0}^{\pi}\int_{0}^{r^{*}}\left(f(r)\,r^{n-1}\sin^{n-2}\phi_1\sin^{n-3}\phi_2\cdots\sin\phi_{n-2}\right) \\ &drd\phi_1\cdots d\phi_{n-1} + \left|P^{1/2}\right|\int_{0}^{2\pi}\int_{0}^{\pi}\cdots\int_{0}^{\pi}\int_{0}^{r^{*}} f(r)\,r^{n+1} \\ &\left(g_{11}\cos^2\phi_1 + g_{22}\sin^2\phi_1\cos^2\phi_1 + \cdots + g_{nn}\sin^2\phi_1\right. \\ &\left.\sin^2\phi_2\cdots\sin^2\phi_{n-1}\right)\sin^{n-2}\phi_1\cdots\sin\phi_{n-2}drd\phi_1 d\phi_2\cdots d\phi_{n-1}. \end{aligned} \tag{13}$$

In (13), the limits of integration $r^{*} = \sqrt{\gamma}$ were determined by utilizing (10) to calculate the optimal set in terms of spherical coordinates as:

$$r^{*} = \sqrt{\gamma}, \tag{14}$$

where $\gamma \in \mathbb{R}$ is defined as:

$$\gamma = \frac{n\left(e^*_{(s-1,t-1)} - e^*_{(s,t-1)}\right)}{tr\left[QP\right]}. \tag{15}$$

3.1.1 The Case of *n* Even

The first inner integral (i.e., the integral with respect to r) in (13) can be evaluated for the case where n is even as follows:

$$\begin{aligned} c_{1e}(\gamma) &= \left(\frac{1}{(2\pi)^{n/2}|P|^{1/2}}\right)\int_0^{\sqrt{\gamma}} r^{n-1}\exp\left\{-\frac{1}{2}r^2\right\}dr \\ &= \left(\frac{2^{\frac{n}{2}-1}\left(\frac{n}{2}-1\right)!}{(2\pi)^{\frac{n}{2}}|P|^{1/2}}\right)\left(1-\left\{\sum_{\substack{j=2\\ j\text{ even}}}^{n}\left\{\frac{\left(\frac{1}{2}\right)^{\frac{j}{2}-1}}{\left(\frac{j}{2}-1\right)!}\right\}\gamma^{\frac{j}{2}-1}\right\}\right)\exp\left\{-\frac{1}{2}\gamma\right\}. \end{aligned} \tag{16}$$

In a similar manner, the second inner integral in (13) can be evaluated as:

$$\begin{aligned} c_{2e}(\gamma) &= \left(\frac{1}{(2\pi)^{n/2}|P|^{1/2}}\right)\int_0^{\sqrt{\gamma}} r^{n+1}\exp\left\{-\frac{1}{2}r^2\right\}dr \\ &= \left(\frac{2^{\frac{n}{2}}\left(\frac{n}{2}\right)!}{(2\pi)^{\frac{n}{2}}|P|^{1/2}}\right)\left(1-\left\{\sum_{\substack{j=2\\ j\text{ even}}}^{n+2}\left\{\frac{\left(\frac{1}{2}\right)^{\frac{j}{2}-1}}{\left(\frac{j}{2}-1\right)!}\right\}\gamma^{\frac{j}{2}-1}\right\}\right)\exp\left\{-\frac{1}{2}\gamma\right\}. \end{aligned} \tag{17}$$

For the case where n is even, the expressions in (16) and (17) can be used to rewrite (13) as:

$$\begin{aligned} e^*_{(s,t)} = e^*_{(s-1,t-1)} &- \left(e^*_{(s-1,t-1)} - e^*_{(s,t-1)}\right)\left|P^{1/2}\right|c_{1e}(\gamma) \\ &\times\int_0^{2\pi}\int_0^{\pi}\cdots\int_0^{\pi}\left(\sin^{n-2}\phi_1\cdots\sin\phi_{n-2}\right)d\phi_1\cdots d\phi_{n-2}d\phi_{n-1} \\ &+ c_{2e}(\gamma)\left|P^{1/2}\right|\int_0^{2\pi}\int_0^{\pi}\cdots\int_0^{\pi}\left(g_{11}\cos^2\phi_1 + g_{22}\sin^2\phi_1\cos^2\phi_1\right. \\ &+\cdots+\left.g_{nn}\sin^2\phi_1\sin^2\phi_2\cdots\sin^2\phi_{n-1}\right)\sin^{n-2}\phi_1\cdots \\ &\quad\sin\phi_{n-2}d\phi_1 d\phi_2\cdots d\phi_{n-1}. \end{aligned} \tag{18}$$

The first set of integrals in (18) can be evaluated with respect to $\phi_1, \ldots, \phi_{n-1}$ as:

$$\int_0^{2\pi}\int_0^{\pi}\cdots\int_0^{\pi}\left(\sin^{n-2}\phi_1\cdots\sin\phi_{n-2}\right)d\phi_1\cdots d\phi_{n-2}d\phi_{n-1}=\frac{(2\pi)^{\frac{n}{2}}}{2^{\frac{n}{2}-1}\left(\frac{n}{2}-1\right)!}, \tag{19}$$

where the fact that $(n-2)\ (n-4)\cdots\ (4)\ (2)\ =\ 2^{\frac{n}{2}-1}\left(\frac{n}{2}-1\right)!$ was utilized. Similarly, the last integral of (18) can be evaluated as:

$$\begin{aligned}&\int_0^{2\pi}\int_0^{\pi}\cdots\int_0^{\pi}\left(g_{11}\cos^2\phi_1+\cdots+g_{nn}\sin^2\phi_1\sin^2\phi_2\cdots\sin^2\phi_{n-2}\sin^2\phi_{n-1}\right)\\&\qquad\qquad\times\sin^n\phi_1\sin^{n-1}\phi_2\cdots\sin^3\phi_{n-2}\sin^2\phi_{n-1}d\phi_1 d\phi_2\cdots d\phi_{n-2}d\phi_{n-1}\\&\qquad\qquad=\operatorname{tr}(QP)\left(\frac{(2\pi)^{\frac{n}{2}}}{2^{\frac{n}{2}}\left(\frac{n}{2}\right)!}\right),\end{aligned} \tag{20}$$

where the facts that $\operatorname{tr}(QP)=(g_{11}\ +g_{22}+\cdots+g_{nn})$ and $(n)\ (n-2)\cdots\ (4)\ (2) = 2^{\frac{n}{2}}\left(\frac{n}{2}\right)!$ were utilized. Thus, after substituting (16), (17), (19), and (20) into (18), $e^*_{(s,t)}$ can be expressed as:

$$\begin{aligned}e^*_{(s,t)}=&e^*_{(s-1,t-1)}-\left(e^*_{(s-1,t-1)}-e^*_{(s,t-1)}\right)\left(\frac{(2\pi)^{\frac{n}{2}}c_{1e}(\gamma)}{2^{\frac{n}{2}-1}\left(\frac{n}{2}-1\right)!}\right)\\&+\operatorname{tr}(QP)\left(\frac{(2\pi)^{\frac{n}{2}}c_{2e}(\gamma)}{2^{\frac{n}{2}}\left(\frac{n}{2}\right)!}\right).\end{aligned} \tag{21}$$

3.1.2 The Case of *n* Odd

Similarly to the case where n is even, the integral expression in (13) can be evaluated as follows for the case where n is odd:

$$\begin{aligned}e^*_{(s,t)}=&e^*_{(s-1,t-1)}-\sqrt{\frac{2}{\pi}}\left[\left(e^*_{(s-1,t-1)}-e^*_{(s,t-1)}\right)\left(\frac{2^{\frac{n-3}{2}}\left(\frac{n-3}{2}\right)!\,(2\pi)^{\frac{n}{2}}}{(n-2)!}\right)c_{1o}(\gamma)\right]\\&+\operatorname{tr}(QP)\left(\frac{(2\pi)^{\frac{n}{2}}\,2^{\frac{n-1}{2}}\left(\frac{n-1}{2}\right)!}{n!}\right)c_{2o}(\gamma),\end{aligned} \tag{22}$$

where $c_{1o}(\gamma)$ and $c_{2o}(\gamma)$ are explicitly defined as:

$$
\begin{aligned}
c_{1o}(\gamma) &= \int_0^{\sqrt{\gamma}} r^{n-1} f(r)\, dr \\
&= \left(\frac{(n-2)!}{(2\pi)^{\frac{n}{2}}\, 2^{\frac{n-3}{2}} \left(\frac{n-3}{2}\right)! \, |P|^{1/2}} \right) \left(\sqrt{2\pi} \left[\Phi\left(\sqrt{\gamma}\right) - \Phi(0) \right] \right. \\
&\quad \left. - \left\{ \sum_{\substack{j=3 \\ j \text{ odd}}}^{n} \left\{ \frac{2^{\frac{j-3}{2}} \left(\frac{j-3}{2}\right)!}{(j-2)!} \right\} \gamma^{\frac{j-2}{2}} \right\} \exp\left\{ -\frac{1}{2}\gamma \right\} \right)
\end{aligned}
\tag{23}
$$

and

$$
\begin{aligned}
c_{2o}(\gamma) &= \int_0^{\sqrt{\gamma}} r^{n+1} f(r)\, dr \\
&= \left(\frac{n!}{2^{\frac{n-1}{2}} \left(\frac{n-1}{2}\right)! \, (2\pi)^{\frac{n}{2}} \, |P|^{1/2}} \right) \left(\sqrt{2\pi} \left[\Phi\left(\sqrt{\gamma}\right) - \Phi(0) \right] \right. \\
&\quad \left. - \left\{ \sum_{\substack{j=3 \\ j \text{ odd}}}^{n+2} \left\{ \frac{2^{\frac{j-3}{2}} \left(\frac{j-3}{2}\right)!}{(j-2)!} \right\} \gamma^{\frac{j-2}{2}} \right\} \exp\left\{ -\frac{1}{2}\gamma \right\} \right)
\end{aligned}
\tag{24}
$$

respectively. In (23) and (24), $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard Gaussian random variable with zero-mean and unit variance, which is defined as $\Phi(b) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} \exp\left\{-\frac{1}{2}\zeta^2\right\} d\zeta$ for any $\zeta \in \mathbb{R}$.

The VOI-based encoder policy in (2) and the decoder policy can be used along with the estimation error recursion formulas given in (21) and (22) to optimize estimation error over an N-step decision horizon, using a communication channel with $M < N$ uses.

4 Source Process is Gauss–Markov

In this section, the optimal estimation technique outlined in the previous section will be applied to a system for which the source data is generated via a *Gauss–Markov process.*

In the case where the source process is Markov driven by an I.I.D. Gaussian process $\{w_k\}$ with zero mean, the source data is generated by the following model:

$$b_k = Ab_{k-1} + w_{k-1}, \tag{25}$$

where b_k, $w_k \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$. The current value of the state b_k depends only on the value of the state at the previous time step (i.e., Markov property). If r_k time steps have passed since the last transmission was received, then the current value of the state b_k will depend on the value of the state r_k time steps in the past, i.e., b_{k-r}. If data were transmitted in the previous time step, then $r_k = 1$, and the current value b_k of the source data is given explicitly by (25). If $r_k > 1$, a linear regression can be performed on (25) to determine the current value of the state b_k in terms of b_{k-r} as:

$$b_k = A^r b_{k-r} + A^{r-1} w_{k-r} + A^{r-2} w_{k-r+1} + \cdots + A^1 w_{k-2} + A^0 w_{k-1}. \tag{26}$$

To simplify the notation in the following analysis, let r denote the number of time units passed since the last transmission of a source output at time step k. So in the presence of noise $\{w_k\}$, the estimation error will increase with the number of time steps passed since the use of the channel for transmission. Based on (26), the expectation $E\left[b_k \mid b_{(k-r)}\right]$ (i.e., the expected value of b_k after r missed transmissions) can be expressed as:

$$E\left[b_k \mid b_{(k-r)}\right] = A^r b_{k-r}, \tag{27}$$

where the fact that w_k has zero mean was utilized. Based on (27), it is apparent that the mean value of b_k varies with the number of missed transmissions r. Thus, the distribution of b can be expressed as:

$$b \sim \mathcal{N}\left(A^r b_{k-r}, P_r\right), \tag{28}$$

where $P_r \in \mathbb{R}^{n\times n}$ denotes the covariance of b_k after r missed transmissions.

4.1 Covariance Matrix Calculation

For the source process given in (25), the expected value of b and the covariance matrix will change with the number of missed transmissions r. In this section, the general formula for the covariance matrix based on r missed transmissions will be derived.

The covariance matrix P_r is defined as:

$$P_r \triangleq E\left[(b - E[b])(b - E[b])^T\right]. \tag{29}$$

In the n dimensional case, P_r can be expressed in matrix form as:

$$P_r = \begin{bmatrix} \sigma_{1r}^2 & \sigma_{1r}\sigma_{2r} & \cdots & \sigma_{1r}\sigma_{nr} \\ \sigma_{1r}\sigma_{2r} & \sigma_{2r}^2 & \cdots & \sigma_{2r}\sigma_{nr} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1r}\sigma_{nr} & \cdots & \sigma_{(n-1)r}\sigma_{nr} & \sigma_{nr}^2 \end{bmatrix}, \tag{30}$$

where σ_{ir}^2 and $\sigma_{ir}\sigma_{jr}\ \forall\ i, j = 1, \ldots, n$ are defined as:

$$\sigma_{ir}\sigma_{jr} \triangleq E\left[\left(b_{ik} - A^r b_{i(k-r)}\right)\left(b_{jk} - A^r b_{j(k-r)}\right)\right], \tag{31}$$

where (27) was utilized, and b_{ik}, $b_{i(k-r)}$ denote the ith element of b at time steps k and $k-r$, respectively. After utilizing (26) and performing the necessary multiplications, (31) can be used to express the (i, j)th element of the covariance matrix as:

$$\begin{aligned} \sigma_{ir}\sigma_{jr} \triangleq E\Bigg[\Bigg(&\sum_{m=1}^{n} A_{i,m}^{r-1} w_{m(k-r)} + \sum_{m=1}^{n} A_{i,m}^{r-2} w_{m(k-r+1)} \\ &+ \cdots + \sum_{m=1}^{n} A_{i,m}^{1} w_{m(k-2)} + \sum_{m=1}^{n} A_{i,m}^{0} w_{i(k-1)}\Bigg) \\ &\times \Bigg(\sum_{m=1}^{n} A_{j,m}^{r-1} w_{m(k-r)} + \sum_{m=1}^{n} A_{j,m}^{r-2} w_{m(k-r+1)} \\ &+ \cdots + \sum_{m=1}^{n} A_{j,m}^{1} w_{m(k-2)} + \sum_{m=1}^{n} A_{j,m}^{0} w_{j(k-1)}\Bigg)\Bigg], \end{aligned} \tag{32}$$

$\forall\ i, j = 1, \ldots, n$, where the notation $A_{i,m}^{x}$ represents the (i, m)th element of the matrix A^x, and $w_{q(p)}$ denotes the qth element of the vector w at time p. After multiplying the parenthetic polynomial terms in (32) and taking the expected values, the variances can be expressed as:

$$\sigma_{ir}^2 = \sum_{q=1}^{r}\left(\sum_{m=1}^{n} A_{i,m}^{q-1}\sigma_m\right)^2 \tag{33}$$

and the covariances can be expressed as:

$$\sigma_{ir}\sigma_{jr} = \sum_{q=1}^{r}\left(\sum_{m=1}^{n} A_{i,m}^{q-1}\sigma_m \sum_{p=1}^{n} A_{j,p}^{q-1}\sigma_p\right), \tag{34}$$

for $i, j = 1, \ldots, n$, where the fact that the noise w is I.I.D. and was utilized.

Note that if $r = 1$, the variances and covariances reduce to σ_i^2 and $\sigma_i \sigma_j\ \forall\ i, j = 1, \ldots, n$, respectively. In the scalar case, $n = 1$, and (33) can be used to show that the variance after r steps without transmissions can be calculated as [12]:

$$\sigma_r^2 = \left(\sum_{k=1}^{r} A^{2(k-1)} \right) \sigma_b^2, \tag{35}$$

where $A \in \mathbb{R}$, σ_r^2 denotes the variance of b after r time steps of no transmissions, and σ_b^2 denotes the variance of b after only a single time step without a transmission (i.e., when $r = 1$).

4.2 Optimal Observation Set

For the case where the source data is generated via the Gauss–Markov process in (25) and (26), the encoder and decoder policies can be derived in a manner very similar to the case where the source data is a Gaussian random variable. The main difference in the Gauss–Markov case is that the mean and covariance of the vector b can change at each time step based on the number of missed transmissions r, as expressed in (27), (30), (33), and (34).

The PDF in the Gauss–Markov case can be expressed as:

$$f(b) = \left(\frac{1}{(2\pi)^{\frac{n}{2}} |P_r|^{1/2}} \right) \exp \left\{ -\frac{1}{2} (b - \mu)^T P_r^{-1} (b - \mu) \right\}, \tag{36}$$

where $\mu = A^r b_{k-r}$, and P_r is defined in (30), (33), and (34). To derive the optimal observation set $\mathcal{J}_{(r_k, s_k, t_k)}$, the optimal cost-to-go is formulated as:

$$e^*_{(r,s,t)} = \min_{\mathcal{J}_{(s,t)}} \left\{ e^*_{(1,s-1,t-1)} - \left(e^*_{(1,s-1,t-1)} - e^*_{(r+1,s-1,t-1)} \right) \int_{b \in \mathcal{J}^c} f(b)\, db \right.$$
$$\left. + \int_{b \in \mathcal{J}^c} (b - \mu)^T Q (b - \mu) f(b)\, db \right\}. \tag{37}$$

The notation $e^*_{(r,s,t)}$ is used in this case to represent the optimal estimation error, since the estimation error depends on the three parameters: r, s, and t in the Gauss–Markov case. To facilitate the following analysis, a variable transformation is defined as:

$$b = P_r^{1/2} x + \mu, \qquad x = P_r^{-1/2} (b - \mu), \qquad db = \left| P_r^{1/2} \right| dx. \tag{38}$$

By using the linear transformation in (38), the PDF in (36) can be expressed as in (7), and the optimal estimation error can be expressed as in (9). Thus, following a procedure identical to that used in the Gaussian case, the integration region minimizing the estimation error can be calculated as:

$$r^{*2} = x^{*T}x^{*} = \frac{n\left(e^{*}_{(1,s-1,t-1)} - e^{*}_{(r+1,s,t-1)}\right)}{\mathrm{tr}\left(QP_r\right)}, \tag{39}$$

where $e^{*}_{(1,s-1,t-1)}$ and $e^{*}_{(r+1,s,t-1)}$ denote the cumulative average estimation error (i.e., optimal cost-to-go) after 1 missed transmission and $r+1$ missed transmissions, respectively. After utilizing (38), the optimal region in (39) can be expressed as:

$$\left(b^{*} - A^{r}b_{k-r}\right)^{T} P^{-1}\left(b^{*} - A^{r}b_{k-r}\right) = \gamma, \tag{40}$$

where the fact that $\mu = A^{r}b_{k-r}$ was utilized, and γ is defined as:

$$\gamma \triangleq \frac{n\left(e^{*}_{(1,s-1,t-1)} - e^{*}_{(r+1,s,t-1)}\right)}{tr\left(QP_r\right)}. \tag{41}$$

Hence, as in the case where the source data is purely Gaussian, the region minimizing the estimation error in the Gauss–Markov case is defined by an n dimensional *hyper-ellipsoid*. Unlike the Gaussian case, however, the center (i.e., μ) and shape (i.e., P_r) of the hyper-ellipsoidal region vary with the number of missed transmissions r.

4.3 Estimation Error Recursion

In a manner similar to the Gaussian source data case, the expression given in (37) can be transformed using (38) to express the optimal estimation error as in (9). Thus, following a procedure identical to that given in Sect. 3, the estimation error recursion formula for the Gauss–Markov case can be obtained as follows.

4.3.1 The Case of *n* Even

$$\begin{aligned} e^{*}_{(r,s,t)} = {} & e^{*}_{(1,s-1,t-1)} - \left(e^{*}_{(1,s-1,t-1)} - e^{*}_{(r+1,s,t-1)}\right) \\ & \times \left(1 - \left\{\sum_{\substack{j=2 \\ j \text{ even}}}^{n} \left\{\frac{\left(\frac{1}{2}\right)^{\frac{j}{2}-1}}{\left(\frac{j}{2}-1\right)!}\right\} \gamma^{\frac{j}{2}-1}\right\} \exp\left\{-\frac{1}{2}\gamma\right\}\right) \\ & + \mathrm{tr}\left(QP_r\right)\left(1 - \left\{\sum_{\substack{j=2 \\ j \text{ even}}}^{n+2} \left\{\frac{\left(\frac{1}{2}\right)^{\frac{j}{2}-1}}{\left(\frac{j}{2}-1\right)!}\right\} \gamma^{\frac{j}{2}-1}\right\} \exp\left\{-\frac{1}{2}\gamma\right\}\right). \end{aligned} \tag{42}$$

4.3.2 The Case of *n* Odd

$$
\begin{aligned}
e^*_{(r,s,t)} &= e^*_{(1,s-1,t-1)} - \left(e^*_{(1,s-1,t-1)} - e^*_{(r+1,s,t-1)}\right) \\
&\quad \times \left(\left[2\Phi\left(\sqrt{\gamma}\right) - 1\right] - \left\{ \sqrt{\frac{2}{\pi}} \sum_{\substack{j=3 \\ j \text{ odd}}}^{n} \left\{ \frac{2^{\frac{j-3}{2}} \left(\frac{j-3}{2}\right)!}{(j-2)!} \right\} \gamma^{\frac{j-2}{2}} \right\} \exp\left\{-\frac{1}{2}\gamma\right\} \right) \\
&\quad + \operatorname{tr}\left(QP_r\right) \left(\left[2\Phi\left(\sqrt{\gamma}\right) - 1\right] - \left\{ \sqrt{\frac{2}{\pi}} \sum_{\substack{j=3 \\ j \text{ odd}}}^{n+2} \left\{ \frac{2^{\frac{j-3}{2}} \left(\frac{j-3}{2}\right)!}{(j-2)!} \right\} \gamma^{\frac{j-2}{2}} \right\} \right. \\
&\qquad \left. \exp\left\{-\frac{1}{2}\gamma\right\} \right). \qquad (43)
\end{aligned}
$$

5 Limited Average Transmission Frequency

The optimal estimation strategy presented in the previous sections can be applied to a system that is restricted to a fixed average number of transmissions per time. In this section, we approach this problem using an infinite horizon approach, where the objective is to maintain an average transmission frequency of ω/k [transmissions per time step].

5.1 Optimal Observation Set

Consider an optimal observation set that is independent of (s, t). Since the objective is based on a fixed average transmission frequency, the optimal observation set also remains fixed with size proportional to the average transmission frequency restriction.

For purely Gaussian source data, the optimal observation set can be calculated by setting the integral of the PDF in (4) equal to the desired no transmit frequency value and solving for the integration region. For example, if the average transmission frequency restriction is 25% (in other words, if the communication rate is restricted to an average of 1 transmission per 4 time steps), then a 25% probability of data transmission will be enforced. Over an infinite horizon, this algorithm will result in an average transmission frequency of 25%.

Let ω denote the fraction indicating the transmission frequency (e.g., $\omega = 0.5$ indicates 50% transmission frequency). By integrating the PDF in (4) and setting the result equal to $1 - \omega$ (i.e., the desired NT frequency), the following region is obtained:

$$b^T P^{-1} b = \ln \left\{ \frac{1}{\omega^2} \right\}. \tag{44}$$

By transforming the expression in (44) to spherical coordinates, integration limits can be determined as given in (14), where

$$\gamma = \ln \left\{ \frac{1}{\omega^2} \right\}.$$

An identical procedure as that in Sect. 3.1 can then be followed to calculate the estimation error recursion formula.

6 Simulation Results

Numerical simulations were created to test the performance of the proposed optimal estimation technique for the cases where the source data is generated via a purely Gaussian random process and via a Gauss–Markov process. For each simulation, a lookup table containing the optimal cost-to-go at each instant (s, t) was generated offline. The lookup table is used along with (21) or (22) for the purely Gaussian case, and with (42) and (43) for the Gauss–Markov case to calculate the optimal cost-to-go at each time step. For clarity of presentation, the simulation results presented in this chapter were obtained using 2-D source data; however, the 2-D case effectively serves to illustrate the capability of this estimation technique to estimate incomplete multidimensional data. It is a trivial task to extend the 2-D results to n-D.

6.1 Source Process is Gaussian

For the purely Gaussian simulation, the source data is generated via a zero mean standard Gaussian random process with a PDF defined as:

$$f\,(b) = \frac{1}{2\pi\,|P|^{1/2}} \exp\left\{-b^T P^{-1} b\right\}, \tag{45}$$

where the constant covariance matrix P is given as:

$$P = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}. \tag{46}$$

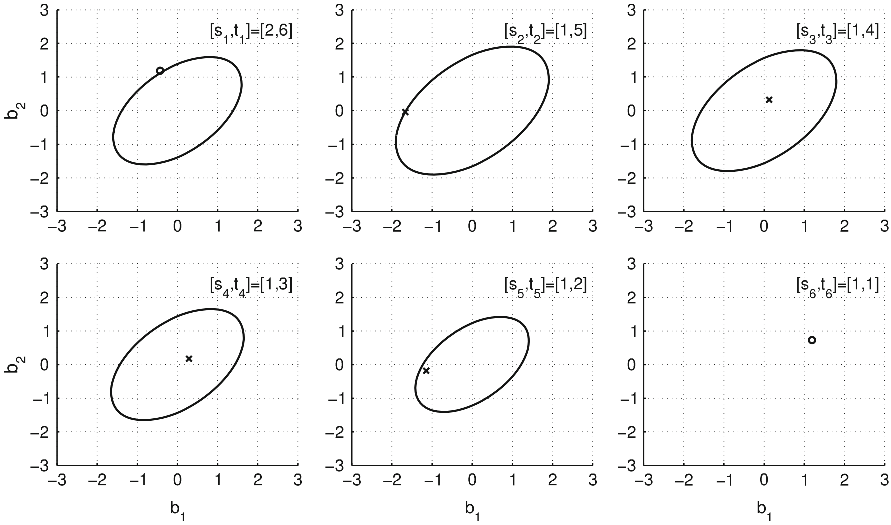


Fig. 2 The region minimizing the estimation error (the elliptical region $\mathcal{J}^{c*}$) and the source data point b_k at each time step for the case where $(M, N) = (2, 6)$

The initial conditions used in the simulation are:

$$e^*_{(t,t)} = 0, \qquad e^*_{(0,t)} = NT, \qquad \forall\, t > 0.$$

Figures 2–4 summarize the results of the numerical simulation for the purely Gaussian source data case. In each of the six plot windows in Figs. 2 and 3, the point indicating the current value of the 2-D vector b_k is denoted as an 'o' or an '×', where 'o' indicates that the data will be transmitted, and '×' indicates that the data will not be transmitted. Toward the top right corner of each of the six plots in Figs. 2 and 3, the $[s_k, t_k]$ value indicates the value of s and t before the decision to transmit or not is made at that time step. Figure 2 shows the optimal 2-D elliptical region $\mathcal{J}^{c*}$ (i.e., see (11)) at each time step for the case where there are $N = 6$ time steps in the decision horizon and $M = 2$ transmission opportunities. The final plot for $[s_6, t_6] = [1, 1]$ shows no ellipse because the optimal observation set includes the entire range space (indicating that the decision is to transmit regardless of the value of b_k).

Figure 3 shows the optimal 2-D region $\mathcal{J}^{c*}$ at each time step for the case where $N = 6$ and $M = 4$. Since there are $4 > 2$ transmission opportunities in this case, $\mathcal{J}^{c*}$ is significantly smaller for the case shown in Fig. 3 in comparison to the case in Fig. 2. The reduced size of the optimal region $\mathcal{J}^{c*}$ is equivalent to an increase in the size of the optimal observation set $\mathcal{J}^*$. The fact that $\mathcal{J}^*$ increases with the number of transmission opportunities M agrees with the heuristic notion that transmitting as much data as possible should reduce the overall estimation error.

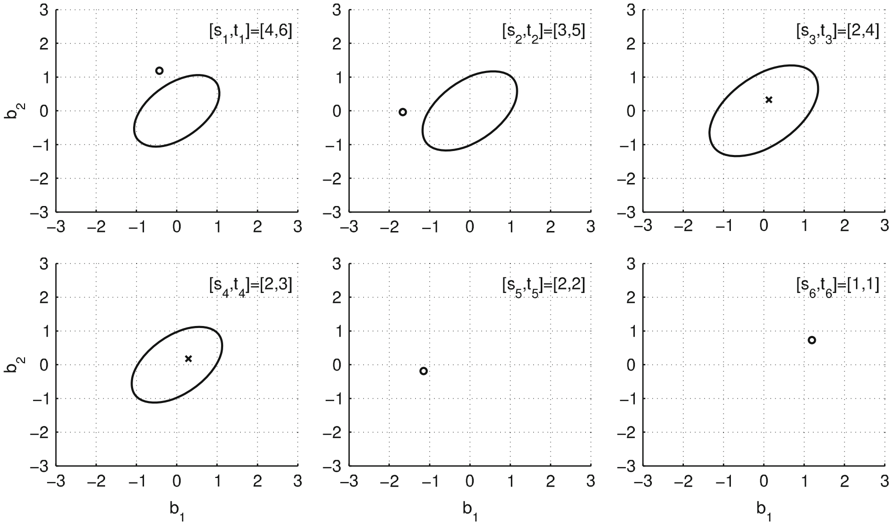


Fig. 3 The region minimizing the estimation error (the elliptical region $\mathcal{J}^{c*}$) and the source data point b_k at each time step for the case where $(M, N) = (4, 6)$

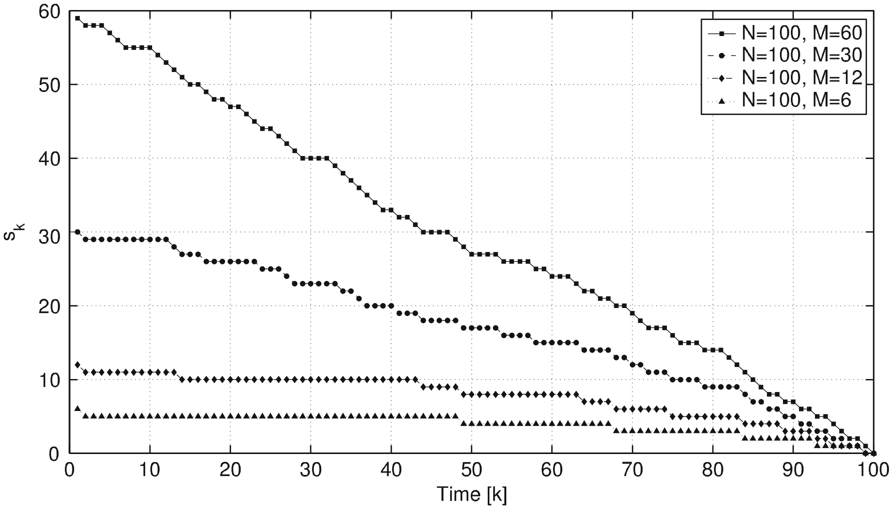


Fig. 4 Number of transmission slots remaining s_k versus time step k for $N = 100$

In Fig. 4, the transmission patterns are shown for four different 2-D scenarios (i.e., for $M = 60$, 30, 12, and 6) over a decision horizon of $N = 100$. Figure 4 illustrates that this optimal estimation algorithm naturally follows transmission patterns to optimize the cost-to-go. This can be understood by noting that the pattern approaches a straight line as $M \to N$ (if $M = N$, the decision would always be to transmit, and the pattern would be a line of slope $\frac{\Delta s_k}{\Delta k} = -1$).

6.2 Source Process is Gauss–Markov

For the 2-D Gauss–Markov case, the source data is generated via (25), where the process matrix $A \in \mathbb{R}^{2\times 2}$ is defined as:

$$A \triangleq \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & 0.9 \end{bmatrix}$$

and $\{w_k\}$ is an I.I.D. Gaussian process with zero mean. The PDF of $b \in \mathbb{R}^2$ can be expressed as:

$$f(b) = \frac{1}{2\pi \left|P_r\right|^{1/2}} \exp\left\{-b^T P_r^{-1} b\right\}, \tag{47}$$

where the r-dependent covariance matrix $P_r \in \mathbb{R}^{2\times 2}$ is given as:

$$P_r = \begin{bmatrix} \sigma_{1r}^2 & \sigma_{1r}\sigma_{2r} \\ \sigma_{1r}\sigma_{2r} & \sigma_{2r}^2 \end{bmatrix}. \tag{48}$$

The elements of the covariance matrix in (48), are defined as:

$$\sigma_{ir}\sigma_{jr} = \sum_{q=1}^{r} \left(\sum_{m=1}^{2} A_{i,m}^{q-1} \sigma_m \sum_{p=1}^{2} A_{j,p}^{q-1} \sigma_p \right) \tag{49}$$

$\forall\ i, j = 1, 2$, where σ_m and σ_p $\forall\ m, p = 1, 2$ are elements of the covariance matrix P_r for $r = 1$, which is given in (46). The initial conditions used in the Gauss–Markov simulation are:

$$e^*_{(r,t,t)} = 0,$$
$$e^*_{(r,0,t)} = \sum_{m=r}^{r+t-1} \sum_{q=1}^{m} \left(\sum_{k=1}^{2} A_{i,k}^{q-1} \sigma_k \sum_{p=1}^{2} A_{j,p}^{q-1} \sigma_p \right)$$

$\forall\ i, j = 1, 2, \forall\ t > 0$.

Figure 5 shows the results from the simulation in the Gauss–Markov case. The dependence of the mean and covariance on the number r of missed transmissions results in the center and shape of the ellipses as shown in Fig. 5 to vary with the time step k. Specifically, the means at each time step $1 \leq k \leq 6$ are (see (27))

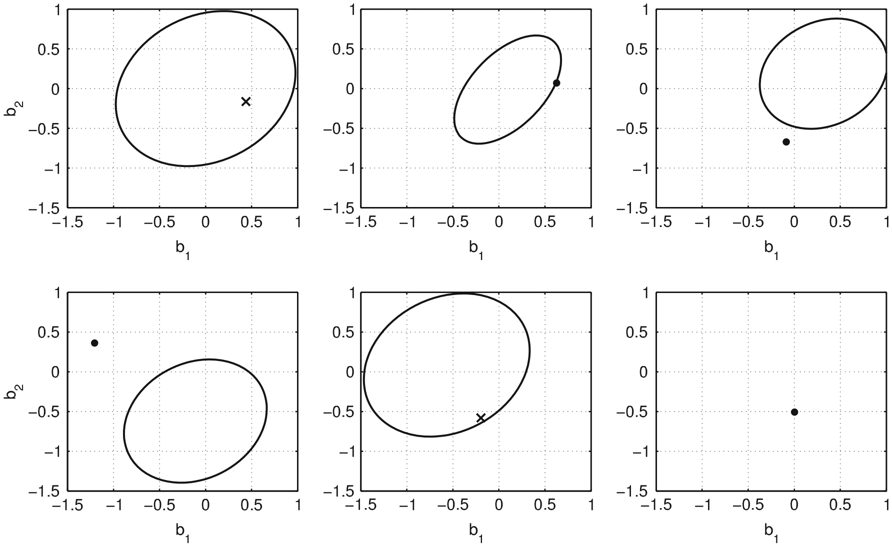


Fig. 5 The region (the ellipse) minimizing the estimation error and the source data point b_k ('o' or '×') at each time step for the case Gauss–Markov case with $(M, N) = (4, 6)$

$E\left[b_k \mid b_{(k-r)}\right] = \begin{bmatrix} 0 & 0.10 & 0.32 & -0.11 & -0.57 & -0.27 \\ 0 & -0.01 & 0.19 & -0.62 & 0.09 & -0.04 \end{bmatrix}$, and the covariance matrices for $r = 1, 2$ are $P_1 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$ and $P_2 = \begin{bmatrix} 0.30 & 0.19 \\ 0.19 & 0.41 \end{bmatrix}$, respectively. The value of r is 2 for time steps $k = 2$ and $k = 6$ since there were no transmissions for $k = 1$ or $k = 5$. The simulation results clearly show that the size of the optimal observation set increases with the number of missed transmissions r. This agrees with the heuristic idea that a greater number of transmissions would be required to reduce the overall estimation error in the presence of increased uncertainty.

6.3 Benefit of Cost-Weighting

To test the effect of implementing a cost-weighting matrix, many trials were carried out to compare the performance of the VOI-based estimation algorithm with and without cost-weighting in the observer policy. Table 1 shows the results from both of these cases over 100, 500, 1,000, and 5,000 trials. In all trials, $(N, M) = (6, 4)$, the covariance matrix P for the bivariate Gaussian randomly generated source data was selected as:

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix} \tag{50}$$

Table 1 Effect of cost-weighting on average estimation error

# of Trials	e^{π} Measure	μ_k Policy	μ_k^Q Policy
100	Mean	0.7593	0.5770
	Median	0.7068	0.3181
500	Mean	0.7899	0.7656
	Median	0.7249	0.3673
1000	Mean	0.7648	0.7310
	Median	0.7104	0.3582
5000	Mean	0.7756	0.7542
	Median	0.7057	0.3602

and the cost-weighting matrix Q was selected as:

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}. \tag{51}$$

The encoder policy for the case where cost-weighting is not used can be expressed as:

$$\mu_k = \begin{cases} b_k & \text{if } b_k \in \mathcal{J}_{(s_k,t_k)}, \\ \text{NT} & \text{if } b_k \in \mathcal{J}^c_{(s_k,t_k)}, \end{cases}$$

where the optimal observation set $\mathcal{J}_{(s_k,t_k)}$ is defined as:

$$\mathcal{J}_{(s_k,t_k)} = \left\{ b_k \mid b_k^T P^{-1} b_k \leq \gamma \right\}. \tag{52}$$

The encoder policy for the case where cost-weighting is used can be expressed as:

$$\mu_k^Q = \begin{cases} b_k & \text{if } b_k \in \mathcal{J}_{Q(s_k,t_k)}, \\ \text{NT} & \text{if } b_k \in \mathcal{J}^c_{Q(s_k,t_k)}, \end{cases}$$

where the optimal observation set $\mathcal{J}_{Q(s_k,t_k)}$ is defined as:

$$\mathcal{J}_{Q(s_k,t_k)} = \left\{ b_k \mid b_k^T G^{-1} b_k \leq \gamma \right\}, \tag{53}$$

where $G \triangleq Q^{1/2} P Q^{1/2}$, and γ is defined in (15).

From the data in Table 1, the inclusion of the cost-weighting clearly results in a reduction in the average estimation error e^{π}. In all cases the mean of the average estimation error (i.e., the mean of e^{π}) over all of the trials is slightly less when the μ_k^Q observer policy is utilized. Moreover, the median of e^{π} over all trials is significantly reduced when the μ_k^Q observer policy is used, being reduced by approximately 55% in the most extreme case.

Figures 6 and 7 show the effect on the transmission patterns for a single trial of the $(M, N) = (4, 6)$ case. Heuristically, these figures show that the overall effect

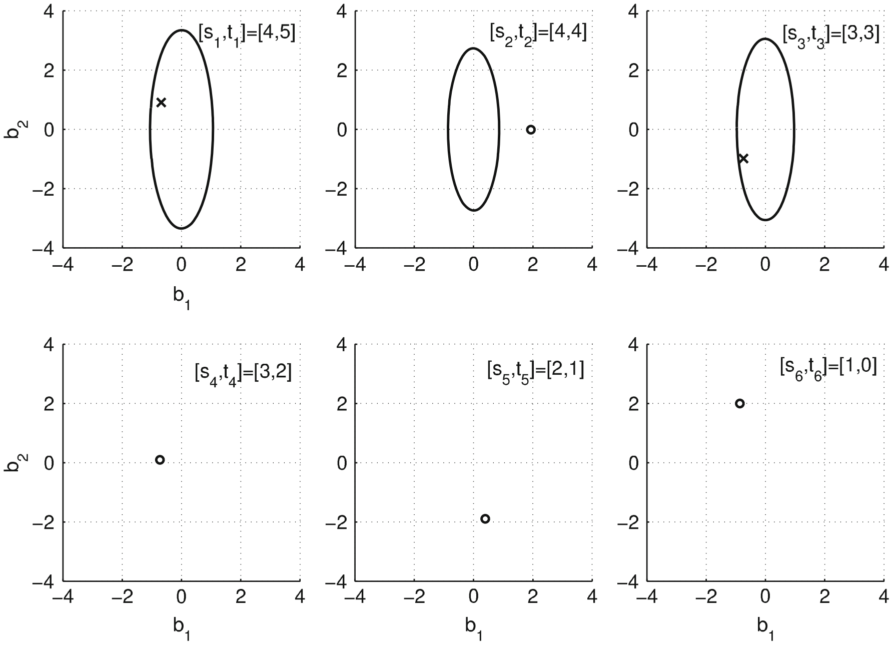


Fig. 6 Transmission pattern for the $(M, N) = (4, 6)$ case using the μ_k observer policy (i.e., without cost-weighting)

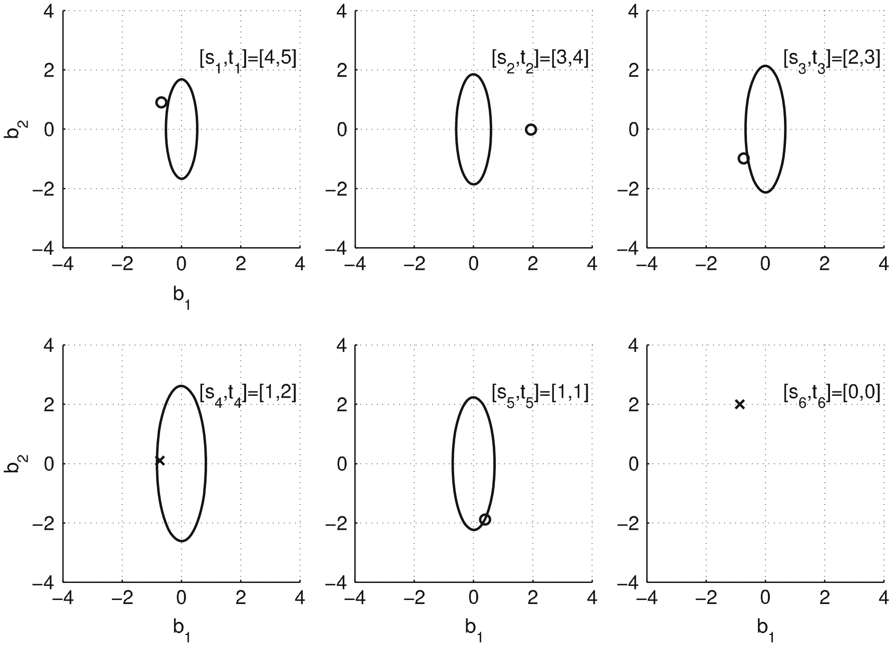


Fig. 7 Transmission pattern for the $(M, N) = (4, 6)$ case using the μ_k^Q observer policy (i.e., with cost-weighting)

of including the cost-weighting matrix Q in the observer policy is to increase the overall size of the optimal observation set, resulting in an alteration of the overall transmission pattern.

7 Conclusion

Optimal encoder policies are developed for a sensor that takes multidimensional measurements of some true state process. The criterion function to be minimized consists of a weighted cumulative quadratic estimation error under the constraint that over a time-window of length, N, only $M < N$ measurements can be transmitted to a remote estimator. We show how the optimal policy is to transmit a measurement only if it lies outside of a certain hyper-ellipsoid, whose shape and size depend on the number of remaining channel-uses, the number of remaining time-steps, and the statistics of the measurement process. We show how dynamic programming can be used to find the optimal hyper-ellipsoid in a Gaussian scenario, as well as in a Gauss–Markov process, where the sensor is subject to Gaussian noise.

We also incorporate an arbitrary (positive-definite) weighting matrix that allows the user to specify which elements of the state vector are most valuable to estimate accurately. This is a first step in extending this research to address the larger issues of informational value and relevance when this problem is set in a closed-loop context (i.e., when the estimate is used by a control system to generate an input signal that alters the trajectory of the state that is being measured). In previous (scalar-valued) work, there was no place for consideration of which dimensions of the sensor data might be most useful to control the system. It is likely that some form of the separation principle holds in the scalar case, and that the encoder–estimator policies are invariant regardless of how the estimate is used in the control system. In the multidimensional case we consider, however, it is not clear that the encoder–estimator polices can safely ignore the nature of the control law since some dimensions of estimation error might be much more critical for successful system control than others.

References

1. M. S. Andersland. On the optimality of open-loop LQG measurement scheduling. *IEEE Transactions on Automatic Control*, 40(10):1796–1799, 1995.
2. J. Baillieul and P. J. Antsaklis. Control and communication challenges in networked real-time systems. *Proceedings of the IEEE*, 95(1):9–28, Jan. 2007.
3. D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific, 1995.
4. E. Biglieri and M. Elia. Multidimensional modulation and coding for band-limited digital channels. *IEEE Transactions on Information Theory*, 34(4):803–809, July 1988.

5. J. H. Braslavsky, R. H. Middleton, and J. S. Freudenberg. Feedback stabilization over signal-to-noise ratio constrained channels. *IEEE Trans. Automat. Contr.*, 52(8):1391–1403, Aug. 2007.
6. C. De Persis. n-Bit stabilization of n-dimensional nonlinear systems in feedforward form. *IEEE Trans. Automat. Contr.*, 50(3):299–311, Mar. 2005.
7. N. Elia and J. N. Eisenbeis. Limitations of linear remote control over packet drop networks. In *IEEE Conference on Decision and Control*, 5152–5157, Atlantis, Paradise Island, Bahamas, Dec. 2004.
8. X. Feng and K. A. Loparo. Active probing for information in control systems with quantized state measurements: A minimum entropy approach. *IEEE Trans. Automat. Contr.*, 42(2):216–238, 1997.
9. N. Gans and J. W. Curtis. A performance bound for decentralized moving horizon estimation. In *Dynamics of Information Systems: Theory and Applications*, Gainesville, FL, Jan. 2009.
10. V. Gupta, T. H. Chung, B. Hassibi, and R. M. Murray. On a stochastic sensor selection algorithm with applications in sensor scheduling and sensor coverage. *Automatica*, 42:251–260, 2006.
11. J. P. Hespanha, P. Naghshtabrizi, and Y. Xu. A survey of recent results in networked control systems. *Proceedings of the IEEE*, 95(1):138–162, Jan. 2007.
12. O. C. Imer. *Optimal estimation and control under communication network constraints*. Ph.d. thesis, University of Illinois at Urbana-Champagne, 2005.
13. O. C. Imer and T. Basar. Optimal estimation with limited measurements. *International Journal of Systems, Control and Communications*, 2(1–2):5–29, 2010.
14. S. Kittipiyakul, P. Elia, and T. Javidi. High-SNR analysis of outage-limited communications with bursty and delay-limited information. *IEEE Transactions on Information Theory*, 55(2):746–763, Feb. 2009.
15. D. Liberzon and J. P. Hespanha. Stabilization of nonlinear systems with limited information feedback. *IEEE Trans. Automat. Contr.*, 50(6):910–915, Jun. 2005.
16. G. M. Lipsa and N. C. Martins. *Certifying the optimality of a distributed state estimation system via majorization theory.* Technical report, Institute for Systems Research Technical Reports, 2009.
17. X. Liu and A. Goldsmith. Kalman filtering with partial observation losses. In *IEEE Conference on Decision and Control*, 4180–4186, Atlantis, Paradise Island, Bahamas, Dec. 2004.
18. L. Meier, J. Peschon, and R. Dressler. Optimal control of measurement subsystems. *IEEE Transactions on Automatic Control*, 12:528–536, 1967.
19. G. N. Nair and R. J. Evans. Stabilization with data-rate-limited feedback: Tightest attainable bounds. *Systems and Control Letters*, 41:49–56, 2000.
20. G. N. Nair, R. J. Evans, I. M. Y. Mareels, and W. Moran. Topological feedback entropy and nonlinear stabilization. *IEEE Trans. Automat. Contr.*, 49(9):1585–1597, Sept. 2004.
21. G. N. Nair, F. Fagnani, S. Zampieri, and R. J. Evans. Feedback control under data rate constraints: An overview. *Proceedings of the IEEE*, 95(1):108–137, Jan. 2007.
22. I. R. Petersen and A. V. Savkin. Multi-rate stabilization of multivariable discrete-time linear systems via a limited capacity communication channel. In *IEEE Conference on Decision and Control*, 304–309, Orlando, FL, Dec. 2001.
23. B. Sinopoli, L. Schenato, M. Francescetti, K. Poolla, M. I. Jordan, and S. S. Sastry. Kalman filtering with intermittent observations. *IEEE Trans. Automat. Contr.*, 49(9):1453–1464, Sept. 2004.
24. S. C. Smith and P. Seiler. Estimation with lossy measurements: Jump estimators for jump systems. *IEEE Trans. Automat. Contr.*, 48(12):2163–2171, Dec. 2003.
25. S. Tatikonda and S. Mitter. Control under communication constraints. *IEEE Trans. Automat. Contr.*, 49(7):1056–1068, July 2004.
26. R. Touri and C. N. Hadjicostis. Stabilisation with feedback control utilising packet-dropping network links. *IET Control Theory Appl.*, 1(1):334–342, Jan. 2007.
27. Y. Q. Wang, H. Ye, and G. Z. Wang. Fault detection of NCS based on eigendecomposition, adaptive evaluation and adaptive threshold. *Int'l Journal of Control*, 80(12):1903–1911, 2007.

28. W. S. Wong and R. W. Brockett. Systems with finite communication bandwidth constraints–part I: State estimation problems. *IEEE Trans. Automat. Contr.*, 42(9):1294–1299, Sept. 1997.
29. W. S. Wong and R. W. Brockett. Systems with finite communication bandwidth constraints–II: Stabilization with limited information feedback. *IEEE Trans. Automat. Contr.*, 44(5): 1049–1053, May 1999.
30. H. Yue and H. Want. Minimum entropy control of closed-loop tracking errors for dynamic stochastic systems. *IEEE Trans. Automat. Contr.*, 48:118–122, 2003.

Information Patterns in Discrete-Time Linear-Quadratic Dynamic Games*

Meir Pachter and Khanh D. Pham

Abstract Information – who knows what, when – plays a critical role in game theory, and, in particular, in dynamic games. Thus, dynamic game theory is an ideal vehicle for exploring the interplay of dynamics and information. We confine our attention to discrete-time Linear-Quadratic Dynamic Games (LQDGs) which have the distinct advantage of readily being amenable to analysis without having to overcome conceptual and technical difficulties, closed-form results are possible, and one is in tune with modern digital signal processing techniques. In this chapter a hierarchy of discrete-time LQDGs are characterized by a sequence of information patterns which increase in complexity is analyzed and an insight into the Dynamics of Information Systems is obtained.

Keywords Dynamic Games • Information pattern • Linear-Quadratic Control

1 Introduction

Information – who knows what, when – plays a critical role in game theory and as such, game theory is an ideal vehicle for exploring the interplay of dynamics and information. In static matrix games, the complete lack of information on

[*] The views expressed in this chapter are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or US Government.

M. Pachter (✉)
Air Force Institute of Technology, Department of Electrical and Computer Engineering,
Wright-Patterson AFB, OH 45433, USA
e-mail: meir.pachter@afit.edu

K.D. Pham
Air Force Research Laboratory, Space Vehicles Directorate, Kirtland AFB, NM 87117, USA
e-mail: khanh.pham@kirtland.af.mil

V.L. Boginski et al. (eds.), *Sensors: Theory, Algorithms, and Applications*,
Springer Optimization and Its Applications 61, DOI 10.1007/978-0-387-88619-0_5,

the adversary's action immediately leads to randomization, or, in the parlance of game theory, mixed strategies. In dynamic games, namely, games that evolve over time, information plays a critical role. In this chapter we consider a hierarchy of Linear-Quadratic Dynamic Games (LQDGs). To fix ideas, the discussion centers on disturbance rejection in control systems and an active noise cancelation/disturbance rejection scenario [1] using digital signal processing is considered. The control system is excited by a disturbance signal: the control signal is u and the disturbance signal is v; the rejection of the disturbance v is addressed. Control is employed to actively counteract the effects of the disturbance on the system's output, and therefore a dynamic game formulation of the problem on hand is called for. The goal of disturbance rejection is to be achieved using state feedback. At the same time, the energy of the feedback control signal is applied for the purpose of disturbance rejection which should remain within reasonable bounds. Hence, a Linear-Quadratic Dynamic Game (LQDG) formulation is employed.

Continuous-time LQDGs, namely, linear-quadratic differential games, are perhaps the best understood dynamic games and have been extensively researched – we refer the reader to the monograph [2] and the recent book [3] published. Since we are interested in digital signal processing, we confine our attention to discrete-time LQDGs. In discrete-time the dynamic programming equation contains product terms between the decision variables, which complicates the solution compared to the continuous-time analogue. In this chapter the deterministic, finite-dimensional, *discrete-time*, zero-sum, LQDG is carefully analyzed. Minimal necessary and sufficient conditions for the existence of a solution to the *discrete-time* LQDG and its complete closed-form solution are provided. Furthermore, the theory is expanded to also include the presence of a random disturbance.

The disturbance rejection paradigm serves as a vehicle for investigating the impact of information on the solution of dynamic games. In this chapter a hierarchy of deterministic and stochastic LQDGs characterized by information patterns which increase in complexity is analyzed and an insight into the dynamics of information systems is obtained.

1.1 Discrete-Time LQDG

The dynamics are linear:

$$x_{k+1} = Ax_k + Bu_k + Cv_k, \quad x_0 \equiv x_0, \quad k = 0, 1, \ldots, N-1. \tag{1}$$

The state $x_k \in R^n$, the control of the minimizing player $u \in R^{m_u}$ and the control of the maximizing player $v \in R^{m_v}$; the dynamics matrix A is a $n \times n$ matrix and the minimizing and maximizing players' input matrices B and C are $n \times m_u$ and $n \times m_v$, respectively. The planning horizon is N. An infinite planning horizon is also considered.

The payoff functional is quadratic:

$$J\left(\{u_k\}_{k=0}^{N-1}, \{v_k\}_{k=0}^{N-1}; x_0\right) = x_N^T H x_N + \sum_{k=0}^{N-1} \left(x_{k+1}^T Q x_{k+1} + u_k^T R_u u_k - v_k^T R_v v_k\right), \tag{2}$$

where H and Q are real symmetric $n \times n$ matrices and the minimizing and maximizing players' control efforts weighting matrices R_u and R_v are typically real symmetric and positive definite $m_u \times m_u$ and $m_v \times m_v$ matrices, respectively. It is oftentimes stipulated that also the state penalty matrices H and Q be positive definite, or, at least, positive semi-definite; we shall relax these assumptions as much as possible.

The disturbance signal is v and the controller-generated signal u aims to minimize its effect on the system's output, say $y = \sqrt{Q}x$, while at the same time keeping the control energy expenditure small. The goal of the disturbance v is diametrically opposed. Hence, in our formulation the player/controller whose control variable is u_k strives to minimize the payoff functional J, the player/disturbance whose control variable is v_k strives to maximize the payoff functional J, and a zero-sum dynamic game is solved.

Concerning the information pattern: Both players are cognizant of the problem's parameters, namely, the dynamics (A, B, C), the payoff functional's parameters (H, Q, R_u, R_v), the planning horizon N, and, during the game, at time k, $k = 0, 1, \ldots, N-1$, both players have access to the system's state x_k.

In this chapter, necessary and sufficient conditions for the existence of a solution, closed form solution and explicit formulae for the optimal strategies and the payoff of the deterministic *discrete-time* LQDG, and LQDGs where random disturbances are present, are provided. The optimal strategies for both players, the controller, and the deleterious disturbance, are explicitly derived; of course, we are mostly interested in the minimizing player's/controller action, which is conducive to minimizing the cost functional, and thus, rejection of the disturbance. Furthermore, minimal conditions for the existence of optimal solutions are provided.

The chapter is organized as follows. The main result, that is, the explicit, closed-form, solution of the deterministic discrete-time LQDG (1)–(2) is given in Sect. 2. The reader is referred to [4] where the proof of the main result is provided. Building on the results of Sect. 2, the simplest possible stochastic LQDGs are addressed in Sect. 3. Novel stochastic LQDGs where the players have access to Nature's input and the players' strategies are state feedback–stroboscopic strategies, including a stochastic LQDG with an asymmetric information pattern, are analyzed. In Sect. 4 an illustrative example of a disturbance rejection scenario, where the discrete-time LQDG paradigm is invoked, is analyzed and some fine points germane to dynamic games with an infinite planning horizon are discussed. Concluding remarks are made in Sect. 5.

2 Main Result

A rigorous approach to the disturbance rejection problem entails its formulation as a dynamic game. Thus, the method of Dynamic Programming (DP) is invoked. It is conducive to the rigorous derivation of necessary and sufficient conditions for the existence of a solution to the dynamic game (1) and (2), and the optimal disturbance rejection strategy can be derived. It does however turn out that the solution of the *discrete-time* linear-quadratic dynamic game is more involved than its continuous time/differential game analogue. At the same time, digital signal processing entails discrete-time and real-time operation, and therefore it is imperative to obtain an explicit, closed form, solution of the LQDG. The main result, whose proof is given in [4], is as follows.

Theorem 1. *A necessary and sufficient condition for the existence of a solution to the discrete-time zero-sum LQDG with dynamics (1), cost functional (2), and complete state information is:*

$$R_u + B^T P_k B > 0 \tag{3}$$

and

$$R_v > C^T P_k C \tag{4}$$

$\forall\ k = 1, \ldots, N-1$, *where the real, symmetric, matrices* P_k *are the solution of the difference equation, where*

$$\begin{aligned} P_{k+1} = A^T \Big\{ & P_k - P_k \Big[B S_B^{-1}(P_k) B^T + B S_B^{-1}(P_k) B^T P_k C (R_v - C^T P_k C)^{-1} C^T \\ & + C(R_v - C^T P_k C)^{-1} C^T P_k B S_B^{-1}(P_k) B^T \\ & + C(R_v - C^T P_k C)^{-1} C^T P_k B S_B^{-1}(P_k) B^T P_k C (R_v - C^T P_k C)^{-1} C^T \\ & + C(C^T P_k C - R_v)^{-1} C^T \Big] P_k \Big\} A + Q, \\ P_0 = H + Q,\ & k = 0, \ldots, N-1. \end{aligned} \tag{5}$$

In (5), the matrix function

$$S_B(P_k) \equiv B^T P_k B + R_u + B^T P_k C (R_v - C^T P_k C)^{-1} C^T P_k B.$$

In addition, the problem parameters must satisfy the conditions:

$$R_u + B^T (Q + H) B > 0 \tag{6}$$

and

$$R_v > C^T (Q + H) C. \tag{7}$$

The value of the LQDG is:

$$V_0(x_0) = x_0^T(P_N - Q)x_0 \tag{8}$$

and the players' optimal strategies are the linear state feedback control laws

$$\begin{aligned} u_k^*(x_k) &= -S_B^{-1}(P_{N-k-1})B^T \\ &\quad \times \Big[I + P_{N-k-1}C(R_v - C^T P_{N-k-1}C)^{-1}C^T \Big] P_{N-k-1}A \cdot x_k \end{aligned} \tag{9}$$

and

$$\begin{aligned} v_k^*(x_k) &= (R_v - C^T P_{N-k-1}C)^{-1}C^T \Big\{ I - P_{N-k-1}BS_B^{-1}(P_{N-k-1})B^T \\ &\quad \times \Big[I + P_{N-k-1}C\left(R_v - C^T P_{N-k-1}C\right)^{-1}C^T \Big] \Big\} P_{N-k-1}A \cdot x_k. \end{aligned} \tag{10}$$

2.1 Discussion

In discrete-time the dynamic programming equation contains product terms between the control variables, which complicated the solution compared to the continuous-time analogue. Using the Schur complement concept [5] and introducing the matrix function $S_B(\cdot)$ allowed us to obtain the explicit, closed form, solution of the LQDG (1)–(2).

The Riccati equation lies at the heart of Linear-Quadratic control. To put things in perspective, we note that the continuous-time analogue of the difference equation (5) is the much simpler Differential Riccati Equation (DRE) from the classical theory of continuous-time zero-sum LQDGs/differential games,

$$\begin{aligned} \dot{P} &= A^T P + PA - P(B^T R_u^{-1} B - C^T R_v^{-1} C)P + Q, \\ P(0) &= H,\ 0 \le t \le T. \end{aligned}$$

In conformity with the well developed theory of continuous-time LQDGs/ differential games [2, 3], the difference equation (5) will be referred to as the Difference Riccati Equation (DRE). One can appreciate the complexity of the recursion (5), brought about by discrete-time action, and yet, the solution of the matrix DRE (5) is a crucial step towards obtaining the solution of the discrete-time LQDG. Note, however, that similar to the continuous-time case, for the calculation of S_B^{-1} and the propagation of P_k in discrete-time, the computation of only a $r_u \times r_u$ and a $r_v \times r_v$ matrix inverse is required and in the important special case of single input systems, this boils down to a scalar division. Furthermore, it is important to recognize that the DRE (5) need not be solved in real-time. The recursion (5) can be solved ahead of time and the matrices P_k can be stored for real time control action according to the optimal disturbance rejection control law (9).

Note that, contrary to the continuous-time case, conditions (3) and (4) for the existence of a solution do not automatically mandate that the players' real and symmetric control effort weighing matrices R_u and R_v be positive definite. Whereas in continuous-time the players' control effort weights must satisfy $R_u \geq 0$, $R_v \geq 0$ and for nonsingular optimal control the conditions are $R_u > 0$, $R_v > 0$, these are not necessary conditions in discrete-time.

We are interested in disturbance rejection and we are therefore exclusively interested in the minimizing player's action u^*, which is given in (9). In the parlance of game theory, it is a pure strategy: a linear state feedback control law is employed. The strategy (9) is a Nash strategy and should the minimizing player unilaterally deviate from this strategy he will be penalized by incurring a higher cost. Furthermore, the value function $V_0(x_0)$ yields the optimal cost, as a function of the initial state x_0: since a zero-sum game is solved and a saddle point is obtained, the Nash strategy is also a *security* strategy and the value function (8) provides a *guarantee* to the minimizing player, that is, as long as the minimizing player sticks with the strategy (9), the actually realized cost will be at most as high as the value $V_0(x_0)$ of the game, and, most likely, if the maximizing player, a.k.a., the disturbance v, deviates from its optimal play, it might actually be lower. This will indeed be the case – the disturbance is a disadvantage because, unlike the controller, in reality it does not have access to the system's state:

$$J\left(\{u_k^*(x_k)\}_{k=0}^{N-1}, \{v_k\}_{k=0}^{N-1}; x_0\right) \leq V_0(x_0) = x_0^T(P_N - Q)x_0$$
$$\forall\ v_k,\ k = 0, \ldots, N-1.$$

The discrete-time Riccati equation (5) can be written as:

$$\begin{aligned} P_{k+1} = A^T\Big\{&P_k - P_k\Big[I + P_kC\Big(R_v - C^TP_kC\Big)^{-1}C^T\Big]^T BS_B^{-1}(P_k)B^T \\ &\times \Big[I + P_kC\Big(R_v - C^TP_kC\Big)^{-1}C^T\Big]P_k \\ &+ P_kC\Big(R_v - C^TP_kC\Big)^{-1}C^TP_k\Big\}A + Q, \\ P_0 = H + Q,\ \ &k = 0, \ldots, N-1 \end{aligned}$$

and since $S_B > 0$, the following inequality holds.

$$P_{k+1} \leq A^TP_kA + A^TP_kC(R_v - C^TP_kC)^{-1}C^TP_kA + Q,\ \ P_0 = H + Q,$$
$$k = 0, \ldots, N-1.$$

If, in addition, $\forall\ k = 0, \ldots, N-1$ the matrices P_k are invertible, an application of the Matrix Inversion Lemma [6] renders the above inequality in the compact form:

$$P_{k+1} \le A^T (P_k^{-1} - CR_v^{-1}C^T)^{-1} A + Q,\ \ P_0 = H + Q,\ \ k = 0, \ldots, N-1.$$

Let Π_k be the solution of the difference equation:

$$\Pi_{k+1} = A^T (\Pi_k^{-1} - CR_v^{-1}C^T)^{-1} A + Q,\ \ \Pi_0 = H + Q,\ \ k = 0, \ldots, N-1.$$

We prove by induction that

$$P_k \le \Pi_k,\ k = 0, \ldots, N.$$

Thus, $P_0 = \Pi_0$; we assume that $P_k \le \Pi_k$ and we show that this implies $P_{k+1} \le \Pi_{k+1}$. Indeed, Proposition 8.8.5 from [9] yields $P_k \le \Pi_k \to P_k^{-1} \ge \Pi_k^{-1}$. Therefore $P_k^{-1} - CR_v^{-1}C^T \ge \Pi_k^{-1} - CR_v^{-1}C^T$ and reapplying Proposition 8.8.5 from [9] gives $(P_k^{-1} - CR_v^{-1}C^T)^{-1} \le (\Pi_k^{-1} - CR_v^{-1}C^T)^{-1}$. Hence

$$\begin{aligned} \Pi_{k+1} &= A^T \left(\Pi_k^{-1} - CR_v^{-1}C^T\right)^{-1} A + Q \\ &\ge A^T \left(P_k^{-1} - CR_v^{-1}C^T\right)^{-1} A + Q \\ &\ge P_{k+1}. \end{aligned}$$

The result that P_k is bounded by the solution Π_k of the difference equation is somewhat similar to the Gronwall–Bellman inequality.

3 Three Stochastic LQDGs with Complete State Observation

So far, the deterministic LQDG was considered. The case where the dynamics are perturbed by Gaussian process noise is now addressed – we refer to Linear-Quadratic Gaussian Dynamic Games (LQGDGs). The LQGDGs' dynamics are:

$$x_{k+1} = Ax_k + Bu_k + Cv_k + \Gamma w_k,\ \ x_0 \equiv x_0,\ \ k = 0, 1, \ldots, N-1. \quad (11)$$

The process noise input matrix Γ is a $n \times m_w$ matrix and the process noise

$$w_k \sim N(0, W),$$

where W is a $m_w \times m_w$ real, symmetric, and a positive definite matrix.

Building on the results of Sect. 2, a hierarchy of three stochastic LQGDGs characterized by increasingly complex information patterns is now addressed and their explicit solution is provided in [4].

3.1 Stochastic LQDG with Complete State Observation 1

The payoff is now a random number and therefore the payoff functional is:

$$J\left(\{u_k\}_{k=0}^{N-1}, \{v_k\}_{k=0}^{N-1}; x_0\right)$$
$$= E_w\left[x_N^T H x_N + \sum_{k=0}^{N-1}\left(x_{k+1}^T Q x_{k+1} + u_k^T R_u u_k - v_k^T R_v v_k\right)\right]. \qquad (12)$$

The expectation is taken over the process noise sequence, where $w \equiv \{w_0, \ldots, w_{N-1}\}$.

As before, the players have complete state information. Thus, the simplest LQGDG is solved. We shall refer to LQGDG 1.

The following holds.

Theorem 2. *Consider the stochastic control system (11) and payoff functional (12). The planning horizon is N. Assume that at decision time k the players have access to the state x_k, that is, the players have complete state information. In addition, assume the deterministic LQDG (1)–(2) has a solution for the planning horizon N. The players' optimal strategies in the LQGDG 1 with complete state information are given by the solution of the deterministic LQDG provided by Theorem 1, namely, the state feedback control laws (9) and (10). However, the value function of the zero-sum game (11) and (12), namely, the players' expected payoff, contains an additional term:*

$$V_0^{(1)}(x_0) = x_0^T(P_N - Q)x_0 + p_N^{(1)}.$$

The real symmetric matrices P_k are calculated according to the recursion (5) for the deterministic case, and having obtained the sequence $\{P_k\}_{k=0}^N$, the new scalar sequence $p_k^{(1)}$ is calculated according to

$$p_k^{(1)} = Trace\left(\Gamma^T\left(\sum_{i=0}^{k-1} P_i\right)\Gamma W\right), \quad k = 1, \ldots, N \qquad (13)$$

so that the value function can be obtained. In (5) and (13), $P_k := P_k + Q$.

As in the deterministic case, conditions (3) and (4) are necessary and sufficient for the existence of a solution to the LQGDG 1 on the finite planning horizon N.

3.2 Stochastic LQDG with Complete State Observation 2

A somewhat nonconventional stochastic dynamic game is considered.

The LQGDG (11) and (12) is revisited. As before, the players have complete state information. In addition, at decision time k the players also have access to Nature's random input w_k, $k = 0, \dots, N-1$; we here refer to a stroboscopic information pattern. Stochastic optimal control problems where the realization of the random variable is known at decision time are encountered in a sequential inspection problems [7]. Stochastic Linear-Quadratic Dynamic Games with such an information pattern and where Nature's input is Gaussian will be referred to as LQGDG 2.

The following holds.

Theorem 3. *Consider the stochastic control system (11) and payoff functional (12). The planning horizon is N. Assume that at decision time k the players have access to the state x_k and the random disturbance w_k. In addition, assume the deterministic LQDG (1)–(2) has a solution for the planning horizon N. In the LQGDG 2 the players' optimal strategies are given by the state feedback/stroboscopic control laws:*

$$\begin{aligned} u_k^*(x_k, w_k) = &- S_B^{-1}(P_{N-k-1}) B^T \\ &\times \left[I + P_{N-k-1} C \left(R_v - C^T P_{N-k-1} C \right)^{-1} C^T \right] \\ &\times P_{N-k-1}(A x_k + \Gamma w_k) \end{aligned} \tag{14}$$

and

$$\begin{aligned} v_k^*(x_k, w_k) = &\left(R_v - C^T P_{N-k-1} C \right)^{-1} C^T \Big\{ I - P_{N-k-1} B S_B^{-1}(P_{N-k-1}) B^T \\ &\times \left[I + P_{N-k-1} C \left(R_v - C^T P_{N-k-1} C \right)^{-1} C^T \right] \Big\} \\ &\times P_{N-k-1}(A x_k + \Gamma w_k). \end{aligned} \tag{15}$$

The value function of the zero-sum game (11)–(12) with the stroboscopic information pattern, namely, the players' expected payoff, is:

$$V_0^{(2)}(x_0) = x_0^T (P_N - Q) x_0 + p_N^{(2)}.$$

The real symmetric matrices P_k are calculated according to the recursion (5) for the deterministic LQDG, and having obtained the sequence $\{P_k\}_{k=0}^{N}$, the new scalar sequence $p_k^{(2)}$ is calculated according to

$$
\begin{aligned}
p_{k+1}^{(2)} = p_k^{(2)} + Trace\Big(\Gamma^T \Big\{P_k - P_k\Big[BS_B^{-1}(P_k)B^T \\
&+ BS_B^{-1}(P_k)B^T P_k C\Big(R_v - C^T P_k C\Big)^{-1} C^T \\
&+ C\Big(R_v - C^T P_k C\Big)^{-1} C^T P_k BS_B^{-1}(P_k)B^T \\
&+ C\Big(R_v - C^T P_k C\Big)^{-1} C^T P_k BS_B^{-1}(P_k)B^T P_k C\Big(R_v - C^T P_k C\Big)^{-1} C^T \\
&+ C\Big(C^T P_k C - R_v\Big)^{-1} C^T\Big]P_k\Big\}\Gamma W\Big),\ p_0^{(2)} = 0,\ \ k = 0, \ldots, N-1.
\end{aligned}
\tag{16}
$$

The solution of the recursion (16) entails a straight summation. In the recursions (5) and (16) $P_k := P_k + Q$.

As in the deterministic case, conditions (3) and (4) are necessary and sufficient for the existence of a solution to the LQGDG 2 on the finite planning horizon N.

Concerning the new scalar difference (16): Comparing the matrix expression whose Trace is calculated in (16) and the R.H.S. of (5), we observe that if the dynamics matrix A is nonsingular – as in discrete-time control systems derived from discretized continuous-time control systems – then the difference (16) can be written in the compact form:

$$
p_k^{(2)} = \mathrm{Trace}\left(\Gamma^T (A^{-1})^T \left[\left(\sum_{i=1}^{k} P_i\right) - kQ\right] A^{-1}\Gamma W\right),\ k = 1, \ldots, N. \tag{17}
$$

3.3 *Stochastic LQDG with Complete State Observation 3*

The following stochastic dynamic game with an asymmetric information pattern is now considered.

The LQGDG (11) and (12) is revisited. As before, the players have complete state information. In addition, at decision time k the P player, namely, the minimizing player whose control input is u_k, also has access to Nature's random input w_k, $k = 0, \ldots, N-1$. The E player, namely, the maximizing player whose control input is v_k does not have access to w_k. Thus, E's information pattern is as in Sect. 3.1 whereas P's information pattern is as in Sect. 3.2; evidently, player P has an information advantage: player P knows everything that player E knows, and more. Notwithstanding the asymmetric information pattern, the LQGDG 3 is tractable.

The results are summarized in as follows.

Theorem 4. *Consider the stochastic control system (11) and payoff functional (12). The planning horizon is N. Assume that at decision time k the player P has access to the state x_k and the random disturbance w_k whereas player E has access to the state x_k only. In addition, assume the deterministic LQDG (1)–(2) has a solution for the planning horizon N. In the LQGDG 3 P's optimal strategy is given by the state feedback/stroboscopic control law:*

$$\begin{aligned} u_k^*(x_k, w_k) = &- S_B^{-1}(P_{N-k-1})B^T \\ &\times \left\{ \left[I + P_{N-k-1}C\left(R_v - C^T P_{N-k-1}C\right)^{-1} C^T \right] P_{N-k-1}A \cdot x_k \right. \\ &\left. + P_{N-k-1}\Gamma \cdot w_k \right\} \end{aligned} \tag{18}$$

and E's optimal strategy is given by the state feedback control law (10). The LQGDG 3 is a zero-sum game and its value function is:

$$V_0^{(3)}(x_0) = x_0^T(P_N - Q)x_0 + p_N^{(3)}. \tag{19}$$

The real symmetric matrices P_k are calculated according to the recursion (5) for the deterministic LQDG, and having obtained the sequence $\{P_k\}_{k=0}^N$, the new scalar sequence is calculated according to the recursion:

$$\begin{aligned} p_{k+1}^{(3)} &= p_k^{(3)} + Trace\Big(\Gamma^T \Big[P_k + P_k B S_B^{-1}(P_k) \\ &\qquad (R_u + B^T P_k B) S_B^{-1}(P_k) B^T P_k \Big] \Gamma W \Big), \\ p_0^{(3)} &= 0,\ k = 0, \ldots, N-1. \end{aligned} \tag{20}$$

As in the deterministic case, conditions (3) and (4) are necessary and sufficient for the existence of a solution to the zero-sum LQGDG 3 on the finite planning horizon N.

The strategy (18) is intuitively appealing – compare to the P player's strategy in the LQGDG 1 where neither player had access to the random input, (9) in Sect. 3.1, and the P player's strategy in the LQGDG 2, (14) in Sect. 3.2, where both players were privy to Nature's input.

Concerning the three stochastic games' value function, the following is derived:

Corollary 5. *The value functions of the LQGDG 1, LQGDG 2, and LQGDG 3 satisfy the inequality*

$$V_k^{(3)}(x_k) < min\left\{V_k^{(1)}(x_k), V_k^{(2)}(x_k)\right\} \;\; \forall\, x_k \in R^n, \;\; \forall\, k = 1, \ldots, N. \tag{21}$$

In other words, the sequence $p^{(3)}$ is dominated by the sequences $p^{(1)}$ and $p^{(2)}$, namely,

$$p_k^{(3)} < min\left\{p_k^{(1)}, p_k^{(2)}\right\} \;\; \forall\, k = 1, \ldots, N \tag{22}$$

and the value function differences are not state dependent, that is,

$$V_k^{(1)}(x_k) - V_k^{(3)}(x_k) = p_k^{(1)} - p_k^{(3)} \equiv const. \;\; \forall\, x_k \in R^n \tag{23}$$

and

$$V_k^{(2)}(x_k) - V_k^{(3)}(x_k) = p_k^{(2)} - p_k^{(3)} \equiv const. \;\; \forall\, x_k \in R^n. \tag{24}$$

Proof. All three dynamic games, namely, LQGDG 1, LQGDG 2, and LQGDG 3, are zero-sum games. The information advantage of player P, the minimizing player, over player E, the maximizing player, in the LQGDG 3 compared to the LQGDG 2 yields the inequality

$$p_k^{(3)} < p_k^{(2)} \;\; \forall\, k = 1, \ldots, N.$$

Similarly, the advantage player P enjoys in the LQGDG 3 compared to the LQGDG 1, or, alternatively, the fact that player E, the maximizing player, is better off playing the LQGDG 1 compared to the LQGDG 3, and the fact that these three games are all zero-sum games, yields the additional inequality

$$p_k^{(3)} < p_k^{(1)} \;\; \forall\, k = 1, \ldots, N$$

wherefrom (21) and (22) follow. □

Similar results can be obtained when the roles are reversed and player E has the information advantage. New scalar sequences $p^{(1)}$, $p^{(2)}$, and $p^{(3)}$ must be calculated. The new sequences will satisfy the inequality

$$p_k^{(3)} > \max\left\{p_k^{(1)}, p_k^{(2)}\right\} \;\; \forall\, k = 1, \ldots, N. \tag{25}$$

4 Example

The concepts and algorithms developed in Sects. 2 and 3 are illustrated using the vehicle of a scalar discrete-time LQDG.

Without loss of generality, let $b = 1$ and $c = 1$. Let $q = 0$, that is, one is exclusively interested in suppressing the effect of the disturbance at the terminal time. The problem parameters are the time constant a, the players' control effort weights r_u and r_v, and the terminal state penalty h.

4.1 *Deterministic Control*

The simple scalar dynamics are:

$$x_{k+1} = ax_k + u_k + v_k \tag{26}$$

and the payoff functional is:

$$J\left(\{u_k\}_{k=0}^{N-1}, \{v_k\}_{k=0}^{N-1}; x_0\right) = hx_N^2 + \sum_{k=0}^{N-1}\left(r_u u_k^2 - r_v v_k^2\right). \tag{27}$$

We first calculate the Schur complement [5]:

$$S_B(P_k) = \frac{(r_v - r_u)P_k + r_u r_v}{r_v - P_k} \;\; (>0).$$

We insert the $S_B(P_k)$ expression into the P_k recursion formula (5) and obtain the scalar DRE:

$$P_{k+1} = a^2 \frac{P_k}{1 + (\frac{1}{r_u} - \frac{1}{r_v})P_k}, \;\; P_0 = h, \;\; k = 0, \ldots, N-1. \tag{28}$$

Admittedly, in this example where both players' control variables are scalars ($m_u = 1$ and $m_v = 1$), the use of the Schur complement merely allowed us to avoid the inversion of a 2×2 matrix.

The cost guarantee is:

$$V_0(x_0) = P_N x_0^2$$

and (9) yields the optimal disturbance rejection strategy,

$$u_k^*(x_k) = -a \frac{P_{N-k-1}}{r_u + \left(1 - \frac{r_u}{r_v}\right)P_{N-k-1}} x_k, \;\; k = 0, \ldots, N-1.$$

Concerning the existence of a solution to the scalar discrete-time LQDG (26)–(27): according to Theorem 1, the solution of the DRE (28) must satisfy

$$-r_u < P_k < r_v, \quad \forall\, k = 1, \ldots, N-1.$$

In addition, the problem parameters must satisfy

$$-\, r_u < h < r_v. \tag{29}$$

From (28) we immediately conclude that choosing the state penalty term $h > 0$ and the control effort penalty terms

$$r_v > r_u > 0$$

guarantees that $P_k > 0 \;\forall\; k = 0, 1, \ldots$.

We will assume that the problem parameters satisfy the above conditions. Indeed, the requirement that $r_u < r_v$ means that the minimizer's energy cost is lower than the maximizer's energy cost – in other words the minimizing player is "more energetic" than the disturbance, and therefore can overcome the action of the disturbance. Hence, $P_k > 0 > -r_u$ and we only need to worry about $P_k < r_v \;\forall\; k = 0, \ldots, N-1$.

If, in addition, the time constant $-1 \leq a \leq 1$ so that the control system, which is excited by the disturbance, is stable, then P_k is monotonically decreasing and therefore $h < r_v$ guarantees $P_k < r_v \;\forall\; k = 0, \ldots, N-1$. Hence, the maximal length of the planning horizon $N_{\max} = \infty$. In addition, $0 < P_{k+1} < P_k < h$ $\forall\; k = 0, 1, \ldots$.. If however $\mid a \mid > 1$, that is, the open-loop system is not stable, then a finite maximal planning horizon might exist – in other words, it is possible that $N_{\max} < \infty$; in which case the DRE (28) has a finite escape time.

We now proceed to solve the DRE (28). Note that

$$\frac{1}{P_{k+1}} = \frac{1}{a^2}\frac{1}{P_k} + \frac{1}{a^2}\left(\frac{1}{r_u} - \frac{1}{r_v}\right),$$

provided that $P_k \neq 0$. Setting

$$\Pi_k := \frac{1}{P_k}$$

yields the first-order linear difference equation:

$$\Pi_{k+1} = \frac{1}{a^2}\Pi_k + \frac{1}{a^2}\left(\frac{1}{r_u} - \frac{1}{r_v}\right), \quad \Pi_0 = \frac{1}{h},$$

whose solution yields the sequence:

$$P_k = \begin{cases} \dfrac{a^{2k}}{\frac{1}{h} + \frac{a^{2k}-1}{a^2-1}\left(\frac{1}{r_u} - \frac{1}{r_v}\right)} & \text{if } \mid a \mid \neq 1 \\ \dfrac{1}{\frac{1}{h} + \left(\frac{1}{r_u} - \frac{1}{r_v}\right)k} & \text{if } a = \pm 1. \end{cases} \tag{30}$$

This, in turn, allows us to obtain the optimal disturbance rejection strategy:

$$u_k^*(x_k) = \begin{cases} -\dfrac{a^2-1}{a} \dfrac{a^{2(N-k)}}{\frac{r_u}{h}(a^2-1) + \left(1 - \frac{r_u}{r_v}\right)(a^{2(N-k)} - 1)}\, x_k & \text{if } \mid a \mid \neq 1 \\ -a \dfrac{1}{\frac{r_u}{h} + \left(1 - \frac{r_u}{r_v}\right)(N-k)}\, x_k & \text{if } a = \pm 1. \end{cases}$$

The optimal disturbance rejection strategy is linear in the state and is of the form:

$$u_k^*(x_k) = K^*(k; N) \cdot x_k.$$

The optimal gain K^* is time-dependent and is a function of the planning horizon N. Thus, at time $k = 0$ the gain is:

$$K^*(0; N) = \begin{cases} -(a^2-1) \dfrac{a^{2N-1}}{\frac{r_u}{h}(a^2-1) + \left(1 - \frac{r_u}{r_v}\right)(a^{2N} - 1)} & \text{if } \mid a \mid \neq 1 \\ -a \dfrac{1}{\frac{r_u}{h} + \left(1 - \frac{r_u}{r_v}\right)N} & \text{if } a = \pm 1, \end{cases}$$

and at time $k = N - 1$ the optimal terminal gain is:

$$K^*(N-1; N) = -a \frac{1}{1 + \frac{r_u}{h} - \frac{r_u}{r_v}} \quad \forall\, a \in R^1,$$

provided that $N \le N_{\max}$ – see (34) in the sequel.

The dynamics parameter a plays an important role in the solution of the LQDG (26)–(27). In general, and not just at the terminal time $k = N - 1$, the gain K^* is continuous in the parameter a as $\mid a \mid \to 1$.

Concerning the cost guarantee: (30) allows us to calculate P_N so that the value function:

$$V_0(x_0) = \begin{cases} \dfrac{a^{2N}}{\frac{1}{h} + \frac{a^{2N}-1}{a^2-1}\left(\frac{1}{r_u} - \frac{1}{r_v}\right)} x_0^2 & \text{if } \mid a \mid \neq 1 \\ \dfrac{1}{\frac{1}{h} + \left(\frac{1}{r_u} - \frac{1}{r_v}\right)N} x_0^2 & \text{if } a = \pm 1. \end{cases} \tag{31}$$

The value of the game is continuous in the parameter a as $\mid a \mid \to 1$ and is nonnegative $\forall\, a \in R^1$.

When the time constant $\mid a \mid > 1$, we must determine $N_{\max}$. To this end, invoke condition (4) for the existence of a solution:

$$P_k = \frac{a^{2k}}{\frac{1}{h} + \frac{a^{2k}-1}{a^2-1}\left(\frac{1}{r_u} - \frac{1}{r_v}\right)} < r_v \;\; \forall\, k = 1, \ldots, N_{\max} - 1,$$

that is, $\forall\, k = 1, \ldots, N_{\max} - 1$ the following must hold:

$$\left(a^2 - \frac{r_v}{r_u}\right) a^{2k} < (a^2 - 1)\frac{r_v}{h} + 1 - \frac{r_v}{r_u}.$$

Now, $\frac{r_v}{r_u} > 1$ and it is readily verifiable that if

$$1 <\mid a \mid \leq \sqrt{\frac{r_v}{r_u}}, \tag{32}$$

then even though the open-loop system is unstable,

$$N_{\max} = \infty.$$

If however the parameter

$$\mid a \mid > \sqrt{\frac{r_v}{r_u}} \;\; (> 1), \tag{33}$$

then

$$N_{\max} = \left\lfloor \frac{1}{2} \frac{\log \dfrac{(a^2-1)\frac{r_v}{h} + 1 - \frac{r_v}{r_u}}{a^2 - \frac{r_v}{r_u}}}{\log \mid a \mid} \right\rfloor, \tag{34}$$

the notation $[\cdot]$ in (34) designates the largest integer less than or equal to the expression in the square brackets.

When the problem parameters are s.t. $\mid a \mid > \sqrt{\frac{r_v}{r_u}}$, the infinite horizon LQDG (26)–(27) does *not* have a solution; this, notwithstanding the fact that the Discrete Algebraic Riccati Equation (DARE) of the LQDG (26)–(27) has a solution: in fact, the DARE has two solutions:

$$P = 0, \tag{35}$$

and

$$P = (a^2 - 1)\frac{r_u r_v}{r_v - r_u}. \tag{36}$$

In summary, when the LQDG's control effort weighing parameters $r_v > r_u > 0$, $r_v > h > 0$ and the dynamics parameter $\mid a \mid \leq 1$, then $N_{\max} = \infty$ and the DARE's applicable solution is given by (35); and if the dynamics parameter $1 < \mid a \mid \leq \sqrt{\frac{r_v}{r_u}}$, also in this case $N_{\max} = \infty$, however the DARE's applicable solution is given by (36). In both cases, $\mid a \mid < 1$, and $1 < \mid a \mid \leq \sqrt{\frac{r_v}{r_u}}$, the infinite horizon LQDG (26)–(27) has a solution, the respective solutions (35) and (36) of the DARE apply, and the DRE's solution sequence P_k converges to the applicable DARE solution for all "initial" conditions $0 < h < r_v$. Indeed, $P_N \to P$ when $N \to \infty$, where – see also (35) and (36):

$$P = \begin{cases} \dfrac{r_u r_v}{r_v - r_u}(a^2 - 1) & \text{if } \sqrt{\frac{r_v}{r_u}} \geq \mid a \mid > 1 \\ 0 & \text{if } \mid a \mid \leq 1, \end{cases} \tag{37}$$

so that the value function

$$V_0(x_0) = \begin{cases} \dfrac{r_u r_v}{r_v - r_u}(a^2 - 1)x_0^2 & \text{if } \sqrt{\frac{r_v}{r_u}} \geq \mid a \mid > 1 \\ 0 & \text{if } \mid a \mid < 1. \end{cases}$$

Hence, the disturbance rejection controller's steady state optimal control law is:

$$u_k^*(x_k) = \begin{cases} -\dfrac{r_v}{r_v - r_u}\left(a - \dfrac{1}{a}\right)x_k, \ k = 0, 1, \ldots & \text{if } \sqrt{\frac{r_v}{r_u}} \geq \mid a \mid > 1 \\ 0 & \text{if } \mid a \mid < 1, \end{cases}$$

and the dynamics of the closed-loop system which is excited by the disturbance are:

$$x_{k+1} = \begin{cases} \dfrac{1}{r_v - r_u}\left(\dfrac{1}{a}r_v - a r_u\right) x_k + v_k, & \text{if } \sqrt{\frac{r_v}{r_u}} \geq \mid a \mid > 1 \\ a x_k + v_k & \text{if } \mid a \mid < 1. \end{cases} \tag{38}$$

If the maximizing player plays optimally, the dynamics are:

$$x_{k+1} = \begin{cases} \dfrac{1}{r_v - r_u}\left(a r_v - \dfrac{1}{a}r_u\right) x_k + u_k, & \text{if } \sqrt{\frac{r_v}{r_u}} \geq \mid a \mid > 1 \\ a x_k + u_k & \text{if } \mid a \mid < 1. \end{cases} \tag{39}$$

If both players play optimally, the closed-loop dynamics are:

$$x_{k+1} = \begin{cases} \dfrac{1}{a} x_k, & \text{if } \sqrt{\frac{r_v}{r_u}} \geq \mid a \mid > 1 \\ a x_k & \text{if } \mid a \mid < 1. \end{cases} \tag{40}$$

It is readily verifiable that always, in (38) the coefficient

$$\left| \frac{1}{r_v - r_u}\left(\frac{1}{a}r_v - a r_u\right) \right| < 1,$$

which guarantees stability. Thus, the minimizer's optimal action brings about stable dynamics. In the same vein, it is readily verifiable that in (39), the coefficient

$$\left| \frac{1}{r_v - r_u}\left(a r_v - \frac{1}{a}r_u\right) \right| > 1,$$

that is, the maximizer's optimal action tends to destabilize the system – which serves the disturbance well. However, when both players play optimally, the closed-loop dynamics (40) are stable: the more energetic minimizing player ($r_u < r_v$) is able to overcome the maximizer's action and always enforce stability. Thus, during optimal play, and when $\mid a \mid < 1$, the trajectory is:

$$x_k = a^k x_0, \quad k = 0, 1, \ldots,$$

the control energy expanded by the minimizing and maximizing players is:

$$\begin{aligned} E_u &\equiv r_u \sum_{k=0}^{\infty} u_k^2 \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} E_v &\equiv r_v \sum_{k=0}^{\infty} v_k^2 \\ &= 0, \end{aligned}$$

respectively; and since $\lim_{N\to\infty} x_N = 0$, the value function $V_0(x_0) = 0 \;\forall\; x_0 \in R^1$. When $\sqrt{\frac{r_v}{r_u}} \geq\mid a \mid> 1$, the trajectory during optimal play is:

$$x_k = a^{-k} x_0, \quad k = 0, 1, \ldots$$

and the optimal controls are:

$$u_k = -\frac{r_v}{r_v - r_u}\left(a - \frac{1}{a}\right) a^{-k} x_0$$

and

$$v_k = \frac{r_u}{r_v - r_u}\left(a - \frac{1}{a}\right) a^{-k} x_0.$$

The calculation of the players' control efforts, E_u and E_v, entails the summation of convergent geometric series. We evaluate

$$E_u = \frac{r_u r_v^2}{(r_v - r_u)^2}(a^2 - 1)x_0^2$$

and

$$E_v = \frac{r_u^2 r_v}{(r_v - r_u)^2}(a^2 - 1)x_0^2$$

and since $\lim_{N\to\infty} x_N = 0$, we directly calculate the value of the game

$$\begin{aligned} V_0(x_0) &= E_u - E_v, \\ &= \frac{r_u r_v}{r_v - r_u}(a^2 - 1)x_0^2, \end{aligned}$$

as expected.

For finite planning horizons N, $P_N > 0$ always; in other words, the value of the game is always positive.

When the LQDG's parameters $r_v > r_u > 0$, $r_v > h > 0$, however the time constant $\mid a \mid > \sqrt{\frac{r_v}{r_u}}$, then $N_{\max} < \infty$ and is given by (34); the DRE (28) exhibits a finite escape time, and the DARE's solutions (35) and (36) do *not* apply to the solution of the LQDG (26)–(27).

In general, when the LQDG formulation is used to model pursuit-evasion scenarios, it is assumed that the minimizing player's/pursuer's control effort penalty matrix:

$$0 < R_u < R_v.$$

This is tantamount to saying that the pursuer's/minimizer's control cost is lower than the evader's/maximizer's, so that the pursuer is more energetic and he will close in on the evader/reduce the miss distance. In light of this observation, in the disturbance rejection scenario it would be interesting to consider the situation where the opposite is the case, that is, in our scalar example $0 < r_v < r_u$. This, in itself, will not preclude the existence of a solution, but will cause the maximum planning horizon to be finite, that is, $N_{\max} < \infty$ always.

4.2 Steady State Control

The following discussion concerns steady state control action. The infinite planning horizon case $N \to \infty$ requires special care [8].

The applicable parameter range is $\mid a \mid \leq \sqrt{\frac{r_v}{r_u}}$ (> 1) – the dynamics parameter $a = \pm 1$ requires special attention.

The steady state disturbance rejection strategy is:

$$u_k^*(x_k) = \begin{cases} -\dfrac{r_v}{r_v - r_u}\left(a - \dfrac{1}{a}\right) x_k, \; k = 0, 1, \ldots & \text{if } \sqrt{\frac{r_v}{r_u}} \geq \mid a \mid > 1 \\ 0, \; k = 0, 1, \ldots & \text{if } \mid a \mid < 1, \end{cases}$$

that is, the optimal control gain is constant:

$$K^*(k; \infty) = \begin{cases} -\dfrac{r_v}{r_v - r_u}\left(a - \dfrac{1}{a}\right), \; \forall\, k = 0, 1, \ldots & \text{if } \sqrt{\frac{r_v}{r_u}} \geq \mid a \mid > 1 \\ 0, \; \forall\, k = 0, 1, \ldots & \text{if } \mid a \mid < 1. \end{cases}$$

In particular, the initial gain at time zero is:

$$K^*(0;\infty) = \begin{cases} -\dfrac{r_v}{r_v - r_u}\left(a - \dfrac{1}{a}\right) & \text{if } \sqrt{\frac{r_v}{r_u}} \geq \mid a \mid > 1 \\ 0 & \text{if } \mid a \mid < 1. \end{cases}$$

Alternatively, the initial gain at time $k = 0$ can be obtained by setting $k = 0$ in the gain formula for the case of a finite planning horizon of length N and letting $N \to \infty$. We calculate

$$\lim_{N\to\infty} K^*(0;N) = K^*(0;\infty),$$

as expected.

Consider now the terminal gain. We are interested in the gain applied at the time instant $k = N - 1$ when $N \to \infty$. The optimal gain formula derived for the finite planning horizon case yields:

$$K^*(N-1;N-1) = -a\frac{1}{1 + \frac{r_u}{h} - \frac{r_u}{r_v}} \ \ \forall \ \mid a \mid \leq \sqrt{\frac{r_v}{r_u}}, \ N = 1, 2, \ldots.$$

Thus, when $N \to \infty$

$$K^*(\infty;\infty) = -a\frac{1}{1 + \frac{r_u}{h} - \frac{r_u}{r_v}} \ \ \forall \ \mid a \mid \leq \sqrt{\frac{r_v}{r_u}},$$

and it is not equal to the steady-state gain at time $k \to \infty$: the terminal (at time $k = \infty$) steady-state strategy is *not* the limit of the finite planning horizon strategy when the planning horizon $N \to \infty$.

This dichotomy is not germane to games only and it also arises in one sided linear-quadratic optimal control problems. In this respect, it is instructive to momentarily digress and consider the infinite horizon optimal control problem from first principles.

The dynamics are:

$$x_{k+1} = ax_k + u_k, \ \ x_0 \equiv x_0, \ \ k = 0, 1, \ldots,$$

and the cost functional is

$$J = hx_\infty^2 + \sum_{k=0}^{\infty} u_k^2.$$

Since we are interested in steady-state action we consider linear control laws of the form:

$$u_k = -Kx_k,$$

and we determine the optimal gain K^*.
The closed-loop dynamics are:

$$x_{k+1} = (a - K)x_k, \;\; x_0 \equiv x_0, \;\; k = 0, 1, \dots,$$

and therefore the state evolves according to

$$x_k = (a - K)^k x_0$$

and the control

$$u_k = -K(a - K)^k x_0.$$

Hence, the expanded control energy is:

$$\begin{aligned} E &= \sum_{k=0}^{\infty} u_k^2 \\ &= K^2 x_0^2 \sum_{k=0}^{\infty} (a - K)^{2k}. \end{aligned}$$

Feasible controls are in l_2: the control energy is finite provided the gain K satisfies

$$a - 1 < K < a + 1,$$

whereupon $x_\infty = 0$ and the cost

$$J(x_0; K) = E = \frac{K^2}{1 - (a - K)^2}\, x_0^2.$$

Hence, the optimal gain

$$K^* = a - \frac{1}{a}.$$

The optimal gain K^* is positive and satisfies the inequality $a - 1 < K^* < a + 1$ if the dynamics parameter

$$\mid a \mid > 1.$$

The optimal (minimal) cost is:

$$J^* = a^2 - 1 \ (>0).$$

Obviously, if

$$| \, a \, |< 1,$$

the gain

$$K^* = 0$$

and the optimal cost

$$J^* = 0.$$

Concerning $| \, a \, |= 1$: If

$$a = 1$$

and

$$0 < K < 2,$$

the cost

$$J(x_0; K) = \frac{K^2}{2K - K^2} \, x_0^2 \ \ (>0).$$

The optimal gain must satisfy

$$0 < K^* < 2$$

and one must solve the optimization problem:

$$\min_{0<K<2} \ \frac{K}{2 - K}.$$

The above cost function is monotonically increasing in K on the interval $0 < K < 2$: if $K = \epsilon$, $0 < \epsilon \ll 1$, the cost $J(x_0; \epsilon) \approx \frac{1}{2}\epsilon$. If however $K = 0$ the cost $J(x_0; 0) = hx_0^2$ (> 0). A minimum does not exist.

Similarly, if

$$a = -1$$

and $-2 < K < 0$, the cost:

$$J(x_0; K) = -\frac{K^2}{2K + K^2}\, x_0^2 \;\; (>0).$$

The optimal gain must satisfy

$$-2 < K^* < 0$$

and one must solve the optimization problem:

$$\min_{-2<K<0} \; -\frac{K}{2+K}.$$

The above cost function is monotonically decreasing on the interval $-2 < K < 2$: if $K = -\epsilon$, $0 < \epsilon \ll 1$, the cost $J(x_0;\epsilon) \approx \frac{1}{2}\epsilon$. However, if $K = 0$, the cost $J(x_0; 0) = hx_0^2$ (> 0). A minimum does not exist.

In summary, the solution of the infinite-horizon optimal control problem is as follows. The optimal control law is:

$$u_k^*(x_k) = K^* x_k,$$

where the optimal gain

$$K^* = \begin{cases} a - \dfrac{1}{a} & \text{if} \quad |\, a \,| > 1 \\ \quad 0 & \text{if} \quad |\, a \,| < 1, \end{cases}$$

the optimal (minimal) cost

$$J^*(x_0) \equiv 0 \;\; \forall x_0 \in R^1$$

and if $a = \pm 1$ an optimal solution does not exist, however by using the gains $K = \epsilon$ if $a = 1$ and $K = -\epsilon$ if $a = -1$ one can make the cost $J^* \to 0$.

In the LQDG, the steady state optimal control law for disturbance rejection is:

$$u_k^*(x_k) = \begin{cases} -\dfrac{r_v}{r_v - r_u}\Big(a - \dfrac{1}{a}\Big)x_k,\; k = 0, 1, \ldots & \text{if } \sqrt{\frac{r_v}{r_u}} \geq |\, a \,| > 1 \\ \qquad 0 & \text{if } |\, a \,| < 1 \\ \text{Does not exist} & \text{if } |\, a \,| = 1. \end{cases}$$

There is a "gap" when $|\, a \,| = 1$ in both the optimal control and LQDG problems. Note however that if $|\, a \,| = 1$, a small gain controller will yield a cost arbitrarily close to the absolute minimum of zero.

4.3 Stochastic Control: LQGDG 1

As in Sect. 4.1, we assume that the parameters of the LQGDG 1 satisfy $r_v > r_u > 0$ and $r_v > h > 0$. In view of the analysis in Sect. 4.1 – see, e.g., (28) – the sequence (13) is:

$$p_k^{(1)} = \begin{cases} \Gamma^2 W \sum_{i=0}^{k-1} \dfrac{a^{2i}}{\frac{1}{h} + \frac{a^{2i}-1}{a^2-1}\left(\frac{1}{r_u} - \frac{1}{r_v}\right)} & \text{if } \mid a \mid \neq 1\ ;\ \ k = 1, \ldots, N-1 \\ \Gamma^2 W \sum_{i=0}^{k-1} \dfrac{1}{\frac{1}{h} + \left(\frac{1}{r_u} - \frac{1}{r_v}\right) i} & \text{if } a = \pm 1\ ;\ \ k = 1, \ldots, N-1. \end{cases}$$

If the dynamics parameter $\mid a \mid \leq 1$ then the increment in the $p_k^{(1)}$ difference equation $\to 0$. Moreover, if the dynamics parameter satisfies $\mid a \mid < 1$, the sequence $p_k^{(1)}$ converges. We therefore conclude that if the open-loop control system is asymptotically stable, the payoff functional in the infinite planning horizon case is bounded and an optimal (Nash) solution to the LQGDG 1 exists. The calculation of the value of the game requires the summation of the following series.

$$V_0^{(1)}(x_0) = \sum_{k=0}^{\infty} \frac{a^{2k}}{\frac{1}{h} + \frac{a^{2k}-1}{a^2-1}\left(\frac{1}{r_u} - \frac{1}{r_v}\right)} \Gamma^2 W x_0^2 \quad \left(< \frac{1}{1-a^2} h \Gamma^2 W x_0^2\right) \quad \forall\ x_0 \in R^1.$$

If the dynamics parameter $a = \pm 1$, the sequence $p_k^{(1)}$ does not converge. Although the players' "optimal" strategies are the steady-state strategies $u_k \equiv 0$ and $v_k \equiv 0$, the payoff functional in the infinite planning horizon case is *not* bounded.

If the dynamics parameter $\mid a \mid > 1$, the increment in the $p_k^{(1)}$ sequence $\to (a^2 - 1) \frac{r_u r_v}{r_v - r_u} \Gamma^2 W\ (> 0)$ so that asymptotically the sequence $p_k^{(1)}$ is an arithmetic progression and the sequence $p_k^{(1)}$ does not converge. If the dynamics parameter satisfies (32), that is, $1 \leq \mid a \mid \leq \sqrt{\frac{r_v}{r_u}}$, then although the open-loop system is unstable, steady state "optimal" strategies exist, however, since the sequence $p_k^{(1)}$ is not bounded the value function in the LQGDG 1 is not bounded; in other words, when the dynamics parameter $\mid a \mid > 1$, a solution to the LQGDG 1 does not exist. This is due to the action of the persistent random disturbance w_k.

4.4 Stochastic Control: LQGDG 2

As in Sect. 4.1, we assume that the parameters of the LQGDG 2 satisfy $r_v > r_u > 0$ and $r_v > h > 0$. In view of the analysis of the example in Sect. 4.1 – see, e.g., (30) – the $p^{(2)}$ recursion is:

$$p_{k+1}^{(2)} = \begin{cases} p_k^{(2)} + \dfrac{a^{2(k+1)}}{\frac{1}{h} + \frac{a^{2(k+1)}-1}{a^2-1}\left(\frac{1}{r_u} - \frac{1}{r_v}\right)} \left(\dfrac{\Gamma}{a}\right)^2 W, \;\; p_0^{(2)} = 0, & k = 0, \ldots, N-1 \\ & \text{if } \mid a \mid \neq 1 \\ p_k^{(2)} + \dfrac{1}{\frac{1}{h} + \left(\frac{1}{r_u} - \frac{1}{r_v}\right)(k+1)} \Gamma^2 W, \;\; p_0^{(2)} = 0, & k = 0, \ldots, N-1 \\ & \text{if } a = \pm 1. \end{cases}$$

Hence, if the dynamics parameter $\mid a \mid \leq 1$ then the increment in the $p^{(2)}$ recursion $\rightarrow 0$. Moreover, the sequence $p_k^{(2)}$ converges, provided that the dynamics parameter satisfies $\mid a \mid < 1$, that is, the open loop system is asymptotically stable. We therefore conclude that if the open-loop system is asymptotically stable the payoff functional in the infinite planning horizon case is bounded and calculating the value of the game requires the summation of the following series.

$$V_0^{(2)} = \sum_{k=0}^{\infty} \frac{a^{2(k+1)}}{\frac{1}{h} + \frac{a^{2(k+1)}-1}{a^2-1}\left(\frac{1}{r_u} - \frac{1}{r_v}\right)} \left(\frac{\Gamma}{a}\right)^2 W \;\; \left(< \frac{1}{1-a^2} h \Gamma^2 W\right).$$

If the dynamics parameter $a = \pm 1$, the sequence $p_k^{(2)}$ does not converge and although the players' "optimal" strategies are the steady-state strategies $u_k \equiv 0$ and $v_k \equiv 0$, in the infinite planning horizon case the payoff functional is not bounded – an optimal solution of the LQGDG 2 does *not* exist.

If the dynamics parameter $\mid a \mid > 1$, the increment in the sequence $p_k^{(2)} \rightarrow (a^2-1)\frac{r_u r_v}{r_v - r_u}\Gamma^2 W \;(> 0)$ so that asymptotically the sequence $p_k^{(2)}$ is an arithmetic progression. Hence, if the dynamics parameter satisfies (32) then the increment in the $p^{(2)}$ recursion $\rightarrow (a^2-1)\frac{r_u r_v}{r_v - r_u}\Gamma^2 W$ and, although in this case of an unstable open-loop system steady state "optimal" strategies exist, similar to the LQGDG 1, the performance functional in the LQGDG 2 is not bounded. Thus, if the planning horizon $N \rightarrow \infty$ and if the dynamics parameter $\mid a \mid \geq 1$ an optimal solution of the LQGDG 2 does *not* exist.

Finally, in our example the following relationship holds.

$$p_k^{(1)} = a^2 p_k^{(2)} + \Gamma^2 W \left[h - \frac{a^{2k}}{\frac{1}{h} + \frac{a^{2k}-1}{a^2-1}} \left(\frac{1}{r_u} - \frac{1}{r_v}\right)\right] \;\; \forall\, k = 1, \ldots, N.$$

Hence, if $\mid a \mid < 1$,

$$p^{(1)} = a^2 p^{(2)} + h \Gamma^2 W$$

and therefore if the open-loop system is asymptotically stable the values in the infinite planning horizon case satisfy the relationship

$$V^{(1)} = a^2 V^{(2)} + h\Gamma^2 W.$$

4.5 Stochastic Control: LQGDG 3

As in Sect. 4.1, we assume that the parameters of the LQGDG 3 satisfy $r_v > r_u > 0$ and $r_v > h > 0$. The scalar $p_k^{(3)}$ in the LQGDG 3 – see (20) – is:

$$\begin{aligned} p_k^{(3)} &= \Gamma^2 W \\ &\quad \times \sum_{i=0}^{k-1} P_i \frac{(r_u + P_i)[P_i^3 + (r_u^2 - 2r_v)P_i^2 + 2r_v(r_v - r_u)P_i + r_u r_v^2] + r_u^2 P_i^2}{[r_v(r_u + P_i) - r_u P_i]^2}, \\ &\quad k = 1, \ldots, N, \end{aligned}$$

where the sequence P_i is specified by (30).

5 Conclusion

In this chapter a hierarchy of discrete-time Linear-Quadratic Dynamic Games is analyzed, in the context of the application of the theory of dynamic games to optimal disturbance rejection using digital signal processing. In discrete-time the dynamic programming equation contains product terms between the decision variables, which complicates the solution compared to the continuous-time analogue. With a view to facilitate the application of the theory of dynamic games to digital signal processing, and, in particular, disturbance rejection, minimal necessary and sufficient conditions for the existence of a solution are established, the complete solution of discrete-time LQDGs is worked out, and explicit results are obtained. A hierarchy of three zero-sum stochastic LQDGs characterized by information patterns which increase in complexity is analyzed and an insight into the dynamics of information systems is obtained. It is shown that if the deterministic LQDG has a solution for the planning horizon N then also the stochastic LQGDGs 1–3 have a solution on this planning horizon. Moreover, while the LQDG value function and the value functions of the LQGDGs’ 1, 2, 3 are all different, the difference in the value functions is not state dependent. The results are illustrated in the example discussed in Sect. 4, where some fine points concerning dynamic games are highlighted. The infinite horizon LQGDGs are investigated and it is shown that while steady state optimal strategies are possible even in the case where the open loop system is not asymptotically stable, the value function is not bounded and a solution does not exist. We do however point out that the inclusion of a temporal discount factor $0 \leq \lambda < 1$ in the cost functional will render the stochastic games' value function finite.

Appendix A: Planning Horizon

The planning horizon N is chosen by the system designer, subject to the constraint:

$$N \leq N_{\max},$$

where $1 \leq N_{\max} \leq \infty$ is exclusively determined by the problem parameters. Indeed, in order for $1 \leq N_{\max}$, the problem parameters must satisfy the minimal conditions (6) and (7). We will assume that this is indeed the case so that $1 \leq N_{\max}$ and the following discussion is not vacuous.

Using the theory developed in Sect. 2, the maximal length of the planning horizon, $N_{\max}$, is determined as follows.

Algorithm 1

Step 0.) Assume that conditions (6) and (7) hold.

Step 1.) Iterate the recursion (5) for P_k $\forall$ $k = 0, \ldots, K-1$; if $K > 1$, assume that conditions (3) and (4) held $\forall$ $k = 1, \ldots, K-1$: set

$$N_{\max} = K$$

and calculate P_K.

Step 2.) Check conditions (3) and (4) for $k = K$. If conditions (3) and (4) hold, proceed to solve the recursion (5) for P_{K+1} and set

$$N_{\max} = K + 1$$

Set $k = K + 1$ and return to the verification Step 2.), where conditions (3) and (4) are checked.

When condition (3) and/or condition (4) fail, the process stops – at which time a finite $N_{\max}$ is established.

The DRE (5) is a nonlinear difference equation. When $N_{\max} < \infty$ the DRE has a "finite escape time" and the LQDG (1)–(2) does *not* have a solution for planning horizons $N > N_{\max}$. However, if the process can be continued ad infinitum – for example, $\{P_k\}_{k=0}^{\infty}$ is a bounded and monotonic sequence of real, symmetric, positive definite matrices, as is the case in convex LQ optimal control with $H \geq 0$ and $Q \geq 0$ – then $N_{\max} = \infty$. When $N_{\max} = \infty$, the LQDG with dynamics (1), cost functional (2), and complete state information, has a solution $\forall$ $N \geq 1$.

The determination of the longest possible planning horizon, $N_{\max}$, s.t. a solution of the discrete-time LQDG (1)–(2) exists, is an interesting problem in its own right. However, one is especially interested in the infinite horizon case where $N_{\max} = \infty$, namely, the steady state solution of the DRE. That the solution of the DRE can be extended to $N \to \infty$ is not at all clear in dynamic games. This comes somewhat as a surprise, for in one sided optimization problems, that is, in linear-quadratic control and estimation/Kalman filtering, this is achievable under relatively mild assumptions, e.g., $0 \leq Q$ and $0 \leq H$.

Extending the planning horizon of the LQDG to infinite time is most desirable, for then, the control action entails linear *constant* gain action – in other words, the system is linear time invariant – a most desirable state of affairs in real-time control, digital signal processing, and filtering. Now, steady state disturbance rejection hinges first and foremost on the solution – see, e.g., (5) – of the following, fairly complex, algebraic equation: when $N_{\max} = \infty$, $\exists$ P s.t. $\lim_{k\to\infty} P_k = P$ and the real, symmetric $n \times n$ matrix P is the solution of the algebraic matrix equation:

$$\begin{aligned} P = A^T \{P - P \big[&BS_B^{-1}(P)B^T + BS_B^{-1}(P)B^T PC(R_v - C^T PC)^{-1}C^T \\ &+ C(R_v - C^T PC)^{-1}C^T PBS_B^{-1}(P)B^T \\ &+ C(R_v - C^T PC)^{-1}C^T PBS_B^{-1}(P)B^T PC(R_v - C^T PC)^{-1}C^T \\ &+ C(C^T PC - R_v)^{-1}C^T\big] P\} A + Q. \end{aligned} \tag{41}$$

The LQDG's value function is then

$$V_0(x_0) = x_0^T (P - Q)x_0 \tag{42}$$

and the optimal strategies are constant gains, that is, Linear Time Invariant (LTI) control laws. Specifically, the optimal disturbance rejection signal is then generated by the constant gain Linear Time Invariant (LTI) control law:

$$u_k^*(x_k) = -S_B^{-1}(P)B^T \left[I + PC\left(R_v - C^T PC\right)^{-1} C^T\right] PA \cdot x_k \tag{43}$$

and

$$\begin{aligned} v_k^*(x_k) = \left(R_v - C^T PC\right)^{-1} C^T \{&I - PBS_B^{-1}(P)B^T \\ &\times \left[I + PC(R_v - C^T PC)^{-1}C^T\right]\} PA \cdot x_k. \end{aligned}$$

In conformity with the well developed theory of continuous-time LQDGs/linear-quadratic differential games [2,3], the algebraic matrix equation (41) will be referred to as the *Discrete* Algebraic Riccati Equation (DARE). Its off-line numerical solution is required for the calculation of the controller's $m_u \times n$ gain matrix and the implementation of steady-state real-time disturbance rejection control – see, e.g., (43). The gain matrix is calculated ahead of time, implementation is easy, and during real-time operation the execution time is bounded.

Note that if $Q = 0$ then $P = 0$ is a solution of the DARE (41). However, a real, symmetric, solution P of the DARE (41), if it exists, is not necessarily unique. The "correct" solution, namely, the solution which is applicable to the LQDG (1)–(2), is the specific solution P of the DARE s.t. the sequence P_k generated by the difference equation (5), initialized according to

$$P_0 = H + Q,$$

satisfies

$$\lim_{k\to\infty} P_k = P.$$

Hence, according to Theorem 1, the following must hold.

Proposition 6. *A candidate solution P of the DARE (41), and which is applicable to the LQDG (1)–(2), must satisfy the conditions:*

$$R_u + B^T PB > 0$$

and

$$R_v > C^T PC.$$

Thus the following is derived.

Corollary 7. *If $Q = 0$ and $R_u > 0$ and $R_v > 0$, $P = 0$ is a candidate solution of the DARE (41).*

In addition, the following holds.

Proposition 8. *A real symmetric nonsingular matrix P is a solution of the DARE (41) if it satisfies the inequality*

$$P \le A^T \left(P^{-1} - CR_v^{-1}C^T\right)^{-1} A + Q. \tag{44}$$

Moreover, if $Q = 0$, the dynamics matrix A is nonsingular and the candidate solution $P > 0$, the following holds.

$$P^{-1} - AP^{-1}A^T \le CR_v^{-1}C^T. \tag{45}$$

Proof. Equation (44) directly follows from the inequality from Sect. 2 concerning the solution of the DRE (5),

$$P_{k+1} \le A^T \left(P_k^{-1} - CR_v^{-1}C^T\right)^{-1} A + Q, \;\; P_0 = H + Q, \;\; k = 0, \dots, N-1$$

and an application of Proposition 8.8.5 from [9] (A and B are positive definite matrices and $A \le B \to B^{-1} \le A^{-1}$) yields (45). □

In dynamic games it is not always possible to extend the solution of the DRE, and consequently, the solution of the LQDG, to $N \to \infty$. $N_{\max}$ is obtained by exercising Algorithm. Even so, and irrespective of whether $N_{\max}$ is finite or infinite, in the context of disturbance rejection an interesting corollary of Theorem 1 is as follows.

Corollary 9. *When the the initial state $x_0 = 0$, the value of the game, that is, the minimizing player's cost guarantee, is always 0 $\forall N$, provided that the planning horizon $N \leq N_{max}$.*

Proof. Follows from (8). □

In other words: consider an initially quiescent dynamical system, excited by a persistent disturbance with bounded energy. The disturbance rejection strategy in Theorem 1 provides a zero cost guarantee, irrespective of the length of the planning horizon N – provided that $N \leq N_{\max}$. Thus, the following holds.

- If the planning horizon $N \leq N_{\max}$ and the initial state $x_0 = 0$, the guaranteed cumulative effect of the disturbance on the output is equal to the energy of the disturbance minus the expanded control energy, that is, the applied control energy directly diminishes the effect of the disturbance.

When the planning horizon $N \to \infty$, that is, the DRE (5) does not have a finite escape time, the players' optimal strategies reach a steady state. Specifically, the P player's optimal steady state strategy is given by (43), where the $n \times n$ real symmetric matrix P is a solution of the DARE (41). At the same time, since $P_k \to P$, the scalar sequence $p_k^{(1)}$ approaches an arithmetic progression. Thus, in the stochastic LQGDG 1 the sequence $p_k^{(1)}$ is monotonic and, in general (if $P \neq 0$), it does *not* reach a steady state. Hence, whereas in the deterministic LQDG the value function is finite, in the stochastic LQGDG 1 the value function is in general not bounded; in other words, an optimal solution does not exist. However, the inclusion of a temporal discount factor $0 \leq \lambda < 1$ in the cost functional will render the value function finite.

The special case where $Q = 0$ and the open-loop system is asymptotically stable is particularly interesting. The applicable steady state solution of the DARE (41) is then $P = 0$ and therefore the optimal strategies are $u_k \equiv 0$ and $v_k \equiv 0$. The value function is finite and a solution to the LQGDG 1 exists. The value of the LQGDG 1 is:

$$V_0^{(1)}(x_0) = p^{(1)} = \text{Trace}\left(\Gamma^T \left(\sum_{k=0}^{\infty} P_k\right) \Gamma W\right) = \text{Const.} \quad \forall\, x_0 \in R^n.$$

Finally, a word of caution concerning the infinite planning horizon is in order. While the existence of a solution to the DARE is paramount, as pointed out by Mageirou in [8], care must be exercised concerning the interpretation of the players' steady-state strategies – see the discussion in Sect. 4.2 concerning the asymptotic properties of the value of the game.

Appendix B: Discounting the Future

In infinite horizon LQDGs, LQGDGs, or in optimal control problems, the value function is not always finite; this is particularly true in the stochastic case. To guarantee a finite value function and the existence of an optimal solution, and motivated by the practice in economics to “discount the future”, the cost functional (2) is modified as follows:

$$J\left(\{u_k\}_{k=0}^{N-1}, \{v_k\}_{k=0}^{N-1};\, x_0\right)$$
$$= \lambda^N x_N^T H x_N + \sum_{k=0}^{N-1} \lambda^k \left(x_{k+1}^T Q x_{k+1} + u_k^T R_u u_k - v_k^T R_v v_k\right), \qquad (46)$$

where the “prevailing interest rate” parameter $0 < \lambda < 1$.

B1 Deterministic LQDG

Theorem 1 is modified as follows.

Theorem 10. *Necessary and sufficient conditions for the existence of a solution to the discrete-time zero-sum LQDG with dynamics (1), cost functional (46), and complete state information are (3) and (4) where the real, symmetric, matrices P_k are the solution of the difference equation:*

$$\begin{aligned} P_{k+1} = \lambda A^T \{ & P_k - P_k \left[BS_B^{-1}(P_k)B^T + BS_B^{-1}(P_k)B^T P_k C(R_v - C^T P_k C)^{-1} C^T\right. \\ & + C(R_v - C^T P_k C)^{-1} C^T P_k BS_B^{-1}(P_k)B^T \\ & + C(R_v - C^T P_k C)^{-1} C^T P_k BS_B^{-1}(P_k)B^T P_k C(R_v - C^T P_k C)^{-1} C^T \\ & \left. + C(C^T P_k C - R_v)^{-1} C^T\, \right] P_k \,\} A + Q, \\ P_0 = \lambda H + Q, & \;\; k = 0, \ldots, N-1. \end{aligned} \qquad (47)$$

As in Theorem 1, the matrix function

$$S_B(P_k) \equiv B^T P_k B + R_u + B^T P_k C(R_v - C^T P_k C)^{-1} C^T P_k B.$$

In addition, the problem parameters must satisfy the conditions:

$$R_u + B^T(Q + \lambda H)B > 0 \qquad (48)$$

and

$$R_v > C^T(Q + \lambda H)C. \qquad (49)$$

The value of the LQDG is:

$$V_0(x_0) = \frac{1}{\lambda} x_0^T (P_N - Q) x_0 \tag{50}$$

and the P and E players' optimal strategies are the linear state feedback control laws (9) and (10), respectively.

B2 Stochastic control: LQGDG 1

The payoff functional is:

$$\begin{aligned} &J\left(\{u_k\}_{k=0}^{N-1}, \{v_k\}_{k=0}^{N-1};\, x_0\right) \\ &\quad = E_w\left(\lambda^N x_N^T H x_N + \sum_{k=0}^{N-1} \lambda^k \left(x_{k+1}^T Q x_{k+1} + u_k^T R_u u_k - v_k^T R_v v_k\right)\right). \end{aligned} \tag{51}$$

The expectation is taken over the process noise sequence, where $w \equiv \{w_0, \ldots, w_{N-1}\}$.

The following holds.

Theorem 11. *Consider the stochastic control system (11) and payoff functional (51). The planning horizon is N. Assume that at decision time k the players have access to the state x_k, that is, the players have complete state information. In addition, assume the deterministic LQDG (1)–(2) has a solution for the planning horizon N. The players' optimal strategies in the LQGDG 1 with complete state information are given by the solution of the deterministic LQDG provided by Theorem 10, namely, the state feedback control laws (9) and (10). However, the value function of the zero-sum game (11) and (51), namely, the players' expected payoff, contains an additional term:*

$$V_0^{(1)}(x_0) = \frac{1}{\lambda} x_0^T (P_N - Q) x_0 + p_N^{(1)}.$$

The real symmetric matrices P_k are calculated according to the recursion (47) for the deterministic LQDG, and having obtained the sequence $\{P_k\}_{k=0}^N$, the new scalar sequence $p_k^{(1)}$ is calculated according to

$$p_k^{(1)} = Trace\left(\Gamma^T \left(\sum_{i=0}^{k-1} \lambda^{k-1-i} P_i\right) \Gamma W\right), \quad k = 1, \ldots, N, \tag{52}$$

so that the value function can be obtained. In (47) and (52), $P_k := \lambda P_k + Q$.

As in the deterministic case, conditions (3) and (4) are necessary and sufficient for the existence of a solution to the LQGDG 1 on the finite planning horizon N.

Concerning the infinite horizon problem: the DARE is:

$$\begin{aligned}
P = \lambda A^T \Big\{ P - P \Big[& B S_B^{-1}(P) B^T + B S_B^{-1}(P) B^T P C (R_v - C^T P C)^{-1} C^T \\
& + C (R_v - C^T P C)^{-1} C^T P B S_B^{-1}(P) B^T \\
& + C (R_v - C^T P C)^{-1} C^T P B S_B^{-1}(P) B^T P C (R_v - C^T P C)^{-1} C^T \\
& + C (C^T P C - R_v)^{-1} C^T \Big] P \Big\} A + Q. \qquad (53)
\end{aligned}$$

The LQDG's value function is then

$$V_0(x_0) = \frac{1}{\lambda} x_0^T (P - Q) x_0 + p^{(1)}, \qquad (54)$$

where

$$p^{(1)} = \frac{1}{1-\lambda} \text{Trace}\left(\Gamma^T P \Gamma W \right) \qquad (55)$$

and P is the applicable solution of the DARE (53).

Remark 1. The following holds.

$$\begin{aligned}
p^{(1)} = & \text{Trace}\left(\Gamma^T \left(\sum_{i=0}^{\infty} \lambda^{k-1-i} P_i \right) \Gamma W \right) \\
= & \frac{1}{1-\lambda} \text{Trace}\left(\Gamma^T P \Gamma W \right). \qquad (56)
\end{aligned}$$

Similar results are obtained for the stochastic LQGDGs 2 and 3.

In conclusion, the temporal discount factor $0 < \lambda < 1$ renders the value of the infinite horizon stochastic games finite and the existence of a solution is exclusively predicated on the DRE (47) not having a finite escape time. The reader is directed to the discussion in Sect. 4.2 concerning the interpretation of the meaning of an infinite horizon game's value and the attendant optimal strategies.

References

1. Tokhi, O. and S. Veres: *Active Sound and Vibration Control*, IEE, London, 2002
2. Basar, T. and P. Bernhard: *H^∞ – optimal control and related minimax design problems: a dynamic game approach*, Birkhhauser, Boston, 2008
3. Engwerda, J.: *LQ Dynamic Optimization and Differential Games*, Wiley, 2005

4. M. Pachter and K. Pham: *Discrete-Time Linear-Quadratic Dynamic Games* (to appear) Journal of Optimization Theory and Applications, Vol. 146, No. 1, July 2010, pp. 151–179.
5. Zhang, Fuzhen: *The Schur Complement and Its Applications,* Springer, 2005. ISBN 0387242716.
6. Pachter, M.: *The LQG Game Against Nature, in Advances in Dynamic Games and their Applications*, P. Bernhard, V. Gaitsgory and O. Pourtallier editors, Birkhauser 2009, pp. 443–453.
7. Brogan, W. L.: *Modern Control Theory,* Prentice-Hall, 1985, pp. 78.
8. E. F. Mageirou: *Values and Strategies for Infinite Time Linear Quadratic Games*, IEEE Trans. on AC, August 1976, pp. 547–550.
9. Bernstein, D. S.: *Matrix Mathematics*, Princeton University Press, 2005, pp. 271.

The Design of Dynamical Inquiring Systems: A Certainty Equivalent Formalization

Laura Di Giacomo and Giacomo Patrizi

Abstract Dynamical systems include measuring sensor inputs of phenomena to yield accurate predictions of the evolving sensor outputs or to determine optimal control management policies based on sensor data. The input and output sets of the system may be generalized and transformed with respect to the sets of sensors available and formal deductive methods and chaos theory may be formulated to obtain Dynamical Inquiring Systems over a horizon to yield solutions which will be precise and be certainty equivalent to the future results of the phenomenon.

The aim of this chapter is to present a formalization of Mathematical Systems Theory to demonstrate the theoretical basis of nonlinear dynamical chaotic systems solved by simultaneous estimation and optimal control processes and to present accurate predictions based on generalized sensor data of many forms both in input and output such as dynamic malfunctioning of systems including engineering, medical, economic, and environmental inquiring systems.

1 Introduction

Inquiringsystems belong to the category of decision systems called teleological, i.e., goal seeking behavior [1, 6] and their design require accurate model formulations [13] even in the presence of significant undetermined aspects [54] by considering deterministic chaotic, instead of stochastic effects over limited time horizons to ensure that the simultaneous estimation and optimization process is accurate [38].

L. Di Giacomo • G. Patrizi (✉)
Dipartimento di Statistica, Sapienza Universita’ di Roma, Italy
e-mail: g.patrizi@caspur.it

V.L. Boginski et al. (eds.), *Sensors: Theory, Algorithms, and Applications*,
Springer Optimization and Its Applications 61, DOI 10.1007/978-0-387-88619-0_6,

Positivist design of inquiring systems was considered by Descartes [9] to formulate the man–machine system so that it could have been used to certify the truth or falsity of singular and clear proposals. The theocratic presupposition consisting in the statement "cogito ergo sum" and its deductive formalization confirms that a positivist realist position was held [6]. The Spinoza inquiring system [46], instead, forms an interpretative methodology [13], as many paradigms may be chosen to perceive the essence of the phenomenon [6]. The Leibnitz proposal is a 'characteristica universalis', a formal universal language to express mathematical, scientific, and derivable representations of phenomena (as indicated in [24]) examined for its logical structure by Gödel, see [8].

System Analysis [53] and General Systems consisted of a rational formulation to specify a priori empirical processes [32] to consider dynamic Input–Output systems as representations of phenomena, characterized by a hierarchy of multilevel systems structure. The inputs and outputs consist, at a conventional infimal and supremal strata, of the appropriate sensor sets, where a sensor is a device that measures a physical quantity and converts it into a signal which can be read by an observer or by an instrument. System Analysis further is characterized by the assumptions of similarity of all systems in a neo-kantian methodological approach, which are required to justify the correctness of models that are proposed [51]. However, these principles limit the methodology to a theory driven realist conception [6] and the System Analysis methodology.

A sensor is predominantly indicated as a physical measurement recording instrument or the measure itself, again usually limited to measures of light, temperature, radiation or many other forms, considered as realist observable term, which must satisfy a number of conditions defined by the theory in which it is applied to ensure that the term conforms with the condition of the theory [18].

Formulating a dynamic system which will result sufficiently precise and certainty equivalent requires to be suitably generalized by removing the limitations conditions of System Analysis and those of sensor input data, since either input elements may not be observable, nor satisfy a realist condition, nor be defined over an interval or ratio scale, but can be extended to nominal and ordinal measurement scales [43]. A dynamic input–output system may be suitably generalized by considering a nonlinear chaotic dynamical structure in the form of a dynamic inquiring system as an instrumental neo-positivist formulation [13], rendered certainty equivalent [10] to distinguish it from realist interpretations or positivist or neo-kantian methodologies.

The aim of this chapter is to formulate inquiring systems consisting of algorithms to estimate and to determine the optimal solution of a nonlinear dynamical chaotic system which can be used to determine predictions of financial crises, vulnerability of structures under seismic events, dynamic malfunctioning of systems and networks in both humans and in machines and many other similar problems and informative systems and summarize the proof of the derivation and precise solutions.

The outline of this chapter is as follows. In Sect. 2 the characterization of the dynamical chaotic system to be formulated will be examined and in Sect. 3 the algorithm schema will be described and its convergence will be proved. In Sect. 4 various implementations will be discussed to show that accurate predictions have been formulated and optimal controls have been achieved in engineering, medical, economic, and environmental inquiring systems, based on generalized sensor data of phenomena. A further section on concluding comments will finalize this chapter.

2 Dynamical Chaos

Empirical Process models may be considered as Data Driven Models rather than Theory Driven Models through an Instrumental methodological approach, to ensure that the derivations of the model are logically correct and the pursued applications are adequate. In Theory Driven modeling, a realist approach is used to formulate implementations, which may not have been formally derived, since they rely on expert opinion and/or anecdotical considerations [13], to align the phenomenon with the accepted explanations and current scientific view [4]. Empirical processes necessarily require nonlinear dynamical systems to ensure that meaningful implementations can be carried out accurately [11].

A model is syntactically correct if it can be cast as a formal system, composed of formal definitions, axioms (or assumptions), derived theorems, incorporating, if need be, other required axiom systems (mathematics, statistics, numerical analysis, etc.). Axioms of the theory must be proved to be complete, independent, and consistent [3]. All propositions derived in the system can be checked for logical consistency and correctness in their derivation. This assures that, if the derivations are also consistent, the application is syntactically correct and will not lead to any contradictory results. In computer science this requirement is known as 'Mathematical Verification'. This may not guarantee that the policy is useful or applicable, since excessive simplifications or unwarranted assumptions may have been introduced to ensure syntactical correctness. Thus the semantical adequacy of the syntactically correct model must be evaluated. A model is semantically adequate if the results of all known legitimate applications of the phenomenon can be reproduced by the model within a given level of accuracy, which is specified a priori, so that it can be considered provisionally valid, until an exception is verified, which will require appropriate extensions of the formulation.

Mathematical System Theory essentially deals with the study of the dynamical relationships of systems under various conditions, more general than those that define difference and differential equation systems [52]. A Dynamical System is a precise mathematical object, and given the flows of the activities of the phenomenon, the input–output relationships must be estimated by appropriate methods. Not every relationship can be modeled by Mathematical System Theory, since a representation that is nonanticipatory is required, while the condition that the functionals are sufficiently smooth, which was previously required, may be wavered. Dynamical Systems may be defined at a high level of generality [11, 27].

Definition 1. A Dynamical System is a composite mathematical object defined by the following axioms:

1. There is a given time set T, a state set X, a set of input values U, a set of acceptable input functions $\Omega = \omega : \Omega \to U$, a set of output values Y and a set of output functions $\Gamma = \gamma : \Gamma \to Y$.
2. (Direction of time). T is an ordered subset of the reals.
3. The input space Ω satisfies the following conditions.
   (a) (Nontriviality). Ω is nonempty.
   (b) (Concatenation of inputs) An input segment $\omega_{(t_1,t_2]}$, $\omega \in \Omega$ restricted to $(t_1, t_2] \cap T$. If $\omega, \omega' \in \Omega$ and $t_1 < t_2 < t_3$ there is a $\omega'' \in \Omega$ such that $\omega''_{(t_1,t_2]} = \omega_{(t_1,t_2]}$ and $\omega''_{(t_2,t_3]} = \omega'_{(t_2,t_3]}$.
4. There is a state transition function $\varphi : T \times T \times X \times \Omega \to X$ whose value is the state $x(t) = \varphi(t; \tau, x, \omega) \in X$ resulting at time $t \in T$ from the initial state $x = x(\tau) \in X$ at the initial time $\tau \in T$ under the action of the input $\omega \in \Omega$. φ has the following properties:
   (a) (Direction of time). φ is defined for all $t \geq \tau$, but not necessarily for all $t < \tau$.
   (b) (Consistency). $\varphi(t; t, x, \omega) = x$ for all $t \in T$, all $x \in X$ and all $\omega \in \Omega$.
   (c) (Composition property). For any $t_1 < t_2 < t_3$ there results:

$$\varphi(t_3; t_1, x, \omega) = \varphi(t_3; t_2, \varphi(t_2; t_1, x, \omega), \omega)$$

   for all $x \in X$ and all $\omega \in \Omega$.
   (d) (Causality). If $\omega, \omega' \in \Omega$ and $\omega_{(\tau,t]} = \omega'_{(\tau,t]}$ then $\varphi(t; \tau, x, \omega) = \varphi(t; \tau, x, \omega')$.
5. There is a given readout map $\eta : T \times X \to Y$ which defines the output $y(t) = \eta(t, x(t))$. The map $(\tau, t] \to Y$ given by $\sigma \mapsto \eta(\sigma, \varphi(\sigma, \tau, x, \omega))$, $\sigma \in (\tau, t]$, is an output segment, that is the restriction $\gamma_{(\tau,t]}$ of some $\gamma \in \Gamma$ to $(\tau, t]$.

The following mathematical structures in Definition 1 will be indicated by:

- the pair $(t, x),\ t \in T,\ x \in X \quad \forall t$ is called an event,
- the values of the state transition function $\varphi(x_t, u_t)$ is called an orbit or trajectory,
- the values of the input function $\Omega = \omega$ which satisfies 3(a)–(b) may indicated as a sensor, if the sensor is a physical measure on an interval or ratio scale which satisfy the conditions to correspond the inputs of the system,
- the sequential values of the readout map $y(t) = \eta(t, x(t))$ is called a time series.

The input space must be nontrivial, so that an initial value of Ω is different from zero, but all the remaining values may be null, which would scarcely define a sensor, although such an experiment is legitimate and produce events. Further, for systems so indicated by Definition 1 the only information that is observable is the time series of system values of the readout map.

Any time series may be represented by a Dynamical System according to Definition 1 and may be characterized as a regular system with or without random disturbances or as a chaotic dynamical system [28]. The dynamical system may not be acquired isomorphically with the representation adopted, since the time series does not necessarily contain information for all states of the system, since some states may be hidden. It may be impossible in principle to reconstruct the state of a stochastic dynamical system from the observed time series, and it is conceivable that the complete history track of the time series is insufficient to determine the present state of the system uniquely [16], so nonlinear nonstationary transformation may be required, which may be neither observable nor realistic, but constructible.

For hyperbolic dynamical systems with small dynamical noise, a shadowing property holds, which means that close to the trajectory of the noisy system there is a trajectory of a noiseless system. A number of shadowing procedures for more general dynamical systems with noise have been proposed [26]. If the shadowing property does not hold for the system considered a noise-free dynamical system representation can still be estimated by using the skeleton property as proposed [5]. The reproduction the features of the original time series will require appropriate modifications of the skeleton by adding dynamical noise features.

An attractor associated with a chaotic motion in state space is not a simple geometrical object as a finite number of points, a closed curve or a torus. It may not be a smooth surface, and it may possess fractal dimensions [33]. The reconstructed attractor has some geometrical structure, endowed with a measure related to the relative frequencies with which different parts of the attractor are visited. The dynamics of a deterministic system can be reconstructed up to a smooth parameter transformation, and the dynamics so characterized. Dimensions and entropies can be used to characterize the attractor and is invariant under the reconstruction procedure. The sensitivity of a system to the perturbation to the initial state of the system, so that chaotic evolution might occur, may be determined through the calculation of the Lyapunov exponents of the nonlinear system [28, 35]. The main algorithm consists in embedding the observations in a suitable dimensional space, which may be determined iteratively, to use the observations obtained to reconstruct the dynamics of the attractor [47]. The Jacobian of the reconstructed dynamics is then used to calculate the Lyapunov exponents of the unknown dynamics [17].

The theory of dynamical systems has been formulated by considering deterministic nonlinear stationary time series, while empirical processes may be subject to nondeterministic influences and are often nonstationary. Special methods can be applied to reduce a given process to the required form, but these procedures are limited and not generally applicable [16]. The reconstruction theorem will not apply in the presence of dynamical noise, unless suitable transformations are introduced as in an Instrumental Data Driven approach.

A chaotic dynamical system may be represented by applying an instrumental approach and a syntactically correct derivation and the model should result semantically adequate over the past time interval. The synergies and suitable inputs are determined by the estimation process as required to attain the necessary desired precision.

The accuracy of a proposed model is verified experimentally by comparing the prediction obtained with the effective outcome. Various procedures have been defined to determine which system alternatives should be considered, depending on the characteristics of the orbits of the time series [20, 28, 48].

A Chaotic Dynamical System is characterized by the following conditions [23]:

(1) the initial conditions influence the state of the system so that each point in such a system is arbitrarily closely approximated by other points with significantly different orbits,
(2) the evolution of the system over time is topologically mixing if any given point or region of the phase space will eventually overlap with any other given region,
(3) periodic orbits are dense, which requires that every point in the state space is approached arbitrarily closely by periodic orbits.

The properties or proposed axioms should be syntactically correct, which implies that they are independent [3], so logical contradictions in the derivations must not occur [13]. However, it has been shown that property (2) implies properties (1) and (3) [50], while properties (2) and (3) imply property (1) [19], which is often indicated as the butterfly effect [29]. Consequently, as there is no precise definition for a chaotic solution to a Dynamical System [28] because it cannot be represented through standard mathematical functions, since it is aperiodic, the theory driven models cannot be implemented, although such systems can be characterized with certain identifiable characteristics as a bounded steady-state behavior that is not an equilibrium, periodic, or quasi-periodic solution [33].

Adopting an Instrumental Data Driven approach and emphasizing syntactical correctness and semantic adequacy of the representation, suitable models of the time series considered can be derived and permit the formulations of an inquiring system. It should be stressed that the approach formulated in the following section will derive an algorithm to achieve this goal, abstracting from any realist theory formulation dependent on positivist or neo-kantian methodologies [13].

3 Properties and Convergence of the Algorithm

Consider an empirical process consisting of a given time series of measurements to be estimated $\hat{y}_t (t = 1, 2, \ldots, T)$, where it is assumed that $y_t \in R$ is a historic time series. Let $x_t \in R^m$ be an m-dimensional vector of state variables of the dynamic process whose dimension must be estimated. Let $u_t \in R^q$ be a q-dimensional vector of control variable for period t, which may be identically null, constituting in this case an autonomous system. It is desired to determine suitable functional forms $\varphi : R^m \to R^m$ and $\eta : R^m \to R$ and a set of suitable coefficients $\Theta \in R^p$ where $p \geq 2m$ which may be much larger than the dimension of the state space, because the maps may be nonlinear. The optimal control vectors $\{u_t | t = t + 1, \ldots, \mathcal{T}\}$ are to be determined over the desired horizon. Also $v_t = \hat{y}_t - y_t$ is a residual stochastic process to be determined, where y_t is the estimated readout value of the time system represented by the system, which will be as close as possible to the historic time series values.

The aim of this section is to describe the algorithm to solve these problems, to prove that the statistical conditions will hold and provide correct solutions, and to prove that the algorithm will converge under general conditions. Finally, it will be shown that if the properties of such a system hold, then a dynamical inquiry system has been formulated, and it will provide certain equivalent results to phenomena.

3.1 Description of the Algorithm

The dynamical system may be represented as follows:

$$x_{t+1} = \varphi(x_t, u_t : \theta_1), \tag{1}$$

$$y_{t+1} = \eta(x_t : \theta_2), \tag{2}$$

where a functional form, state space vectors must be determined appropriately. The historical series of the system may be represented as:

$$\{\varphi(x_0, u_0 : \theta_1), \varphi(\varphi(x_0, u_0 : \theta_1), u_1 : \theta_1), \ldots, \varphi(\varphi(\ldots,), u_1 : \theta_1)\} \tag{3}$$

in obvious notation. Without loss of generality, it is less complex in terms of notations to indicate each element of the vector given in (3) in the original version x_t. Suppose that the dimension of the state space is m, so m initial state variables must be determined, while the number of parameters that must be determined will depend on the nonlinear structure of the functional space. If it is supposed that the number of parameters of nonlinear functional form is q, then more than $q + m$ values of the time series are required to estimate the system.

As the system to be determined will be overfitted, residual errors may occur, so let $v_t = \hat{y}_t - y_t$ $(t = 1, 2, \ldots, T)$. It is desired to minimize the error between the actual and the estimated values.

The number of periods to predict will depend on the characteristics of the system. Suppose that there are T historical time periods values of the process and it is desired to predict the values of the next $\mathcal{T} - T$ periods, where $\mathcal{T} > T$. The mathematical programming problem to determine the dynamical systems can be formulated as follows:

$$\text{Min} \quad J = \sum_{t=T+1}^{\mathcal{T}} (\hat{y}_t - y_t)^2 \tag{4}$$

$$\text{s.t.} \quad x_{t+1} = \varphi(x_t, u_t : \theta_1) \tag{5}$$

$$\hat{y}_{t+1} = \eta(x_t : \theta_2) \tag{6}$$

$$v_t = \hat{y}_t - y_t \quad t = 1, 2, \ldots, T \tag{7}$$

$$\frac{1}{T}\sum_{t=1}^{T} v_t = 0 \tag{8}$$

$$\frac{1}{T}\sum_{t=1}^{T} v_t^2 \le k_v \tag{9}$$

$$-\epsilon_2 \le \frac{1}{T}\sum_{t=1}^{T} v_t v_{t-1} \le \epsilon_2 \tag{10}$$

$$, \dots\dots\dots\dots\dots\dots$$

$$-\epsilon_{2s+2} \le \frac{1}{T}\sum_{t=1}^{T} v_t v_{t-s} \le \epsilon_{2s+2} \tag{11}$$

$$\frac{1}{2} g_v^T \Psi \left(\Psi^T \Psi\right)^{-1} \Psi^T g_v - \frac{T}{2} \le \chi^2_{1-\alpha:p-1} \tag{12}$$

$$-\epsilon_{2r+1} \le \frac{1}{T}\sum_{t=1}^{T} v_t^{2r+1} \le \epsilon_{2r+1} \quad r = 3, 4, \dots \tag{13}$$

$$\frac{1}{T}\sum_{t=1}^{T} v_t^{2r} \le \frac{2r!}{r!2^r}\sigma_v^{2r} \quad r = 3, 4, \dots \tag{14}$$

$$x_i \in X, y_i \in Y, v_i \in V, x_0. \tag{15}$$

The dynamical system is estimated by the system of equations between (5) and (15). For autonomous systems a suitable cost function may be considered with respect to the readout and control values by minimizing the objective function (4).

A number of statistical conditions must be satisfied, which are set up as constraints of the optimization problem so that the parameters to be determined are defined implicitly by the problem. The solution of the mathematical programming problem, which is nonconvex as the constraints are nonlinear and nonconvex, yields estimates of the parameters such that estimated values of the time series are as close as possible to the values of the time series. All the available information is applied, and the uncertainty of the estimates and the data fit is reduced to the maximum extent possible. Thus the estimates of the parameters satisfy all the statistical conditions, so they are the 'best' possible in a 'technical' sense [30], which will always be determined, if the mathematical programming problem is feasible.

Given that a time series with $T > (q + m)$ elements is available, then the state space variables may be determined by solving the nonlinear terms indicated in (3) substituted in the system (4)–(15), which will hold under mild conditions.

Theorem 1 ([47]). *Let $X \in \mathbf{R}^n$ be a compact set of dimension m. For pairs (φ, η), $\varphi : X \rightarrow X$ a smooth bijective map and $\eta : X \rightarrow \mathbf{R}$ a smooth function. It is a generic property that the map $\Psi_{(\varphi,\eta)} : X \rightarrow \mathbf{R}^{2m+1}$ is an embedding.*

The statistical conditions that must be satisfied by the system [25, 30] are:

1. the parameter estimates are unbiased,
2. the parameter estimates are consistent,
3. the parameter estimates are asymptotically efficient,
4. the residuals have minimum variance,
5. the residuals are unbiased (have zero mean),
6. the residuals have a non informative distribution (usually, the Gaussian distribution). If the distribution of the residuals is informative, the extra information could be used to reduce the variance of the residuals or their bias to obtain better estimates.

These properties are ensured by solving the problem including the constraints (8)–(15).

Theorem 2. *Let the constrained minimization problem (4)–(15) have an optimal solution, then the residuals* $\{v_t | t \in \{1, 2, \ldots, T\}\}$ *have zero mean, are serially uncorrelated and homoscedastic with finite minimum variance.*

The theorem states that conditions 4 and 5 hold for the model and the data can always be satisfied. The next lemma, and in particular the corollaries which follow, prove that conditions 2 and 1, given the results of Theorem 2, hold, so the estimates of the parameters are consistent and unbiased.

Lemma 1 ([7]). *Any rational function or power of a rational function of the sample moments, converges in probability to a constant obtained by substituting throughout the corresponding population moments, provided that the latter exists and that the resulting expression is well defined.*

Corollary 1. *Let the constrained minimization problem (4)–(15) have an optimal solution, with minimum values for the variances of the residuals, as the sample size increases, then the constraints (8)–(12) will tend to their constant population values.*

Corollary 2. *If the constrained minimization problem (4)–(15) has an optimal solution, the solution* θ_n^* *is an unbiased estimator of the population value.*

The constraints (13)–(14) are sample moments of the probability distribution function of the residuals which are made to assume given values in terms of the variance σ^2 and its higher powers. These constraints enforce the residuals to have a noninformative distribution, here a Gaussian, a fact reinforced by the next result.

Theorem 3 ([30]). *Let the constrained minimization problem (4)–(15) have a solution and let the regression function and its derivatives up to the third order with regard to all arguments be bounded; then* $\left(\tilde{\theta} - \theta\right)\sqrt{n}$ *is normally distributed as the sample size* $n \to \infty$.

The conditions 6 and 3 will hold in all cases that the constrained minimization problem (4)–(15) has a solution, as the next theorem shows.

Theorem 4 ([30]). *Let the constrained minimization problem (4)–(15) have a solution, then the estimator* $\tilde{\theta}$ *is asymptotically efficient.*

In traditional Theory Driven modeling procedures, the functional form is assumed to be known and suitable parameters must be estimated, which should satisfy the statistical conditions indicated above [25,30] to be valid. Thus the optimal control solution of the system (4)–(15) will dominate the Maximum Likelihood solution of any Theory Driven model specification of the dynamic system estimation of the time series considered.

Theorem 5. *Let the maximum likelihood solution of the theory driven formulation of the dynamic system have a unique solution and consequently all the asymptotic properties of nonlinear least squares estimates are met [25] then this solution will be equal to the solution of the constrained optimization problem (4)–(15), but not conversely .*

Proof. The constrained optimization problem will determine a global minimum to the objective function, which satisfies the statistical conditions necessary for a solution to be unbiased. If the Maximum Likelihood unique solution of the theory driven formulation exists, then it must be identical to the optimal control solution. Otherwise the maximum Likelihood solution is biased, because one or more statistical conditions are not satisfied and then the value of the Maximum Likelihood function may be less than the global solution found, because not all the constraints are respected, or the variance of the Maximum Likelihood solution is larger than the global minimum value, because theory driven considerations have imposed additional conditions, which limits the value determinable. In these cases the Theory Driven solution is either incorrect because the statistical conditions are not respected, or the found solution has a value of the variance, which is larger than the one of the optimal control solution. □

The feasibility of the system may be ensured by relaxing the coefficients on the right hand side of the constraints that are infeasible. By modifying these coefficients and also carrying out binary search techniques on those that are feasible, and recursing on the functional form of the system, better and better fits can be obtained. At each iteration, the best combination of the parameters and functional forms are derived by solving the optimization problem, and by subsequent iterations the objective function, if possible, is reduced. A global minimum will be determined when all subsequent recursions yield infeasible solutions.

3.2 Mathematical Convergence of the Algorithm

The general convergence of the system may be summarily presented by expressing the system (4)–(15) in the following way:

$$\text{Min} \quad Z = f(w) \qquad f : R^n \rightarrow R, \tag{16}$$

$$g(w) \geq 0 \qquad g : R^n \rightarrow R^p, \tag{17}$$

$$h(w) = 0 \qquad h : R^n \rightarrow R^q. \tag{18}$$

The proposed algorithm consists in defining a quadratic approximation to the objective function, a linear approximation to the constraints and determining a critical point of the approximation by solving a linear complementarity problem, as given in [37].

Expanding the functions in a Taylor series, at the given iteration point w^k, the equality constraints may be eliminated simply by converting them into $p + 1$ inequality constraints. Thus:

$$h(w) = h\left(w^k\right) + \nabla h\left(w^k\right)\left(w - w^k\right) \geq 0, \tag{19}$$

$$-e_q^T h(w) = -e_q^T\left(h\left(w^k\right)\right) + \nabla h\left(w^k\right)\left(w - w^k\right) \geq 0. \tag{20}$$

A set of trust region constraints can be imposed on the problem as a system of linear inequalities centered around the iteration point, to limit the change in the possible solution:

$$Dx + d \geq 0, \tag{21}$$

where $D \in R^{n\times n}$ is a suitable matrix which may be changed at every iteration and $d \in R^n$ a suitable vector. These can be included in the inequalities, so the problem to be solved iteratively is:

$$\begin{aligned}\text{Min}\ \ f(x) = {} & f\left(x^k + e_n\zeta\right) + \nabla f\left(x^k + e_n\zeta\right)\left(x - x^k\right)\\ & + \frac{1}{2}\left(x - x^k\right)^T \nabla^2 f\left(x^k + e_n\zeta\right)\left(x - x^k\right)\end{aligned} \tag{22}$$

$$\text{s.t.}\ \ g(x) = g\left(x^k + e_n\zeta\right) + \nabla g\left(x^k + e_n\zeta\right)\left(x - x^k\right) \geq 0 \tag{23}$$

$$x \geq 0, \tag{24}$$

where $g : R^n \rightarrow R^{n+p+q+1}$. The local solution of the mathematical programming problem (22)–(24) leads to one of three cases. If the point is inside the trust region, then it is an approximate stationary point. If the point is on a trust region constraint and the point is feasible while a reduction in the objective function has occurred, the point is taken as the new starting point and a new iteration is commenced. Otherwise, if the new point is infeasible, the trust region is reduced. Finally if there has been an increase in the objective function, the trust region is enlarged and the iteration is repeated, with suitable safeguards to provoke a reduction in the objective function.

The problem can be written, without loss of generality as given in (22)–(24) and the Kuhn–Tucker points for this problem will be given by determining suitable solutions to the following nonlinear complementarity problem:

$$F(z) \geq 0, \tag{25}$$

$$z \geq 0, \tag{26}$$

$$z^T F(z) = 0. \tag{27}$$

This problem can be written equivalently as a variational inequality, since both are defined over a convex set,

$$F(z)^T(y-z) \geq 0. \tag{28}$$

Consider the application $F : R^n \rightarrow R^n$ and expand it in a Taylor series around a point $z' \in R^n$ to get:

$$F(z) = F\left(z'\right) + \nabla F\left(z'\right)\left(z-z'\right), \tag{29}$$

then for any $\varepsilon_1 > 0$ there exists a scalar $r > 0$ such that:

$$\left\| F(z) - F\left(z'\right) + \nabla F\left(z'\right)\left(z-z'\right)\right\| \leq \varepsilon_1 \left\|z-z'\right\|, \quad \forall \left\|z-z'\right\| \leq r, \tag{30}$$

as it has been proved in [15].

Thus, in a small neighborhood, the approximation of the nonlinear complementarity problem by a linear complementarity problem will be sufficiently accurate, so that instead of solving system (25)–(27), the linear complementarity system approximation can be solved. Recall that by construction, the subspace of the Euclidean space is bounded and closed, so the nonlinear complementarity problem can be solved by its linear complementarity approximation. Every linear complementarity problem can be solved, or a solution can be shown not to exist by solving an appropriate parametric linear programming problem in a scalar variable [37]. The algorithm will find the solution of the linear complementarity problem, if such a solution exists, such that $\|x\| \leq \alpha$ for some constant $\alpha > 0$, or declare that no solution satisfying this bound exists. In this case the bound can be increased. The convergence of the algorithm can now be demonstrated. Consider a point $x' \in R^n$ such that $F(x') \geq 0$ and therefore feasible. Determine a neighborhood, as large as possible, which can be indicated by:

$$Q = \left\{z \mid \left\|z-z'\right\| \leq r\right\}, \tag{31}$$

where r is the coefficient defined above in (30). Suppose that the acceptable tolerance to our solution is ε_2 so that if $(z^*)^T F(z^*) \leq \varepsilon_2$ then the solution is accepted. In this case, impose that:

$$\varepsilon_1 r \leq \frac{\varepsilon_2}{\alpha}. \tag{32}$$

The local convergence of the algorithm is established in the following theorem.

Theorem 6. *If the linear complementarity problem has a solution z^* where all the trust region constraints are not binding, then such a solution is also a solution to the nonlinear complementarity problem (25)–(27) for which $F(z^*) \geq 0$ and $(z^*)^T F(z^*) \leq \varepsilon_2$.*

Theorem 7. *The objective function is bounded below by zero for the nonlinear optimization problem (4)–(15) and the problem has a feasible solution so that a sequence of feasible solutions is determined which exhibit strictly decreasing values of the objective function through modifications of the trust region. If there exists a feasible solutions x^* which satisfies the hypotheses of Theorem 6 but no other feasible solution can be determined with a strictly lower objective function satisfying the same hypotheses, then that solution x^* is a global minimum of the problem.*

Proof. By Theorem 6 each solution to the LCP is an approximate solution to the nonlinear complementarity problem (25)–(27). A connected set exists such that the nonlinear complementarity problem has a solution within the trust region or on a constraint.

From this sequence choose a subsequence such that the value of the objective function, as given by (4). Since the objective function is bounded not all the solutions can lie on some trust region constraint, so a solution of the nonlinear complementarity problem which lies within the trust region constraints must eventually be determined, if such a solution exists. Let this solution be a local minimum to the nonlinear optimization problem. By repeating this procedure the global minimum will be determined. □

The transient function, readout function, and state vectors yield estimates of the time series very close to the actual values of the time series by construction, but nonzero random variables may also be present with zero mean value and finite variance.

Therefore, consider the time series composed of the residuals of the estimation of the system above, which is a sequence of random variables, say $z_t (t = 1, 2, \ldots, T)$ and apply the system (4)–(15) specialized to time series of the residuals. A dynamic chaotic nonlinear function can be estimated, determined, and constructed for the time series that has a null mean value over the horizon of a finite variance and is stationary. Moreover, u_t is taken to be identically zero and without loss of generality, the objective function is given as:

$$\text{Min} \quad J = \sum_{t=T+1}^{\mathcal{T}} (\hat{y}_t - y_t)^2 \,. \tag{33}$$

The derivations given above demonstrate that the system can be estimated and in particular by Theorem 1 the required embedding can be determined, irrespective of the possibility of formulating a Chaotic system as a theory driven model, so this system can be taken to be homomorphic to the sequence of residuals which are random variables by construction, and all the conditions of the theorems are satisfied, so the estimated system can be taken to be a chaotic system and these dated predicted chaotic values over the immediate future can be added to the predictions formulated on the basis of the initial system, yielding aggregate predicted values of the phenomenon over a close immediate future [54, 55].

3.3 Certainty Equivalence of Solutions

A sufficiently general representation of a dynamic system may be formulated, with a slight abuse of notation, in the following way:

$$x_{t+1} = \varphi(x_t, u_t), \tag{34}$$

$$y_t = \eta(x_t). \tag{35}$$

Dynamical systems are based on intermediary set of states and transition functions, by applying the simultaneous estimation and optimization algorithm to determine the state set X and the transition function [45]. Under appropriate conditions the representation of the system will result unique, as it is indicated below.

Definition 2. Given two states x_{t_0} and $\hat{x}_{t_0}$ belonging to systems S and $\hat{S}$ which may not be identical, but have a common input space Ω and output space Y, the two states are said to be equivalent if and only if for all input segments $\omega_{[t_0,t)} \in \Omega$ the response segment of S starting in state x_{t_0} is identical with the response segment of $\hat{S}$ starting in state $\hat{x}_{t_0}$, that is,

$$x_{t_0} \cong \hat{x}_{t_0} \Leftrightarrow \eta(t, \varphi(x_{t_0}, \omega_{[t_0,t)})) = \hat{\eta}(t, \hat{\varphi}(\hat{x}_{t_0}, \omega_{[t_0,t)})) \quad \forall t \in T, t_0 \leq t, \forall \omega_{[t_0,t)} \in S, \hat{S}. \tag{36}$$

The systems S and $\hat{S}$ may be two representations of a phenomenon.

Definition 3. A system is in a reduced form if there are no distinct states in its state space which are equivalent to each other.

Suppose the actual system is in a reduced form.

Definition 4. Systems S and $\hat{S}$ are equivalent $S \equiv \hat{S}$ if and only if to every state in the state space of S there corresponds an equivalent state in the state space of $\hat{S}$ and vice versa.

Definition 5. Simple and multiple experiments involve different sets of input/output pairs:

- A simple experiment is an input/output pair $(u_{[t_0,t)}, y_{[t_0,t)})$, that is, given the system in an unknown state an input $u_{[t_0,t)}$ is applied over the interval of time (t_0, t) and the output $y_{[t_0,t)}$ is observed.
- A multiple experiment of size M consists of M input/output pairs $(u^i_{[t_0,t)}, y^i_{[t_0,t)})$, $i = 1, 2, \ldots, M$, where on applying on the ith realization of the M systems the input $\left(u^i_{[t_0,t)}\right)$, the ith output $y^i_{[t_0,t)}$ is observed.

While a simple experiment is thought as a stimulus and response experiment, multiple experiments are more complex and define multi-determined reactions, such as synergisms of the system.

Definition 6. A system is simply (multiply) observable at state x_{t_0} if and only if a simple experiment (a multiple experiment) permits the determination of that state uniquely.

Definition 7. Equivalence of dynamic systems can be distinguished:

- Two systems are simply equivalent if it is impossible to distinguish them by any simple experiment.
- Two systems are multiply equivalent if it is impossible to distinguish them by any multiple experiment.

Theorem 8 ([27]). *If two systems are multiply equivalent then they are equivalent.*

Definition 8. A system is initial-state determinable if the initial state x_0 can be determined from an experiment on the system started at x_0.

Theorem 9 ([27]). *A system is in reduced form if and only if it is initial-state determinable by an infinite multiple experiment.*

Definitions 5–8 and the results given in Theorems 8 and 9 formally justify the possibility of defining one or more representations of the dynamical system. The distinction between systems that are simply equivalent and multiply equivalent is crucial, as comparative static or equilibrium models will be simply equivalent, while for the analysis of dynamical systems which are multiple equivalent allow to compare different representations and determine the optimal trajectory for the system.

Thus consider the following definition.

Definition 9. An ex ante solution is a solution formed in a given period t based on anticipated outcome of future activities maturing in period $t + 1$ and consists of the forecast of the optimal values of the control variables.

Definition 10. An ex post solution is a solution formed in a given period t, regarding outcomes of activities maturing in period $t + 1$ which are assumed known (with fore-knowledge), in period t so as to determine the optimal values of the control variables in period $t + 1$.

Let a phenomenon be represented by a multiply equivalent dynamic system in an initial-state determinable and the state in period $t + 1$ be predicted in period t. The state at $t + 1$ may be obtained from a representation of another copy of such a system in the period $t + 1$. The state of period $t + 1$ predicted in period t on the basis of the knowledge in the same period, will be equivalent to the state at period $t + 1$ on the basis of the knowledge at period $t + 1$, which is indicated as the ex post state.

Theorem 10. *An ex ante multiply equivalent Dynamic system in an initial-state is equivalent to an ex post system, also a multiply equivalent dynamic system in an initial-state.*

Proof. By Theorem 8 the ex ante dynamical system is multiply equivalent and its initial state is in reduced form by Theorem 9.

Furthermore by the same reasoning the ex post dynamical system is also multiply equivalent and its initial state is in reduced form. □

By Definitions 3 and 4 they are state equivalent.

Corollary 3. *A Certainty Equivalent solution consists of an optimal solution to an optimization problem determined as an ex ante solution in the control variables u^* which are also optimal ex post.*

The state of the systems are equivalent if the state transition functions do not exhibit significant random variation. The eventual random variation will be negligible, but the expected value of the disturbances in both cases will assume null values while the variance of the processes may be positive, but the autocorrelation and cross-correlation must be zero and the processes are stationary.

This formulation is analogous to the properties derived earlier by different methods [44, 49].

No limitations have been enforced for the output equations (35). The random disturbances of these forms may vary. The resulting outcome, as an output variable may differ due to disturbances, although optimal solution values of the control variables will be identical in the two processes.

4 Implementations

The objective of this section is to describe a number of implementations to show how the the algorithm may be used to design inquiring systems to determine accurate predictions of phenomena, how to use the algorithms in an inferential mode and to achieve optimal control of phenomena.

4.1 Economic Implementations

The study of financial quotations may be useful to inquire in the status of the economy and to determine the economic dynamics of certain sectors of the economy.

To this end, certain key quotations are often used as predictors, such as VIX index, various interest rates and the foreign exchange quotations. A realist observable sensor may be a single quotation, say at closing time or a certain set of quotations, but this leads to difficulties of realism and to observe them continuously. Thus the concept of a sensor becomes more complex and should be in all cases defined precisely.

Typical questions to be posed in inquiry is the prediction of a set of future events similar to other historical events, see Theorem 7, and appropriate classification of similar events in suitable classes. It is important in such implementations to define precisely the characteristics of attributes common to what is considered in

Table 1 Weekly Predictions five weeks ahead for financial future indices

Week	SPX (1)		Stoxx (2)		Nikkei (3)		VIX (4)	
Date	Actual	Prediction	Actual	Prediction	Actual	Prediction	Actual	Prediction
08/13	1190.16	1190.08	2792.47	2792.68	11735.06	11732.36	20.42	20.55
08/20	1161.97	1161.94	2728.47	2728.25	11445.54	11444.93	22.87	23.84
08/27	1184.93	1184.94	2886.97	2887.70	10811.37	11167.05	20.56	19.71
09/03	1133.58	1133.41	2898.36	2898.14	10729.60	10715.34	25.85	24.94
09/10	1085.78	1085.68	2897.37	2898.14	10395.10	10518.82	31.84	31.06
09/24	965.80	965.24	2940.94	2941.19	9637.60	9559.98	37.75	42.48
10/01	1040.94	1040.62	2978.50	2978.48	9766.75	9777.52	32.32	31.84
10/08	1071.38	1071.72	2993.54	2993.48	10143.39	10206.37	35.12	33.47

a similarity set and those that differ between sets. Also given a set of objects, an inquiring system should determine, which attributes from pertinent lists must be considered to include these objects in a suitable identifiable similarity set.

A major crisis followed the terrorist attack on New York and Washington on September 11th 2001. The New York Stock Exchange was closed for one week and the predictions during that period are important. The period considered extends from July 6th 2001 for 20 weeks. A second crisis followed on September 15th 2008 when the bankruptcy protection was filed by Lehman Brothers. To each period of crisis, ten prior periods were added to the periods to be used in predictions since at least 5 periods are required to initialize the predictions. For a number of future index contracts the quotations were predicted 5 weeks ahead. Twenty periods were transferred in the training set and placed at the end of the series to form a verification set. There are two strong discontinuities: the first because of the transfer of the 20 periods and the second occurs at the start of the verification set, but this discontinuity is partially annulled because of the use of additional 5 periods as indicated above [12].

We consider the following futures indices contract quotations:

1. The Standard and Poor 500 common stock index, (SPX).
2. The Dow Jones Euro stock index, consisting of 50 stocks expressed in Euros, (SX5E).
3. Nikkei 225 Stock Average, (NKY).
4. Chicago Board Options Exchange: Volatility Index, (VIX)

Predictions were formulated weekly after closing time on Friday, to indicate the closing quotation prediction on the next Monday evening for 5 subsequent Mondays. The predictions that had been made 5 weeks before are given in the row corresponding to the forecast date in Table 1 for the four future indices considered.

These indices tend to vary analogously over long intervals of time, except that VIX tends to vary in the opposite fashion. For the initial 3 weeks reported the 4 indices seem to confirm this generalized behavior. For the predictions for 09/03 there is a 4% reduction in SPX index, no marked reductions in the European indices, a fall less than 3.2% in the Nikkei index and an increase of over 6% for the VIX index.

Table 2 Weekly predictions 5 weeks ahead for future indices for Lehman Brothers' bankruptcy

Week	SPX (1)		Stoxx (2)		Nikkei (3)		VIX (4)	
Date	Actual	Prediction	Actual	Prediction	Actual	Prediction	Actual	Prediction
08/26	1290.47	1292.26	3312.41	3312.12	12797.54	12876.65	18.78	18.81
09/02	1287.83	1282.89	3365.63	3365.75	12936.81	12834.38	20.65	20.64
09/08	1249.50	1242.20	3185.83	3184.53	12359.93	12626.53	22.22	23.05
09/15	1250.92	1255.28	3151.17	3151.08	12028.45	11626.53	25.66	31.10
09/22	1255.37	1213.25	3253.52	3253.77	12037.89	12113.62	32.40	32.08
09/29	1209.07	1098.27	3156.46	3155.75	10883.25	11892.44	36.92	34.76
10/06	1097.56	897.56	3113.82	3113.84	10817.27	10935.46	45.12	45.00
10/13	912.75	944.02	2421.87	2415.31	8407.94	8288.18	69.95	68.26

Consider Table 1, comparative analysis of the quotations as between the European, Japanese, and US indices suggests that a financial crisis, limited to the US was apparent in the period between 09/10 and 09/24 so that the predictions 5 weeks before were available by 08/13 and 08/27 as at that time predictions would have been made for the Monday 17th September. The data cannot imply that a terrorist attack was being planned and then effected, but rather it is obvious the predicted quotations can be tested for statistical significance in potential abnormal stock movements which will be partially recuperated in 10/01 and 10/08. Hence, some abnormal events were foreseen by some experts in the US, less so in Japan and not at all in Europe.

It is not the trajectories of any particular series or sensor which can indicate the financial crisis but the analysis of a set of quotations in time together with the country, type of quotation and relative movements. It is unlikely that such an analysis may be classed as the study of a set of sensors, but rather firstly the design, secondly the analysis of an inquiring system. This requires a study of the semantical adequacy of the design, once its syntactical correctness has been established.

In Table 2 similar results as in the previous event can be evidenced for the latter period. The bankruptcy was filed on 15th of September 2008, but little effect seems to have occasioned in the markets for three weeks, except for a marked increment in the VIX index quotations. Instead at the start of the fourth week (13th October 2008) a decrease over 20–30% in the quotations was predicted which in fact occurred. From the scanty data, it would seem that a dynamic effect was in action.

Once again it is apparent that the inquiring system works well in the predictive aspect, especially as these predictions were formulated 5 weeks before the events, which occurred. Difficult inferential processes should be studied to determine if speculative effects occurred in the interval from September 15th to October 13th or if the restrictions demanded on the credit available had a delayed effect.

4.2 *Medical Implementations*

In-vitro fertilization is widely used to treat infertility, although the rate of success of treatment is low, with an average overall birth rate per cycle of treatment of

Table 3 Embryo classification results: precision in the verification sample, 2 classes, 5 moments per instance, see reference [41]

	Mean	s.e.e.	Best	Worst	Number
Verification Results	0.8406	0.003645	0.9383	0.7531	801

13.9%, [41]. Various factors have been identified as affecting this rate, such as the age group of the women concerned, the duration of infertility, the usage of donor's eggs and perhaps the treatment center, a uterine factor, and other subsidiary aspects. All these measurements and evaluations could be considered as sort of sensors of the pathology, but the results were very limited, so typically a completely different approach was adopted by designing an appropriate inquiry system based on the properties of the images of the embryos [31].

The procedure used photographs of 801 embryos at the 4 cell stage, taken for 40–50 h after fertilization and before transfer, by placing each embryo under a microscope (Inverted Microscope Olympus IX 70), with a camera (videocamera JVC TK-C401EG). Characteristics are defined from these images, by regarding each image as an array of intensities of shades of gray, from which the frequency distribution of the gray tones for each picture, in the horizontal and vertical direction can be calculated. This pixel-intensity profile indicates the homogeneity of the image, or whether it has many dark and light spots and how they are distributed. From these distributions a certain number of moments (polynomial functions of central tendency and spread) can be calculated, to express the shape of the distribution in a standardized way. Thus a suitable number of these parameters 10, 20, or 30 can be used to define the pattern vector for each image. The results here are given only for the 10 element pattern vector (5 moments in each coordinate direction), as this sample proved to be the most accurate, but one can see [21] for other results.

In a training set the outcome class could initially only be assigned with certainty in the case of single embryo transfers, or when all the transferred embryos produced births, or when no births occurred as a result of that transfer. These are termed sure instances. Multiple transfers involving more births may be troublesome. Generally it is difficult to determine which one of a set of embryos transferred is responsible for the birth. Thus if 3 embryos are transferred and one baby is born, the embryo that has given rise to the birth must be determined.

The classification algorithm is applied to these sure instances and consists of a nonlinear complementarity algorithm [40] which is just a variant of the algorithm described and determines an estimate of the classifier which can be improved by enlarging the classification sample. Once a classifier has been obtained on the sure instances, all the rest of the embryos can be classified on the basis of this classifier. This will assign a class label to every embryo in the given set and suitable constraints must be given to ensure that the correct number of embryos of the given type are assigned in training.

Better results can be obtained by enlarging the training set [40].

Analogous results have been obtained in various medical implementations in which the input data files must be considered general inputs, defined appropriately for every implementation, so all these applications confirm the utility of designing inquiry systems according to this instrumental methodology. Consider, for instance the diagnosis of Alport Syndromes [39], a diagnostic classification of geriatric disorders [42], a classification of numerous well known pathologies [34], the analysis of dynamic medical treatment of pathologies, such as ECG data [2] and solving large protein secondary structure classification problems [40].

4.3 Environmental Implementations

Following a major seismic catastrophy in Southern Italy on November 23, 1980 a sample of the damages incurred to buildings were accurately classified so that more than 37,863 buildings were examined, consisting of over 57,000 dwellings, 257,000 rooms and over 44,000 other constructions [14] which must be considered a very large set of sensors to measure this phenomenon.

Damages were incurred in a municipality S.Angelo dei Lombardi destroying 20.06% of the buildings and 45.3% were heavily damaged, while in a nearby township the percentages were respectively 3.22% and 22.5%. Clearly many factors must be involved if the phenomenon is so varied and it was evidenced that the difference in damages cannot be imputed to random elements [14]. The sample consisted of 41 out of 639 municipalities. Eight levels of damages were considered, defined explicitly from ‘no damage’ to ‘partially destroyed’ and finally ‘destroyed’ and categories each group was aggregated in 4 major categories: ‘building material’, ‘height of building’, etc., for each building in the sample.

The association present in the resulting data set can be analyzed by cross-classified categorical methods [22] which result in a nonlinear estimation problem and can be solved in the traditional ways, or the algorithm described above can be applied.

The relative frequencies of damages differ greatly from one municipality to another, even if they are close by, because of many factors such as building material and the height of the buildings, and many other factors indicated in [14], so preventive security policies should be enacted. Through this analysis the vulnerability of each building can be estimated, and preventive structural modifications can be enacted. For instance, constructions of buildings of loose stone should be kept of low height and possibly enforced though elastic supports and additional floors should never be allowed on such type of structures, which in fact were allowed by Italian authorities.

Such an inquiry system allows to accurately modify the most potentially dangerous constructions and is a typical implementation of the design of an inquiry system under an instrumental methodology, and the interested reader can examine additional relevant implementations [36].

5 Conclusions

The design of inquiring systems provides a useful platform to determine accurate predictions and to classify or recognize or acquire knowledge about objects both in an inferential mode or in a taxonomic fashion.

The acquisition of information should not depend on a priori affirmations or on anecdotical suggestions but on the knowledge that can be grasped from data. This instrumental approach is purely provisionally correct, essentially the results can always be generalized and improved, but the objective of the instrumental approach is to determine the best possible solution to the system, if possible, without falling back to unobservable relationships.

References

1. R. L. Ackoff and F. E. Emery. *On Purposeful Systems*. Aldine, Chicago, 1972.
2. F. Bartolozzi, A. De Gaetano, E. Di Lena, S. Marino, L. Nieddu, and G. Patrizi. Operational research techniques in medical treatment and diognosis: A review. *European Journal Of Operational Research*, 121:435–466, 2000.
3. E. W. Beth. *Foundations of Mathematics*. North-Holland, Amseterdam, 1959.
4. R. B. Braithwaite. *Scientific Explanation: A Study of the Function of Theory, Probability, and Law in Science*. Cambridge University Press, Cambridge, 1953.
5. C. Cheng and H. Tong. On consistent non-parametric order determination and chaos. *Journal of R. Statistical Soc., series B*, 54:427–449, 1992.
6. C. W. Churchman. *The Design of Inquiring Systems: Basic Concepts of Systems and Organization*. Basic Books, New York, 1971.
7. H. Cramer. *Mathematical Methods in Statistics*. Princeton University Press, Princeton, 1945.
8. J. W. Dawson Jr. *Logical Dilemmas: The Life and Work of Kurt Gödel*. A. K. Peters., Wellesley MA, 1997.
9. René Descartes. *Oevres édition Charles Adam, Paul Tannery, Lépold Cerf*. Vrin-CNRS (édition de référence (11 volumes), Paris, 1964–1974.
10. L. Di Giacomo, E. Di Lena, G. Patrizi, L. Pomaranzi, and F. Sensi. C.a.s.s.a.n.d.r.a. computerized analysis for supply chain distribution activity. In L. Bertazzi, M. G. Speranza, and J. Van Nunen, editors, *Innovations in Distribution Logistics*. Springer, Berlin, 2009.
11. L. Di Giacomo and G. Patrizi. Dynamic nonlinear modelization of operational supply chain systems. *Journal of Global Optimization*, 34:503–534, 2006.
12. L. Di Giacomo and G. Patrizi. Optimal dynamic nonlinear prediction methods for management of financial instruments. Technical report, Dipartimento di Statistica, Probabilita e Statistiche Applicate, Universita di Roma, La Sapienza, Rome, 2006.
13. L. Di Giacomo and G. Patrizi. Methodological analysis of supply chain management applications. *European Journal of Operational Research*, 207:249–257, 2010.
14. L. Di Sopra and G. Patrizi. The application of o.r. techniques for the prediction and understanding of damages caused by seismic events. *European Journal of Operational Research*, 28:180–195, 1987.
15. J Dieudonné. *Fondaments d'Analyse*. Gauthiers Villars, Paris, 1960, vol. 1.
16. C. Diks. *Nonlinear Time Series Analysis*. World Scientific, Singapore, 1999.
17. J.-P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Review of Modern Physics*, 57:617–656, 1985.
18. F. Suppe (ed.). *The Structure of Scientific Theories*. University of Illinois Press, Urbana, 1974.

19. S. N. Elaydi. *Discrete Chaos*. Chapman and Hall CRC press, London, 1999.
20. G. A. Gottwald and I Melbourne. Testing for chaos in deterministic systems with noise. *Physica D*, 212:100–110, 2005.
21. G. Grimaldi, C. Manna, L. Nieddu, G. Patrizi, and P. Simonazzi. A diagnostic decision support system and its application to the choice of suitable embryos in human assisted reproduction. *Central European Journal Of Operational Research*, 10(1):29–44, 2002.
22. S. Haberman. *The Analysis of Frequency Data*. The University of Chicago press, Chicago, 1974.
23. B. Hasselblatt and A. Katok. *A First Course in Dynamics: with a Panorama of Recent Developments*. University Press, Cambridge, 2003.
24. J. Hintikka. *Lingua Universalis vs. Calculus Ratiocinator. An ultimate presupposition of Twentieth-century philosophy*. Kluwer, Boston, 1997.
25. R. I. Jennrich. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statisitcs*, 40:633–643, 1969.
26. K. Judd. Forecasting with imperfect models, dynamically constrained inverse problems, and gradient descent agorithms. *Physics D*, 237:216–232, 2008.
27. R. E. Kalman, P. L. Falb, and M. A. Arbib. *Topics in Mathematical System Theory*. McGraw-Hill, New York, 1969.
28. H. Kantz and Th. Schreiber. *Nonlinear Time Series Analysis*. University Press (2nd Edition), Cambridge, 1997.
29. E. Lorenz. Deterministic non-periodic flow. *Journal of the Atmospheric Sciences*, 20:130–141, 1963.
30. E. Malinvaud. *Méthodes Statistiques de l' économétrie*. Dunod, Paris, 3eme ed., 1978.
31. C. Manna, G. Patrizi, A. Rahman, and H. Sallam. Experimental results on the recognition of embryos in human assisted reproduction. *Reproductive BioMedicine Online (www.rbmonline.com/Article/1170)*, 8(4):460–469, 2004.
32. J. G. Miller. *Living Systems*. McGraw-Hill, New York, 1978.
33. A. H. Nayfeh and B. Balachandran. *Applied Nonlinear Dynamics*. Wiley, New York, 1995.
34. L. Nieddu and G. Patrizi. Formal properties of pattern recognition algorithms: A review. *European Journal of Operational Research*, 120:459–495, 2000.
35. P. Pardalos and V. A. Yatsenko. Optimization approach to the estimation and control of lyapunov exponents. *Journal of Optimization Theory and Applications*, 128:29–48, 2006.
36. G. Patrizi. Model based selection of data arrays for inferences on large surveys. In R. Coppi and S. Bolasco, editors, *Multiway Data Analysis*, pp. 521–530, Amsterdam, 1989. North-Holland.
37. G. Patrizi. The equivalence of an lcp to a parametric linear program with a scalar parameter. *European Journal of Operational Research*, 51:367–386, 1991.
38. G. Patrizi. *S.O.C.R.A.t.E.S. s*imultaneous *o*ptimal *c*ontrol by *r*ecursive and *a*daptive *e*stimation *s*ystem: Problem formulation and computational results. In M. Lassonde, editor, *Optimization and Approximation, Vth International Conference on Approximation and Optimization in the Carribean*, pp. 245–253. Physika- Verlag, Heidelberg, 2001.
39. G. Patrizi, G. Addonisio, C. Giannakakis, A. Onetti Muda, Gr. Patrizi, and T. Faraggiana. Diagnosis of alport syndrome by pattern recognition techniques. In P. M. Pardalos, V. L. Boginski, and A. Vazacopoulos, editors, *Data Mining in Biomedicine*, pp. 209–230. Springer, Berlin, 2007.
40. G. Patrizi and C. Cifarelli. Solving large protein secondary structure classification problems by a nonlinear complementarity algorithm with {0,1} variables. *Optimization and Software*, 22: 25–49, 2007.
41. G. Patrizi, C. Manna, C. Moscatelli, and L. Nieddu. Pattern recognition methods in human assisted reproduction. *International Transactions in Operational Research*, 11:365–379, 2004.
42. G. Patrizi, Gr. Patrizi, L. Di Cioccio, and C. Bauco. Clinical analysis of the diagnostic classification of geriatric disorders. In P. M. Pardalos, V. L. Boginski, and A. Vazacopoulos, editors, *Data Mining in Biomedicine*, pp. 231–260. Springer, Berlin, 2007.
43. J. Pfanzagl. *Theory of Measurement*. Physica-Verlag, Wien, 1971.

44. H. A. Simon. Dynamic programming under uncertainty with a quadratic criterion function. *Econometrica*, 24: 74–81, 1956.
45. T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, Englewood Cliffs, N.J., 1989.
46. Baruch Spinoza. *Tractatus Theologico-Politicus*. Henricum Künraht, Hamburg, 1670.
47. F. Takens. Detecting strange attractors in turbulance. In D. A. Rand and L.S. Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, pp. 366–381. Springer, New York, 1981.
48. F. Takens. Invariants related to dimension and entropy. *Ats do 13 Colloquio Brasileito de Matematica, Istituto de Matematica Pura e Applicada, Rio de Janeiro*, pp. 1–23, 1983.
49. H. Theil. A note on certainty equivalence in dynamic programming. *Econometrica*, 25: 346–349, 1957.
50. M. Vellekoop and R. Berglund. On intervals transitivity = chaos. *The American Mathematical Monthly*, 101: 353–355, 1994.
51. L. von Bertalanffy. *General Systems Theory*. Braziller, New York, 1974.
52. J. Warga. *Optimal Control of Differential and Functional Equations*. Academic Press, New York, 1972.
53. G: M. Weinberg. *An Introduction to general System Theory*. Wiley, New York, 1975.
54. Charlotte Werndl. Are deterministic descriptions and indeterministic descriptions observationally equivalent? *Studies in History and Philosophy of Modern Physics*, 40: 232–242, 2009.
55. Charlotte Werndl. What are the new implications of chaos for unpredictability? *The British Journal for the Philosophy of Science*, 60: 195–220, 2009.

Part III
Sensors in Real-World Applications

Sensors in Transportation and Logistics Networks

Chrysafis Vogiatzis

Abstract Transportation engineering and logistics have been utilizing sensor networks for statistical analysis and data collection for years. In the last decades, due to the increased interest in sensor networks for optimization techniques, advancements have been made in attempts to provide on the fly algorithms that adapt to an ever-changing world. This chapter aims to give useful insight and present the latest developments in this growing branch of optimization and operations research.

1 Introduction

In the last decades, the economics of transportation have become a growing importance for corporations, municipalities, and commuters. Due to the large congestion levels every day users have to face lot of trouble in traveling for long durations, and it is good for which being able to find even an approximate shortest path is vital. However, the analyses that was provided to us with very useful observations and insight would never be possibly correct and accurate if it were not from an automated system of data collection used around the transportation networks.

Since 1950 [1] the need for data collection to observe the levels of congestion or the condition of links in the network has appeared. Researchers installed simple or sophisticated sensors in order to measure the volume of the traffic flow on specific links, the average vehicle velocity, and the time required to traverse a street among others.

Nowadays, the majority of the vehicles using the network are equipped with routing guidance (usually GPS devices). In addition to that, the increased wireless

C. Vogiatzis (✉)
Center for Applied Optimization, Department of Industrial and Systems Engineering
University of Florida, Gainesville, FL 32611, USA
e-mail: chvogiat@ufl.edu

V.L. Boginski et al. (eds.), *Sensors: Theory, Algorithms, and Applications*,
Springer Optimization and Its Applications 61, DOI 10.1007/978-0-387-88619-0_7,

capacities of cell phones and PDA devices in the last few years have made tracking and wireless sensing a common everyday phenomenon. The attempts to combine these increased capabilities with practical transportation problems are going to be described in detail in the following sections.

Moreover, it is important to note that especially in the United States, the appearance of high congestion levels is noticed more often than ever. This comes as a result of the large number of commuters that use the traffic network on an everyday basis [2]. It can also be seen that public transportation users have shrunk to insignificant levels while automobile commuters never cease to increase since 1950 [1]. It can be seen in the work of O' Toole [2] that the last few years, average vehicle speed is rapidly decreasing, while energy consumption and CO_2 levels are increasing. It becomes evident that it is now more than ever that optimal routing and traffic assignment techniques need to be applied.

1.1 Chapter Outline

This chapter is outlined as follows. In Sect. 2, a presentation of sensor-based methods for collecting useful data for analysis in the fields of transportation will be given. Section 3 will then present the most recent advancements in the infamous problems of vehicle routing and traffic assignment. In that section, special attention will be given to modern GPS systems and intelligent vehicle sensors. Lastly, in Sect. 4, the conclusions of the author along with future research possibilities will be presented.

2 Data Collection and Statistical Analyses

Among the most important and vital ongoing research in any transportation system is the development and implementation of *Intelligent Transportation Systems*. Systems of this kind can be used to reduce the human factor as far as traffic assignment and vehicle routing are concerned. In addition to that, optimization of traffic flow in intersections will also result in less congestion and accidents [3].

2.1 Overview of Sensor Methods

First of all, let us focus on the tracking methods that have been put to practice in the last decade. These algorithms have extensively investigated computer vision as a means of vehicle tracking. A seminal contribution in the field of traffic monitoring has been presented by Peterfreund [4] in 2002. The author focuses on the snakes active contour models, introduced by Kass, Witkin and Terzopoulos [5] and proposes a stochastic *velocity snake model* in order to track vehicles in intersections and traffic lights. Equally used is the method of Gardner and Lawton [6] which is common in traffic images.

However, all the above methods are difficult to put into practice when, because of congestion, there is a large number of vehicles at a given intersection at any time. That is the insight behind the Kamijo et al. approach [3] where each vehicle is tracked individually even if in these congestion levels, certain vehicles might be not completely within the sensor/camera range. This phenomenon is referred to as occlusion and has been a major obstacle in traffic monitoring.

2.2 Vehicle Detection Sensors

First of all, let us start by providing the reader with a summary of the major types of sensors that are currently available for vehicle detection, as they were noted by Luz and Mimbella [11].

- *Inductive Loop*:
  One of the intrusive sensors, most usually installed on the pavement surface. They can also be installed underneath the monitored road by tunneling under the surface of the street. Even though they represent a technology that is greatly understood and very accurate, their installation and maintenance is hard since it requires the disruption of flow for a big time.
- *Magnetometer*:
  Similar to the inductive loop, a magnetometer is a flexible sensor that can be used for a variety of purposes. Unfortunately, it is prone to erroneous measurements when the traffic is heavy and, hence, it has become unattractive. In addition to that, another drawback that a magnetometer inherits from an inductive loop is that it involves a huge cost of installation and maintenance.
- *Active and Passive Infrared*:
  Active and passive infrared sensors can be used in combinations in order to cover all measurements of the vehicles on a given road segment. An active infrared sensor can measure accurately the vehicle position and class, while it can also provide us with an estimate on the speed. A passive, on the other hand, can be used for speed measurements. Both categories of infrared sensors share a common disadvantage: they are susceptible to false measurements when visibility is limited (i.e., less than 20 feet) because of weather conditions.
- *Ultrasonic*:
  Ultrasonic permit multiple lane monitoring at the same time. This is the main advantage that makes them attractive to use. However, they present problematic behavior when functioning under sudden temperature changes and air turbulences.
- *Acoustic*:
  Similarly to ultrasonic sensors, for apparent reasons, acoustic sensors can also monitor more than one lane at a time with accuracy. Their major problems have to do with cold temperatures and specific traffic patterns. For example they are

not recommended for use in large city crossroads where the phenomenon of stop and go traffic is usual.

- *Video image processing*:
  A video image processing sensor is, in most cases, the ideal candidate. Provides monitoring for as many lanes as the video image capturer can take care of and it is easy to install and maintain without disrupting the normal vehicular flow. A major drawback is the increasing complexity at decoding the sensor measurements into actual data that can be used. Also, as far as requirements for normal installation are concerned, it needs to be ensured that they are installed at a high point, usually 50–60 feet above the monitored road.

2.3 Accident Detection Through Image Processing

The fundamental idea of Kamijo et al. [3] is the application of a stochastic relaxation algorithm in the detection process. This is vital for the practice, since vehicles have a series of very distinct characteristics, including but not limited to their appearance, their direction and their color. Given that in an intersection, congestion levels are higher and occlusion occurs often, it is impractical to adopt a simple contour method for tracking.

For the modeling purposes of their algorithm, the tracking of a single vehicle is transformed into a labeling method, where each pixel is assigned (i.e., labelled) to a specific vehicle at any time. After all pixels have been labelled, it is part for the deductive method to be applied in order to obtain an initial, but accurate, object mapping.

There are five major components to the deductive process:

- *Initialization*:
  The background image is set by using a 20 min image sequence. Also, the entrance points to the intersection are set by defining slits where new intensities of the image (i.e., incoming traffic) appear.
- *Generation of new vehicles*:
  Whenever along the slits defined, a new intensity is observed, a new ID is assigned to the incoming vehicle. All the pixels sharing that intensity are assigned the same ID.
- *Vehicle vector estimation*:
  At each block of time, the similarity of the vehicle motion is estimated by:

$$(x(t+1), y(t+1)) = (x(t) + u(t), y(t) + v(t)) \tag{1}$$

  and the motion vector is approximated as the most frequent motion vector of all pixels assigned with the same label,

$$D = \sum_{0 \le d_i \le 8, 0 \le d_j \le 8} |I(i+di+u_i, j+dj+v_j; t+1) - I(i+di, j+dj; t)|. \tag{2}$$

  In this case, $I(x, y; t)$ is the intensity of pixel (x, y) at time t.

- *Vehicle region update*:
  All the vehicle blocks are updated at time $t + 1$ compared to time t as per the motion vector obtained in the previous step. If the new intensity difference is smaller than a threshold then the vehicle is considered to be out of the intersection. In the case where the difference is bigger, then it is assumed that a neighboring vehicle is hiding a part of the vehicle tracked.
- *Vehicle blocks division*:
  At the entrance point of the camera (the slits defined during initialization), simultaneous arrivals may result in characterizing multiple vehicles as one. This is the reason why it is checked always if the motion vectors obtained by a labelled vehicle match. If that is not the case then the pixels corresponding to the same label are divided in order to distinguish between different vehicles.

The last part of this methodology consists of the stochastic relaxation in order to assign certain pixels to more than one vehicles, because of occlusion. The authors then extend the MRF model to include and deal with not only images as in the work by Geman and Geman [7] and by Andrey and Tarroux [8], but also time-axis distribution. Their spatio-temporal MRF model provides an estimation of the current object map based on the information provided in the previous object map along with the current and previous images. After applying the stochastic relaxation procedure to their methodology, the traffic monitoring system for detecting accidents is ready to be examined and tested with the results being successful in a specific intersection, but can be generalized to each topology and geometry [3].

2.4 Sensor Networks for Traffic Monitoring

Before the recent advancements in image processing and data extraction, the use of inductive loop detectors was and still remains the most common way to collect information on traffic conditions mainly because of their high levels of accuracy and reliability. However, their installation and their maintenance requires a significant down-time of the arc being updated and hence, they become exceedingly expensive. That is the main reason why in the last years, more sophisticated sensors are being used, such as surveillance cameras, microwave radars, ultrasound, and infrared sensors. However, these sensors are less reliable, even though the methods have become more involved over the years, and they also are costly.

Therefore, numerous scientific efforts have been made in order to incorporate cheaper, wireless sensors in the existing infrastructure for statistical and monitoring purposes. The approaches that will be focused upon in this subsection are the ones by Coleri et al. [9, 10].

In the work of Coleri et al. [9], the Traffic-Dot sensor model consists of the following major components:

Table 1 Notation for the model of Coleri et al. [10]

Notation	Definition
$G = (V, E)$	The graph representing the sensor network, with $V = \{1\} \cup V_s \cup V_r$ symbolizing the nodes. Node 1 is the access point, $V_s = [2, N]$ are the sensor nodes and $V_r = [N + 1, M]$ the relay nodes. If nodes i and j are within transmission range then $(i, j) \in E$.
g_i	Rate of packets per unit time that node $i \in [1, N]$ can transmit.
p_s	Energy spent in sensor when obtaining packets in one packet.
$p_{tx,ij}$	Energy spent for the transmission of a packet from node i to node j per time unit
f_{ij}	Average time required to receive packets at node j from node i.
e_i	The battery energy of each of the parts of the nodes $i \in [1, M]$

- processor,
- radio,
- magnetometer, and
- battery.

The magnetometer (magnetic sensor) is used for vehicle detection. The power consumption of such a sensor based model is very small, making it an ideal candidate for traffic measurements and data collection for statistical analysis. Its accuracy is also very high, reaching the 97% accuracy that is achieved by inductive loop detectors. In order to increase efficiency and battery utilization the authors have proposed the following linear programming model [10], that is described in (3)–(9), with the notation given in Table 1.

$$\min \quad \sum_{i=1}^{M} e_i \tag{3}$$

$$\text{s.t.} \quad \sum_j f_{ij} - \sum_j f_{ji} = g_i, \qquad \forall i \in [2, N] \tag{4}$$

$$t_d \left(\sum_j p_{tx,ij} f_{ij} + \sum_j p_{rx} f_{ji} + p_s g_i \right) \leq e_i, \qquad \forall i \in [2, N] \tag{5}$$

$$\sum_j f_{ij} - \sum_j f_{ji} = 0, \qquad \forall i \in [N + 1, M] \tag{6}$$

$$t_d \left(\sum_j p_{tx,ij} f_{ij} + \sum_j p_{rx} f_{ji} \right) \leq e_i, \qquad \forall i \in [N + 1, M] \tag{7}$$

$$f_{ij} \geq 0, \qquad \forall i, j \in [1, M] \tag{8}$$

$$e_i \geq 0, \qquad \forall i \in [1, M]. \tag{9}$$

The above linear program has the limitation that it constrains the relay sensors to be in a fixed position in the network. In reality, it is desired to be able to determine the optimal placement of the relay nodes, thus altering dynamically the topology of the network. The resulting formulation is presented in (10)–(18).

$$\min \quad \sum_{i=1}^{M} e_i \tag{10}$$

$$\text{s.t.} \quad \sum_j f_{ij} - \sum_j f_{ji} = g_i, \qquad \forall i \in [2, N] \tag{11}$$

$$t_d \left(\sum_j p_{tx,ij} f_{ij} + \sum_j p_{rx} f_{ji} + p_s g_i \right) \le e_i, \qquad \forall i \in [2, N] \tag{12}$$

$$\sum_j f_{ij} - \sum_j f_{ji} = 0, \qquad \forall i \in [N+1, M] \tag{13}$$

$$t_d \left(\sum_j p_{tx,ij} f_{ij} + \sum_j p_{rx} f_{ji} \right) \le e_i, \qquad \forall i \in [N+1, M] \tag{14}$$

$$p_{tx,ij} = p_{tx}(d(i, j)), \qquad \forall i, j \in [1, M] \tag{15}$$

$$d(i, j)^2 = |l_i - l_j|^2, \qquad \forall i, j \in [1, M] \tag{16}$$

$$f_{ij} \ge 0, \qquad \forall i, j \in [1, M] \tag{17}$$

$$e_i \ge 0, \qquad \forall i \in [1, M]. \tag{18}$$

However, the resulting formulation as can be seen by the reader is a nonlinear, nonconvex problem and hence, the authors propose an approximate algorithm for its solution. Their proposed algorithm is then tested through simulation with remarkable results. Their novel approach in energy management for sensors in a wireless network can result in a more efficient method of data collection in traffic engineering, as can be seen by their work at the Traffic-Dot [9].

3 Vehicle Routing and Traffic Assignment

Vehicle routing and traffic assignment problems are two of the most important problems in transportation engineering and planning. Many attempts to solve them in a dynamically changing environment such as the modern urban grid have appeared in literature, however it is in the last few years with the boom of smartphones and sensors that there exist the tools to tackle them successfully.

Online algorithms [12] use recent information as feedback to alter their solutions in order to remain optimal at every instance. For these algorithmic approaches to function effectively reliable information needs to be provided in a fast and efficient way. With the recent increase of guidance devices, such as GPS, and cheap wireless sensors, the possibility to obtain these data appeared.

3.1 Sensor-Based Robotic Vehicle Routing Complexity

Before generalizing the vehicle routing problem to realistic, practical applications involving decision makers in the urban environment, it is important to review the

Table 2 The notation for the formulation of Sharma et al. [18]

Notation	Definition
Q	The square environment where agents are allowed to move of area A.
$(O, D)_i$	The origin–destination pair of agent i.
$t_{0,i}$	The time when an agent is dispatched in the network, $t_{0,i} \geq 0$.
T_i	The time an agent requires in the network until it reaches its destination.
γ_i	The time-dependent path that agent i is following, $\gamma : [0, T_i] \rightarrow Q$.
$v_i(t)$	The velocity of agent i, $v_i(t) \leq v_{\max}$.
$C_i(t)$	The exclusion zone C of an agent at time t.

robotic vehicle sensor based routing. This is a common application in automated control systems in industry [13–15] and, thus, there have been attempts to model and optimize their behavior.

The results of this scientific research has provided insight to researchers as far as online guidance through intelligent automobile systems equipped with sensors are concerned. Usually, in robotic applications, instead of an optimal shortest path for all the vehicles involved, researchers are interested in limiting the selection to a set of available paths and selecting the best among those. This approach has been studied by Inalhan et al. [16] where fixed routes were given to the vehicles and by Gerkey et al. [17], where a fixed roadmap with arcs and nodes was employed.

The above approaches, even though they are practical and present good results, are not theoretically guaranteeing optimality. So, that led Sharma et al. [18] to research the time complexity of this sensor-based vehicle routing problem with no limitations on the selection of routes to the agents utilizing the network. For their work, previous research on communication between robotic systems [19] was considered and customized. The notation for the setup of the problem is presented in Table 2.

The exclusion zone C is a disk of a nonconstant radius centered at the position of the agent, where there can be no other vehicle. If there is another vehicle at the same time, then a conflict is said to appear and new routing has to occur. The disk is defined in (19). As it can be seen, the radius is dependent on the velocity of the agent i at that time t. More specifically, a conflict occurs and there exists a time t_c such that:

- agents i and j are active at time t_c and
- $C_i(t_c) \cap C_j(t_C) \neq \emptyset$.

$$C_i(t) = \{z \in R^2 : ||z - x_i(t)|| \leq r_0 + k||v_i(t)||\}. \tag{19}$$

The objective of the sensor based vehicle routing problem with robotic agents is to find an optimal *routing policy*. As such we define a mapping:

$$\pi : (O, D) \rightarrow (t_0, T, \gamma),$$

which is safe, that is, no conflict occurs at any time. The authors then define T_π to be the time when all agents have safely arrived at their destinations according to policy π. Then, the time complexity is defined as:

$$T^*(O, D) = \inf_{\text{safe}\,\pi} T_\pi(O, D).$$

This formulation leads to interesting theoretical results when it comes to the upper and lower bounds in the time complexity of the problem. There are two cases that were examined:

- Best case scenario.
- Average case scenario.

In the best case scenario, the $(O, D)_i$ pairs are selected so as to minimize the total time required for all agents to remain in the environment. The authors go ahead and prove the following lemma in that case.

Lemma 1. *For any set of n (O, D) pairs, such that the average distance between origin and destination points is $\bar{L}$, the time complexity of the problem is $\Omega\left(\sqrt{n}\bar{L}\right)$.*

They also proceed to prove the following lemma for an upper bound on the time complexity of the best case.

Lemma 2. *For any $n \in N$, $\exists$ n (O, D) pairs such that the time complexity is $O\left(\sqrt{n}\bar{L}\right)$, where $\bar{L}$ represents the average distance of any two origin–destination points.*

Combining these two lemmas, the authors obtain a very important theorem on the time complexity of the sensor based vehicle routing problem as formulated above. However, the most important aspect of their work is yet to follow with the average case scenario study, where the origin–destination pairs are no longer arbitrarily selected, but are random. In that case, the following lemma is proved by the authors on a lower bound of the time complexity.

Lemma 3. *The time complexity of the problem when the set of n origin–destination pairs is randomly selected from the uniform distribution is, with high probability, $\Omega\left(\sqrt{n}\right)$.*

Next, the authors present their algorithmic framework that terminates in $O(\sqrt{n})$ time, hence concluding that with high probability the time complexity of the problem is $O(\sqrt{n})$. Sensor based fully automated vehicles are used for transportation purposes in large facilities [20] or depot centers [21] and, hence, the result of the time complexity of their optimal routing problem is of significant importance in the fields of logistics.

3.2 *Intelligent Vehicle Routing Through a Centralized Highway System*

The previous results may apply only to supply chain networks and to the optimization of the automated procedures regarding storing, handling, and shipping, however inspired a number of researchers in generalizing the notions in the large-scale, real-life traffic system. The idea of receiving feedback that provides the tripmaker with information on the state of the roads that they are planning to use is not new at all.

Television and radio channels spend time on informing the audiences which streets should be avoided, where the users are experiencing normal traffic flows or if there has been an accident. In the last years, guidance devices, such as the GPS routing system, provide the possibility to obtain traffic data in real-time [22].

As was mentioned in the previous subsection, many attempts to model realistically the problem for real vehicular flows emanate from the robotic world and a fully automated and controlled system of vehicles. That was, also, the insight of Baskar, De Schutter and Hellendoorn, who first proposed an hierarchical traffic control system for intelligent vehicles [23, 24] and then adapted its formulation to obtain a mathematical programming problem in order to come up with the optimal routing of the vehicles [25].

The framework, described in Baskar et al. [23] consists of the following elements:

1. the vehicle controllers, which control the speed and steering of the vehicles by receiving orders from the platoon controllers;
2. the platoon controllers, which take care of the merges and splits of platoons and the vehicle to vehicle distances by receiving control commands from the roadside controllers;
3. the roadside controllers, which are in charge of a segment of the whole network in the infrastructure;
4. and the higher-level controllers, which coordinate the whole network and supervise all other controllers.

Using this infrastructure, the authors focus on optimal routing to each platoon that is currently in the network. The notation that is used is given in the following table.

The simple linear model that corresponds to the framework described before is given in (20)–(23). The notation that is used is given in Table 3.

$$\min\ J_{\text{links}} = \sum_{(o,d)\in O\times D}\ \sum_{l\in L_{o,d}} x_{l,o,d}\,\tau_l T \tag{20}$$

$$\text{s.t.} \sum_{l\in L_o^{\text{out}}\cap L_{o,d}} x_{l,o,d} = D_{o,d}\,, \quad \forall o\in O, \forall d\in D \tag{21}$$

Table 3 Notation used in the autonomous vehicle routing by Baskar et al. [25]

Notation	Definition
O	Origin nodes
D	Destination nodes
I	Internal nodes
V	The set of all nodes, $V = O \cup I \cup D$
L	The set of links in the network
(o, d)	One origin–destination pair, $(o, d) \in O \times D$
$L_{o,d}$	The set of links that belong to a route connecting o to d
$D_{o,d}$	The demand of the pair (o, d)
C_l	The capacity of the link $l \in L$
v_l	The speed on link $l \in L$
τ_l	The travel time on link $l \in L$
L_v^{in}	The set of all links incoming to node v
L_v^{out}	The set of all links leaving node v
$x_{l,o,d}$	Decision variables denoting the flows for every pair $(o, d) \in O \times D$
T	The simulation period

$$\sum_{l \in L_v^{\text{in}} \cap L_{o,d}} x_{l,o,d} = \sum_{l \in L_v^{\text{out}} \cap L_{o,d}} x_{l,o,d}, \quad \forall v \in, \forall (o, d) \in O \times D \tag{22}$$

$$\sum_{(o,d) \in I_{od,l}} x_{l,o,d} \leq C_l, \quad \forall l \in L. \tag{23}$$

The model is simple to understand since it involves only the flow balance in the network and the capacity constraints for each of the links. The objective function in (20) is a measure of the time that the vehicles have to spend while traveling in the network.

In order to more realistically model the problem, the authors then proceed to include queues that can be formed at the entries of the infrastructure. So now the model can be rewritten as:

$$\min J_{\text{links}} + J_{\text{queue}} = \sum_{(o,d) \in O \times D} \sum_{l \in L_{o,d}} x_{l,o,d} \tau_l T \tag{24}$$

$$+ \sum_{(o,d) \in O \times D} \frac{1}{2} \left(D_{o,d} - F_{o,d}^{\text{out}} \right) T^2$$

$$\text{s.t.} \sum_{l \in L_v^{\text{in}} \cap L_{o,d}} x_{l,o,d} = \sum_{l \in L_v^{\text{out}} \cap L_{o,d}} x_{l,o,d}, \quad \forall v \in, \forall (o, d) \in O \times D \tag{25}$$

$$\sum_{(o,d) \in I_{od,l}} x_{l,o,d} \leq C_l, \quad \forall l \in L \tag{26}$$

$$\sum_{l \in L_o^{\text{out}} \cap L_{o,d}} x_{l,o,d} \leq D_{o,d}, \quad \forall o \in O, \forall d \in D \tag{27}$$

$$F_{o,d}^{\text{out}} = \sum_{l \in L_o^{\text{out}} \cap L_{o,d}} x_{l,o,d}. \tag{28}$$

J_{queue} is a measure of the time spent by the vehicles in the queues formed in the origin nodes. In order to come up with an estimate for these measures, the authors note that the queue size increases with time with a rate of $D_{o,d} - F_{o,d}^{\text{out}}$. Hence, at the end of the simulation the total length is $\left(D_{o,d} - F_{o,d}^{\text{out}}\right) T$ and the average can be calculated as $\frac{1}{2}(D_{o,d} - F_{o,d}^{\text{out}})T$. That is how the term of J_{queue} is computed in the objective function above. It is important to note that once more the mathematical program obtained is linear.

Even in this last model, the approach is highly unrealistic. It is not a valid assumption that the demands are static, but have to be considered dynamic in order to accommodate most practical applications. In order to do so, a discretization of the time spent on each link is introduced, with the elementary measurement of T_s. This can be written more clearly as:

$$\tau_l = \kappa_l T_s, \quad \text{where } \kappa_l \in Z^+. \tag{29}$$

Letting $q_{o,d}(k)$ be the partial queue length of vehicles traveling from o to d at time k, i.e., $t = kT_s$, and by assuming that the network is initially empty, i.e., $q_{o,d}(k) = 0$ and $x_{l,o,d}(k) = 0$ for $k \leq 0$ we can now have for each of the origin nodes o:

$$\sum_{l \in L_o^{\text{out}} \cap L_{o,d}} x_{l,o,d}(k) \leq D_{o,d}(k) + \frac{q_{o,d}(k)}{T_s} \quad \forall d \in D, \tag{30}$$

and by definition $D_{o,d}(k) = 0$ for $k \geq K$. Now, by considering the fact that every vehicle on link l will reach the end of the link after κ_l time segments, we obtain that

$$\sum_{l \in L_v^{\text{in}} \cap L_{o,d}} x_{l,o,d}(k - \tau_l) = \sum_{l \in L_v^{\text{out}} \cap L_{o,d}} x_{l,o,d}(k), \quad \forall v \in I \quad \forall (o,d) \in O \times D. \tag{31}$$

Also for every link, we have the capacity constraints, however taking into consideration the time we are in. So (25) is now transformed in the dynamic case into

$$\sum_{(o,d) \in I_{od,l}} x_{l,o,d}(k) \leq C_l, \quad \forall l \in L. \tag{32}$$

The important part of this modeling approach is the description of the queues formed. The flow is given by a similar constraint to the one presented in (27), which however is transformed in order to accommodate the time factor into

$$F^{\text{out}}_{o,d}(k) = \sum_{l \in L^{\text{out}}_o \cap L_{o,d}} x_{l,o,d}(k). \tag{33}$$

Therefore, the queue length is increasing linearly with the rate of $D_{o,d}(k) - F^{\text{out}}_{o,d}(k)$ for the time interval $[kT_s, (k+1)T_s)$ and we get the following equation for the queue length:

$$q_{o,d}(k+1) = \max\left(0, q_{o,d}(k) + \left(D_{o,d}(k) - F^{\text{out}}_{o,d}(k)\right) T_s\right). \tag{34}$$

There exist two cases which can be distinguished for the determination of the time $J_{\text{queue},o,d}(k)$ that a vehicle has to spend in the queue formed at an origin o:

a. The queue length becomes zero while the interval $[kT_s, (k+1)T_s)$.
b. The queue length remains positive in the same interval.

Let us consider the second case. Defining the time after kT_s at which the queue length becomes zero as:

$$T_{o,d}(k) = \frac{q_{o,d}(k)}{F^{\text{out}}_{o,d}(k) - D_{o,d}(k)}, \tag{35}$$

J_{queue} can now be estimated as:

$$J_{\text{queue},o,d}(k) = \begin{cases} \frac{1}{2}(q_{o,d}(k) + Q_{o,d}(k+1))T_s & \text{for the first case} \\ \frac{1}{2}q_{o,d}(k)T_{o,d}(k) & \text{for the second case.} \end{cases} \tag{36}$$

In general now, we have

$$J_{\text{queue}} = \sum_{k=0}^{K_{\text{end}}-1} \sum_{(o,d)\in O\times D} \sum_{l\in L_{o,d}} J_{\text{queue},o,d}(k) \tag{37}$$

and

$$J_{\text{links}} = \sum_{k=0}^{K_{\text{end}}-1} \sum_{(o,d)\in O\times D} \sum_{l\in L_{o,d}} x_{l,o,d}(k)\kappa_l T_s^2. \tag{38}$$

So, finally the mathematical formulation becomes

$$\min(J_{\text{links}} + J_{\text{queue}}) \tag{39}$$

$$\text{s.t.}\ \ (30)\text{–}(34). \tag{40}$$

This model is for the second case a nonlinear, nonconvex, and nonsmooth problem. As such, this problem is hard to solve and hence, the authors present an approximate solution algorithm. Either way, by transforming the above problem into a mixed integer linear program, there exist several solvers that can solve it efficiently [26]. For this transformation to take place, the properties of Bemporad and Morari [27] can be used:

Property 1. $[f \leq 0] \iff [\delta = 1]$ is true iff

$$\begin{cases} f \leq M(1-\delta) \\ f \geq \epsilon + (m-\epsilon)\delta, \end{cases}$$

where ϵ is a small positive number.

Property 2. $y = \delta f$ is equivalent to

$$\begin{cases} y \leq M\delta \\ y \geq m\delta \\ y \leq f - m(1-\delta) \\ y \geq f - M\delta. \end{cases}$$

Then, the term of $F^{\text{out}}_{o,d}(k)$ from (34) can be eliminated, thus

$$q_{o,d}(k+1) = \max\left(0, q_{o,d}(k) + \left(D_{o,d}(k) - \sum_{l \in L_o^{\text{out}} \cap L_{o,d}} x_{l,o,d}(k)\right) T_s\right), \quad (41)$$

which still is nonlinear. However, by letting $D_{\max,o,d}$ be the maximum demand for $(o,d) \in O \times D$, $F_{\max,o,d}$ be the maximum feasible flow (i.e., $F_{\max,o,d} = \sum_{l \in L_o^{\text{out}} \cap L_{o,d}} C_l$) and $q_{\max,o,d}$ be the maximum queue length formed from origin o to destination d and equal to $D_{\max,o,d} T_s K_{\text{end}}$, then two new parameters can be defined as:

$$m^{\text{low}}_{o,d} = -F_{\max,o,d} T_s \quad (42)$$

$$m^{\text{upp}}_{o,d} = q_{\max,o,d} + D_{\max,o,d} T_s, \quad (43)$$

hence the following always stands:

$$m^{\text{low}}_{o,d} \leq q_{o,d}(k) + (D_{o,d}(k) - \sum_{l \in L_o^{\text{out}} \cap L_{o,d}} x_{l,o,d}(k)) T_s \leq m^{\text{upp}}_{o,d}. \quad (44)$$

Now, by introducing the binary variables $\delta_{o,d}(k)$ as:

$$\delta_{o,d}(k) = \begin{cases} 1 \ \text{ iff } \ q_{o,d}(k) + \left(D_{o,d}(k) - \sum_{l \in L_o^{\text{out}} \cap L_{o,d}} x_{l,o,d}(k)\right) T_s \geq 0 \\ 0 \quad \text{otherwise.} \end{cases} \quad (45)$$

So, applying Property 1, constraint (41) takes the form of

$$q_{o,d}(k+1) = \delta_{o,d}(k)\left(q_{o,d}(k) + D_{o,d}(k) - \sum_{l \in L_o^{\text{out}} \cap L_{o,d}} x_{l,o,d}(k)\right) T_s. \quad (46)$$

3.3 Vehicle Routing and Traffic Monitoring Using Personal Sensors

As personal sensors we denote these devices that can provide on the fly feedback and information, such as smartphones, Personal Digital Assistants (PDA), and GPS systems. The main driving idea is that the number of the traffic network users that also owns one of the aforementioned devices has significantly increased in the last years, making the propagation of information easier.

This has been the insight of Thiagarajan et al. [28] in creating a prototype application that gathers the information required through mobile phones and provides online routing based on recent traffic trends. In their work, they present real-time traffic monitoring system tracking the vehicle trajectories using a hidden Markov chain.

In general, three are the major pillars of a high quality online routing algorithm:

1. Accuracy:
   The time estimations and the congestion levels that are used to reroute and guide vehicles in the network have to be close enough to represent the real situation.
2. Energy efficiency:
   It is an important note that a smart phone battery is consumed much faster when an algorithm that requires access to online data is executed continuously. Therefore, a trade-off between the sampling time of the data and a high quality solution needs to be agreed.
3. Time efficiency:
   The algorithm applied needs to be fast and efficient. An algorithm that is too computationally expensive may present optimal routes, however it is unrealistic to assume that it can provide on the fly routes when needed.

However, the above objectives of an online routing algorithm present a number of challenges, the most difficult and common of which are mentioned below:

- Map matching of the trace to the road segment it corresponds to [29, 30].
- Time estimation of specific segments in a route.
- Accuracy is energy consumptive. Sampling GPS has been shown to be [31] far more expensive in terms of energy that WiFi sampling, which unfortunately is less accurate.

For the tracking algorithm proposed, the authors use a hidden Markov chain. By that, they imply that the positions at each sampling period of time are known, however the road segments (i.e., the transitions between positions) are unknown. Given a set of known positions over the time of the vehicle movement, the goal is to detect the maximum likelihood road segments that were used. The algorithm that they proposed can be summarized as follows:

- Compute **transmission probabilities**.
- Compute **emission probabilities**.
- Employ the **Viterbi** decoding algorithm.
- **Bad zone** detection and removal.

In order to compute the transition probabilities, the following notions need to be considered. First of all, there exists a probability that the car will still be in the same road segment for the next sampling period. Also, a car can only change road segments if there exists an intersection between the segment it was on and the segment it is observed to be on. Last, there are limits on the vehicle speeds that prohibit the car to go extremely fast in any given road segment. Mathematically, the notions above are summarized in (47)–(49) representing the probability p for a vehicle whose position at sampling time $t-1$ is i while at sampling time t is j:

$$\text{If} \quad i=j, \quad p=\epsilon. \tag{47}$$

$$\text{If } j \text{ and } i \text{ share no intersection,} \quad p=0. \tag{48}$$

$$\text{If } i \text{ and } j \text{ share an intersection,} \quad p=\epsilon \text{ or } p=0. \tag{49}$$

Equation (47) defines the probability that a car is still found in the same road segment and ϵ is defined as:

$$\epsilon \leq \frac{1}{d_{\max}+1}.$$

The third equation prohibits effectively the vehicle to move extremely fast at any road segment. If a vehicle is detected at position i at time $t-1$ and then at time t is found at j, then the algorithm computes the time it would normally take the vehicle to traverse this route. If that implies that the car is traveling at a speed that is greater than the threshold speed S_{outlier} defined at 200 mph, then the probability is set to be equal to 0; otherwise it is equal to ϵ.

The next step of the tracking algorithm involves the emission probabilities of the model. The emission probability notion is employed to cover the fact that it is possible for a point to be observed from a road segment that is close but is not necessarily the closest one. So, using a Gaussian function with zero mean N, the emission probability of the road segment i at position l is defined as:

$$N(\text{dist}(i,l)),$$

where $\text{dist}(i,l)$ is the Euclidean distance. The variance of N is dependent on the sensor that produced the position and, hence, different variances are used for WiFi and GPS position sampling.

The most important component of the technique applied is the Viterbi decoding algorithm [29] which finds the most likely sequence of hidden states, i.e., road segments, that the vehicle is required to pass through. Last in the sequence, after having obtained a valid route by applying the Viterbi algorithm, the "bad zones" are detected and removed from the route. That way, the authors can ensure that the route is as realistic as possible and they can use it to obtain useful information on the traffic status and the time required to traverse these arcs in real time.

Overall, the algorithm presented was applied to real data, with important results, including the facts that:

- Using WiFi localization, 90% of the routes predicted were within a 10–15% of the optimal route. GPS localization presented optimal results with high accuracy over a sampling period of 30 s.
- A hybrid algorithm employing the 30 s GPS sample with WiFi localization in between has an improved performance over the two methods mentioned above, however the gains are much less than the energy consumptions.
- Using GPS localization over a sampling period of 20 s outperforms the hybrid approach.

4 Conclusions and Future Work

The recent boom that smart phones have seen in the last few years has made cheap sensors available to a number of users. Especially when it comes to transportation systems, there is now the possibility of collecting information fast through the wireless networks that support these phones. This insight drove a number of researchers like in [28] to investigate the methods that feedback can be derived from these devices and provided to sophisticated algorithms.

As it is easy to see by the patents submitted by Resende and Pardalos [32], using sensor-based algorithms or deploying smartphones to collect information and optimize real-time problems is becoming more widely used. An important part of these algorithms that would benefit the trip makers' decisions would be the incorporation of historical data of the day or the period in the prediction model. In order to do so, sophisticated time-series approaches [33] and/or kernel regression machine learning can be applied to the model. Data mining techniques [34] can also provide us with useful remarks on the traffic behavior throughout a time period.

Another component of these algorithms that needs to be improved significantly is energy consumption. Nowadays, it is known that tracking and routing devices are expensive and use up a significant amount of battery. Therefore, it is not only important to provide travelers with reliable, on-the-fly algorithms for routing, but also algorithms that use up as little energy as possible. These are the major directions for future research that will optimize the procedures of using sensors in routing and traffic assigning. If we were to improve these conditions, then the algorithms discussed in this chapter would certainly be much more accessible to a number of users.

References

1. National Transportation Statistics. http://www.bts.gov/publications/national_transportation_statistics/.
2. O' Toole R. Gridlock: Why we're stuck in traffic and what to do about it, *CATO Institute*, Washington, D.C., 2009.
3. Kamijo, S. and Matsushita, Y. and Ikeuchi, K. and Sakauchi, M. Traffic monitoring and accident detection at intersections. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 1, No 2, pp. 108–118, 2002.
4. Peterfreund N. Robust tracking of position and velocity with Kalman snakes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No 6, pp. 564–569, 2002.
5. Kass M., Witkin A. and Terzopoulos D. Snakes: Active contour models, *International journal of computer vision*, Vol. 1, No 4, pp. 321–331, Springer, 1988.
6. Gardner W.F. and Lawton D.T. Interactive model-based vehicle tracking, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 18, pp. 1115–1121, Nov. 1996.
7. Geman S. and Geman D. Stochastic relaxation, Gibbs distribution ad the bayesian restoration of images, *IEEE Trans. Pattern Anal. Machine Intell*, Vol. PAMI-6, pp. 721–741, June 1984.
8. Andrey P. and Tarroux P. Unsupervised segmentation of Markov random field modeled textured images using selectionist relaxation, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 20, Mar. 1998.
9. Coleri S., Cheung S.Y. and Varaiya P. Sensor networks for monitoring traffic, *Allerton conference on communication, control and computing*, pp. 32–40, Citeseer, 2004.
10. Coleri S. and Varaiya P. Optimal placement of relay nodes for energy efficiency in sensor networks, *IEEE International Conference on Communications, ICC '06*, Vol. 8, pp. 3473–3479, IEEE, 2006.
11. Luz E.Y. and Mimbella A. Summary of vehicle detection and surveillance technologies in intelligent transportation systems, *The Vehicle Detector Clearinghouse*, Southwest Technology Development Institute at New Mexico State University, Fall 2007.
12. Karp R.M. On-line algorithms versus off-line algorithms: How much is it worth to know the future, *Proceedings of IFIP 12th World Computer Congress*, Vol. 1, pp. 416–429, Madrid, Spain, 1992.
13. Bullo F., Cortés J. and Martinez S. *Distributed control of robotic networks*, Princeton University Press, 2009.
14. Qu Z. *Cooperative Control of Dynamical Systems: Applications to Autonomous Vehicles*, Springer Verlag, 2009.
15. Dixon C. and Frew E.W. tMaintaining optimal communication chains in robotic sensor networks using mobility control, *Mobile Networks and Applications*, Vol. 14, no. 3, pp. 281–291, Kluwer Academic Publishers, 2009.
16. Inalhan G., Stepanovic D.M. and Tomlin C.J. Decentralized optimization, with application to multiple aircraft coordination, *Proceedings of IEEE Conference on Decision and Control*, 2002.
17. Gerkey B.P. and Mataric M.J. Sold!: Auction methods for multi-robot coordinatio, *IEEE Transactions on Robotics and Automation*, Vol. 18, no. 5, pp 758–768, 2002.
18. Sharma V., Savchenko M., Frazzoli E. and Voulgaris P. Time complexity of sensor-based vehicle routing, *Robotics: Science and Systems*, pp. 297–304, Citeseer, 2005.

19. Klavins E. Communication complexity of multi-robot systems, *Proceedings of Fifth International Workshop on the Algorithmic Foundations of Robotics*, Nice, France, 2002.
20. MacDuffie J.P. and Krafcik J. Integrating technology and human resources for high-performance manufacturing: Evidence from the international auto industry, *Transforming organizations*, pp. 209–226, Oxford University Press, 1992.
21. Vivaldini K.C.T., Galdames J.P.M., Bueno T.S., Araújo R.C., Sobral R.M., Becker M. and Caurin G.A.P. Robotic forklifts for intelligent warehouses: Routing, path planning, and auto-localization, *2010 IEEE International Conference on Industrial Technology (ICIT)*, pp. 1463–1468, 2010.
22. Nadeem T., Dashtinezhad S., Liao C. and Iftode L. TrafficView: traffic data dissemination using car-to-car communication, *ACM SIGMOBILE Mobile Computing and Communications Review*, Vol. 8, no. 3, pp. 6–19, ACM, 2004.
23. Baskar L.D., De Schutter B. and Hellendoorn H. Hierarchical traffic control and management with intelligent vehicles, *Proceedings of the 2007 IEEE Intelligent Vehicles Symposium (IV'07)*, pp. 834–839, Istanbul, Turkey, June 2007.
24. Baskar L.D., De Schutter B. and Hellendoorn H. Traffic management for intelligent vehicle highway systems using model-based predictive control, *Proceedings of the 88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009, Paper 09-2107.
25. Baskar L.D., De Schutter B. and Hellendoorn H. Optimal routing for intelligent vehicle highway systems using mixed integer linear programming, *Proceedings of the 12th Symposium on Transportation Systems*, Redondo Beach, California, pp. 569–575, Sept. 2009.
26. Pardalos P.M. and Resende M.G.C. *Handbook of Applied Optimization*, Oxford University Press, Oxford, UK, 2002.
27. Bemporad A. and Morari M. Control of systems integrating logic, dynamics and constraints, *Automatica*, Vol. 35, no. 3, pp. 407–427, 1999.
28. Thiagarajan, A., Ravindranath, L., LaCurts, K., Madden, S., Balakrishnan, H., Toledo, S. and Eriksson, J. VTrack: accurate, energy-aware road traffic delay estimation using mobile phones, *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, pp. 85–98, ACM, 2009.
29. Hummel B. Map matching for vehicle guidance. In *Dynamic and Mobile GIS: Investigating Space and Time*, CRC Press: Florida, 2006.
30. Krumm J., Letchner J. and Horvitz E. Map matching with travel time constraints. In *SAE World Congress*, 2007.
31. Gaonkar S., Li J. Choudhury R.R., Cox L. and Schmidt A. Micro-blog:sharing and querying content through mobile phones and social participation. In *MobiSys*, pp. 174–186, 2008.
32. Williams B.M. and Hoel L.A. *Modeling and forecasting vehicular traffic flow as a seasonal stochastic time series process*, Pennsylvania State University, Pennsylvania Transportation Institute, 1999.
33. Cook D.J. and Holder L.B. Graph-based data mining, *Intelligent Systems and their Applications*, Vol. 15, no. 2, pp. 32–41, IEEE, 2000.
34. Hirsch M., Pardalos P.M. and Resende M.C. Sensor registration in a sensor network by continuous GRASP, *Proceedings of IEEE Military Communications Conference*, Washington, D.C., Oct. 2006.

Study of Mobile Mixed Sensing Networks in an Automotive Context

Animesh Chakravarthy, Kyungyeol Song, Jaime Peraire, and Eric Feron

Abstract Mixed sensing mobile networks comprise of mobile sensors that have different sensing capabilities. We look at such sensor networks in an automotive context; wherein automobiles with two levels of sensing (and consequently with two different dynamics) are 'mixed' among one another. The two levels of sensing considered are local, near-neighbor information sensing; and advance, far-ahead information sensing. We look for conditions governing the way the two types of sensors should be mixed (i.e., required minimum number and distribution of the far-ahead information sensing vehicles in a mixed N-vehicle string) in order to meet certain performance objectives. In this regard, two types of models are considered – microscopic models (using ODEs) governing individual vehicle behavior; and macroscopic models (using PDEs) governing average behavior of groups of vehicles. The performance objective that we address is related to the safety of the overall network, and depends on the type of model being adopted – thus in the microscopic model, the performance metric is one of achieving zero collisions, in conditions where there otherwise would have been multi-vehicle collisions; while in the macroscopic model, the metric is one of weakening the shock waves that otherwise would have existed.

A. Chakravarthy (✉)
Wichita State University, Wichita, KS, USA
e-mail: animesh.chakravarthy@wichita.edu

K. Song
McKinsey Corporation, Seoul, South Korea
e-mail: drsky@alum.mit.edu

J. Peraire
Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: peraire@mit.edu

E. Feron
Georgia Tech, Atlanta, GA, USA
e-mail: feron@gatech.edu

V.L. Boginski et al. (eds.), *Sensors: Theory, Algorithms, and Applications*,
Springer Optimization and Its Applications 61, DOI 10.1007/978-0-387-88619-0_8,

1 Introduction

The concept of coordinating the use of sensors of different types, so as to enhance the overall performance, is well known. Such a concept could be applicable in a variety of ways, and in a variety of scenarios. Complimentary filtering is possibly the most well-known version of this concept, wherein one combines the use of sensors that have good characteristics at low frequencies with sensors that have good characteristics at high frequencies, in order to measure a single quantity that can vary over a broad frequency range. In the context of mobile sensor networks, the concept of 'mixing' different types of sensors could be applicable in the format shown in Fig. 1. The figure shows a network of N mobile sensors, comprising of two types of sensors schematically depicted in orange and green. This differentiation is performed on the basis of two features – (a) their respective dynamics and (b) their ability to communicate. The orange sensors share a common set of dynamics (denoted by $\dot{x} = f_1(x, u)$); while the green sensors share a different set of dynamics (denoted by $\dot{x} = f_2(x, u)$), with x and u representing the state and input vectors respectively, of the entire sensor network. Additionally, the orange sensors have access to state information of their immediate neighbors, as well as the ability to communicate with other orange sensors in the network (say through wireless communication), while the green sensors have access only to the state information of their immediate neighbors. It is reasonable to consider each orange sensor as being inherently more expensive than each green sensor (due to the wireless communication abilities required by the former). With a view to keeping the total number of expensive sensors as low as possible, a key issue then is that for given N, f_1, and f_2, what should be the *minimum* number of orange sensors in the network, and how should these orange sensors be *distributed* (i.e., placed in the network), so that a given performance metric of the overall mobile sensor network is satisfied?

The answer to the above question obviously depends on the *type* of mobile sensor network under consideration, as well as the *performance metric* of the network, that is of interest to us. In this chapter, the network we consider is one of automobiles on the highway; and the performance metric we consider is one related to the occurrence of pileup crashes (i.e., multiple vehicle collisions or shock waves on the highway).

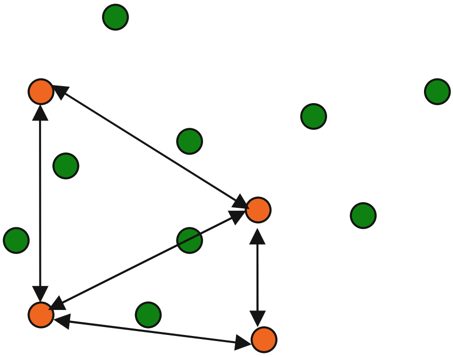


Fig. 1 Mobile sensor network comprising of sensors of two types. The orange sensors have the ability to communicate wirelessly among themselves; while the green sensors do not. The two types of sensors also have different dynamics

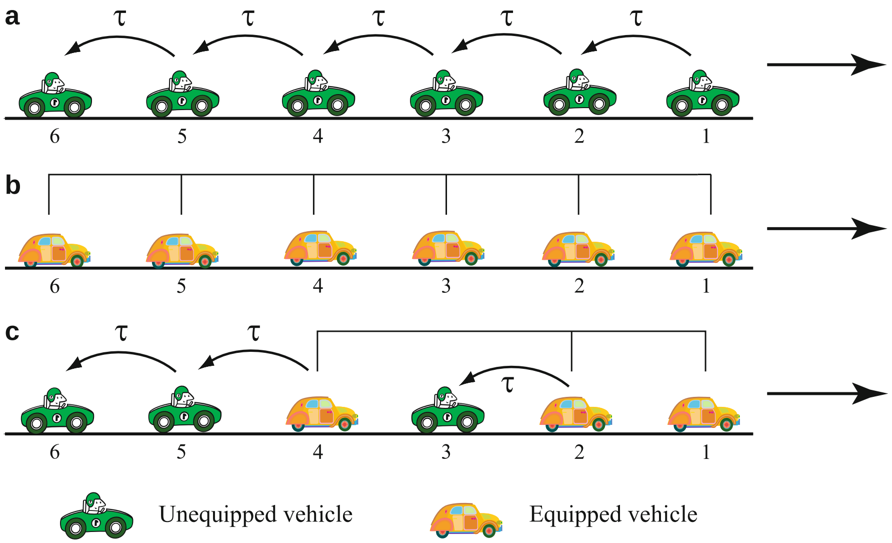


Fig. 2 Propagation of slowdown information in cases (**a**), (**b**), and (**c**)

Rear end collisions are a major cause of multiple car crashes, especially during bad weather conditions [1–4]. The cause for such crashes is that each driver gets warned of an impending slowdown ahead, only when the brake-lights of the car immediately in front of him/her, turn on. This is particularly true during poor visibility conditions, and/or while driving behind a large vehicle, when a driver is unable to look several cars ahead, as he/she otherwise normally would have. Thus each car-driver unit is able to sense only the relative velocity and inter-vehicle distance to the car immediately in front of him/her. So, if we consider a platoon of cars traveling on a single lane, and the lead car executes an abrupt deceleration, this information is propagated from car to car in a staggered fashion (Fig. 2a), as the brake-lights of each car come on, one after the other. There is an associated delay τ for each car as the information propagates through the line of cars, (τ comprises of the time it takes for each driver to realize that the front car's brake-lights are on, and to react with a corresponding deceleration that turns his/her own brake-lights on). Thus if car 1 (i.e., the lead car) poses a hazard by a sudden deceleration that turns on its brake-lights at time $t = 0$, then the kth car $(k > 2)$ senses the slowdown ahead at $t = (k - 2)\tau$, and turns on his/her own brake-lights $(k - 1)\tau$ seconds after the first generation of the hazard. Thus the driver's reaction time τ gets continuously accumulated as the information propagates through the line of cars. Consequently, cars with higher position number k in the platoon are more likely to crash. This illustrates that this mode of transmission of information of a slowdown (from car to car as in Fig. 2a) is too slow, and since drivers further behind in the platoon sense the hazard later than those ahead of them, they consequently have lesser time to react. Car pile-up crashes are the consequence of this delayed sensing.

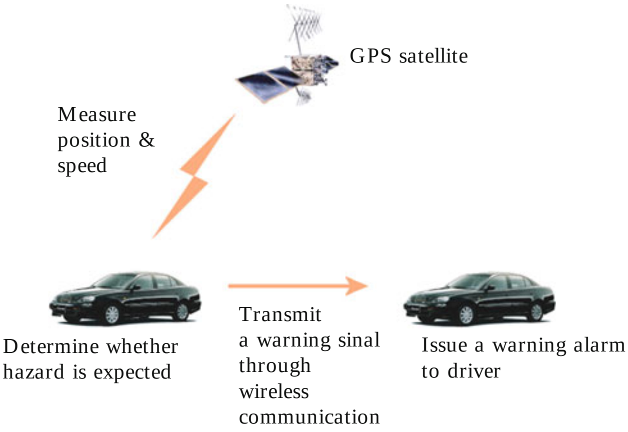

GPS satellite
Measure position & speed
Determine whether hazard is expected
Transmit a warning sinal through wireless communication
Issue a warning alarm to driver

Fig. 3 Schematic representation

The concept of inter-vehicle communication has been discussed in the literature, e.g., in [5,6,8–12]. Researchers discuss concepts such as vehicular ad-hoc networks (VANETS) through which information can propagate in a wireless mode. In particular, one can equip cars with a slowdown warning system [15–18]. A car equipped with such a system has the ability to (a) Automatically transmit a slowdown warning signal whenever it decelerates abruptly, or its velocity becomes dangerously low for highway driving conditions, and (b) Receive a slowdown warning signal, and alert the driver accordingly, if it deems the signal to be relevant. A schematic representation of this slowdown warning system concept (involving the use of GPS) is shown in Fig. 3. Since wireless communication delays are of the order of milliseconds [32], they can be considered small compared to human reaction time delays; therefore information of a slowdown is sensed near simultaneously by all the cars within the communication range (as in Fig. 2b). This allows sufficient time for the car drivers to react appropriately to avoid the crash.

We now study the influence of having only a few equipped vehicles in this multi-vehicle system. In this case, information is propagated as in Fig. 2c. The figure shows a schematic representation of some equipped vehicles, scattered among the unequipped vehicles. Thus if car 1 transmits a warning signal at time $t = 0$, then since cars 2 and 4 are equipped with the warning system, both of them receive the warning signal at $t = 0$ and thus sense the hazard almost as soon as it has occurred. Furthermore, the unequipped cars 5 and 6 also sense the hazard since they receive (indirect) information of the slowdown at τ and 2τ, respectively, which contributes significantly to safety improvement, as compared to the case in Fig. 2a when cars 5 and 6 receive the warning information only at $t = 3\tau$ and 4τ, respectively.

Figure 2c is thus the context in which we explore the original schematic depicted in Fig. 1. Each car-driver unit is a sensor in that it can sense the velocity and inter-vehicle distance to its immediate predecessor. The equipped vehicles (that have the ability to communicate wirelessly among themselves, and thus sense advance far-ahead information) represent the orange sensors of Fig. 1. The unequipped vehicles (that do not have the ability for inter-vehicle communication, and can sense only local information) represent the green sensors of Fig. 1. We will explore this mobile sensor network in two modeling contexts:

(a) Microscopic models that involve the use of ODEs (Ordinary Differential Equations) to model the dynamics of the vehicles. We will use linear ODEs in this chapter, and formulate the problem as a single-lane problem [18, 19].

(b) Macroscopic models that involve the use of PDEs (Partial Differential Equations) to model the vehicle dynamics. We will use nonlinear, hyperbolic PDEs in this chapter and formulate the problem as an aggregate multi-lane problem [19, 20].

Note that while microscopic models can detect the presence/absence of collisions between vehicles, the same is not true for macroscopic models. Since macroscopic models govern average or aggregate behavior of groups of vehicles, it is therefore not possible for these models to detect actual vehicular collisions. Macroscopic models can, however, demonstrate shock waves, i.e., discontinuous solutions representing large decreases in average velocity with corresponding large increases in average vehicle density, where the average vehicle density denotes the average number of cars per unit length of the highway. Now since a large decrease in average velocity represents several vehicles having to slam on their brakes, the presence of a shock wave can be construed to be qualitatively similar to the occurrence of a pileup crash. There exists an important analogy between the occurrence of car pile-up crashes and the shock waves occurring in compressible flow dynamics. In compressible fluid flow, a shock wave occurs when an object in the fluid travels faster than the speed at which information of its existence propagates ahead. Likewise, in traffic flow, car pile-ups occur if the cars travel at a speed higher than that at which the information (of the lead car's deceleration) propagates. Shock waves in traffic flow dynamics have been discussed elsewhere, for example, in [35]. In both types of models (microscopic and macroscopic), the equipped and unequipped vehicles will have different dynamics. In the microscopic model case, we will look for conditions governing the minimum number and distribution of equipped vehicles in a N-vehicle network that will guarantee zero collisions (i.e., our performance metric is one of zero collisions); while in the macroscopic model case, we will look for conditions on the equipped vehicle average density (i.e., average number of equipped vehicles per unit length of the highway) that will ensure that the shock wave strength is weakened (i.e., our performance metric is one of velocity change occurring across a shock).

This chapter is organized as follows. Microscopic models of this mobile sensor network (with 'mixed' sensing capabilities) are addressed in Sect. 2, wherein the goal is to arrive at analytical estimates of the requisite minimum number and distribution of equipped vehicles that will guarantee zero collisions. Macroscopic models of this mobile sensor network are addressed in Sect. 3, wherein the goal is to arrive at estimates of the requisite equipped vehicle density to significantly weaken a shock. Finally, Sect. 4 presents the conclusions.

2 Microscopic Models of the Mobile Sensor Network: Simulations and Analysis

In this section, a microscopic model is used to analyze the mixed sensing network. Section 2.1 discusses a finite-state model of this network. Section 2.2 demonstrates simulation results showing the decrease in inter-vehicle spacing along a line of vehicles (subsequent to an abrupt deceleration by the lead vehicle) that leads to a violation (i.e., the presence of collisions) of the desired performance index of the network. It then demonstrates how the presence of a few vehicles with advance sensing capabilities has a beneficial effect on the overall safety of the interconnected system of vehicles. In Sects. 2.3 and 2.4, we conduct a theoretical analysis of the effects indicated by the simulation results of Sect. 2.2.

2.1 Finite-State Model of the Mixed Sensing Network

To account for the change in driver behavior in response to brake-lights and/or advance slowdown warnings, we assumed the driver to be a finite-state system (refer Fig. 4), whereby he could be in one of three different modes. Modeling the driver as a finite-state system has been done before, for example, [23]. Before we go into a discussion of the modes, we briefly introduce some nomenclature.

There are N vehicles in a platoon, numbered $1, 2, \dots, N$, with car 1 being the lead car, and i denoting the ith vehicle. Some of the cars are equipped with the slowdown warning system and thus have advance sensing capabilities, while the others are not. Define:

- E : The set of vehicles that have advance sensing capabilities (since they are equipped with the slowdown warning system).

- U : The set of vehicles that have only local sensing capabilities (since they do not have a slowdown warning system).

- sl_rec(i) = A flag indicating whether a vehicle $i \in E$ (i.e., an equipped car) is currently receiving a slowdown warning. A value of 1 indicates that a warning is being received, while 0 indicates otherwise.

- b_l(i) = A flag indicating whether a vehicle i (which may be equipped or unequipped) currently has its brake-lights on. A value of 1 indicates that its brake-lights are on, while 0 indicates otherwise.
  The velocities of the cars are denoted by $V_1, V_2, \dots, V_N$ while the inter car separations are denoted by $s_{1,2}, s_{2,3}, \dots, s_{N-1,N}$.

At time $t = 0$, it is assumed that all the n cars are traveling at equal velocities, and the inter car separations are all equal. The lead car then suddenly decelerates, and

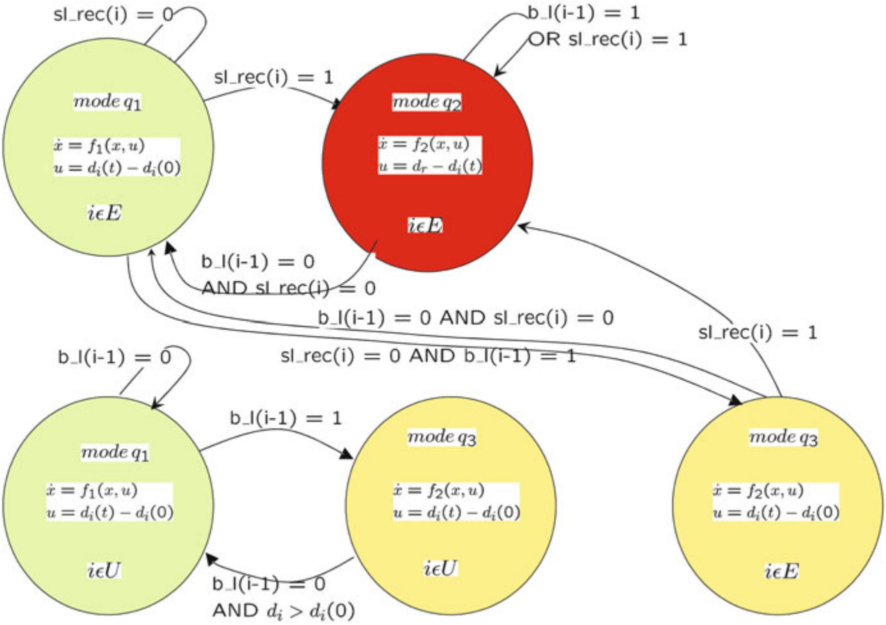


Fig. 4 Finite-state model of a car driver

emits a slowdown warning signal that is received by all the equipped cars behind it. The instant the equipped cars receive this signal, the drivers of these cars go into an alert mode and smoothly increase the distance between themselves and the car immediately in front of them. The unequipped cars, on the other hand, receive no such signal – only when the brake-lights of the car immediately in front of them come on, do these drivers go into an alert mode. However, they do not have the time to increase the distance between themselves and the car in front. We assume that the distance a car maintains to his/her immediate predecessor is equal to the product of the velocity of the following car and a quantity referred to as the time headway.

Thus, at any given time, the driver of the ith (following) car can be in one of three modes, viz., q_1, q_2, or q_3, and he/she transitions from one mode to another, depending on the flags sl_rec(i) and b_l($i-1$). The descriptions of the modes are as follows:

1. Mode q_1: This is the initial mode in which all the drivers reside, when both the flags sl_rec(i) and b_l($i-1$) are zero. This mode is characterized by 'normal' driver dynamics, manifested by 'normal' time delays. We denote the driver dynamics in this mode by $\dot{x} = f_1(x,u)$. A driver can reside in mode q_1 for long periods of time. In this mode, he/she tries to maintain a 'normal' distance between his car and the one immediately ahead of him/her (this distance is characterized by shorter time headways).

2. Mode q_2: Only the drivers in cars with advance sensing capabilities can be in this mode. These drivers transition from mode q_1 to q_2 if and only if they receive sensing information of the hazard ahead. This mode is characterized by (a) Faster driver dynamics, manifested by shorter time delays compared to q_1 (it is denoted by the equation $\dot{x} = f_2(x, u)$) and (b) A higher value of reference distance/time headway $d_{\text{ref}}(i)$, compared to q_1. Mode q_2 is a high alert mode, in which a driver resides for only a short duration, before reverting back to q_1.
3. Mode q_3: The drivers of both the equipped as well as the unequipped cars can reside in this mode. The drivers of the unequipped cars transition from q_1 to q_3 if and only if the brake-lights of the car ahead of them come on, while the drivers of the equipped cars transition from q_1 to q_3 if and only if the brake-lights of the car ahead of them come on, and additionally, they are not in receipt of a slowdown warning. This mode is characterized by the faster driver dynamics, represented by $\dot{x} = f_2(x, u)$ and the 'normal' reference distance (with a shorter time headway). A driver resides in mode q_3 only for a short duration of time.

Each of these modes, and the transitions thereof, are schematically represented in Fig. 4, for equipped and unequipped cars. Note that this figure holds for all following cars and the behavior of the lead car is treated as an input to the inter-connected system.

2.2 Numerical Simulation

Consider a string of vehicles driving on a single-lane highway. We assume that at $t = 0$, they are all driving with equal speeds and equal inter-car distances. The string of cars is modeled as an inter-connected system, with each car-driver system forming one element of the inter-connected system. Each car-driver unit is able to sense only the relative velocity and inter-vehicle distance to his/her immediate predecessor. The acceleration response of the driver of the ith vehicle is modeled by the following equation, presented in [22, 23]:

$$\begin{aligned} \frac{\mathrm{d}v_i(t)}{\mathrm{d}t} &= K_1\left(s_{i-1,i}(t-\tau) - Tv_i(t-\tau)\right) \\ &\quad + K_2\left(v_{i-1}(t-\tau) - v_i(t-\tau)\right), \\ \frac{\mathrm{d}s_{i-1,i}(t)}{\mathrm{d}t} &= v_{i-1}(t) - v_i(t), \end{aligned} \tag{1}$$

where v_i indicates the velocity of the ith vehicle and $s_{i-1,i}$ represents the inter-car distance between the ith and the $(i-1)$th vehicles, with vehicle 1 being the lead vehicle. τ indicates the response delay of each car-driver system. K_2 represents the sensitivity of each driver to the velocity difference between his/her vehicle and the one immediately ahead while K_1 is the sensitivity to the difference between the desired inter-vehicle distance and the true inter-vehicle distance. The desired

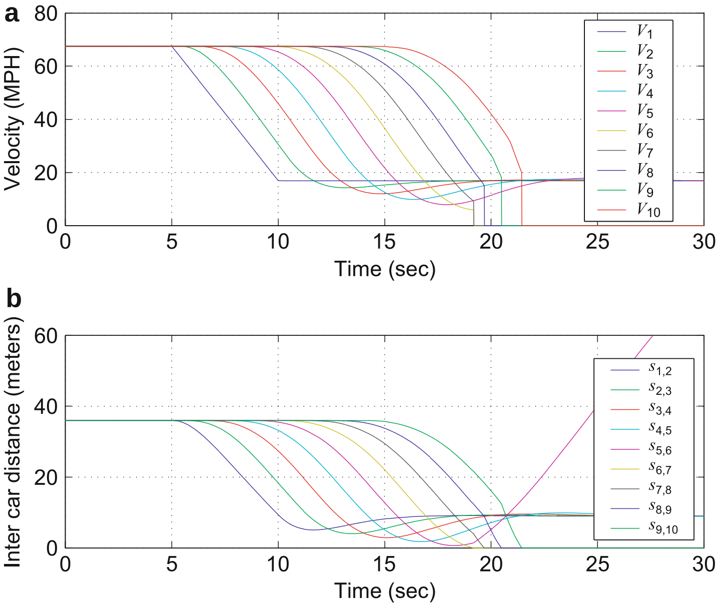


Fig. 5 All vehicles unequipped with advance sensing capability (**a**) Velocity vs. time profiles (**b**) Inter-car separation vs. time profiles

inter-vehicle distance of each driver (to the vehicle ahead) is proportional to his/her own velocity, with the proportionality constant being T (the time headway).

Consider a situation wherein all the vehicles are initially traveling at typical highway speeds of about 30 m/s, (i.e., 67.5 mph), with the inter-vehicle distance being 36 m (i.e., $T = 1.2$ s). We assume τ to be a ball-park value of 0.6 s, and then determine K_1 and K_2 such that it ensures stable, non-oscillatory behavior for each two-vehicle system. (These were obtained using guidelines available, for example, in [25]). At $t = 5$ s, the lead vehicle begins to execute an abrupt deceleration, and decelerates continuously for 5 s. Refer to Fig. 5, which shows the velocity and inter-car distance profiles of 10 vehicles, when the information of the lead vehicle's deceleration is transmitted from vehicle to vehicle, as in Fig. 2a. It can be seen that the values of the minimum vehicle velocity and minimum inter-vehicle distance keep decreasing with increasing vehicle index, until vehicle 6 is rear-ended by vehicle 7, and crashes occur for all the vehicles behind. It can be seen from the figure that if there were more vehicles behind vehicle 10, they too would all collide, thus leading to a pile-up. This is the phenomenon of string instability [13, 24, 31].

String instability refers to the amplification of velocity errors as these errors travel along a string of vehicles. If we define $\epsilon_1(t) = V_1(t) - V_2(t)$, $\epsilon_2(t) = V_2(t) - V_3(t), \ldots, \epsilon_{N-1}(t) = V_{N-1}(t) - V_N(t)$, then a string of vehicles is said to be l_p stable if:

$$||\epsilon_1(t)||_p \leq ||\epsilon_2(t)||_p \leq \cdots \leq ||\epsilon_N(t)||_p, \qquad (2)$$

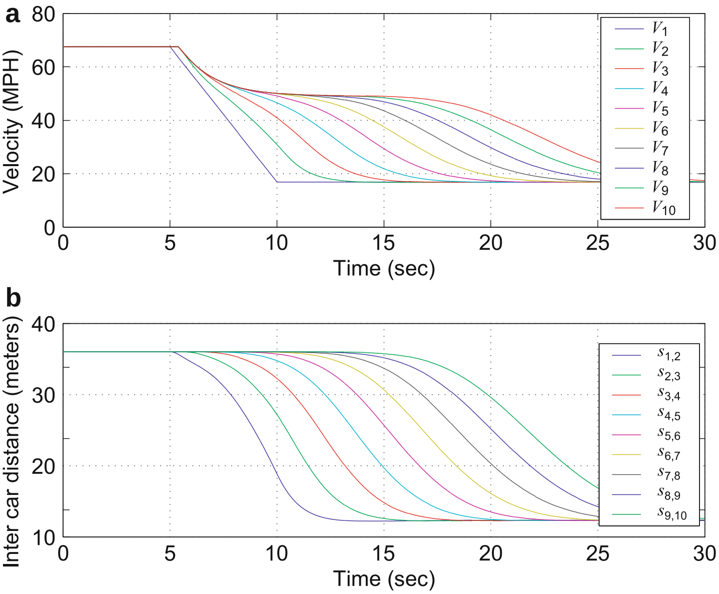


Fig. 6 All vehicles equipped with advance sensing capability (**a**) Velocity vs. time profiles (**b**) Inter-car separation vs. time profiles

where p indicates the pth norm. In the context of this chapter, we are interested in the amplification (or otherwise) of the ∞ norm of the velocity errors, i.e., we are interested in l_∞ string stability. The smaller the time headways, the more severe is the instability; for sufficiently large time headways, string instability does not manifest [31].

Now, consider a scenario when all the 10 vehicles are equipped with the slowdown warning system, and thus all vehicles possess advance sensing capabilities (as shown in Fig. 2b). If the lead vehicle now executes an identical deceleration profile, the trailing vehicles are able to react much earlier (vehicle 10, for example, is able to react τ seconds after the lead vehicle begins to decelerate, as opposed to 9τ seconds that it would otherwise have taken, if all vehicles were unequipped). We make the reasonable assumption that on receipt of the slowdown warning signal, the driver of each equipped vehicle transitions to a slightly lower value of τ than when he was unequipped (in these simulations, we assume $\tau = 0.4$ s for an equipped vehicle – this signifies the increased alertness of the driver on receipt of the warning signal). Furthermore, the driver of each equipped vehicle attempts to increase his/her time headway to the vehicle in front of him/her, in anticipation of the imminent slowdown. The result is shown in Fig. 6, where, on receipt of the warning signal, each driver tries to increase the time headway to the vehicle immediately ahead (from the original $T = 1.2$ s to $T = 1.65$ s). The string instability trend (earlier observed in Fig. 5) is no longer seen.

Now consider a case when only some cars possess long distance sensing capabilities, while other cars possess only local sensing capabilities. It turns out

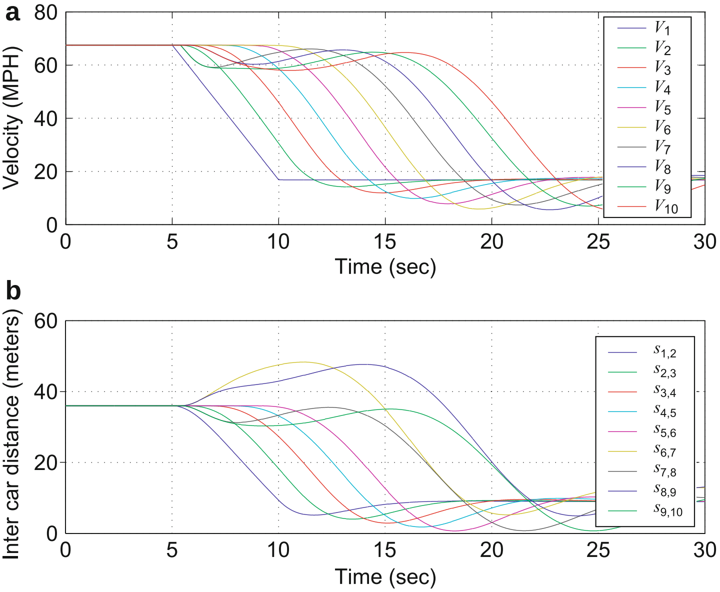


Fig. 7 Only vehicles 1, 7, and 9 equipped with advance sensing capability

however, that in many cases, even if a fraction of the cars are equipped, this can still be sufficient to break the trend of decreasing inter-vehicle spacings as it propagates down the line of vehicles, and this can prevent pile-up crashes. This is illustrated in Fig. 7, where only vehicles 1, 7, and 9 are equipped with advance sensing transmit/receive capabilities. It is seen that after the lead vehicle decelerates, there is an onset of decreasing inter-vehicle spacings from vehicles 2 to 6. However, the fact that vehicle 7 is equipped breaks this trend, and in fact, the minimum value of V_7 (as also x_6) is higher than V_6 (respectively, x_5). Furthermore, since vehicle 8 is unequipped, it re-initiates the trend of decreasing inter-vehicle spacings and therefore the minimum value of V_8 (as also x_7) is indeed lower than that of V_7 (respectively, x_6); yet it is higher than that in the case when vehicle 7 was unequipped (see, Fig. 5). Similarly, since vehicle 9 is equipped, not only is the minimum value of x_8 high enough, but also that of x_9 is higher than what it was when vehicle 9 was unequipped. Consequently, no crashes occur. This shows that it is possible that even if a fraction of the vehicles are equipped, they are able to ensure the safety of not only themselves, but the unequipped vehicles as well.

2.3 *Mathematical Definitions and Safety Conditions*

We use the following notation : F and f represent the same signal (or system) in the frequency and time domain, respectively, i.e., $F(s) = L(f) =$

$\int\limits_0^\infty f(t)\mathrm{e}^{-st}\mathrm{d}t$. s is the Laplace variable. $\|f\|_\infty$ denotes $\sup_{t>0}|f|$, while $\|F\|_\infty$ denotes $\sup_{\omega>0}|F(j\omega)|$. Both $\|f\|_1$ and $\|F\|_1$ denote $\int\limits_0^\infty |f\ (t)|\,\mathrm{d}t$. Consider a string of N vehicles driving on a single-lane with the dynamics of each vehicle defined by:

$$V_i\ (s)\ =\ G_i\ (s)\ V_{i-1}\ (s), \tag{3}$$

for all $i\ \in\ N$, where V_i represents the longitudinal velocity of the ith vehicle. $G_i\ (s)$ is the transfer function connecting the velocity response of the ith vehicle to that of the $(i-1)$th vehicle. While the simulation results of Sect. 2.2 (that use (1)) assume a $G_i(s)$ of the form $G_i(s) = \frac{K_1+sK_2}{s^2\mathrm{e}^{s\tau}+s(K_1T+K_2)+K_1}$, the statements made in this subsection are true for arbitrary $G_i(s)$. We assume that all the vehicles in this inter-connected system are driving with equal initial speeds and equal initial inter-vehicle spacings.

Lemma 1. *For an inter-connected system of vehicles in (3), the fluctuation of the inter-vehicle spacing $\Delta\,(x_i - x_{i+1})$ can be written as:*

$$\frac{\Delta\,(X_i - X_{i+1})}{\Delta\,(X_{i-1} - X_i)} = \left[\frac{1-G_{i+1}}{1-G_i}G_i\right] = \widehat{G}_i. \tag{4}$$

Also, $\Delta\,(x_i - x_{i+1})$ can be written in terms of the velocity of the lead vehicle V_1 as:

$$\Delta\,(x_i - x_{i+1}) = L^{-1}\left\{\left[\prod_{k=2}^{i} G_k\right]\frac{(1-G_{i+1})}{s}V_1\right\}, \tag{5}$$

where x_i is the position of the ith vehicle (x_1 is the position of the lead vehicle) and x_i are measured in a direction such that $x_i(0) > x_{i+1}(0)\forall i$. The condition for a pileup crash can be described in terms of $\widehat{G}_i$, as shown in the following theorem, proved in [18].

Theorem 1. *Consider an inter-connected system of N vehicles governed by (3), with all vehicles driving with equal initial speeds and equal initial inter-vehicle spacing s_0. Then, if the lead vehicle executes an abrupt deceleration, it is guaranteed that there will be no pileup crash if*

1. $\|\widehat{g}_i(t)\|_1 \le 1$ *for all i, and*
2. $\|\Delta\,(x_1 - x_2)\|_\infty \le s_0$

Here, $\widehat{g}_i$ represents the impulse response of the transfer function $\widehat{G}_i$.

We should note that the condition $\|\widehat{g}_i\|_1 > 1$ for some i does not necessarily imply that there will be a pileup crash in the inter-connected system, because $\|\Delta\,(x_i - x_{i+1})\|_\infty$ can be smaller than $\|\Delta\,(x_{i-1} - x_i)\|_\infty$, even when $\|\widehat{g}_i\|_1 > 1$. At the same time however, if $\|\widehat{g}_i\|_1 > 1$, then in the absence of further knowledge of the deceleration profile of the lead vehicle, one cannot *guarantee* the absence of a pile-up crash. (This statement is particularly true for large deceleration magnitudes of the lead vehicle). Note however that the system satisfying the condition in Theorem 1 will never have a pileup crash in any event.

2.4 Effect of Equipped Vehicles

In this section, we investigate the role of the equipped vehicles (that have advance sensing capabilities across a wireless communication link) in mitigating the generation of a pileup crash in a mixed sensing environment. In this context, we therefore assume that, on receipt of a slowdown warning signal, $G_i\,(s)$ can be either $U(s)$ or $E(s)$, i.e.,

$$V_i \;=\; \begin{cases} U(s)V_{i-1}, \text{if } i\text{th vehicle has only local sensing capability} \\ E(s)V_{i-1}, \text{if } i\text{th vehicle has advance sensing capability.} \end{cases} \tag{6}$$

In general, $U(s)$ is characterized by high values of τ accompanied by small values of T; while $E(s)$ is characterized by low values of τ accompanied by high values of T. The differences between $U(s)$ and $E(s)$ (for the driver dynamics presented in (1) are clearly seen in Fig. 8. In this figure, the frequency response of the vehicle dynamics is plotted for different values of T and τ. It can be seen that for a given value of T, as τ increases, it has the effect of increasing the magnitude of the frequency response. At the same time, for a given τ, as T increases, this has the effect of decreasing the frequency response magnitude.

Using the above guidelines, we assume that $U(s)$ and $E(s)$ have the following characteristics.

1. $$\|U(s)\|_\infty > 1, \text{ and } |U(0)| = 1 \tag{7}$$
2. $$\|E(s)\|_\infty = |E(0)| = 1, \text{ and } e(t) > 0 \;\forall t > 0\,,$$

where $e(t)$ is the impulse response of $E(s)$.

If all the vehicles have only local sensing capabilities, i.e., $G_i\,(s) = U(s)\;\forall i$, then $\widehat{G}_i\,(s) = U(s)\;\forall i$ by (4), and therefore, $\|\widehat{g}_i\|_1 = \|u(t)\|_1 > 1\;\forall i$. By the analysis in Sect. 2.3, we have seen that this condition on $\|\widehat{g}_i\|_1$ is indicative of the possibility of occurrence of a pileup crash.

On the other hand, when all the vehicles are equipped with advance sensing capabilities, we have $\|\widehat{g}_i\|_1 = \|e(t)\|_1 = 1\forall i$ (because $\widehat{G}_i\,(s) = E(s)\forall i$), and therefore, $\|\Delta\,(x_i - x_{i+1})\|_\infty \leq s_0$ for all i, which satisfies the condition for zero collisions.

Figure 9a shows an example of a velocity profile, wherein a car decelerates sharply from an initial velocity of 30 m/s and comes to a complete stop over a time span of a little more than one second. The corresponding Fourier Transform of this signal is shown in Fig. 9b, from which it can be seen that there is substantial magnitude content of the signal at the frequency ω_0 shown in Fig. 8.

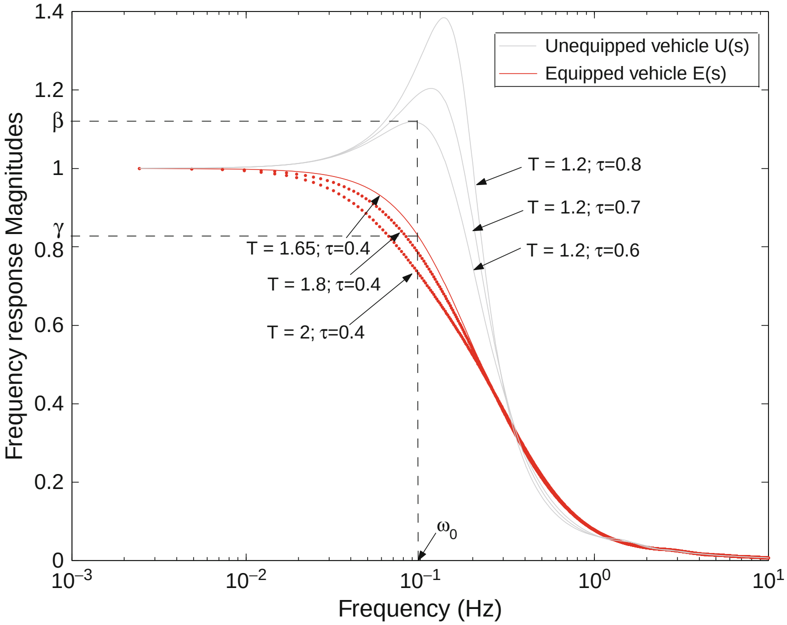


Fig. 8 Frequency responses of $U(s)$ and $E(s)$

Figure 10 shows the impulse responses of $U(s)$ and $E(s)$ for different values of T and τ. Smaller values of T, accompanied by larger values of τ (that characterize $U(s)$) lead to oscillatory impulse responses (with the amplitude of oscillation decreasing with decreasing τ). On the other hand, larger values of T, accompanied by smaller values of τ (that characterize $E(s)$) lead to non-oscillatory responses.

Figures 8 and 10 indicate that in the case of an equipped vehicle, the larger the increase in the time headway T that a driver attains (subsequent to the receipt of a slowdown warning signal), the greater the attenuation that equipped vehicle exerts on the errors propagating through the mixed string of vehicles. This fact is also brought out in the following ten car simulation, in which cars 1, 7, and 9 are equipped. In this simulation, three different scenarios of headway increase are considered, viz., $T = 1.65, 1.8, 2\,$s. It can be clearly seen in Figure 11 that the larger the headway increase, the higher the higher are the minimum values of velocity and inter-car distance (to the car ahead) of the equipped cars.

We now consider a mixed sensing environment in which only a small number of vehicles are equipped. The following theorem, proved in [18] provides a sufficient condition that guarantees the performance metric of this mixed sensing network is satisfied.

Theorem 2. *Consider an inter-connected system of N vehicles governed by (3), with all vehicles driving at equal initial speeds with equal inter-vehicle spacing s_0, and L out of N vehicles are equipped with advance sensing capabilities. Assume*

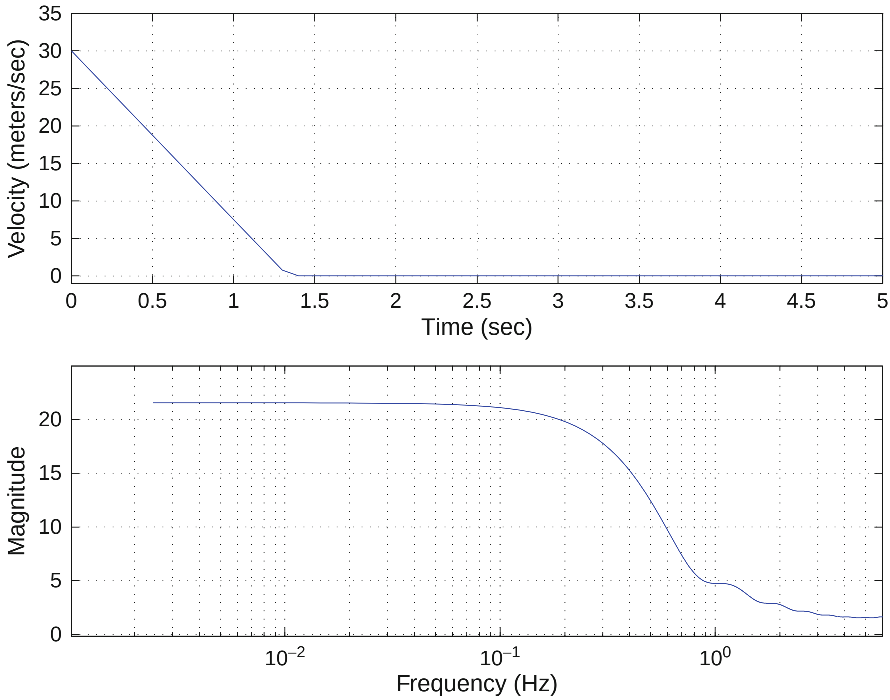


Fig. 9 An example velocity profile and its corresponding frequency content

that the lead vehicle executes an abrupt deceleration, such that when all vehicles are unequipped, collisions are initiated at the n*th vehicle (i.e.,* $\|\Delta(x_i - x_{i+1})\|_\infty > s_0$ *for* $n \leq i \leq N - 1$*). Then, under the same deceleration profile of the lead vehicle, it is guaranteed that there will be zero collisions if*

$$L \geq M, \tag{8}$$

where $M = N - n + 1$ *is the number of vehicles that would have crashed if all vehicles were unequipped.*

In the above theorem, it should be noted that there is no constraint on the distribution of the equipped vehicles within the N vehicle system. In other words, as long as $L \geq N - n + 1$, the performance metric of this mixed sensing network is always satisfied for any distribution of the vehicles with advance sensing capabilities.

The theorem in (8) is a sufficient condition to avoid a pileup crash. In other words, there could be situations where the number of equipped vehicles is smaller than M, but this is still adequate to avoid the pileup crash completely [20].

In order to investigate the performance metric for $L < M$ and derive a condition for no pileup crash that is less conservative condition than (8), we make another assumption on the dynamics of the inter-connected vehicle system and the

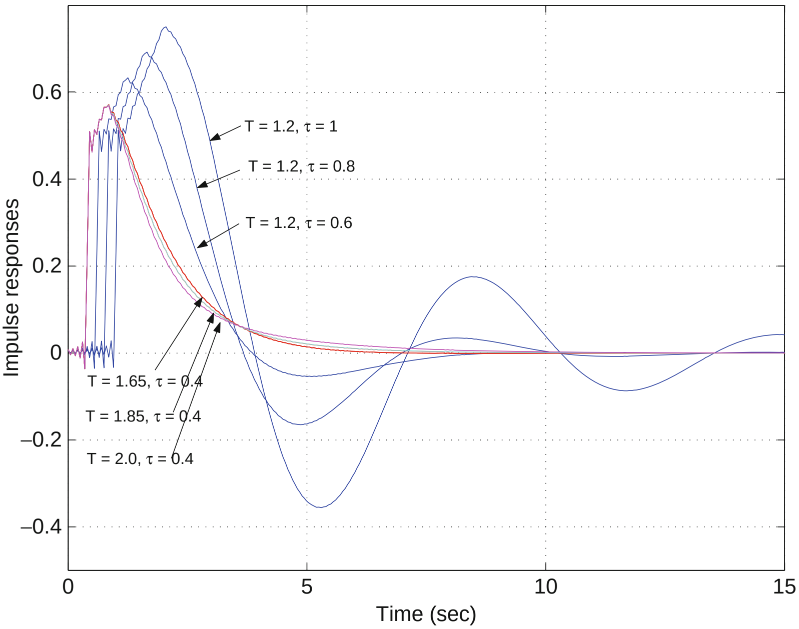


Fig. 10 Impulse responses of $U(s)$ and $E(s)$

deceleration profile of the lead vehicle. That is, we assume that $G_k(s)$ and V_1 satisfy the following inequality,

$$\|\Delta\,(x_i - x_{i+1})\|_\infty \le \alpha \left\| \prod_{k=2}^{i} G_k \right\|_\infty \tag{9}$$

where,

$$\alpha = \left\| \int_0^t (v_1(t) - u(t) * v_1(t))\,\mathrm{d}t \right\|_\infty \tag{10}$$

Here, '$*$' is a convolution integration in the time domain, i.e., $(u * v_1)(t) = \int_0^t u(t-\tau)v_1(\tau)\mathrm{d}\tau$.

Clearly, the condition in (9) is not guaranteed in general. However, in (9) turns out to be true [18] for the $G_k(s)$ represented by (1) and a V_1 representing a typical deceleration, shown in Fig. 13. Using the assumption in (9) allows us to derive a much less conservative condition (for satisfying the performance metric) than the one in (8).

Figure 12 shows the relation between the amplification factor (which is defined as $\frac{\|\Delta(x_2 - x_3)\|_\infty}{\alpha}$), $\|U\|_1$ and $\|U\|_\infty$ for varying values of $\tau = 0.2, 0.4, 0.8, 1$ s. It can be seen that with increasing values of τ, the amplification factor increases (as was also evidenced from Figs. 8 and 10); yet, at the same time, the amplification factor remains consistently lower than $\|U\|_\infty$.

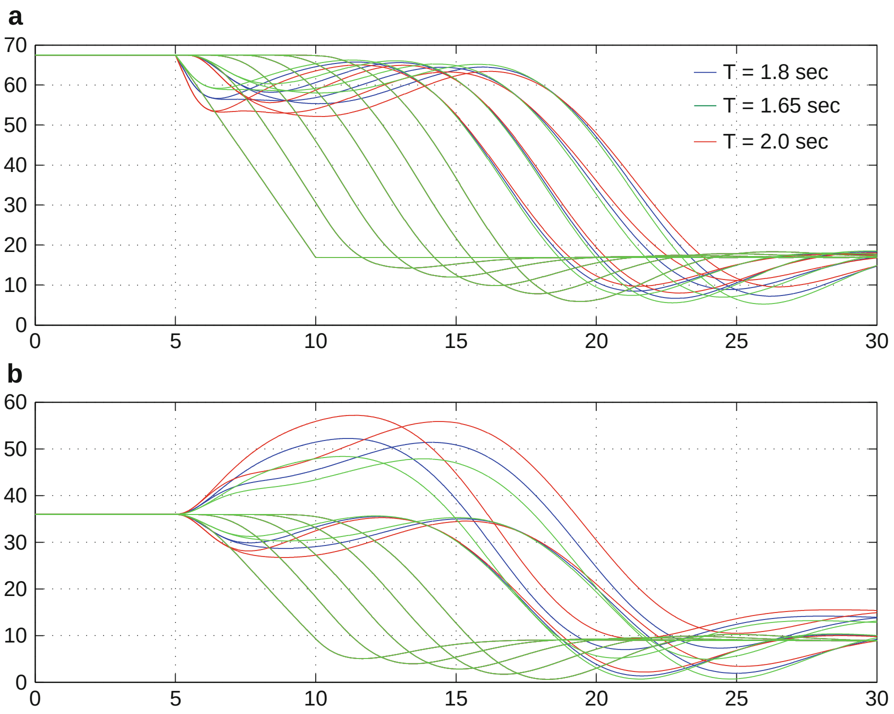


Fig. 11 Effect of varying headway increases of an equipped vehicle on receipt of the slowdown warning signal. **(a)** Velocity (mph) versus Time (s); **(b)** Inter-vehicle spacing (m) versus Time (s)

The less conservative condition (obtained as a consequence of using (9)) is given in the following theorem, proved in [18].

Theorem 3. *Consider N vehicles driving with equal initial speeds and equal inter-vehicle spacing s_0, L out of N vehicles are equipped with advance sensing capabilities, and the lead vehicle decelerates abruptly such that (when all cars are unequipped) collisions are initiated at the nth vehicle, i.e., $\|\Delta\,(x_i - x_{i+1})\|_\infty > s_0$ for $n \leq i \leq N-1$. Let L_k be the number of equipped vehicles between the first and the $(n+k-1)$th vehicle (by definition, $L_{N-n+1} = L$). Under the same deceleration profile of the lead vehicle and the assumption in (9), the performance metric (of zero collisions) of the mixed sensing network is satisfied, if*

$$L_k \geq \lfloor k\lambda \rfloor \; for 1 \leq k \leq N-n+1, \tag{11}$$

where

$$\lambda = \frac{\log\beta}{\log\beta - \log\gamma}. \tag{12}$$

Here, $\beta = \|U(s)\|_\infty = |U\,(j\omega_0)| > 1$, ω_0 is the frequency at which $|U\,(j\omega)|$ is maximum, $\gamma = |E\,(j\omega_0)| < 1$, and $\lfloor x \rfloor$ denotes the smallest integer greater than

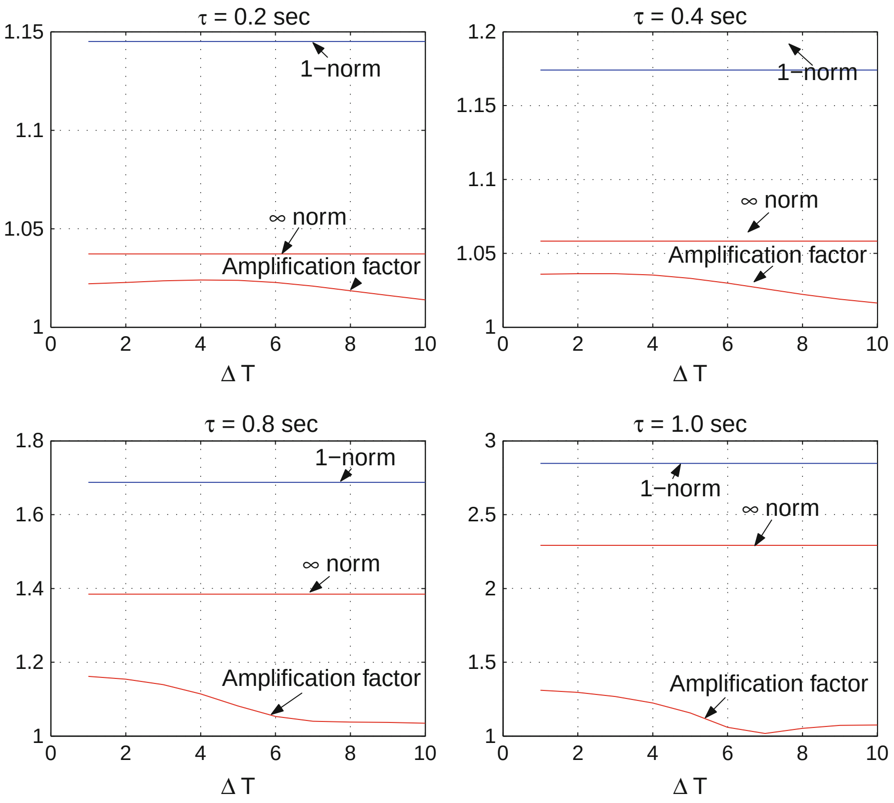


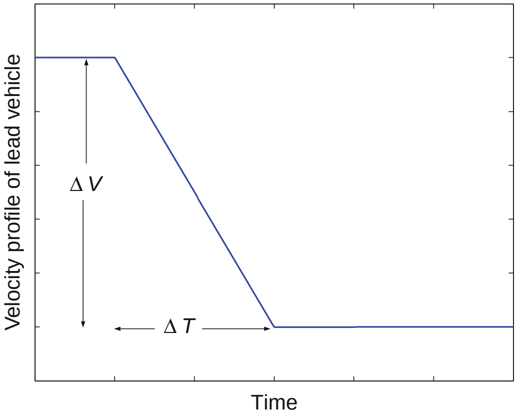
Fig. 12 Effect of varying τ on the relation among $\|U\|_1$, $\|U\|_\infty$, and actual amplification factor

x. (Figure 8 demonstrates β and γ for a representative choice of $U(s)$ and $E(s)$). Therefore, the total number of equipped vehicles L should be greater than $\lfloor M\lambda \rfloor$ to guarantee that there will be no pileup crash, where $M = N - n + 1$ is the number of vehicles that would have crashed if all vehicles were unequipped.

It should be noted that the new condition $L \geq \lfloor M\lambda \rfloor$ to avoid a pileup crash is much less conservative than the condition $L \geq M$, because $\lambda < 1$. For example, the vehicle dynamics used in the simulation in Sect. 2.2 give $\beta = 1.12$ and $\gamma = 0.85$, which yields $\lambda = 0.41$. Therefore, the number of equipped vehicles that will enable the performance metric of the mixed sensing network be satisfied when $N = 100$ and $M = 20$ is $\lfloor 20 \times 0.41 \rfloor = 9$. Furthermore, these 9 vehicles need to be distributed in a manner so that at least $\lfloor 0.41k \rfloor$ equipped vehicles need to be present between the first and the $(n + k - 1)$th vehicle for $1 \leq k \leq N - n + 1$.

It was found that the probability of satisfying the condition in (11) is quite acceptable in most cases. This is shown in Fig. 14 as N varies from 10 to 50 and $\lambda = 0.5$. In this figure, we assume for each N that 20% of vehicles experience collisions when none are equipped (i.e.,M=$\lfloor 0.2N \rfloor$), and we equip $L = \lfloor M\lambda \rfloor$ of

Fig. 13 Typical velocity profile of lead vehicle



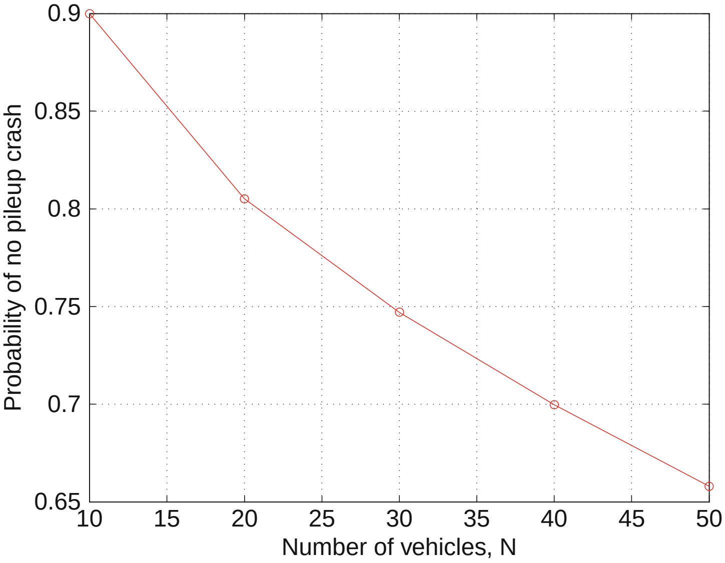


Fig. 14 Probability of achieving performance metric of zero collisions for $L = \lfloor 0.1\, N \rfloor$

the N vehicles. For example, when $N = 20$, M, and L become $M = \lfloor 0.2N \rfloor = 4$, and $L = \lfloor M\lambda \rfloor = 2$, respectively, and the probability of satisfying the condition in (11) can be computed as:

$$\frac{\text{Number of combinations of at least one vehicle equipped} \atop \text{between 1 and 17 and the other one equipped between 1 and 19}}{\text{Number of combinations of 2 out of 20 vehicles equipped}}$$
$$= \frac{(17 \times 18)\,/2}{C(20, 2)} = 0.8053, \tag{13}$$

where $C(n,k)$ represents the number of combinations of n objects taken k at a time. Proceeding in the manner outlined in (13), we can compute the probability of achieving a performance metric of zero collisions for general N, M, and L. It can be seen from Fig. 14 that there is a high probability (above 65%) of satisfying the performance metric if about 10% of the total number of vehicles possess advance sensing capabilities.

3 Macroscopic Models of Mobile Sensor Networks: Simulations and Analysis

We next look at the use of macroscopic models for analyzing our mobile mixed sensor network. The use of macroscopic models for studying traffic flows has a fairly long history. The Lighthill–Whitham–Richards (LWR) model [35, 36] represents the earliest use of macroscopic models to represent traffic flow. The LWR model is basically a first order model that is based on a gas dynamic-like continuity equation. Subsequently, second order models have been developed by Payne–Whitham [37,38] and also Phillips [40]. There has been some controversy in the past about the viability of second order models in general [39], and attempts have been made to address some of them in [41, 44]. Prigogine and Herman [47] developed traffic flow equations based on the Boltzmann equation, which have been further refined by Paveri-Fontana [48]. Based on Paveri-Fontana's equations, Helbing then derived a (gas dynamic based) third order macroscopic traffic model [44] (this model included an equation for the velocity variance), and also a second order traffic model [45], that is anisotropic in nature. Helbing also derived a gas dynamic based two species traffic model where the two species were cars and trucks [45], as also did Hoogendoorn and Bovy [46]. There have also been second order models developed by Aw and Rascle [41] – these models however, are not based on gas dynamic foundations. There have also been papers on analysis of stability in traffic flows [49, 50]. In this chapter, we adapt the Helbing model [45] to a situation wherein the two species comprise vehicles equipped with the ability to receive advance far-ahead information, interspersed with unequipped vehicles that are capable of sensing only local information. We also adapt the Helbing model appropriately in order to account for a finite speed of information propagation among the equipped vehicles. In other words, the communication speed among the equipped vehicles is now no longer assumed to be infinite (as it was in the microscopic model case discussed in the previous section).

Section 3.1 discusses the macroscopic model used and demonstrates simulation results showing initial conditions under which a shock propagates through the vehicle network when all vehicles are capable of only local sensing. Section 3.2 then discusses the macroscopic model used in the 'mixed sensing' situation, and with a finite speed of information propagation among the equipped vehicles; and demonstrates simulation results showing the strength of the shock wave under the same initial conditions, and for varying equipped vehicle densities.

3.1 All Vehicles with Only Local Sensing Capabilities: Model and Simulations

We use the model derived by Helbing [45], which in turn, has been inspired by the gas kinetic based models derived by Prigogine and Herman [47], and Paveri-Fontana [48]. The model is briefly reviewed below. The state space vector of a vehicle α, is defined as $X_\alpha(t) = [x_\alpha(t), v_\alpha(t), v^o{}_\alpha(t)]$, where $x_\alpha(t)$, $v_\alpha(t)$, and $v^o{}_\alpha(t)$ represent the position, velocity, and desired velocity, respectively of particle α. Defining $\tilde{\rho}(x, v, t)$ as the phase space density, the following Boltzmann like equation governs the evolution of $\tilde{\rho}$ in phase space [45, 47, 48]:

$$\begin{aligned}\frac{\partial\tilde{\rho}}{\partial t} + \frac{\partial}{\partial x}(\tilde{\rho}v) + \frac{\partial}{\partial v}(\tilde{\rho}\frac{v^o - v}{\tau}) = {} & \frac{(1-p_1)}{p_2}\int_v^\infty \mathrm{d}w|v-w|\tilde{\rho}(x,w,t)\tilde{\rho}(x,v,t) \\ & -\frac{(1-p_1)}{p_2}\int_0^v \mathrm{d}w|w-v|\tilde{\rho}(x,w,t)\tilde{\rho}(x,v,t) \\ & +\frac{\partial^2(\tilde{\rho}D)}{\partial v^2}. \qquad (14)\end{aligned}$$

In the above, p_1 represents the probability of a vehicle overtaking (by a lane change) a slower vehicle in front, while p_2 represents the effects of finite space interaction between a pair of vehicles. (Helbing assumes $p_1 = p_2$). Two dimensional represents the covariance of the acceleration noise of a vehicle (which is assumed to be gaussian white noise). The dynamics of particle α are governed by the following state space equations:

$$\frac{\mathrm{d}x_\alpha}{\mathrm{d}t} = v_\alpha, \qquad (15)$$

$$\frac{\mathrm{d}v_\alpha}{\mathrm{d}t} = \frac{v_\alpha^o - v_\alpha}{\tau_\alpha} + \Sigma_{\beta\neq\alpha} f_{\alpha\beta} + \xi_\alpha(t), \qquad (16)$$

$$\frac{\mathrm{d}v_\alpha^o}{\mathrm{d}t} = 0, \qquad (17)$$

where $\xi_\alpha(t)$ represents the acceleration noise, for vehicle α, and $f_{\alpha\beta}$ indicates the interaction effect of vehicle β on α. The noise $\xi_\alpha(t)$ is assumed to be gaussian white noise, i.e., it has zero mean with specified co-variance as follows:

$$\langle\xi_\alpha(t)\rangle = 0\forall\alpha, \qquad (18)$$

$$\langle\xi_\alpha(t)\xi_\beta(\tau)\rangle = 2D\delta_{\alpha\beta}\delta(t-\tau). \qquad (19)$$

By taking moments of the above equation, and defining $\rho(x,t) = \int \mathrm{d}v\tilde{\rho}(x,v,t)$ as the average density, $V(x,t) = \langle v\rangle = \frac{\int v\mathrm{d}v\tilde{\rho}(x,v,t)}{\rho(x,t)}$ as the average velocity, $\theta(x,t) = \langle(v - V(x,t))^2\rangle = \frac{\int[v-V(x,t)]^2\mathrm{d}v\tilde{\rho}(x,v,t)}{\rho(x,t)}$ as the velocity variance, the following macroscopic equations are obtained [45, 47, 48]:

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho V)}{\partial x} = 0, \tag{20}$$

$$\frac{\partial(\rho V)}{\partial t} + \frac{\partial(\rho V^2 + \rho\theta)}{\partial x} = \rho \frac{V^{eq} - V}{\tau}. \tag{21}$$

The above hierarchy of equations is closed by assuming that $\theta = A(\rho)V^2$ [45]. Furthermore, $V_e(x, t)$ represents the average equilibrium velocity and is given by [45]:

$$V^{eq}(x, t) = V^o - P(\rho_a)B\rho\tau\theta, \tag{22}$$

where V^o represents the average desired velocity, τ represents the average relaxation time, ρ_a represents the average density computed at the interaction point $x_a = x + \gamma(l + VT)$, with $l = 1/\rho_{\max}$ representing the average vehicle length ($\rho_{\max}$ represents the maximum vehicle density), T denotes the average time headway that vehicles try to maintain in the limit of maximum density, and $\gamma \in [1, 3]$ represents an anticipation factor. The factor $P(\rho)$ that takes into effect both the probability of overtaking, as well as the existence of a finite interaction-free space, is defined as $P(\rho) = \frac{V^o \rho T^2}{\tau A(\rho_{\max})(1-(\rho/\rho_{\max})^2)}$. The factor B that takes into account the anisotropic interaction effects, is given as [45]:

$$B(\delta_v) = \delta_v \frac{\mathrm{e}^{-\delta_v^2/2}}{\sqrt{2\pi}} + (1 + \delta_v^2) \int_{-\infty}^{\delta_v} \mathrm{d}y \frac{\mathrm{e}^{-y^2/2}}{\sqrt{2\pi}}, \tag{23}$$

where $\delta_v = (V - V_a)/\sqrt{\theta + \theta_a}$ with V_a and θ_a representing the average velocity and velocity variance computed at the interaction point x_a. The following values have been assumed for the numerical data (when all vehicles are unequipped) :
Average desired velocity $V_o = 110$ km/h (This corresponds to a highway speed limit that a driver would like to maintain, if the road was empty)
Average relaxation time $\tau = 15$ s
Maximum vehicle density $\rho_{\max} = 160$ vehicles/km/lane
Average time headway $T = 1$ s
$A(\rho)$ which is the density-dependent pre-factor has the profile given in Fig. 15b [45]. Using these values, the following curve representing the variation of average equilibrium velocity with density, is obtained (Fig. 15a) [45].

A good prototype of an initial condition used to test the influence of the slowdown warning system in a mixed equipage scenario, is the Reimann Problem. The Reimann Problem represents an initial condition comprising of a left state and a right state joined by a discontinuity, in each of the dependent variables, with the discontinuity occurring at the same spatial location for both variables. The left states are denoted by ρ_L and V_L, while the right states are denoted by ρ_R and V_R, respectively. Schematically, such a condition is represented as shown in Fig. 16.

In the Reimann Problems that we will consider, we will assume that $\rho_L < \rho_R$ and $V_L > V_R$. It can be seen that a large drop in average velocity, occurring over a short distance (in other words, a large negative spatial velocity gradient) is indicative of a potentially unsafe driving situation. We choose $\rho_L = 15$ vehicles/km/lane and

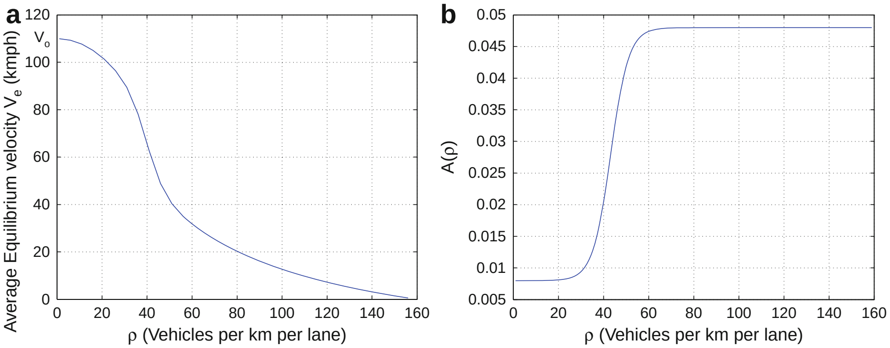
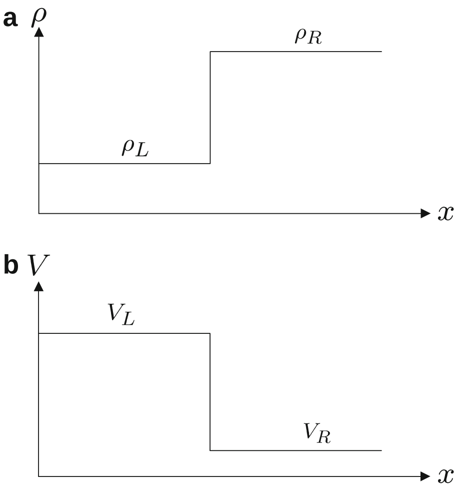


Fig. 15 (**a**) Equilibrium average velocity profile; (**b**) variance pre-factor profile

Fig. 16 Reimann problem



$\rho_R = 140$ vehicles/km/lane. We assume that ρ changes from ρ_L to ρ_R over a length of 200 m, which appears as a shock over a length scale of 10 km. Additionally, we will assume that the left and right states are both in their respective equilibrium. From Fig. 15a, it can be seen that this implies that $V_L = 105.67$ kmph and $V_R = 3.17$ kmph. Using the representation for velocity variance given above, we see that at the microscopic level, such a condition is indicative of a driver having to perform an instantaneous velocity change from an initial value that lies in the velocity probability density function $P(V_L)$ to a final value that lies in the velocity probability density function $P(V_R)$. $P(V_L)$ and $P(V_R)$ are represented in Fig. 17. We use boundary conditions as follows : $\rho(0,t) = \rho(0,0)$; $V(0,t) = V(0,0)$.

Figure 18a then shows the average density and average velocity profiles as a function of space and time, when all the vehicles are unequipped. It can be seen that the initial large negative velocity gradient propagates, almost unattenuated, backward along the highway. The wave speed at which it propagates is found as

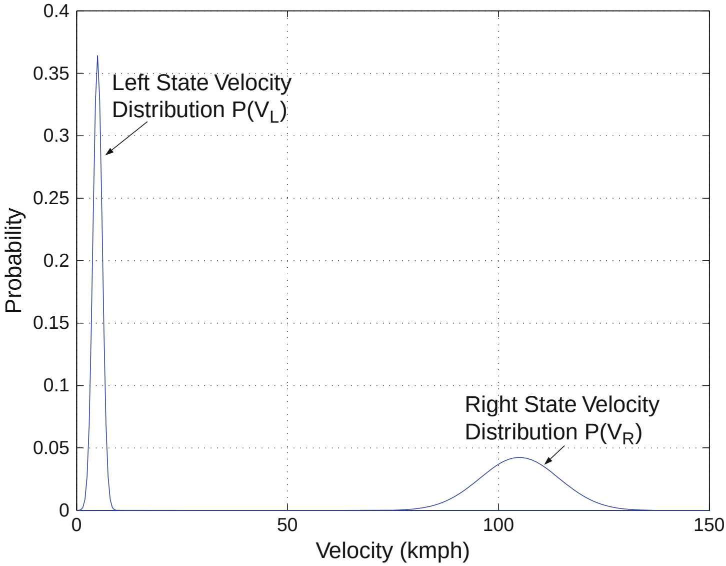


Fig. 17 Equivalent initial conditions at a microscopic level

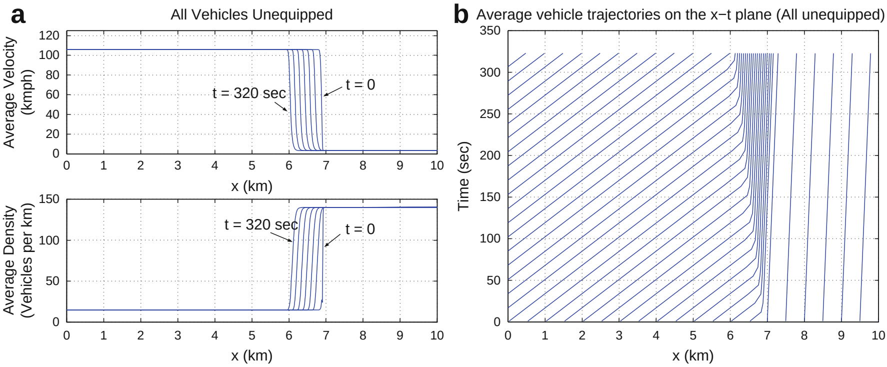


Fig. 18 (**a**) Average velocity and density profiles; (**b**) average vehicle trajectories on the *x*–*t* plane (All vehicles unequipped) for the Reimann Problem

$\frac{\rho_L V_L - \rho_R V_R}{\rho_L - \rho_R} = -9.1$ kmph. Figure 18b shows the average driver trajectories on a space-time plane. On this figure too, the shock-like behavior is clearly seen.

In other words, if we consider a situation when there is a large negative velocity gradient occurring at some point on the highway, then information of the existence of such a gradient is propagated from car to car in a staggered fashion (Figure 2a), as the brake-lights of each car come on, one after the other. This mode of information propagation is often too slow, and as a consequence, these large velocity gradients travel along the highway, mostly unattenuated.

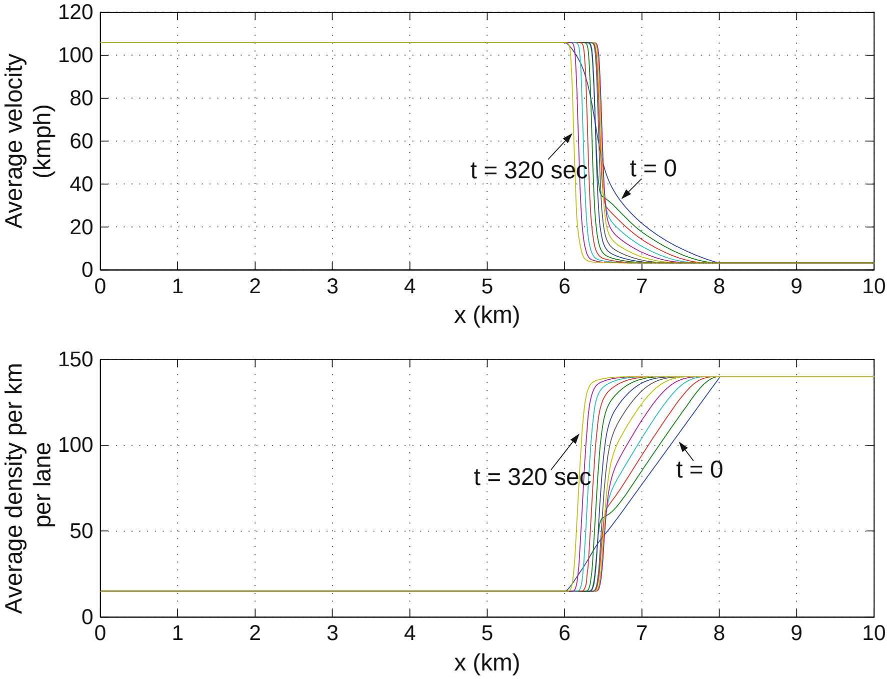


Fig. 19 Average velocity and density profiles (all vehicles unequipped) for a continuous initial condition

The presence of a large negative gradient on an initial velocity condition can also be seen as a large negative perturbation on $\frac{\partial V}{\partial x}$. As can be seen from Fig. 23a, with all vehicles unequipped, $\frac{\partial V}{\partial x}$ attenuates in magnitude initially for a short distance, only very slightly, and then propagates along unattenuated. If we define $||\frac{\partial V}{\partial x}||_\infty = \max_x \frac{\partial V}{\partial x}$ at a given time t, then the time history of $||\frac{\partial V}{\partial x}||_\infty$ is shown in Fig. 23b. In the next section, we will analyze how the same initial condition evolves in a mixed sensing environment.

A second initial condition of interest is one that is initially continuous, but then propagates with time, in a manner such as to eventually form a shock. In other words, the initial (decreasing) average velocity profile steepens with time. It is of interest to see how a partial equipage of the slowdown warning system can help arrest the wave steepening scenario that can exist (when all vehicles are unequipped), and to then parametrize this effect as a function of varying equipage.

For this purpose, we invoke an initial condition with identical left and right states as before, i.e., $\rho_L = 15$ vehicles/km/lane, $\rho_R = 140$ vehicles/km/lane and $V_L = 105.67$ kmph, $V_R = 3.17$ kmph; but instead of joining them by a discontinuity, we now join ρ_L to ρ_R by a gradual transition, so that the average density increases from ρ_L to ρ_R over a span of 2 km. The average velocity varies from V_L to V_R in a manner so that the average velocity is in equilibrium with the average density at each x.

Figure 19 then shows the average density and average velocity profiles as they evolve with time, from the above initial condition. It is seen that the top

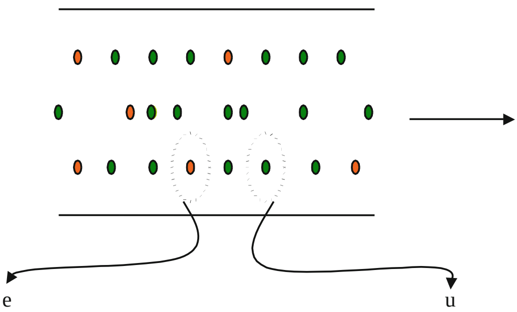


Fig. 20 Schematic multi-vehicle scenario comprising of unequipped and equipped vehicles

portion of the velocity wave (and the bottom portion of the density wave) move forward relative to the highway, i.e., they have positive wave velocity; while the bottom portion of the velocity wave (as also the high density part of the density wave) move backward, with a negative wave velocity. This kind of wave motion (wherein different parts of the wave have wave velocities of opposite signs), leads to further and further steepening of the wave, until eventually a shock is formed, that then moves backward as a whole. The evolution of $||\frac{\partial V}{\partial x}||_\infty$ showing the gradual steepening of the wave is given in Fig. 26b, while Fig. 26a gives the magnitude of ΔV, which represents the velocity change that occurs over the region where the value of $\frac{\partial V}{\partial x}$ is less than -100 kmph/km. It is seen that over a span of approximately 5 min, ΔV increases to almost 100 kmph, which makes it almost identical to the initial condition of the first case we explored.

3.2 Mixed Sensing Network: Model and Simulations

We now intend to test the above two initial conditions in a mixed sensing environment, depicted in Fig. 20. To this end, we assume that at $t = 0$, the average velocity of the equipped vehicles is identical to that of the unequipped vehicles. $\rho_U(x,t)$ and $V_U(x,t)$ are used to represent the average density and average velocity of the unequipped vehicles, while $\rho_E(x,t)$ and $V_E(x,t)$ represent the average density and average velocity of the equipped vehicles. To test the effect of varying equipage, we vary ρ_U and ρ_E, so that $\frac{\rho_E(x)}{\rho_U(x)+\rho_E(x)}$ represents the percentage of equipage at each x, and we keep $\rho_U(x,0) + \rho_E(x,0) =$ a constant which is equal to the density of vehicles when they were all unequipped. In other words, $\rho_{UL}(x,0) + \rho_{EL}(x,0) = \rho_L(x,0)$; $\rho_{UR}(x,0) + \rho_{ER}(x,0) = \rho_R(x,0)$; $V_{UL}(x,0) = V_{EL}(x,0)$; $V_{UR}(x,0) = V_{ER}(x,0)$, where the values for ρ_L, ρ_R, V_L, and V_R correspond to the values when all vehicles were unequipped (as discussed in the previous section).

The following macroscopic equations for the mixed equipage scenario are used [19, 20]:

$$\frac{\partial \rho_U}{\partial t} + \frac{\partial(\rho_U V_U)}{\partial x} = 0, \tag{24}$$

$$\frac{\partial \rho_E}{\partial t} + \frac{\partial(\rho_E V_E)}{\partial x} = 0, \tag{25}$$

$$\frac{\partial(\rho_U V_U)}{\partial t} + \frac{\partial(\rho_U V_U^2 + \rho_U \theta_U^2)}{\partial x} = \frac{V_U^{\text{eq}} - V_U}{\tau}, \tag{26}$$

$$\frac{\partial(\rho_E V_E)}{\partial t} + \frac{\partial(\rho_E V_E^2 + \rho_E \theta_E)}{\partial x} = \frac{V_E^{\text{eq}} - V_E}{\tau}, \tag{27}$$

where $V_E^{\text{eq}}(x,t)$ represents the average equilibrium velocity of the equipped vehicles and is given by:

$$V_E^{\text{eq}}(x,t) = V_E^o - PB_{EU}\rho_U \tau \theta_U - PB_{EE}\rho_E \tau \theta_E, \tag{28}$$

and where $V_U^{eq}(x,t)$ represents the average equilibrium velocity of the unequipped vehicles and is given by:

$$V_U^{\text{eq}}(x,t) = V_U^o - PB_{UE}\rho_E \tau \theta_E - PB_{UU}\rho_U \tau \theta_U. \tag{29}$$

V_U^o and V_E^o denote the average desired velocities of the unequipped and equipped vehicles, respectively. $P = \frac{(\rho_U+\rho_E)T^2 V_{\text{avg}}^o}{\tau A_{\rho_{\max}}(1-((\rho_U+\rho_E)/\rho_{\max})^2)}$ and $V_{\text{avg}}^o = (\rho_U V_U + \rho_E V_E)/(\rho_U + \rho_E)$. θ_U and θ_E represent the velocity variances of the unequipped and equipped vehicles respectively, and it is assumed that $\theta_U = A(\rho_U + \rho_E)V_U^2$ and $\theta_E = A(\rho_U + \rho_E)V_E^2$.

The above equations are similar to the equations used in [45] when the two species of vehicles assumed were cars and trucks, and in that context, it was assumed that the desired vehicles of the cars and trucks remained constant for all time. In our context however, we assume that the equipped vehicles change their desired velocities instantaneously on receipt of the communication wave – we therefore define an additional variable $\gamma(x,t)$ and add in the following additional equations:

$$V_E^o = \gamma(x,t)V_{E\text{final}}^o + (1-\gamma(x,t))V_{E\text{initial}}^o, \tag{30}$$

$$\frac{\partial \gamma}{\partial t} + a\frac{\partial \gamma}{\partial x} = 0, \tag{31}$$

where $\gamma(x,t)$ is a Heaviside step function defined such that $\gamma(x,t) = 0$ for that x (Part of the highway that has not received the communication wave by time t), and $\gamma(x,t) = 1$ for all other x (30).) thus implies that the moment an equipped vehicle at x receives the slowdown warning signal at a time t, its desired velocity changes instantaneously from its initial value $V_{E\text{initial}}^o$ (which is assumed to be the same as V_U^o – the desired velocity of the unequipped vehicles) to a final value of $V_{E\text{final}}^o$ (which is assumed to be approximately equal to the average velocity occurring at the degraded point far ahead, where a hazard has occurred). Equation 31 is a PDE that postulates the evolution of $\gamma(x,t)$ and in which $a < 0$ represents the communication speed. The boundary condition $\gamma(10,t) = 1$ is imposed.

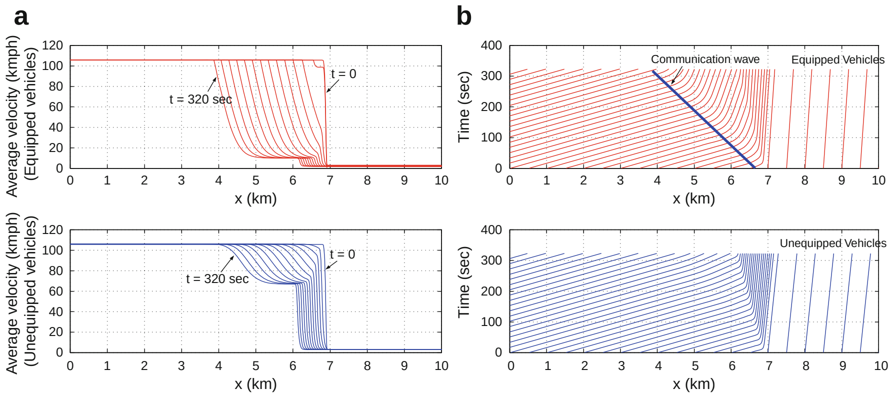


Fig. 21 (**a**) Average velocity profiles; (**b**) average vehicle trajectories of equipped and unequipped vehicles (5% equipage) for the Reimann Problem

We note that alternative formulations are also possible. For instance, if we assume that information of the location of the hazard is also broadcasted to the equipped vehicles (along with the warning signal), then it is reasonable to assume that the driver of the equipped vehicle will adapt his desired velocity (as a function of distance to the hazard) so that he attains his final desired velocity by the time he reaches the location of the hazard. In this case, we could rewrite (30) as:

$$V_E^o = \gamma(x,t)\left[(1-\alpha(x,t))V_{E\text{final}}^o + \alpha(x,t)V_{E\text{initial}}^o\right] + (1-\gamma(x,t))V_{E\text{initial}}^o, \quad (32)$$

where $\alpha(x,t)$ is a function that evolves according to the PDE

$$\frac{\partial \alpha}{\partial t} + V_E\frac{\partial \alpha}{\partial x} = -\frac{V_E}{d_0}, \quad (33)$$

with d_0 representing the average distance of an equipped car to the location of the hazard, when it first received the warning signal, and the initial condition on α is specified such that $\alpha(x,0) = 1$ for all x to the left of the hazard, and $\alpha(x,0) = 0$ for all x to the right of the hazard. The boundary condition on α would be $\alpha(0,t) = 1$. For the purposes of this chapter, we assume that the change in the desired velocity of the equipped vehicle occurs instantaneously, i.e., (30) and (31) are employed.

For the first initial condition, we assume an average communication speed of 25 kmph. Figure 21a shows the average velocity profiles of the equipped and unequipped vehicles respectively (for a 5% equipage scenario). It can be seen that as the communication wave propagates through the equipped vehicles, causing them to slow down, the unequipped vehicles are also forced to slow down earlier than they otherwise would have (they thus receive indirect information of the hazard ahead). The wave velocity of the top portion of the average velocity of the unequipped vehicles has now become negative (it was formerly positive when they had no

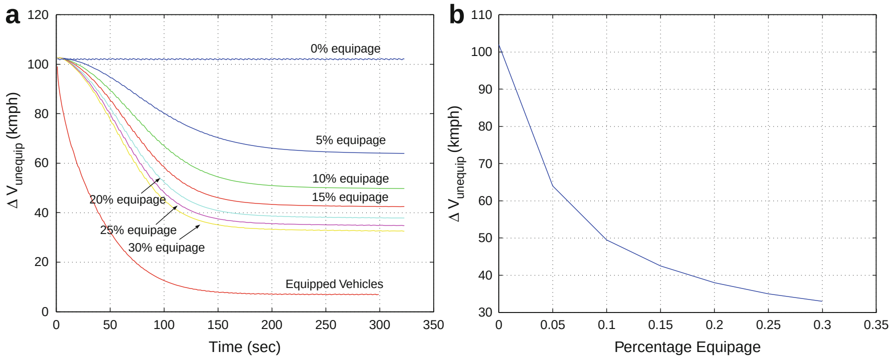


Fig. 22 (**a**) Magnitude of Δv as a function of time; (**b**) magnitude of $\Delta V_{unequip}$ at steady state for different percentages of equipage for the Reimann Problem

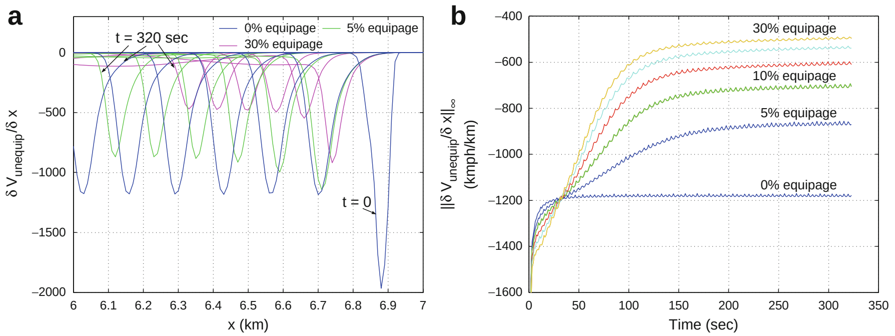


Fig. 23 (**a**) $\frac{\partial V_{unequip}}{\partial x}$ as a function of time; (**b**) $||\frac{\partial V_{unequip}}{\partial x}||_\infty$ for varying equipages for the Reimann Problem

equipped vehicles among their midst); and this in turn has led to a lower magnitude of the average velocity shock experienced by the unequipped vehicles. Figure 21b shows the average vehicle trajectories of the equipped and unequipped vehicles, on a $x - t$ plane. The propagation of the communication wave is also seen.

Figure 22a demonstrates the magnitude of the velocity shock as a function of time, for the different equipages. It is seen from Fig. 22b that the largest reduction in ΔV that can occur with a 5% increase in equipage, occurs in the 0–5% range. With 10% equipage, the velocity shock magnitude in the unequipped vehicles is reduced almost by a factor of one-half, for equipages above 15%, the magnitude of benefit obtained (as measured from the reduction in shock strength of the unequipped vehicles per unit increase in the density of the equipped vehicles), is not significantly increased. This behavior is also manifested in Fig. 23b, which demonstrates $||\frac{\partial V_{unequip}}{\partial x}||_\infty$, as a function of time. Figure 24 demonstrates a manifestation of the same behavior, when viewed in the spatial frequency domain.

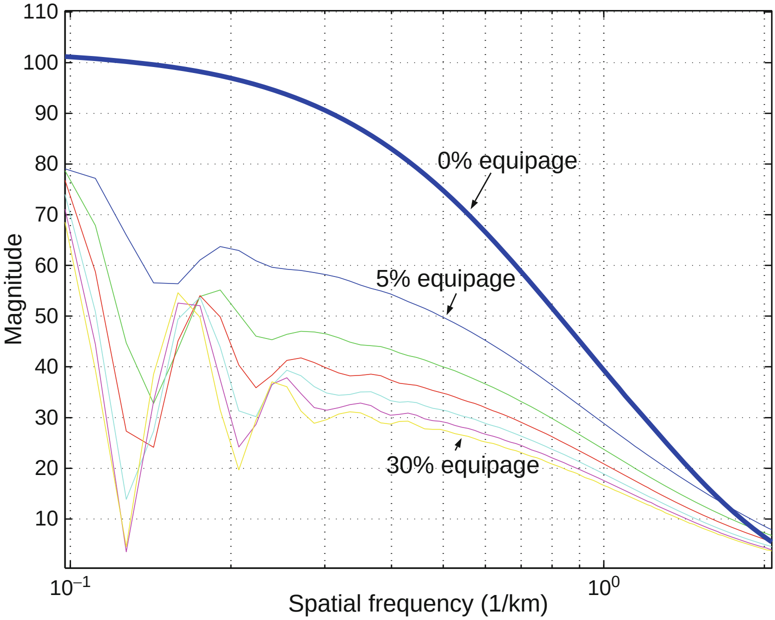


Fig. 24 Spatial frequency content of $\frac{\partial V_{unequip}}{\partial x}$ at $t = 320$ s for different percentages of equipage for the Reimann Problem

After our discussion on the Reimann Problem, we now direct our attention toward the second initial condition studied earlier, i.e., a situation wherein an initially continuous condition, evolved with time, to get progressively steeper and eventually appear like a discontinuity. We then test a scenario wherein we assume that information of the existence of a velocity gradient is made available to the equipped vehicles residing to the left of the point $x = 6$ km, at $t = 0$. Again, the communication wave is assumed to travel at a constant speed of 25 kmph, in the backward direction; this time originating from $x = 6$ km, at $t = 0$. The reason that this case is interesting is because it enables us to see if and how varying the percentage of equipped vehicles can arrest the formation of the discontinuity, before it has developed.

Figure 25 shows the average velocity profiles of the equipped and unequipped vehicles (assuming a 30% level of equipment). It is seen that the top portion of the average velocity (which had positive wave velocity when all vehicles were unequipped, i.e., it was moving forward relative to the highway), now immediately begins to move backward as the communication wave passes through the equipped vehicles. This arrests the wave steepening effect that was present in the case of no equipage; and consequently the equipped vehicles do not experience any abrupt velocity gradient, while the unequipped vehicles experience a significantly reduced magnitude of negative velocity gradient, than they otherwise would have.

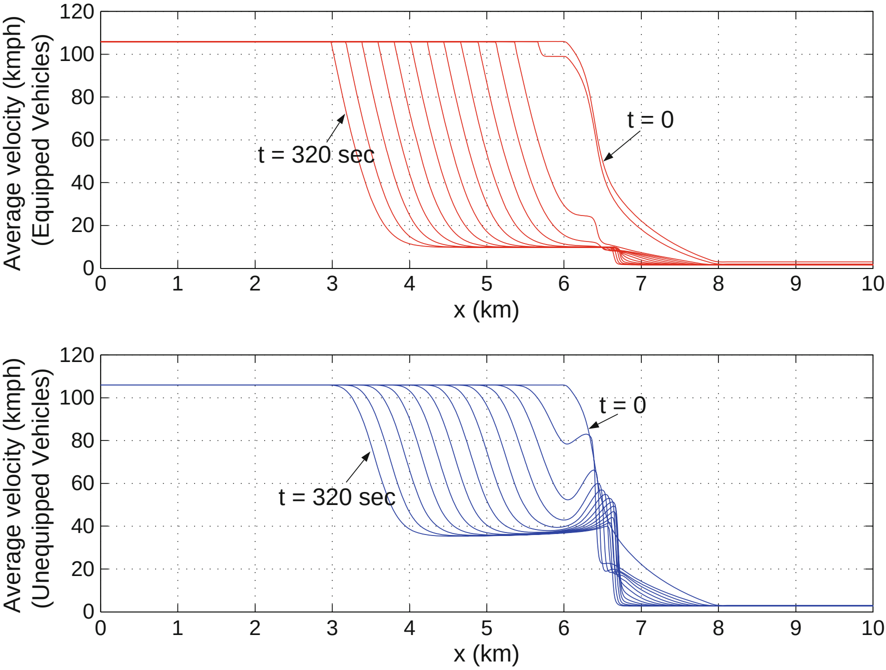


Fig. 25 Average velocity profiles of vehicles in the mixed sensing network (30% equipage) for the continuous initial condition

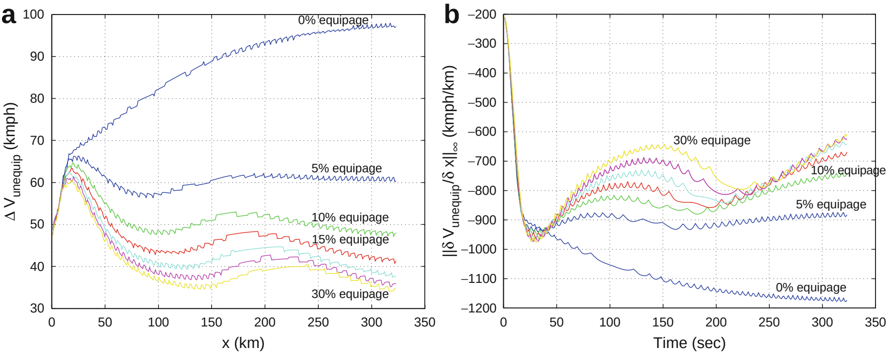


Fig. 26 (**a**) Magnitude of $\Delta V_{unequip}$; (**b**) $||\frac{\partial V_{unequip}}{\partial x}||_\infty$ for varying equipages for the continuous initial condition

Figure 26a shows the magnitude of ΔV for the unequipped vehicles, with ΔV representing the average velocity change of the unequipped vehicles over the region where $\frac{\partial V_{unequip}}{\partial x}$ is smaller than -100 kmph/km. It is seen from Fig. 26a that again a 5% equipage causes greatest reduction in ΔV and that above an equipage of 15%, the benefit obtained per unit increase in percentage equipage, is not significantly greater. The same effect is manifested in Fig. 26b that shows $||\frac{\partial V_{unequip}}{\partial x}||_\infty$.

4 Conclusions

In this chapter, we look at the problem of mobile mixed sensing networks in an automotive context. Specifically, we look at scenarios wherein automobiles with two different levels of sensing capabilities (one type capable of sensing only local, near-neighbor information, and the other type capable of sensing advance, far-ahead information through wireless communication) are scattered (or mixed) among each other. Under these circumstances, we look for conditions on the number and distribution of the equipped vehicles (which are the ones capable of sensing far ahead information) that will satisfy a given performance objective of the overall sensor network. The performance objective considered in this chapter is related to the safety of the overall network; and the safety metrics considered are those of zero collisions (in the microscopic modeling case) and weakened shock waves (in the macroscopic modeling case).

Acknowledgments This work is supported by the National Science Foundation Award CCR-0208831, the Deshpande Center for Technological Innovation Award 009216-015, and the Ford-MIT Alliance Program.

References

1. CNN News, Five seriously hurt in 194-vehicle California pileup, Nov. 3, 2002.
2. CNN News, Massive pile-up near Ga.-Tenn. line kills 4, March 14, 2002.
3. Boston Globe News, Sixty six vehicle California Pileup, April 2, 2004.
4. Boston Globe News, Rains trigger 30 car pileup on I-93, August 4, 2003.
5. J. K. Hedrick, R. Sengupta, Q. Xu, Y. Kang, C. Lee, Enhanced AHS Safety Through the Integration of Vehicle Control and Communication, *California PATH Research Report UCB-ITS-PRR-2003-27*, September 2003.
6. X. Y. Lu, J. K. Hedrick, M. Drew, ACC/CACC - Control Design, Stability and Robust Performance, *Proc. Of the American Control Conference*, May 8–10, 2002, pp. 4327–4332.
7. P. Seiler, A. Pant, K. Hedrick, Disturbance Propagation in Vehicle Strings, *IEEE Transactions on Automatic Control*, Vol. 49, No. 10, October 2004, pp. 1835–1842.
8. A. R. Girard, J. B. de Sousa, J. K. Hedrick, An Overview of Emerging Results in Networked Multi-Vehicle Systems, *Proc. of the 40th IEEE Conference on Decision and Control*, December 2001, pp. 1485–1490.
9. Q. Xu, K. Hedrick, R. Sengupta, J. VanderWerf, Effects of Vehicle-vehicle/roadside-vehicle Communication on Adaptive Cruise Controlled Highway Systems, *Proc. of the 56th IEEE Vehicular Technology Conference*, Vol. 2, September 2002, pp. 1249–1253.
10. S. Kato, S. Tsugawa, K. Tokuda, T. Matsui, H. Fujii, Vehicle control algorithms for cooperative driving with automated vehicles and intervehicle communications, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 3, No. 3, September 2002, pp. 155–161.
11. J. Carbaugh, D. Godbole, R. Sengupta, Tools for safety-Throughput analysis of vehicle automation systems, *Proc. Of the American Control Conference*, Vol. 3, 1997, pp. 2031–2035.
12. J. Carbaugh, D. Godbole, R. Sengupta, Safety and capacity analysis of automated and manual highway systems, *Transportation Research C (6)*, 1998, pp. 69–99.

13. A. Bose, P. A. Ioannou, Analysis of Traffic Flow with Mixed Manual and Semiautomated Vehicles, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 4, No. 4, December 2003, pp. 173–188.
14. A. Bose, P. A Ioannou, Mixed manual/semi-automated traffic: a macroscopic analysis, *Transportation Research Part C* 11 (2003), pp. 439–462.
15. A. Chakravarthy, K.Y. Song, and E. Feron, A GPS-based slowdown warning system for automotive safety, *Proc. of IEEE Intelligent Vehicles Symposium*, June 14–17, 2004, Parma, Italy, pp. 489–494.
16. A. Chakravarthy, K.Y. Song, and E. Feron, A slowdown warning system for automobiles,"*Proc. of IEEE SMC, October 10–13*, 2004, The Hague, Netherlands, pp. 3962–3969.
17. A. Chakravarthy, K.Y. Song, E. Feron, Influence of a slowdown warning system on a multi-vehicle stream, *Proc. of American Control Conference*, June 8–10, 2005, Portland, pp. 2134–2140.
18. A. Chakravarthy, K.Y. Song, E. Feron, Preventing automotive pileup crashes in mixed communication environments, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 10, No. 2, June 2009, pp. 211–225
19. A. Chakravarthy, J. Peraire, E. Feron, Effects of a slowdown warning system in mixed communication environments: A Macroscopic Study, *Proc. of European Control Conference*, July 2–5, 2007, Greece, pp. 1901–1908, ISBN: 978-960-89028-5-5.
20. A. Chakravarthy, *Safety of a multi-vehicle system in mixed communication environments*, Ph.D Thesis, Massachusetts Institute of Technology, 2007.
21. R. E. Chandler, R. Herman, and E. W. Montroll, Traffic dynamics: Studies in car following, *Operations Research*, Vol. 6, pp. 165–184, 1958.
22. J. S. Tyler, The Characteristics of Model Following Systems as Synthesized by Optimal Control, *IEEE Transactions on Automatic Control*, Vol. AC-9, pp. 485–498, 1964.
23. G. O. Burnham, J. Seo and G. A. Bekey, Identification of Human Driver Models in Car Following, *IEEE Transactions on Automatic Control*, Vol. 19, pp. 911–915, 1974.
24. D. Swaroop and J. K. Hedrick, String Stability of Interconnected Systems, *IEEE Transactions on Automatic Control*, pp. 349–357, Vol. 41, No. 3, March 1996.
25. D.T Mcruer and E.S Krendel, Mathematical Models of Human Behavior, *AGARD, AGARD-AG-188*, 1974.
26. R. Shah, C. Nowakowski and P. Green, *U.S Highway Attributes Relevant to Lane Tracking*, University of Michigan, UMTRI Technical Report 98-34, 1998.
27. GPS Errors and Estimating Your Receiver's Accuracy, http://www.edu-observatory.org/gps/gps_accuracy.html
28. T.M Nguyen, J.W Sinko and R.C Galijan, Using Differential Carrier Phase GPS to Control Automated Vehicles, *Proceedings of the 40th Midwest Symposium on Circuits and Systems*, Sacramento, CA, August 1997, pp. 493–496
29. D.M Bevly, J.C Gerdes, C. Wislon and G. Zhang, The Use of GPS Based Velocity Measurements for Improved Vehicle State Estimation, *Proceedings of the American Control Conference*, Chicago, Illinois, June 2000, pp. 2538–2542.
30. Nationwide Differential Global Positioning System Program Fact Sheet, http://www.tfhrc.gov/its/ndgps/02072.html
31. D. Swaroop, String Stability of Interconnected Systems - An application to Automated Highway Systems, *Ph.D Dissertation, University of California at Berkeley*, December 1994.
32. M.T Moreno, D. Jiang and H. Hartenstein, Broadcast Reception Rates and Effects of Priority Access in 802.11-Based Vehicular Ad-Hoc Networks, *Proceedings of VANET 04*, October 2004, pp. 10–18.
33. http://www.toyota-eshowroom.com/consumer/ToyotaNavPrius/toyota_nav.html
34. J.C McCall and M.M. Trivedi, Video based lane estimation and tracking for driver assistance: survey, system and evaluation, *IEEE Transactions on Intelligent Transportation Systems*, March 2006, pp. 20–37.
35. P. I. Richards, Shock Waves on the Highway, *Operations Research*, Vol. 4, No. 42, 1956, pp. 209–229.

36. M. H. Lighthill, G. B. Whitham, On kinematic waves II : A theory of traffic flow on long, crowded roads, *Proc. Of the Royal Society of London Ser. A* 229, pp. 317–345, 1945.
37. H. J. Payne, Models of Freeway Traffic and Control, *Simulation Council, 1971.*
38. G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, 1974.
39. C. Daganzo, Requiem for second-order fluid approximation to traffic flow, *Transportation Research Part B* 29 (4), 1995, pp. 277–286.
40. W.F Phillips, A kinetic model for traffic flow with continuum implication, *Transportation Planning and Technology* 5, 1979, pp. 131–138.
41. A. Aw and M. Rascle, Resurrection of second order models of traffic flow? *SIAM Journal of Applied Math.*, 60(3), 2000, pp. 916–938.
42. D. Helbing, Improved fluid dynamic model for vehicular traffic, *Physics Review E 51*, pp. 3164–3169, 1995.
43. D. Helbing, A. Hennecke, V. Shvetsov and M. Treiber, Micro- and Macro-Simulation of Freeway Traffic, *Mathematical and Computer Modelling* 35, 2002, pp. 517–547.
44. D. Helbing, Gas-kinetic derivation of Navier-Stokes-like traffic equations, *Physical Review E* 53, March 1996, pp. 2366–2381.
45. M. Treiber, A. Hennecke and D. Helbing, Derivation, properties and simulation of a gas-kinetic-based, nonlocal traffic model, *Physical Review E* 59, January 1999, pp. 239–253.
46. S. P. Hoogendoorn and P. H. L. Bovy, Continuum modeling of multiclass traffic flow, *Transportation Research, Part B* 34, 2000, pp. 123–146.
47. I. Prigogine and R. Herman, *Kinetic theory of vehicular traffic*, Elsevier, 1971.
48. S. L. Paveri-Fontana, On Boltzmann like treatments for traffic flow, *Transportation Research Part B*, Vol. 9, pp. 225–235, 1975.
49. Swaroop D. and K. R. Rajagopal, Intelligent Cruise Control Systems and Traffic Flow Stability, *Transportation Research Part C: Emerging Technologies*, 7(6), 1999, pp. 329–352.
50. J. Yi, H. Lin, L. Alvarez and R. Horowitz, Stability of Macroscopic Traffic Flow Modeling through Wavefront Expansion, *Transportation Research Part B: Methodological*, 37(7), 2003, pp. 661–679.

Navigation in Difficult Environments: Multi-Sensor Fusion Techniques

Andrey Soloviev and Mikel M. Miller

Abstract This chapter focuses on multi-sensor fusion for navigation in difficult environments where none of the existing navigation technologies can satisfy requirements for accurate and reliable navigation if used in a stand-alone mode. A generic multi-sensor fusion approach is presented. This approach builds the navigation mechanization around a self-contained inertial navigator, which is used as a core sensor. Other sensors generally derive navigation-related measurements from external signals, such as Global Navigation Satellite System (GNSS) signals and signals of opportunity (SoOP), or external observations, for example, features extracted from images of laser scanners and video cameras. Depending on a specific navigation mission, these measurements may or may not be available. Therefore, externally-dependent sources of navigation information (including GNSS, SoOP, laser scanners, video cameras, pseudolites, Doppler radars, etc.) are treated as secondary sensors. When available, measurements of a secondary sensor or sensors are utilized to reduce drift in inertial navigation outputs. Inertial data are applied to improve the robustness of secondary sensors' signal processing. Applications of the multi-sensor fusion approach are illustrated in detail for two case studies: (1) integration of Global Positioning System (GPS), laser scanner, and inertial navigation; and, (2) fusion of laser scanner, video camera, and inertial measurements. Experimental and simulation results are presented to illustrate performance of multi-sensor fusion algorithms.

A. Soloviev (✉)
Research and Engineering Education Facility, University of Florida, Shalimar, FL 32579, USA
e-mail: soloviev@ufl.edu

M.M. Miller
Air Force Research Laboratory – Munitions Directorate, Eglin AFB, FL 32542, USA
e-mail: mikel.miller@eglin.af.mil

V.L. Boginski et al. (eds.), *Sensors: Theory, Algorithms, and Applications*,
Springer Optimization and Its Applications 61, DOI 10.1007/978-0-387-88619-0_9,

1 Motivation

Many existing and perspective applications of navigation systems would benefit notably from the ability to navigate accurately and reliably in difficult environments. Examples of difficult navigation scenarios include urban canyons, indoor applications, radio-frequency (RF) interference and jamming environments. In addition, different segments of a mission path can impose significantly different requirements on the navigation sensing technology and data processing algorithms. To exemplify, Fig. 1 shows a mission scenario of an autonomous aerial vehicle (UAV).

For this example, the UAV is deployed in an open field; next, the vehicle enters an urban canyon to perform tasks such as surveillance and reconnaissance; and, finally, it returns to the deployment point. To enable operation of the UAV at any point on the flight path, a precision navigation, attitude, and time capability on-board the vehicle is required. Global Navigation Satellite System (GNSS) generally provides satisfactory performance in open fields and suburban areas, but has fragmented availability in urban environments due to satellite blockages by buildings and other obstacles. Feature-based navigation techniques demonstrate a promising potential in dense urban areas where enough navigation-related features can be extracted from images of digital cameras or laser scanners. However, the feature availability can be limited in relatively open areas. A self-contained inertial navigation system (INS) can provide navigation solution at any environment; however, the solution accuracy drifts over time. In a stand-alone mode, none of the existing navigation technologies has a potential to satisfy the requirements for the navigation accuracy, continuity, and availability over the entire duration of the UAV flight. Therefore, multi-sensor fusion techniques are pursued. In other words, to be able to navigate at any environment at any time, it is beneficial to utilize any potential source of navigation-related information.

Example applications that involve navigation in difficult environments include but are not limited to navigation, guidance, and control of autonomous ground

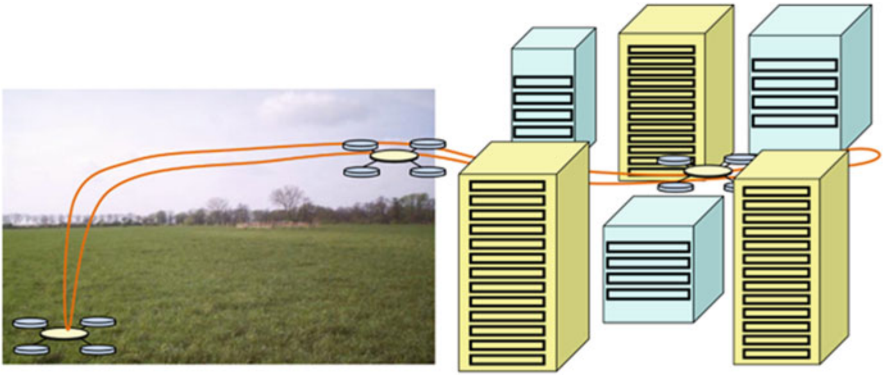


Fig. 1 UAV mission example

vehicle (UGV) and UAV, as well as teams of UGVs and UAVs for urban surveillance and reconnaissance tasks; geographical information system (GIS) data collection for mapping applications on open highways and dense urban environments; indoor search and rescue applications; monitoring of urban infrastructure for situational awareness; and, precise automotive applications such as automated lane keeping. Meter-level to decimeter-level reliable positioning capabilities are generally needed for these application examples. As stated previously, none of the existing navigation technologies can currently satisfy these requirements or has a potential to provide these capabilities in a stand-alone mode.

This chapter discusses multi-sensor fusion approaches for navigation in difficult environments. A generic concept of the multi-sensor fusion is first presented. Next, the chapter exemplifies applications of the generic multi-sensor fusion concept for the development of specific multi-sensor mechanizations. Specifically, integrated Global Positioning System (GPS)/laser scanner/INS and laser scanner/video camera/INS mechanizations are considered.

2 Multi-Sensor Fusion Approach

The generic concept of the multi-sensor navigation utilizes a self-contained inertial navigator as a core navigation sensor. The INS does not rely on any type of external information and as a result can operate in any environment. However, inertial navigation solution drifts over time [1]. To mitigate INS drift, this core sensor is augmented by reference navigation data sources (such as, for example, GPS or a laser scanner). Reference data sources generally rely on external observations or signals that may or may not be available. Therefore, these sources are treated as secondary sensors. When available, secondary sensors' measurements are applied to reduce the drift in inertial navigation outputs. Inertial data are used to bridge over reference sensor outages. In addition, inertial data can be applied to improve the robustness of reference sensor signal processing: for instance, to significantly increase the GPS signal integration interval in order to enable processing of very weak GPS signals and to reduce the susceptibility of GPS to RF interference [2]. Figure 2 illustrates the multi-sensor fusion approach.

Figure 3 provides a more detailed illustration of the interaction between the INS and a secondary navigation sensor.

Secondary sensor generally includes a signal processing part and a navigation solution part. The signal processing part receives navigation-related signals and measures their parameters. For example, GPS receiver tracking loops [3] or open-loop batch estimators [17] measure parameters (pseudoranges, Doppler frequency shift, and carrier phase) of the received GPS signals. Another example is a laser scanner time-of-flight measurement that is directly related to the distance between the scanner and a reflecting object. Measurements of signal parameters are then applied to compute the navigation solution. For example, GPS pseudoranges are used to compute the GPS receiver position [3], changes in distances to reflecting

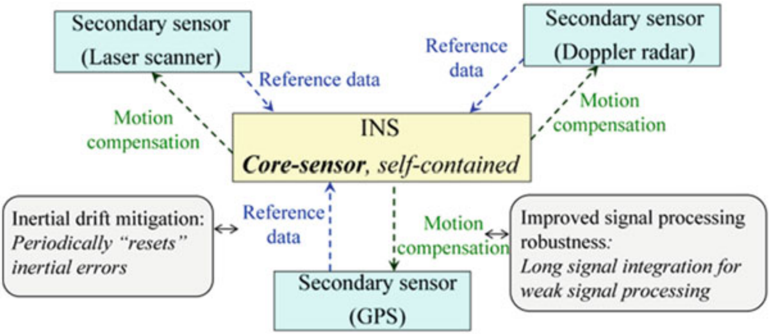


Fig. 2 Generic multi-sensor fusion approach

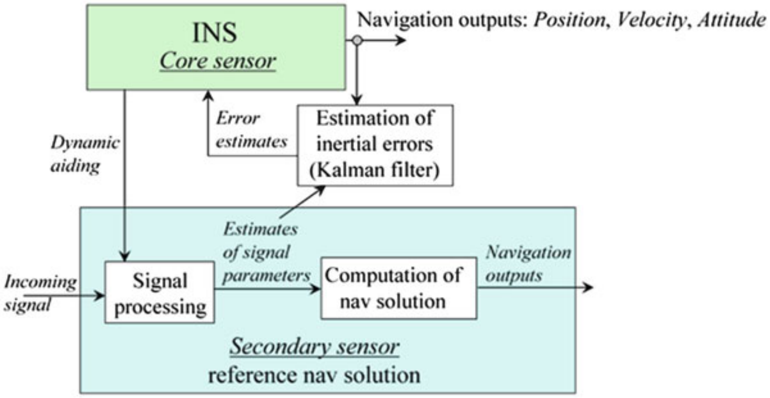

INS
Core sensor
Navigation outputs: Position, Velocity, Attitude
Dynamic aiding
Error estimates
Estimation of inertial errors (Kalman filter)
Estimates of signal parameters
Incoming signal
Signal processing
Computation of nav solution
Navigation outputs
Secondary sensor
reference nav solution

Fig. 3 Data fusion between the INS and a secondary navigation sensor

stationary objects are exploited to compute the change in the position of the laser scanner [4]. Note that the navigation solution can only be computed if a sufficient number of signal measurements is made. For example, at least four pseudorange measurements must be available to compute the GPS-based position; at least two lines must be extracted from an image of a two-dimensional (2D) laser scanner to compute a 2D laser position [4].

The multi-sensor fusion architecture above applies secondary sensor's signal measurements to estimate drifts in inertial navigation outputs. This approach is generally referred to as tight coupling. The use of signal measurements for the INS drift re-calibration is more beneficial as compared to the use of the secondary sensor's navigation solution (this form of sensor fusion is referred to as loose coupling). As opposed to loose coupling, tight coupling still allows for the INS drift mitigation even for those cases where insufficient number of signal measurements is available and the complete navigation solution cannot be derived.

The estimation of the INS drift terms is performed using the mechanism of a complementary Kalman filter. The idea is that a signal parameter can be generally represented as a function of position, velocity, and attitude. This function is computed based on INS navigation outputs and then compared to the actual signal measurement. A discrepancy between the INS-based signal prediction and the signal measurement is used by the complementary Kalman filter mechanization to estimate INS drift terms. A generic Kalman filter measurement observable $\rlap{-}y$ is formulated as follows:

$$\rlap{-}y = \rho(\mathbf{R}_{\mathrm{INS}}, \mathbf{V}_{\mathrm{INS}}, \alpha_{\mathrm{INS}}) - \tilde{\rho} \tag{1}$$

where:

- $\mathbf{R}_{\mathrm{INS}}$ is the INS position vector
- $\mathbf{V}_{\mathrm{INS}}$ is the INS velocity vector
- α_{INS} is the INS attitude (that can be represented as a direction cosine matrix, rotation vector, rotation quaternion, or Euler angles [1])
- $\rho(\mathbf{R}_{\mathrm{INS}}, \mathbf{V}_{\mathrm{INS}}, \alpha_{\mathrm{INS}})$ is the signal value predicted based on the inertial output;
- $\tilde{\rho}$ is the actual signal measurement

The signal measurement herein is denoted by ρ since the tight coupling approach was first applied to the GPS/INS integration case in order to fuse GPS pseudorange measurements and integrated Doppler range measurements with inertial data [5]. For this reason, the INS calibration in the signal measurements' domain is also often referred to as the range-domain data fusion. Obviously, the formulation above is not limited to the GPS/INS integration case and is applicable to any type of navigation-related signal measurements. It is important to mention that video measurements serve as an exclusion for this general rule of the observation formulation. For a vision-based case where the distance to a navigation-related feature is unknown and cannot be measured (for example, a monocular camera case without any prior feature information), image features are related to navigation parameters not in the form of a non-linear function, but rather in the form of a non-linear motion constraint. This constraint is formulated as $f(\mathbf{R}, \alpha, \mathbf{m}, \rho) = 0$, where $\mathbf{m}$ represents feature homogeneous coordinates (i.e., Cartesian coordinates scaled by the image depth) and ρ is the distance to the feature. In this case, non-linear motion constrains and inertial data are used to formulate Kalman filter observables: i.e., $f(\mathbf{R}_{\mathrm{INS}}, \alpha_{\mathrm{INS}}, \tilde{\mathbf{m}}, \rho) = 0$, where $\tilde{\mathbf{m}}$ denotes homogeneous feature coordinates that are extracted from video images.

To implement a complementary Kalman filter, (1) is linearized as using a Taylor series expansion:

$$\rlap{-}y = \rho(\mathbf{R}_{\mathrm{INS}}, \mathbf{V}_{\mathrm{INS}}, \alpha_{\mathrm{INS}}) - \tilde{\rho} \tag{2}$$

$$\approx \rho + \frac{\partial \rho}{\partial \mathbf{R}_{\mathrm{INS}}}\delta \mathbf{R}_{\mathrm{INS}} + \frac{\partial \rho}{\partial \mathbf{V}_{\mathrm{INS}}}\delta \mathbf{R}_{\mathrm{INS}} + \frac{\partial \rho}{\partial \alpha_{\mathrm{INS}}}\delta \alpha_{\mathrm{INS}} - (\rho - n_{\rho}) \tag{3}$$

$$= \frac{\partial \rho}{\partial \mathbf{R}_{\mathrm{INS}}}\delta \mathbf{R}_{\mathrm{INS}} + \frac{\partial \rho}{\partial \mathbf{V}_{\mathrm{INS}}}\delta \mathbf{R}_{\mathrm{INS}} + \frac{\partial \rho}{\partial \alpha_{\mathrm{INS}}}\delta \alpha_{\mathrm{INS}} - n_{\rho} \tag{4}$$

where:

- ρ is the true value of the signal parameter
- n_ρ is the measurement noise
- $\delta\mathbf{R}_{\mathrm{INS}}$, $\delta\mathbf{V}_{\mathrm{INS}}$, and $\delta\alpha_{\mathrm{INS}}$ are the INS position, velocity, and attitude errors, respectively.

Non-linear motion constraints are linearized for the case of video measurements. These linearized formulations are then utilized by standard Kalman filter routines (i.e., prediction, estimation, and covariance updates) to estimate INS error states.

For those cases where multiple measurements are available from secondary navigation sensors, the measurement observation vector is expressed as follows:

$$\left\{\text{ɏ}_l^{(i)}\right\} = \rho_l^{(i)}(\mathbf{R}_{\mathrm{INS}}, \mathbf{V}_{\mathrm{INS}}, \alpha_{\mathrm{INS}}) - \tilde{\rho}_l^{(i)} \tag{5}$$

$$l = 1, \ldots, L \tag{6}$$

$$i = 1, \ldots, I \tag{7}$$

In (5), l is the reference sensor index and i is the measurement index for a particular reference sensor. For example, reference sensors can include GPS, laser scanner, and Doppler radar with corresponding indexes 1, 2, and 3. In this case, ɏ_2^5 represents a 5th measurement observable for laser scanner measurements. Specific formulations of Kalman filter measurement observations are exemplified in Sect. 3 for cases of GPS, laser scanner, and vision reference sensors.

As mentioned previously, while the secondary sensors' measurements are used to improve the inertial accuracy, the INS data can be applied to improve the robustness of the secondary sensor's signal processing. For example, inertial data can be used for robust matching of features between different images of a video camera or a laser scanner. Another example is the use of inertial aiding for GPS signal integration in order to enable tracking of signals that are attenuated by buildings [6].

Secondary navigation sensors that can be applied for the multi-sensor fusion with the INS include:

- **GNSS and partial GNSS:** GPS and GNSS signal measurements can be efficiently integrated with inertial data in open areas and also in difficult GPS signal environments (such as urban areas) where limited GNSS signals may still be available [6, 7].
- **Feature-based navigation sensors:** Examples include image-aided inertial navigation [8] and ladar-aided inertial navigation [4].
- **Beacon-based navigation (including pseudolites)**: If the GPS signal is not adequate for navigation in a particular environment, it is possible to transmit an additional signal or signals that are specifically designed for navigation purposes. If the transmitted signals are similar to GPS signals, then such beacon transmitters are usually called "pseudolites." Examples of beacon-based navigation systems for indoor navigation can be found in [9] and [10].

- **Signals of opportunity (SoOP)**. Signals of opportunity, as defined in this chapter, are radio-frequency (RF) signals that are not intended for navigation. Examples from previous research include digital television [11], analog television [12], and AM radio [13, 14].

The remainder of this chapter illustrates application of the generic multi-sensor fusion approach described above GPS/laser/inertial and vision/laser/inertial integrated mechanizations.

3 Example Multi-Sensor Fusion Mechanizations

3.1 GPS/Laser Scanner/Inertial

Feature-based navigation techniques represent a viable option for navigation in difficult GPS environments. For example, the integrated laser scanner/INS solution was demonstrated to provide sub-meter accurate navigation for dense urban environments where multiple lines can be extracted from scan images [4]. However, the feature-based navigation approach clearly has its limitations. Relatively open streets represent challenging conditions for the line-based navigation due to limited line availability. To illustrate, Fig. 4 shows a scan image recorded on a relatively open street. For this example, only one line is extracted from the image and the system must augment its position computations by the INS coasting option.

In addition, even in dense urban canyons, line geometry observed in scan images can be insufficient to support complete observability of the navigation states. Figure 5 illustrates the case where lines in the scan image are created by two nearly parallel building walls. In this case, lines extracted from laser scans are nearly collinear. This line geometry allows for the estimation of the cross-track position component only: i.e., the vehicle motion component in the direction perpendicular

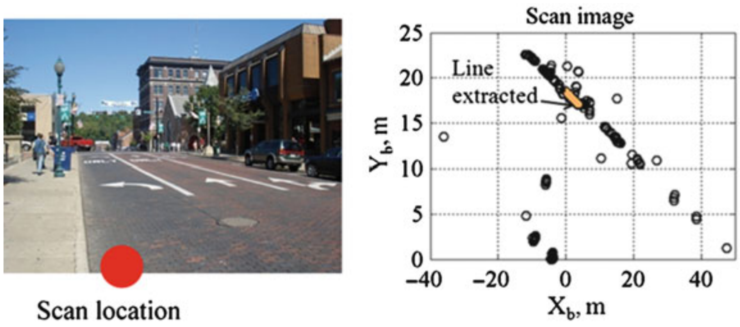


Fig. 4 Example of the laser scan image taken on a relatively open street: only one line is extracted from the image, which is insufficient to compute the position solution

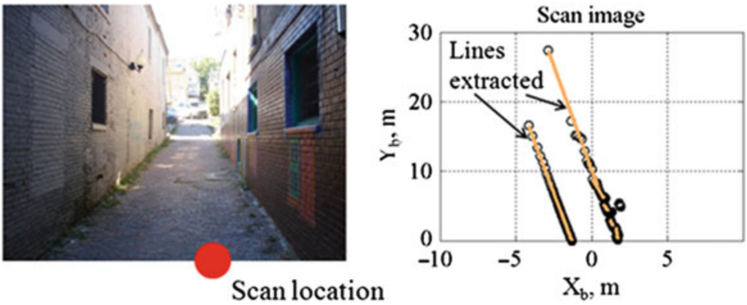


Fig. 5 Example of the laser scan in an urban canyon: nearly collinear lines are extracted, which limits the observability of position states

to the walls can be estimated, while the along-track (parallel to the walls) motion component is unobservable and INS coasting has to be used to compute vehicle displacement in this direction.

The use of GPS can efficiently augment the feature-based navigation approach. Relatively open streets (see, for example, Fig. 4) generally provide enough open sky visibility to track a number of GPS satellite signals that is sufficient for navigation computations. In dense urban canyons where only limited feature geometry is available (see, for example, Fig. 5), a reduced number of satellite signals can be still trackable through limited portions of an open sky. Combining line measurements with these additional GPS measurements can complete the observability of navigation states. To exemplify, the availability of high-elevation satellites that have line-of-sight (LOS) azimuth angles aligned with the direction of an urban canyon shown in Fig. 5 can complete the observability for the along-track position component. In this case, two satellites are required to complete the 2D position and clock solution. One satellite is sufficient for the case where a reliable estimate of the GPS receiver clock bias is available: i.e., the receiver clock bias and clock drift have been previously estimated and the clock stability figures allow for an accurate propagation of the clock state estimates from the estimation time to the current measurement epoch.

Integration of GPS, laser scanner, and inertial data is discussed in the remainder of this subsection. Figure 6 illustrates the integrated system architecture.

This architecture combines GPS, laser scanner, and inertial data for trajectory reconstruction, i.e., the estimation of changes in the user position (or delta position) between successive GPS and laser measurement epochs. An open loop software GPS receiver [17] is exploited for robust carrier phase tracking in difficult environments. The integrated solution utilizes INS navigation outputs to improve the solution robustness. Particularly, inertial data are applied to (1) predict feature displacements between laser scans for robust feature association between scan images, and, (2) computationally adjust a 2D scan plane for tilting of the laser scanner platform.

As shown in Fig. 6, a Kalman filter is used to periodically compute estimates of inertial error states that are then applied to mitigate the drift in INS navigation

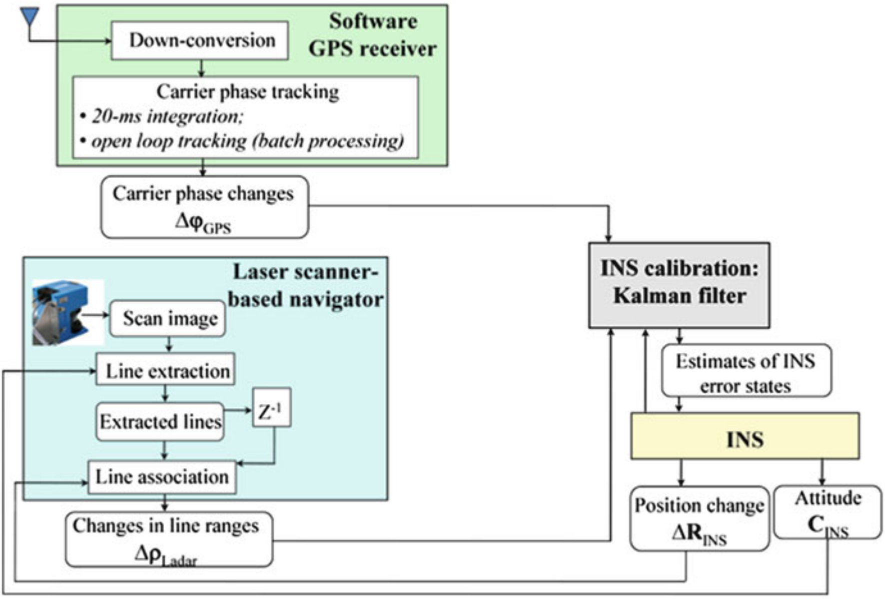

Software GPS receiver
Down-conversion
Carrier phase tracking
• 20-ms integration;
• open loop tracking (batch processing)
Carrier phase changes $\Delta\varphi_{GPS}$
Laser scanner-based navigator
Scan image
Line extraction
Extracted lines
Z^{-1}
Line association
Changes in line ranges $\Delta\rho_{Ladar}$
INS calibration: Kalman filter
Estimates of INS error states
INS
Position change ΔR_{INS}
Attitude C_{INS}

Fig. 6 Integrated GPS/laser scanner/INS system architecture

outputs. Kalman filter observables are formulated in the measurement domain of GPS and laser scanner, which is, as mentioned previously, referred to as range-domain data fusion. Specifically, changes in GPS carrier phase measurements and changes in ranges to lines extracted from scan images serve as filter measurement observables. A complementary filter formulation [15] is utilized for the efficient linearization of system state relations: i.e., filter measurement observables are computed as differences between GPS (and/or laser scanner) measurement observables and INS navigation outputs transformed into the GPS (and/or laser scanner) measurement domain. In addition, a dynamic-state filtering approach [16] is exploited. In this case, the filter implements displacement error states (dynamic states) rather than absolute position error states.

GPS observables of the Kalman filter are based on carrier phase single differences (SDs) between the filter update epochs. The SD approach allows exploiting mm-level accurate carrier phase measurements for trajectory reconstruction without the need to resolve integer ambiguities. Equation (8) formulates the carrier phase SD equation [17]:

$$\Delta\tilde{\varphi}_j = \tilde{\varphi}_j(t_M) - \tilde{\varphi}_j(t_{M-1}) = \Delta r_j + \Delta\delta t_{\mathrm{rcvr}} + \Delta\mathrm{errors}_j + \Delta\eta_j \tag{8}$$

where:

- $\Delta\tilde{\varphi}_j$ is the carrier phase SD for satellite j
- $\tilde{\varphi}_j$ is the carrier phase measurement for satellite j

- $t_M = t_0 + M \cdot \Delta t_{\text{GPS}}$ is the discrete time and Δt_{GPS} is the GPS measurement update interval
- $\Delta r_j = r_j(t_M) - r_j(t_{M-1})$ is the SD of the range r_j between the GPS antenna and satellite j
- $\Delta\delta t_{\text{rcvr}}$ is the SD of the receiver clock bias, or, equivalently, the receiver clock drift accumulated over the Δt_{GPS} interval
- the Δerrors_j term represents changes in deterministic error components of standalone GPS measurements and includes changes in ionospheric and tropospheric delays, changes in the satellite clock bias, and drift components of relativistic corrections
- $\Delta\eta_j$ is the joint noise and multi-path term, which includes carrier noise and multi-path

The SD in satellite/receiver range is expressed as follows:

$$\Delta r_j = SV\text{Doppler}_j - \Delta\text{geometry}_j - \left(\mathbf{e}_j(t_M), \Delta\mathbf{R}\right) \tag{9}$$

where:

- $SV\text{Doppler}_j$ is a change in the range due to the satellite motion along the line-of-sight (LOS)
- $\Delta\text{geometry}_j$ accounts for changes in the relative satellite/receiver geometry
- $\mathbf{e}_j$ is the unit vector pointed from the receiver to the satellite, this vector is generally referred to as the LOS unit vector
- $\Delta\mathbf{R}$ is the receiver position change vector for the interval $[t_{M-1}, t_M]$
- (,) is the vector dot product

Carrier phase SDs are adjusted for the satellite motion terms, geometry terms, and delta error terms prior to their exploitation as Kalman filter measurement observables. For the SD adjustment, satellite motion and geometry terms are computed as follows [17]:

$$\text{SVDoppler}_j = \left(\mathbf{e}_j(t_M), \mathbf{R}_{\text{SV}j}(t_M)\right) - \left(\mathbf{e}_j(t_{M-1}), \mathbf{R}_{\text{SV}j}(t_{M-1})\right) \tag{10}$$

$$\Delta\text{geometry}_j = \left(\mathbf{e}_j(t_M) - \mathbf{e}_j(t_{M-1}), \mathbf{R}(t_{M-1})\right) \tag{11}$$

In (10) and (11):

- $\mathbf{R}_{\text{SV}j}$ is the satellite position vector
- $\mathbf{R}$ is the receiver position vector

For geometry compensation, the receiver position vector $\mathbf{R}$ at the previous update (t_{M-1}) is estimated based on GPS pseudorange measurements. For those cases where not enough pseudorange measurements are available, the position estimate is propagated using inertial data. Note that a sub-hundred-m level accurate position estimate is generally sufficient to support mm-level accuracy in the carrier phase SDs [17]. The satellite position $\mathbf{R}_{\text{SV}j}$ vector is computed from ephemeris data, and the LOS unit vector $\mathbf{e}_j$ is computed based on ephemeris data and the pseudorange-based receiver position estimate. Tropospheric drift terms are compensated based on tropo models [3]. Iono delta errors are normally compensated using dual frequency measurements [3]. However, generally, iono drift terms stay at a mm/s level or less unless ionospheric scintillations are present [17]. Thus, for most operational scenarios, uncompensated iono drift does not significantly influence the accuracy of carrier phase SDs. For this reason, the system mechanization reported in this chapter does not implement iono corrections.

From (8) and (9), carrier phase SDs that are adjusted for the satellite motion, geometry changes, and delta error terms are expressed as follows:

$$\Delta\tilde{\varphi}_j^{\mathrm{adj}}(t_M) = -\big(\mathbf{e}_j(t_M), \Delta\mathbf{R}\big) + \Delta\delta t_{\mathrm{rcvr}} + \Delta\eta_j \tag{12}$$

GPS observables of the Kalman filter are computed as differences between INS-based predicted values of carrier phase SDs and carrier phase SDs that are actually measured by the GPS receiver. For satellite j, this difference is formulated as:

$$u_j^{\mathrm{Kalman}} = \Delta\tilde{\varphi}_j^{\mathrm{INS}} - \Delta\tilde{\varphi}_j^{\mathrm{adj}} \tag{13}$$

where:

$$\Delta\tilde{\varphi}_j^{\mathrm{INS}} = -\left(\mathbf{e}_j(t_M), \Delta\tilde{\mathbf{R}}_{\mathrm{INS}} + \Delta\tilde{\mathbf{C}}_b^N \cdot \mathbf{L}_{\mathrm{IMU}}^{\mathrm{GPS}}\right) \tag{14}$$

In (14), $\Delta\tilde{\mathbf{C}}_b^N = \tilde{\mathbf{C}}_b^N(t_M) - \tilde{\mathbf{C}}_b^N(t_{M-1})$ and $\mathbf{L}_{\mathrm{IMU}}^{\mathrm{GPS}}$ is the lever-arm vector between the inertial measurement unit (IMU) and GPS antenna with the lever-arm vector components being resolved in the body frame. Equations (13) and (14) illustrate the specific implementation of the generic multi-sensor fusion observable, which is formulated by (1), for the case where GPS carrier phase measurements serve as reference measurements for the INS drift estimation.

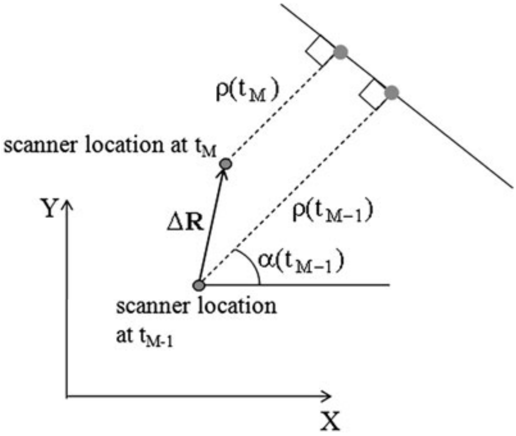
Laser scanner observables of the Kalman filter are computed from navigation-related features observed in scan images. These observations are illustrated in Fig. 7.

For the laser scanner case, lines are chosen as the basis navigation feature [4]. Kalman filter observables are formulated based on changes in line parameters between consecutive scans. As illustrated in Fig. 7, changes in the scanner location between scans create changes in parameters of lines observed in scan images. In Fig. 7, line is represented by its normal point, where a normal point is defined as a perpendicular intersection of the line itself and the line originating from the scanner location. Line normal points are characterized by their polar parameters: range ρ and angle α. From the geometry shown in Fig. 7, position change is related to the line range change as follows [4]:

$$\Delta\tilde{\rho}_k = \tilde{\rho}_k(t_M) - \tilde{\rho}_k(t_{M-1}) \tag{15}$$

$$= -\Delta R_X \cos\big(\alpha_k(t_{M-1})\big) - \Delta R_Y \sin\big(\alpha_k(t_{M-1})\big) + \Delta\varepsilon_k \tag{16}$$

Fig. 7 Laser scanner motion and observed change in line parameters



where:

- $\Delta\tilde{\rho}_k$ is the line range change for line k
- $\tilde{\rho}_k$ is the line range estimated by a line extraction procedure (e.g., using a modified iterative split and merge line extraction algorithm described in reference [18])
- ΔR_X and ΔR_Y are position change components
- α_k is the line angle
- $\Delta\varepsilon_k$ is the noise in estimated line delta range that is due to line extraction errors; these errors generally comprise of laser measurement noise and a texture of a scanned surface

Note equation (15) operates with line parameters that are transformed into the horizontal plane using the laser tilt compensation procedure described in [4]. Lines are observed at the laser scanner body frame. This frame can be tilted due to non-zero pitch and roll angles of the scanner platform. The tilt compensation procedure exploits INS attitude outputs to project lines observed in the scanner body-frame onto a horizontal plane of the navigation frame in order to mitigate the influence of laser tilt angles on the navigation solution accuracy. Thus, (15) relates changes in line parameters with changes in position vector components that are resolved in the axes of the navigation frame. For the system implementation considered in this chapter, navigation is performed at the East-North-Up (ENU) frame. Laser scanner body-frame and body-frame of the IMU are computationally aligned to the axes of the navigation frame during the system initialization stage. IMU body-frame is physically aligned with the laser body-frame and misalignment errors are calibrated during the system initialization.

Below, (15) is reformulated in a vector form:

$$\Delta\tilde{\rho}_k = -\big(\mathbf{n}_k(t_M), \Delta\mathbf{R}\big) + \Delta\varepsilon_k \tag{17}$$

where:

$$\mathbf{n}_k\left(t_{M-1}\right) = \left[\cos\left(\alpha_k\left(t_{M-1}\right)\right)\ \sin\left(\alpha_k\left(t_{M-1}\right)\right)\ 0\right]^T \tag{18}$$

A laser scanner observable of the Kalman filter is computed as a difference between line delta range predicted based on inertial data and a line delta range extracted from scanner measurements. For line p, this is formalized as follows:

$$v_k^{\text{Kalman}} = \Delta\tilde{\rho}_k^{\text{INS}} - \Delta\tilde{\rho}_k \tag{19}$$

where:

$$\Delta\tilde{\rho}_k^{\text{INS}} = -\left(\mathbf{n}_k\left(t_M\right), \Delta\tilde{\mathbf{R}}_{\text{INS}} + \Delta\tilde{\mathbf{C}}_b^N \cdot \mathbf{L}_{\text{IMU}}^{\text{Laser}}\right) \tag{20}$$

and, $\mathbf{L}_{\text{IMU}}^{\text{Laser}}$ is the IMU to laser scanner lever arm resolved in the body-frame. Note that (19) and (20) represent a specific implementation of the generic multi-sensor fusion observable defined by (1) for the case of laser/inertial data fusion.

GPS and laser observables of the Kalman filter are combined into a joint filter measurement vector:

$$\bar{\mathbf{y}}_{\text{Kalman}} = \left[u_1^{\text{Kalman}} \ldots u_J^{\text{Kalman}}\ v_1^{\text{Kalman}} \ldots v_K^{\text{Kalman}}\right]^T \tag{21}$$

As mentioned previously the Kalman filter implements the dynamic-state estimation approach: i.e., the filter does not estimate absolute position; instead, position changes between GPS and laser scanner measurement updates are estimated. The state vector for the dynamic-state filter formulation includes twenty states: delta position error states (three states); velocity error states (three states); attitude error states (three states); delta attitude error states (three states) – these are attitude errors accumulated over the filter update interval; gyro bias states (three states); accelerometer bias states (three states); and, GPS receiver clock states that include delta bias state and drift state. For this state vector, the filter observation matrix ($\mathbf{H}_{\text{Kalman}}$) is derived from the filter observation equations (13), (14), (19), and (20). Elements of this matrix projections of filter states into the filter observation domain as illustrated in Fig. 8.

In Fig. 8:

- $\mathbf{0}$'s are 3×1 zero rows
- $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{d}$ are 3×1 rows that account for the transformation of INS attitude error terms into the filter observables through the lever-arm compensation

These matrices were derived using the approach proposed in [16]. Results of the derivation are as follows:

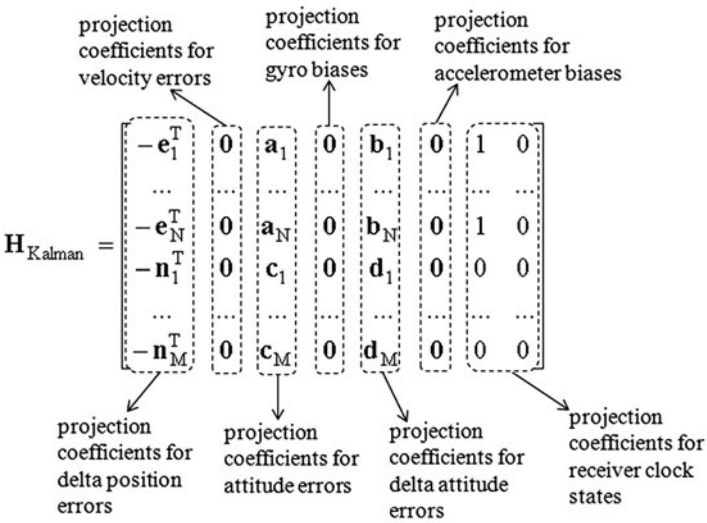

projection coefficients for velocity errors
projection coefficients for gyro biases
projection coefficients for accelerometer biases
$\mathbf{H}_{\text{Kalman}} = \begin{bmatrix} -\mathbf{e}_1^T & \mathbf{0} & \mathbf{a}_1 & \mathbf{0} & \mathbf{b}_1 & \mathbf{0} & 1 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -\mathbf{e}_N^T & \mathbf{0} & \mathbf{a}_N & \mathbf{0} & \mathbf{b}_N & \mathbf{0} & 1 & 0 \\ -\mathbf{n}_1^T & \mathbf{0} & \mathbf{c}_1 & \mathbf{0} & \mathbf{d}_1 & \mathbf{0} & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -\mathbf{n}_M^T & \mathbf{0} & \mathbf{c}_M & \mathbf{0} & \mathbf{d}_M & \mathbf{0} & 0 & 0 \end{bmatrix}$
projection coefficients for delta position errors
projection coefficients for attitude errors
projection coefficients for delta attitude errors
projection coefficients for receiver clock states

Fig. 8 Kalman filter observation matrix: matrix elements define projection coefficients for projecting filter states into the filter observation domain

$$
\begin{aligned}
\mathbf{a}_j &= \left(\mathbf{e}_j \times \left(\Delta\tilde{\mathbf{C}}_b^N \cdot \mathbf{L}_{\text{IMU}}^{\text{GPS}}\right)\right)^T \\
\mathbf{b}_j &= \left(\mathbf{e}_j \times \left(\tilde{\mathbf{C}}_b^N(t_{M-1}) \cdot \mathbf{L}_{\text{IMU}}^{\text{GPS}}\right)\right)^T \\
\mathbf{c}_k &= \left(\mathbf{n}_k \times \left(\Delta\tilde{\mathbf{C}}_b^N \cdot \mathbf{L}_{\text{IMU}}^{\text{Laser}}\right)\right)^T \\
\mathbf{d}_k &= \left(\mathbf{n}_k \times \left(\tilde{\mathbf{C}}_b^N(t_{M-1}) \cdot \mathbf{L}_{\text{IMU}}^{\text{Laser}}\right)\right)^T
\end{aligned} \tag{22}
$$

where:

- × is the vector dot product

The filter measurement noise matrix is a diagonal matrix defined by variances in carrier phase SDs and line delta ranges:

$$\mathbf{R}_{\text{Kalman}} = \operatorname{diag}\left(\sigma^2_{\Delta\varphi_1}, \ldots, \sigma^2_{\Delta\varphi_J}, \sigma^2_{\Delta\rho_1}, \ldots, \sigma^2_{\Delta\rho_K}\right) \tag{23}$$

Carrier phase sigmas ($\sigma_{\Delta\varphi_k}$) are computed as follows:

$$\sigma^2_{\Delta\varphi_j} = \sigma^2_{\varphi_j}(t_M) + \sigma^2_{\varphi_j}(t_{M-1}) \tag{24}$$

$$\sigma_{\varphi_j}(t_M) = \frac{\lambda_{L1}}{2\pi} \cdot \sqrt{\frac{1}{T_{\text{int}}} \cdot 10^{-\frac{C/No_j(t_M)}{10}}} \tag{25}$$

$$j = 1, \ldots, J \tag{26}$$

where:

- λ_{L1} is wavelength of the GPS link 1 ($L1$) carrier (0.19 cm, approximately)
- T_{int} is the GPS receiver signal integration interval
- C/No is the carrier-to-noise ratio that is routinely estimated by the open loop GPS receiver

Laser delta range sigmas ($\sigma_{\Delta\rho_k}$) are calculated based on (27):

$$\sigma^2_{\Delta\rho_k}(t_M) = \sigma^2_{\rho_k}(t_M) + \sigma^2_{\rho_k}(t_{M-1})\ , k = 1, ..., K \tag{27}$$

where:

$\sigma_{\Delta\rho_k}$ are line range sigmas that are estimated using the approach reported in [19].

With the filter states listed above, filter measurements formulated by (13), (14), (19), and (20), and filter matrices represented in Fig. 8 and defined by (22) and (23), the INS drift estimation procedure implements a complementary Kalman filter algorithm for the estimation of the inertial error states and GPS receiver clock states. The filter formulation is similar to the GPS/INS Kalman filter model found in [15].

Performance of the GPS/laser scanner/inertial solution is illustrated below using experimental data collected in real urban environments. A test van was used as a platform for urban data acquisition. The data were acquired in Athens, OH, USA. Figure 9 shows a photograph of the data collection setup.

The setup used to acquire and process live GPS, laser scanner, and inertial urban data includes:

- Laser scanner: SICK LMS-200. A centimeter distance measurement mode was chosen. For this mode, a standard deviation of the laser ranging noise is specified as 5 mm. The maximum measurement range is 80 m. A scan angular range is from 0 to 180° with an angular resolution of 0.5°. A scanner update rate of one scan per 0.4 s was used.
- IMU: Systron Donner DQI IMU. This IMU represents a tactical-grade unit whose main characteristics are summarized in Table 1.
- GPS receivers: A software receiver front-end developed at the Ohio University Avionics Engineering Center was used for collection of raw GPS signal samples [20]. A NovAtel Superstar II receiver provided 1 PPS signal for time-stamping of laser scanner and inertial measurements.
- Synchronization and data collection board: Xilinx Spartan-3 Field Programmable Gate Arrays (FPGA). The board decodes laser and inertial data from corresponding measurement sensors, time stamps the measurements decoded using 1 PPS signal, and then sends time-stamped measurements via a Digilent USB board to a PC that collects the data for post-processing.
- Data processing: Post-processing implemented in Matlab$^{\text{TM}}$.

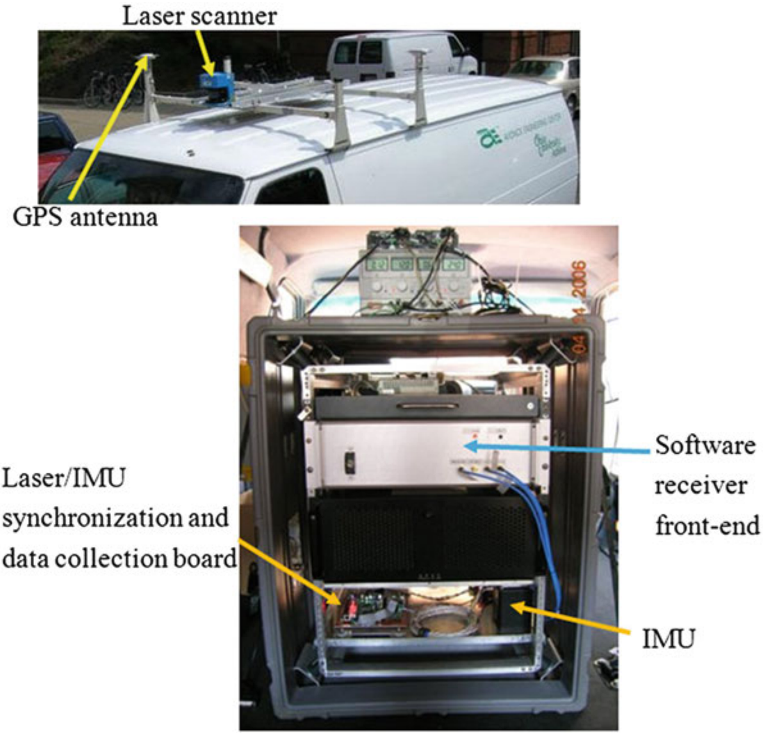

Laser scanner
GPS antenna
Laser/IMU synchronization and data collection board
Software receiver front-end
IMU

Fig. 9 Data collection setup used for the acquisition of GPS/Laser scanner/inertial urban test data

Table 1 IMU characteristics

Parameter	Value
Gyro bias in-run stability	3°/h (sigma)
Gyro noise	$0.035°/\sqrt{\text{h}}$ (sigma)
Gyro scale factor	350 ppm (sigma)
Accelerometer bias in-run stability	200 μg (sigma)
Accelerometer noise	$2\ \mu\text{g}/\sqrt{\text{Hz}}$ (sigma)
Accelerometer scale-factor	350 ppm (sigma)
Sensor axis non-orthogonality	0.5 mrad

Figure 10 shows test trajectory implemented for the first test scenario. This figure shows the vehicle trajectory along with images of the environment taken at select trajectory points. The test trajectory shown is reconstructed by the laser/INS integration that is described in [4].

As it can be seen from the images represented, the test trajectory includes both relatively open streets as well as dense urban canyons. Thus, this trajectory allows testing performance of the integrated solution in a variety of urban environmental conditions.

Figure 11 illustrates complementary availability of GPS and laser measurements for the first test scenario. Availability of GPS measurements is represented by

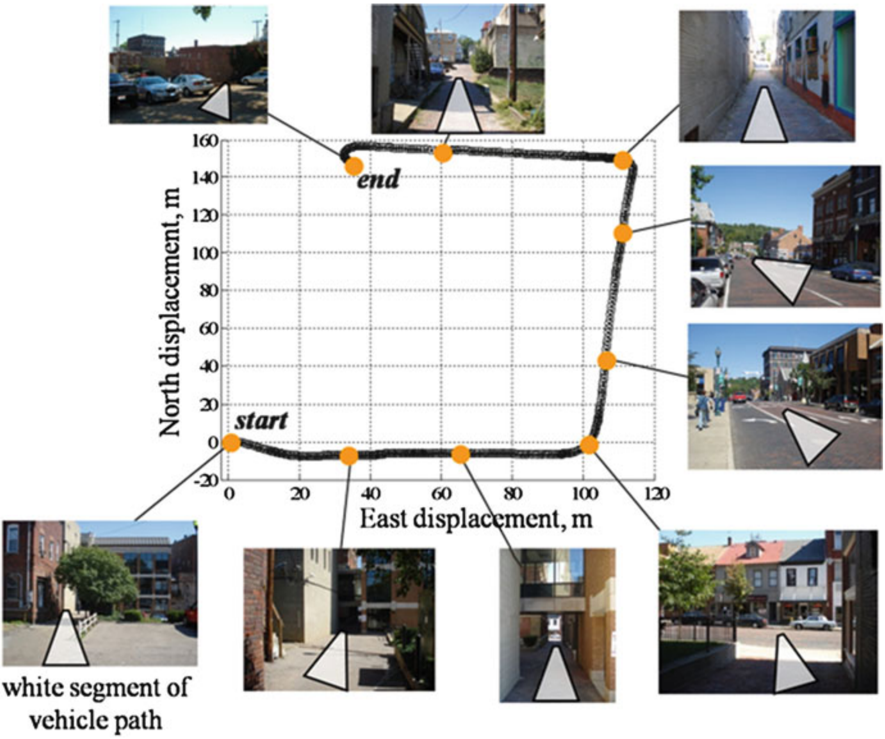


Fig. 10 Test trajectory: trajectory reconstructed by the laser/INS integration is shown along with images of the environment taken at select trajectory points

satellite tracking statuses. A particular satellite is designated by its pseudorandom code number (PRN). The satellite tracking status is shown using a black-to-white color scheme. Black color indicates satellites with a strong C/No (50 dB-Hz or higher). White color is used for satellites with a C/No value below the tracking threshold of 32 dB-Hz: i.e., white color indicates that a satellite is not visible. The black-to-white color scheme is also applied to illustrate the availability of features extracted from scan images. Black color indicates that multiple lines are extracted from scan images and the line geometry allows for complete navigation computations. White color designates cases where no lines are extracted. As shown in Fig. 11, GPS and laser scanner measurements exhibit complementary availability. Limited or no GPS measurements are available in dense canyons where multiple lines are extracted from scan images: see, for example, a portion of the plot that corresponds to the urban canyon image. Vice-versa, a sufficient number of visible satellites is present on open streets where limited or no lines are extracted from laser scans: see, for example, a portion of the plot that corresponds to the open street image.

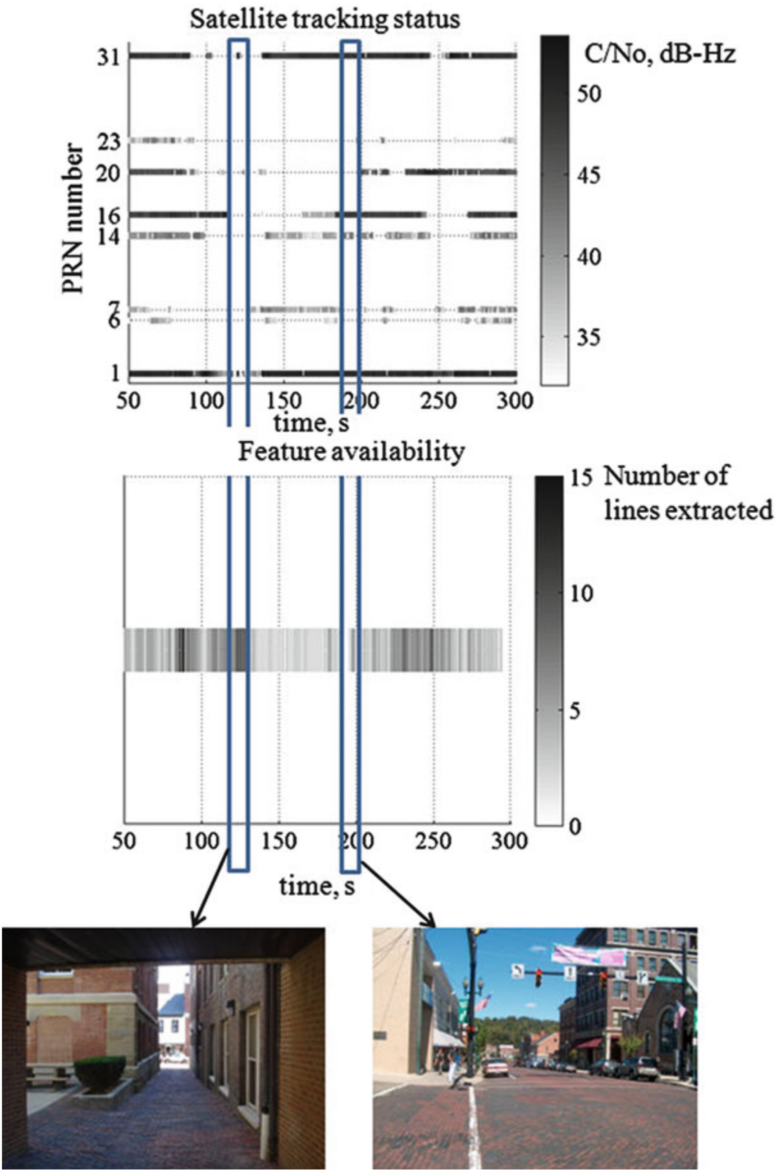


Fig. 11 Complementary availability of GPS and laser scanner measurements

Delta position residuals are applied to characterize the trajectory estimation performance. Residual vector ($\delta\mathbf{R}$) is computed as the difference between the laser/GPS weighted LMS estimate of the delta position ($\Delta\mathbf{R}_{Laser/GPS}$) and inertial

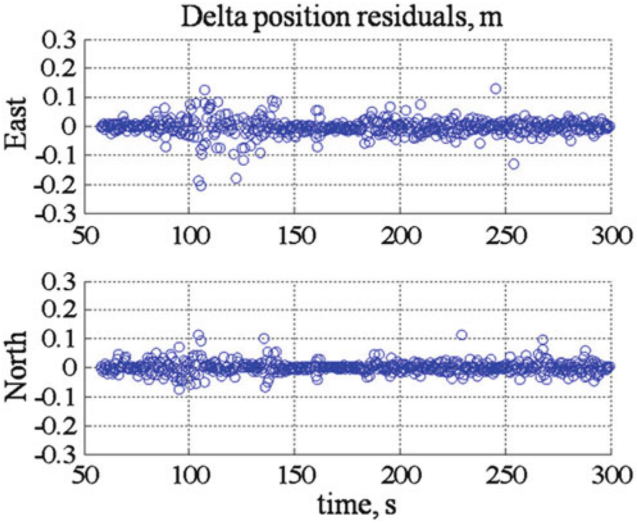


Fig. 12 Delta position residuals for the GPS/laser/INS urban test example

delta position ($\Delta\mathbf{R}_{\mathrm{INS}}$) that is compensated for drift terms using Kalman filter predictions:

$$\delta\mathbf{R} = \Delta\mathbf{R}_{\mathrm{Laser/GPS}} - \Delta\mathbf{R}_{\mathrm{INS}} - \left(\tilde{\mathbf{C}}_{\mathrm{b}}^{\mathrm{N}}(t_M) - \tilde{\mathbf{C}}_{\mathrm{b}}^{\mathrm{N}}(t_{M-1})\right) \cdot \mathbf{L}_{\mathrm{IMU}}^{\mathrm{GPS}} \tag{28}$$

The GPS/Laser LMS delta position solution is computed based on GPS carrier phase SDs and laser line delta ranges without the use of the inertial. LMS estimation details are given in [7]. Essentially, delta position residuals characterize the level of noise in the reconstructed trajectory. This noise is a combined effect of the GPS carrier phase noise, noise in line ranges, and the noise component of INS position drift. Figure 12 shows residual plots for East and North components of delta position.

The standard deviations of East and North residual components are computed as 3.0 cm and 2.0 cm, respectively. For a similar test trajectory, residual standard deviations of the laser/INS integrated implementation (i.e., when GPS observables of the Kalman filter are not used) are at a 7-cm level [4]. GPS/laser/INS residual noise is generally increased when the LMS solution relies primarily on line ranges as compared to cases with good satellite availability. This is due to a higher level of noise in line ranges (caused mostly by the texture of scanned urban surfaces) as compared to the carrier phase noise. For example, an increased residual noise can be observed for the time interval starting at approximately 60 s and ending at approximately 130 s. This interval corresponds to a part of the trajectory that belongs to dense urban environments where limited satellites are available: for illustration of the environment, see the third image from the trajectory start in Fig. 10.

3.2 Vision/Laser Scanner/Inertial Integration

Vision-based navigation techniques serve as a viable option for autonomous, passive navigation, and guidance in GPS-denied environments [8]. However, vision-based methods, including stereo-vision, are known to be brittle to signal noise, particularly in terms of estimating the image depth. Stereo vision suffers from several disadvantages. The first is synchronization: if the stereo system is moving, and the two images are not captured at the same time, the time discrepancy between the two images can invalidate the stereo reconstruction. This timing discrepancy cannot be completely compensated using the relative motion information (provided, for example, by the inertial navigation) due to the unknown image depth. A second problem is "vergence," i.e., the fact that the area of interest must be in the field of view (FoV) of both cameras, and must be focused in both cameras. This typically requires different orientation of the cameras for close targets (e.g., more "cross-eyed") and far targets. Differences in the optical response of the two cameras can also make matching more difficult. Monocular vision methods give a scaled estimate of camera motion and a scaled estimate of scene structure. This scaled estimate is the major drawback of monocular methods, and requires additional sensors or information to recover the correct scale.

The fusion of vision and laser data is applied for efficient resolution of the scale ambiguity and performance enhancement of the navigation accuracy and reliability in the face of image noise. Laser scan data are exploited to initialize three-dimensional (3D) tracking of stationary features and to enhance the feature tracking performance. Example 3D features include planar surfaces (characterized by range and normal vector) and point features (characterized by their Cartesian coordinates). Changes in feature parameters between images are used for navigation. Inertial data are applied to perform robust feature matching between images and coasting in environments where insufficient features are available. INS data can also be used to adjust laser scan range measurements for platform motion in order to compensate for the time discrepancy between camera images and laser scans. The proposed integration concept allows for the initialization of 3D feature tracking based on a limited number of laser scans: two scans are required, after which the system can operate in a completely passive mode. This serves as an important aspect for many surveillance, reconnaissance, and navigation missions.

To improve the robustness of the navigation solution, the vision/laser integrated mechanization is augmented with inertial sensor measurements. INS outputs are exploited for robust matching of the features that are extracted from vision and laser data. The use of INS outputs for feature matching in video images is described in [8]. Details of the INS-based feature matching in laser scan images are discussed in [4]. Inertial data are also used to coast through those cases where a limited number of features is available. Figure 13 shows a high-level diagram of the Vision/Laser/INS integrated mechanization.

Laser scanner measurements are used to initialize the depth estimate of the vision-based features. As stated previously, a very limited number of scans (two

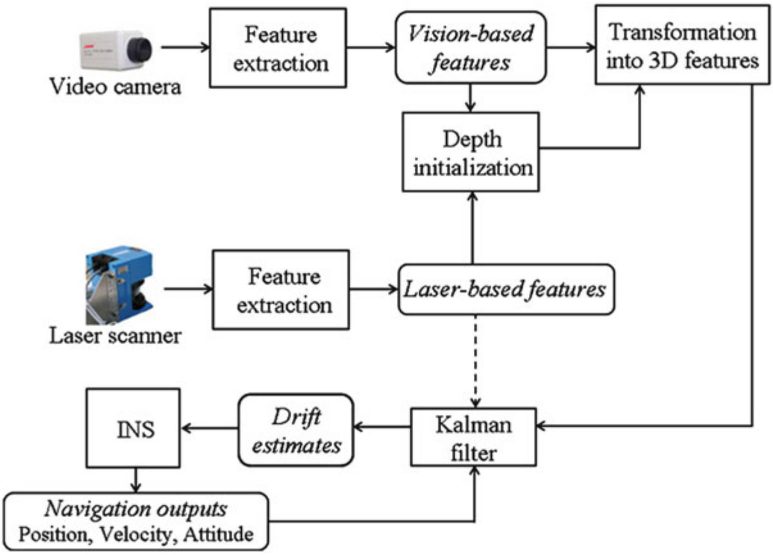


Fig. 13 Vision/Laser/INS integrated solution

scans) is sufficient for the image depth initialization. Following the initialization step, the system can operate in a complete passive mode using vision-based features only. Navigation accuracy can be improved considerably if periodic scans at a limited rate (such as, for example, one scan per 30 s) are applied. This aspect is explained in more details in the feasibility demonstration discussion below. For those cases where the active component of the laser measurement mechanism does not represent a concern (i.e., for the majority of civil use cases), laser scanner can operate continuously to provide additional feature measurements in order to improve the navigation performance.

The vision/laser/INS mechanization shown in Fig. 13 estimates the navigation solution (position, velocity, and attitude) from the INS. Vision and laser feature measurements are applied to estimate inertial error states in order to mitigate drift in inertial navigation outputs. Laser-based observables of the Kalman filter are defined by (17) through (20) above.

Equation (29) formulates vision-based observables of the Kalman filter using the unit sphere representation of image features that supports multi-aperture camera cases [21]:

$$\eta_p^{\text{Kalman}} = \begin{bmatrix} \left(\tilde{\mathbf{n}}_p^{(2)}\right)^T \cdot \Delta\tilde{\mathbf{C}}_N^b \cdot \mathbf{B} \cdot \Delta\tilde{\mathbf{R}}_{\text{INS}} - \left(\tilde{\mathbf{n}}_p^{(1)}\right)^T \cdot \mathbf{B}^T \cdot \Delta\tilde{\mathbf{C}}_b^N \cdot \tilde{\mathbf{n}}_p^{(2)} \cdot \hat{\rho}_p^{(1)} \\ \left(\tilde{\mathbf{n}}_p^{(2)}\right)^T \cdot \Delta\tilde{\mathbf{C}}_N^b \cdot \mathbf{D}_p \cdot \Delta\tilde{\mathbf{R}}_{\text{INS}} - \left(\tilde{\mathbf{n}}_p^{(2)}\right)^T \cdot \Delta\tilde{\mathbf{C}}_N^b \cdot \tilde{\mathbf{n}}_{\perp p}^{(1)} \cdot \hat{\rho}_p^{(1)} \end{bmatrix} \tag{29}$$

$$p = 1, \ldots, P \tag{30}$$

where:

- $\tilde{\mathbf{n}}_p^{(s)}$, $p = 1, \ldots, P$, $s = 1, 2$ is the unit vectors of the feature p (i.e., the unit vector pointed from the center of the camera to the feature p) that is extracted from image s
- $\tilde{\mathbf{n}}_{\perp p}^{(s)}$ is the unit vector that is perpendicular to the feature unit vector; $\hat{\rho}_p{}^{(1)}$ is the estimated feature range for image 1
- $\Delta\tilde{\mathbf{R}}_{\mathrm{INS}}$ and $\Delta\tilde{\mathbf{C}}_b^N$ are INS position and orientation changes between images 1 and 2
- Matrices **B** and **D** are defined as follows:

$$\mathbf{B} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{31}$$

$$\mathbf{D_p} = \begin{bmatrix} 0 & 0 & -\cos\left(\varphi_p^{(1)}\right) \\ 0 & 0 & -\sin\left(\varphi_p^{(1)}\right) \\ \cos\left(\varphi_p^{(1)}\right) & \sin\left(\varphi_p^{(1)}\right) & 0 \end{bmatrix} \tag{32}$$

where:

$\varphi_p^{(1)}$ is the feature spherical azimuth angle as measured from image 1.

Image 1 is generally the image where the feature was first observed and image 2 is the current image. Equation (29) is linearized to support the linear formulation of the complementary Kalman filter.

As stated previously the key motivation for combining vision and laser data is the use of laser scan measurements for the estimation of the unknown depth of video images. The estimation approach is illustrated in Fig. 14.

The depth estimation exploits two images from a video camera and two laser scan images that are acquired at two different locations. It is assumed that the camera and the laser are mounted on the same platform and their measurement frames are collocated. Parameters of features that are extracted from laser scan images and video images at two locations of the platform are used to estimate the image depth. The estimation does not require the camera and the laser to observe the same features: i.e., video and laser features can be completely unrelated to each other. The initialization method discussed in this chapter assumes that point features are extracted from video images and lines are extracted from laser scans. This method can be generalized to include other types of features.

To provide a conceptual illustration of the depth initialization method, a simplified 2D case is considered below. This case is illustrated in Fig. 15.

Camera and laser are placed on the same moving platform and their measurement frames are aligned. The platform moves along the X axis of the XZ navigation

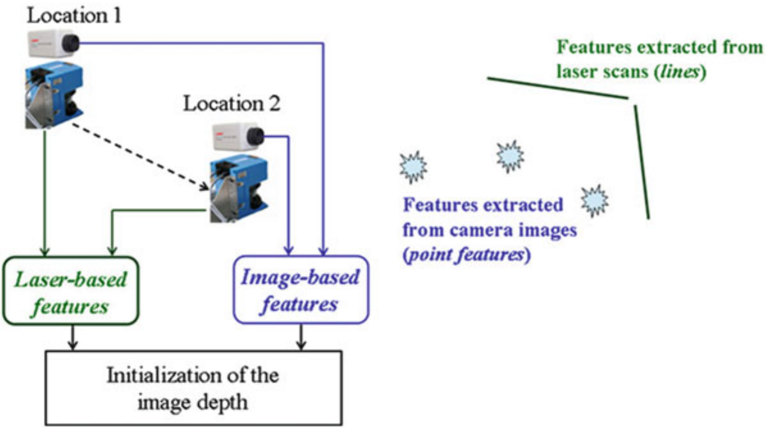

Location 1
Location 2
Laser-based features
Image-based features
Initialization of the image depth
Features extracted from laser scans (*lines*)
Features extracted from camera images (*point features*)

Fig. 14 Image depth initialization method

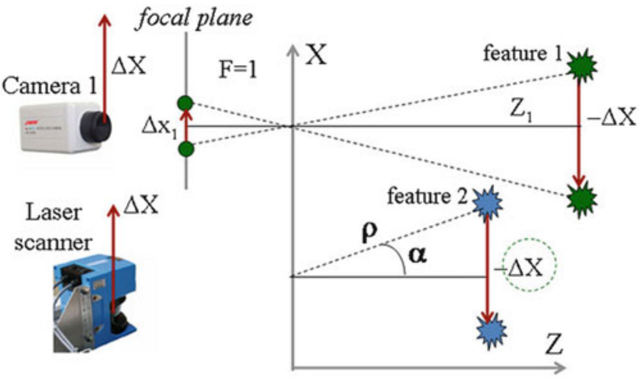


Fig. 15 Illustration of the depth initialization method: simplified 2D case

coordinate frame. The camera optical axis is aligned with the Z axis of the frame. For simplicity, it is assumed that the camera focal length F equals unity. Feature 1 is observed by the camera and feature 2 is observed by the laser scanner. As stated previously, these features may be completely unrelated. The platform displacement $\varDelta X$ transforms into the displacement of the vision-based feature as:

$$\varDelta x^{(1)} = \varDelta X / Z^{(1)} \tag{33}$$

where:

- $Z^{(1)}$ is the unknown depth

The laser scanner observes feature 2. The distance to this feature is directly measured by the scanner: 2D scan points are generally represented by their polar coordinates that include range and polar angle. Polar coordinates of feature 2 that are extracted from two scan images are applied to estimate the platform displacement:

$$\Delta\hat{X} = -\left(\rho_2^{(2)}\sin\left(\alpha_2^{(2)}\right) - \rho_1^{(2)}\sin\left(\alpha_1^{(2)}\right)\right) \tag{34}$$

where:

- $\left(\rho_1^{(2)}, \alpha_1^{(2)}\right)$ and $\left(\rho_2^{(2)}, \alpha_2^{(2)}\right)$ are the range and polar angles of feature 2 in the first and second scan images, accordingly

Platform displacement that is estimated based on scan data is then used to initialize the depth of the image-based feature:

$$\hat{Z}^{(1)} = \Delta\hat{X}/\Delta x^{(1)} \tag{35}$$

A similar approach is applied to estimate the image depth for a general 3D case. Equation (36) relates 3D Cartesian feature coordinates $\mathbf{M}_p$ to its homogeneous coordinates $\mathbf{m}_p$ in the 2D image frame:

$$\mathbf{M}_p^{(1)} = Z_p^{(1)}\mathbf{m}_p^{(1)} \tag{36}$$

$$\mathbf{M}_p^{(2)} = Z_p^{(2)}\mathbf{m}_p^{(2)} \tag{37}$$

$$p = 1, \ldots, P \tag{38}$$

where:

- $Z_k^{(j)}$ is the depth of the kth feature in the jth image

Cartesian feature coordinates in two images are related through the orientation change matrix and the translational vector:

$$\mathbf{M}_p^{(2)} = \Delta\mathbf{C}_b^N\left(\mathbf{M}_p^{(1)} - \Delta\mathbf{R}\right) \tag{39}$$

Substituting (36) into (39) then yields:

$$\mathbf{M}_p^{(2)} = \Delta\mathbf{C}_b^N\left(Z_p^{(1)}\cdot\mathbf{m}_p^{(1)} - \Delta\mathbf{R}\right) \tag{40}$$

Assuming that the optical axis is aligned with the Z-axis of the camera frame, the depth of the feature is related to its Cartesian coordinates as follows:

$$Z_p^{(j)} = \mathbf{n}_Z^T\cdot\mathbf{M}_p^{(j)} \tag{41}$$

where:

- $\mathbf{n}_Z = [0\ 0\ 1]^T$

From (36) through (41), the following relationship can be derived [22]:

$$\mathbf{H}_p \cdot \begin{bmatrix} \Delta\mathbf{R} \\ Z_p^{(1)} \end{bmatrix} = \mathbf{0}_{2\times 1} \tag{42}$$

$$\mathbf{H}_p = \mathbf{A} \cdot \left(\mathbf{I}_{3\times 3} - \Delta\mathbf{C}_b^N \cdot \mathbf{m}_k^{(2)} \cdot \left(\mathbf{n}_Z^T \Delta\mathbf{C}_N^b\right)\right) \cdot \left[\, \mathbf{I}_{3\times 3} \; -\mathbf{m}_p^{(1)} \,\right] \tag{43}$$

$$p = 1, \ldots, P \tag{44}$$

where:

- $\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$
- $\mathbf{I}_{3\times 3}$ is the 3-by-3 unit matrix
- $\mathbf{0}_{2\times 1}$ is a 2×1 zero column

Equation (42) formulates vision-based linear equations for the estimation of the image depth. Note that unknowns include the depth components ($Z_p^{(1)}$, $p = 1, \ldots, P$) and the displacement vector $\Delta\mathbf{R}$. The orientation change matrix is computed from the inertial data and is not estimated based on the laser/vision fusion. This allows for the linear formulation of the depth estimation problem. Note that in this case a correlation between inertial angular errors and initial depth estimates is created. This correlation must be taken into account in the design of the Kalman filter that estimates inertial drift states.

The solution of the equation system (42) is not unique. The displacement vector and depth can only be determined within an ambiguity of a scale-factor γ: i.e., if $\Delta\hat{\mathbf{R}}$ and $\hat{Z}_p^{(1)}$ satisfy the system (42), then $\gamma\Delta\hat{\mathbf{R}}$ and $\gamma\hat{Z}_p^{(1)}$ satisfy this system as well. To remove the scale-factor ambiguity, the system (42) is augmented by laser measurement observables. Line features are extracted from laser images and applied for the image depth initialization. It is assumed herein that lines are created in 2D scan images as a result of the intersection of the horizontal laser scanning plane with vertical planes (such as building walls in urban environments). In this case, the relationship between changes in line parameters and displacement vector components is formulated by (15) above, which, in the matrix form, is expressed as follows:

$$\mathbf{n}_k^T \Delta\mathbf{R} = \rho_k^{(1)} - \rho_k^{(2)}$$

$$\mathbf{n}_k = \left[\cos\left(\alpha_k^{(1)}\right) \quad \sin\left(\alpha_k^{(1)}\right)\right]^T$$

$$k = 1, ..., K \tag{45}$$

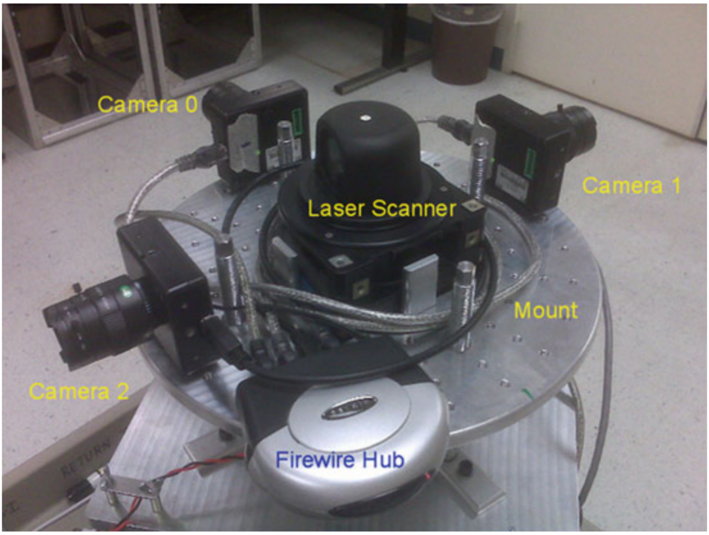


Fig. 16 Data collection setup used for demonstrating the feasibility of the laser/vision depth estimation

Equation (45) provides the laser part of the image depth estimation relations. Combining (45) with the vision-based part defined by (42) yields:

$$\mathbf{H}_p \cdot \begin{bmatrix} \Delta\mathbf{R} \\ Z_p^{(1)} \end{bmatrix} = \mathbf{0}_{2\times 1},\ p = 1, \ldots, P$$

$$\mathbf{n}_k^T \Delta\mathbf{R} = \rho_k^{(1)} - \rho_k^{(2)}, k = 1, \ldots, K \tag{46}$$

Equation (46) defines a linear system of equations for the estimation of image depth. Note that the minimum feature configuration that is required to initialize the image depth based on (46) includes three vision features and one laser feature. In this case, seven equations are available (two per vision feature and one for the laser feature) to estimate six unknowns (three depth values and three components of the displacement vector).

The image depth initialization method was verified with experimental data. Indoor experimental data were collected in hallways of the Air Force Research Laboratory at Eglin Air Force Base. Figure 16 shows a photograph of the data collection setup.

The data collection setup includes a laser scanner and three video cameras. This setup was assembled for the development and verification of the laser scanner/multi-aperture video data fusion. Only one video camera was used for the experiment discussed herein. A straight motion trajectory was implemented. Two laser scans were used to initialize the depth of video images. Following the depth initialization,

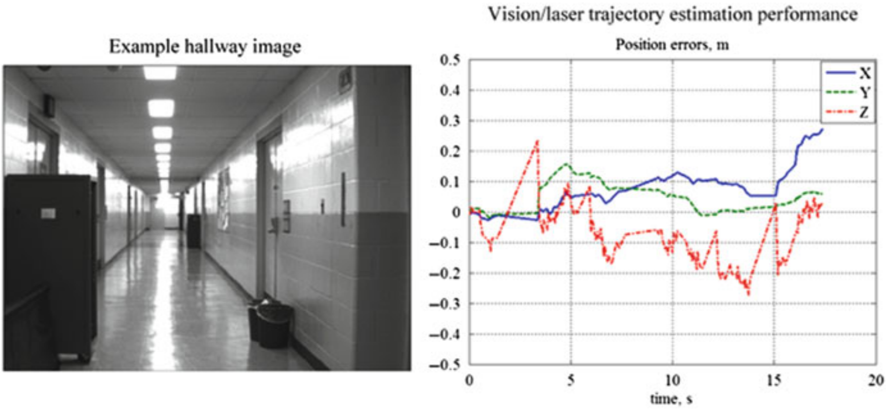


Fig. 17 Example of the vision/laser trajectory estimation performance: only two laser scans are applied to initialize the image depth (at the system start-up and at about 5 s after the beginning of the experiment); following the depth initialization, 3D position components are estimated based on video images only

the displacement vector was estimated based on vision features only without the use of laser scan data. Vision-based displacement estimates were compared to the reference motion trajectory. To construct the reference trajectory, cm-accurate displacement information was estimated from laser scan images as described in reference [4]. Figure 17 exemplifies experimental results.

Experimental data presented show that decimeter-level positioning accuracy that is achieved using video images and two laser scans to initialize the image depth. These results clearly demonstrate the feasibility of the semi-passive navigation approach that uses very limited laser scans (two scans for the example case considered herein) for the initialization of 3D image-based navigation.

Feasibility of the vision/laser/INS mechanization described above was assessed in an indoor simulated environment that was implemented in Matlab. For the simulation, the sensor specifications were implemented as follows:

1. Video: a multi-aperture camera head that includes three video cameras (see Fig. 16 for the illustration of the multi-aperture implementation); the angular separation between the camera optical axis is 120°; for each camera the resolution is 640 × 480, the azimuth FoV is 40°, the elevation FoV is 40°, and, the update rate is 10 Hz;
2. Laser scanner: measurement range is 80 m; angular range: is from 0 to 360°; range noise is 1 cm (std); and, angular resolution is 0.5°;
3. IMU: accelerometer bias is 1 mg; and, gyro drift is 50°/h.

Figure 18 illustrates a 2D horizontal projection of the simulated indoor environment that includes multiple indoor hallways. Hallway walls were simulated as vertical with the wall height equal to 2.5 m. A horizontal motion trajectory with the absolute velocity value of 1 m/s was implemented as shown in Fig. 18.

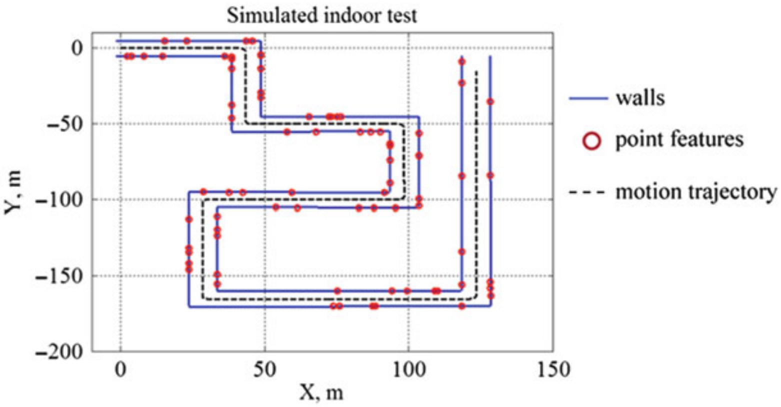


Fig. 18 Simulated indoor environments

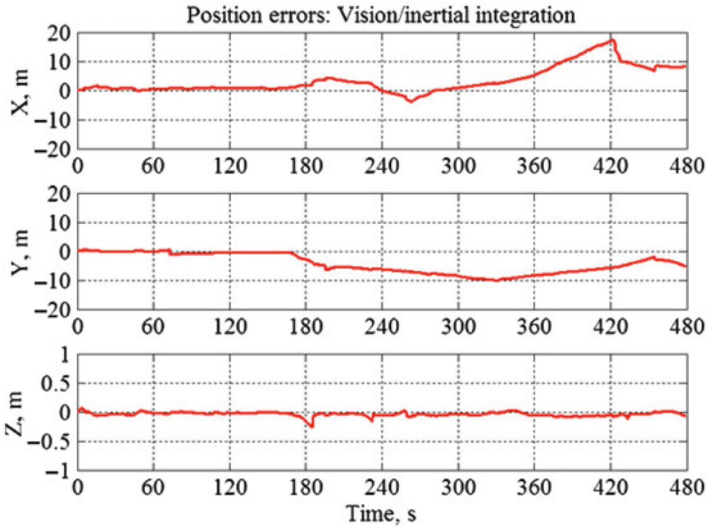


Fig. 19 Performance of the vision/inertial integration for the indoor simulated environment

Figure 19 shows performance of the vision/inertial integration, i.e., the use of laser measurements.

In this case, the feature depth is initialized by observing the feature from two different location of the platform and using inertial displacement and orientation change measurements for the unambiguous estimation of the image depth as described in [21]. Simulation results shown in Fig. 19 indicate that position errors of the vision/INS integration stay in the range from −10 m to 20 m. This level of navigation performance is improved significantly if laser scanner

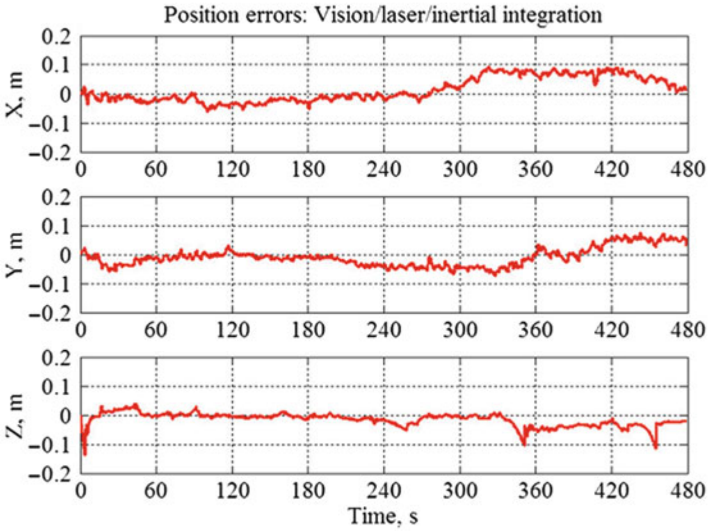


Fig. 20 Performance of the vision/inertial integration for the indoor simulated environment: 1-s laser updates

measurements are incorporated into the solution. Figure 20 illustrates performance of the vision/laser/INS mechanization for the case where a 1-s update rate of laser images is used.

The use of a 1-s laser updates enables cm-accurate estimation of the 3D position vector. Figure 21 shows simulation results for the case where periodic laser scans at a very limited rate are applied. In this case, laser scans are made only once per 30 s.

It is important to note that the system that employs laser scans at this low rate can be still considered as practically passive since, from the practical point of view, a laser scanning at this rate cannot be detected. Positioning accuracy is maintained at a sub-m-level, which provides an order of magnitude performance improvement as compared to the vision/laser implementation. Therefore, for the vision/laser/INS integration it is extremely beneficial to use the system implementation that operates in a semi-passive mode employing periodic scans at a limited scan rate.

4 Summary

Navigating in difficult environments requires the use of multi-sensor fusion techniques. This chapter proposes a generic multi-sensor fusion approach and applies this approach for developing GPS/laser scanner/inertial and vision/laser/inertial integrated mechanizations. Simulated data and data collected in various urban indoor and outdoor environments show that multi-sensor fusion techniques developed demonstrate a significant potential for enabling reliable and accurate navigation capabilities for a variety of challenging navigation scenarios.

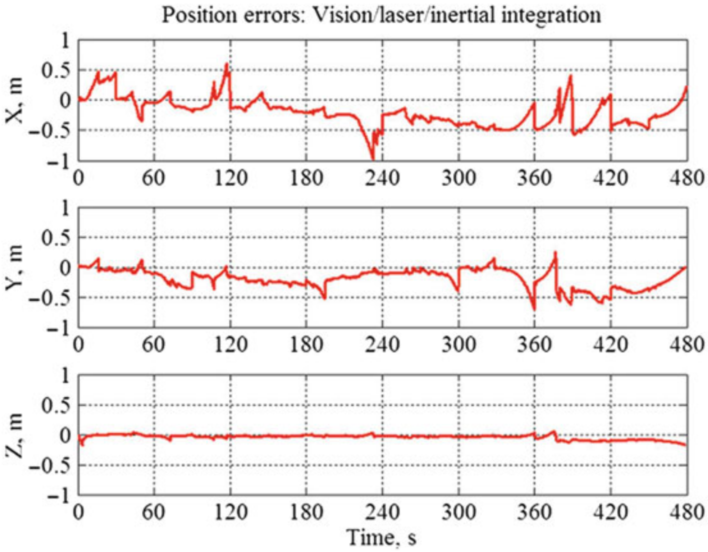


Fig. 21 Performance of the vision/inertial integration for the indoor simulated environment: 30-s laser updates

5 Disclaimer

The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S Government.

References

1. D. H. Titterton, and J. L. Weston, Strapdown Inertial Navigation Technology, Second Edition, The American Institute of Aeronautics and Astronautics, Reston, Virginia, USA and The Institute of Electrical Engineers, Stevenage, UK, 2004.
2. A. Soloviev, S. Gunawardena, F. V. Graas, Deeply Integrated GPS/Low-Cost IMU for Low CNR Signal Processing: Concept Description and In-Flight Demonstration, *NAVIGATION, Journal of the Institute of Navigation*, Vol. 54, No. 1, pp. 1–13, 2008.
3. E. Kaplan, and C. Hegarty (Editors), Understanding GPS: Principles and Applications, 2nd ed. Norwood Massachusetts, USA: Artech House, 2006.
4. A. Soloviev, D. Bates, and F. V. Graas, Tight Coupling of Laser Scanner and Inertial Measurements for a Fully Autonomous Relative Navigation Solution, *NAVIGATION, Journal of the Institute of Navigation*, Vol. 54, No. 3, pp. 189–205, Fall 2007.
5. G. T. Schmidt, R. E. Phillips, INS/GPS Integration Architectures, *NATO RTO Lecture Series*, Spring 2010.
6. A. Soloviev, D. Bruckner, F. V. Graas, and L. Marti, Assessment of GPS Signal Quality in Urban Environments Using Deeply Integrated GPS/IMU, *Proceedings of the Institute of Navigation National Technical Meeting*, San Diego, CA, January 2007.

7. A. Soloviev, Tight Coupling of GPS, Laser Scanner, and Inertial Measurements for Navigation in Urban Environments, *Proceedings of IEEE/ION Position Location and Navigation Symposium*, Monterrey, CA, May 2008.
8. M. Veth, and J. Raquet, Fusion of Low-Cost Imaging and Inertial Sensors for Navigation, *Proceedings on ION GNSS-2006*, Fort Worth, TX, September 2006.
9. J. Barnes, C. Rizos, M. Kanli, D. Small, G. Voigt, N. Gambale, J. Lamance, T. Nunan, C. Reid, Indoor Industrial Machine Guidance Using Locata: A Pilot Study at BlueScope Steel, *Proceedings of 2004 ION Annual Meeting*, San Diego, CA, June 2004.
10. G. Opshaug, and P. Enge, GPS and UWB for Indoor Positioning, *Proceedings of ION GPS-2001*, Salt Lake City, UT, September 2001.
11. M. Rabinowitz, and J. Spilker, The Rosum Television Positioning Technology, *Proceedings of 2003 ION Annual Meeting*, Albuquerque, NM, June 2003.
12. R. Eggert, and J. Raquet, Evaluating the Navigation Potential of the NTSC Analog Television Broadcast Signal, *Proceedings of ION GNSS-2004*, Long Beach, CA, September 2004.
13. T. D. Hall, P. Misra, Radiolocation Using AM Broadcast Signals: Positioning Performance, *Proceedings of ION GPS-2002*, Portland, OR, September 2002.
14. J. McEllroy, J. Raquet, and M. Temple, Use of a Software Radio to Evaluate Signals of Opportunity for Navigation, *Proceedings on ION GNSS-2006*, Fort Worth, TX, September 2006.
15. R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*, 3rd ed., John Wiley & Sons, Inc., New York, 1997.
16. J. L. Farrell, GPS/INS-Streamlined, *NAVIGATION, Journal of the Institute of Navigation*, Vol. 49, No. 4, pp. 171–182, Summer 2002.
17. F. V. Graas and A. Soloviev, Precise Velocity Estimation Using a Stand-Alone GPS Receiver, *NAVIGATION, Journal of the Institute of Navigation*, Vol. 51 No. 4, pp. 283–292, 2004.
18. D. Bates, Navigation Using Optical Tracking of Objects at Unknown Locations, M.S. Thesis, Ohio University, 2006.
19. D. Bates and F. V. Graas, Covariance Analysis Considering the Propagation of Laser Scanning Errors use in LADAR Navigation, *Proceedings of the Institute of Navigation Annual Meeting*, Cambridge, MA, April 2007.
20. S. Gunawardena, Development of a Transform-Domain Instrumentation Global Positioning System Receiver for Signal Quality and Anomalous event Monitoring, Ph.D. Dissertation, Ohio University, June 2007.
21. M. Miller, A. Soloviev, Navigation in GPS Denied Environments: Feature-Based Navigation Techniques, *NATO RTO Lecture Series*, Spring 2010.
22. A. Soloviev, N. Gans, M. Uijt de Haag, Integration of Video Camera with 2D Laser Scanner for 3D Navigation, *Proceedings of the Institute of Navigation International Technical Meeting*, Anaheim, CA, January 2009.

A Spectral Clustering Approach for Modeling Connectivity Patterns in Electroencephalogram Sensor Networks

Petros Xanthopoulos, Ashwin Arulselvan, and Panos M. Pardalos

Abstract Electroencephalography (EEG) is a non-invasive low cost monitoring exam that is used for the study of the brain in every hospital and research labs. Time series recorded from EEG sensors can be studied from the perspective of computational neuroscience and network theory to extract meaningful features of the brain. In this chapter we present a network clustering approach for studying synchronization phenomena as captured by cross-correlation in EEG recordings. We demonstrate the proposed clustering idea in simulated data and in EEG recordings from patients with epilepsy.

1 Introduction

Sensors are devices that measure a physical quantity and transform it into electrical measurement that can be processed by a computer. This very broad definition of sensors includes a vast number of applications in many fields of science and engineering. Electromagnetic sensors, acoustic sensors, movement sensors, light sensors to name a few play a very important role in modern technology and science.

Sensors could be represented as a graph or network with the sensors being the nodes and their interaction between them as edges. A telecommunication network is a common example of a sensor network, where the human voice is converted into electromagnetic pulses and transmitted till they reach the destination device. We

P. Xanthopoulos (✉) • P.M. Pardalos
Center for Applied Optimization, Department of Industrial and Systems Engineering,
University of Florida, Gainesville, FL 32611, USA
e-mail: petrosx@ufl.edu; pardalos@ufl.edu

A. Arulselvan
Technische Universität Berlin, Berlin, Germany
e-mail: arulsel@math.tu-berlin.de

V.L. Boginski et al. (eds.), *Sensors: Theory, Algorithms, and Applications*,
Springer Optimization and Its Applications 61, DOI 10.1007/978-0-387-88619-0_10,

present another kind of sensor network in this study, in which we utilize the time series similarity of sensors in a network. In this case, the connection between two sensors is defined by some time series similarity measure (linear or non-linear).

Graphs are the appropriate mathematical tools to represent and analyze sensor networks. In addition, much research has been carried over in the area of graph theory in mathematics and it is closely related to the mathematical theory of optimization which is well suited to model real life problems such as scheduling, Internet traffic, biology etc. A well-studied problem in the area of graph theory with several applications is the clustering problem, where we identify components or clusters of nodes hidden in a graph based on some well-defined objective function. A more formal definition is provided later. The study of clusters in graphs can provide some very useful insights. For instance, clustering in a call graph could reveal cliques of people who call each other [1] and in a market graph, clustering helps in detecting dependencies between stocks that are less conspicuous otherwise [2].

In this chapter, we will focus on the techniques available in spectral graph clustering and use them to visualize and interpret the information recorded in electroencephalogram (EEG) time series. Networks under investigation are constructed using cross-correlation. Cross-correlation is a well-studied fundamental time-invariant statistical metric for capturing linear connectivity between time series.

The chapter is organized as follows. In Sect. 2 we review the major min-cut formulations used in the spectral graph clustering theory and present the most commonly used algorithms. In Sect. 3 we discuss some of the most commonly used bivariate measures used in connectivity analysis of EEG. In Sect. 4 we illustrate the use of the algorithms in real data and we finally conclude with some discussion and possible future extensions of the current work.

2 Cut Formulations for Graph Clustering

2.1 Graph Preliminaries

In order to introduce the spectral graph clustering techniques we need to define some preliminary notation first. We define a simple undirected graph $G(N, E)$ as a node set, N with an edge set, E, which is set of unordered pair of distinct nodes from N. We define the weighted adjacency matrix W as

$$W = \begin{cases} w(i,j), & (i,j) \in E \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The scope of the chapter is restricted to symmetric adjacency matrices dealing with undirected graphs. If W is not symmetric, either because the graph is directed or because of numerical errors, we will be using $\overline{W} = \frac{W+W^T}{2}$ that possesses the desired property. We will denote the volume d_i of a node a_i as the sum of all the weights of the edges emanating from a_i. That is:

$$d_i = \mathrm{vol}(a_i) = \sum_{a_j \in N} w(i, j). \tag{2}$$

In matrix notation we can write:

$$d = [d_1 \; d_2 \; \cdots \; d_N] = \mathbf{1}^T W, \tag{3}$$

where **1** is the column vector that has all of its elements equal to 1. We will denote the volume of a set of edges S as the sum of the volumes of the nodes contained in the set:

$$\mathrm{vol}(S) = \sum_{a_i \in S} \mathrm{vol}(a_i). \tag{4}$$

Now we are ready to define the most important cut objective function used for spectral graph clustering.

2.2 Min Cut Formulations

A cutset, $cut(A, B)$, is the set of edges between the node sets $A \subseteq N$ and $B \subseteq N$, such that $A \cup B = N$ and $A \cap B = \emptyset$. The minimum cut problem consists in findings such a cut of minimum cost and more formally

$$\min \mathrm{cut}(A, B) = \min_{i,j} \sum_{a_i \in A, a_j \in B} w(i, j). \tag{5}$$

The minimum cut problem is a well-studied problem in the literature and despite the existence of exponential number of cuts there are algorithms that solve this problem efficiently in polynomial time [3].

The problem that arises if we adopt the minimum cut objective function is the lack of robustness to outlier data and the unbalanced cuts produced especially in large graphs. It is easy for one to see that edges with low weight that connect single nodes to the graph are most likely to belong to the optimal cut separating a single node from the rest of the graph. Therefore some outlier nodes and connections can influence substantially the quality of our solution. Also, addition of nodes in an existing network can dramatically change the optimal solution and the set of edges that belong to the cut. Therefore, one needs to use cuts that take into account not only the cost of the cut itself but also the size of the clusters. The first such objective function was proposed in [4–7]

$$R\text{cut}(A, B) = \frac{\text{cut(A, B)}}{|A|} + \frac{\text{cut(A, B)}}{|B|}, \tag{6}$$

where $|A|$ and $|B|$ the cardinality (number of nodes) of the node sets A and B respectively. In this formulation, cuts that produce equal sizes are favored. The problem was proved to be NP-Hard [8]. A heuristic solution for this problem was provided by solving the following relaxed problem.

$$Lq = \lambda q, \tag{7}$$

where $L = diag(\mathbf{1}W) - W$ is the Laplacian matrix of the graph $G(N, E)$. Here, q is relaxed over the set of real numbers as the NP-hardness of the problem was perceived to be a consequence of the restriction on q to take discrete values [8]. The solution to the relaxed problem is given by the second smallest eigenvalue of the Laplacian matrix (the smallest is equal to zero and corresponds to an eigenvector **1**). Now, we divide the nodes into two clusters in the following manner.

$$A = \{a_i | q_i < 0\}, B = \{a_i | q_i > 0\} \tag{8}$$

and the cutset is defined as

$$\text{MC} = \{(i, j) | a_i \in A, a_j \in B\} \tag{9}$$

We illustrate the cut algorithm in a randomly generated graph with 1,000 nodes that has two clusters by construction (strong intra-cluster connection and some weak inter-cluster connections). The adjacency matrix (without loss of generality all weights are equal to (1) is shown in Fig. 1).

We decompose the Laplacian of the above graph and sort the eigenvector that corresponds to the second lowest eigenvalue. The results are shown in Fig. 2.

In [8] Shi & Malik introduced the normalized cut that involves the volume of each cluster instead of the cardinality of the cluster.

$$N\text{cut}(A, B) = \frac{\text{cut}(A, B)}{\text{vol}A} + \frac{\text{cut}(A, B)}{\text{vol}(B)} = \left(\frac{1}{\text{vol}(A) + \text{vol}(B)}\right) \text{cut}(A, B) \tag{10}$$

The problem has been proved to be NP-Hard [8]. It can be proved that this problem is equivalent to minimizing a Rayleigh quotient subject to binary constraints, which makes the problem NP-hard.

$$\min N\text{Cut}(A, B) = \min R(L, y) = \min_y \frac{y^T(D - W)y}{y^T D y} \tag{11}$$

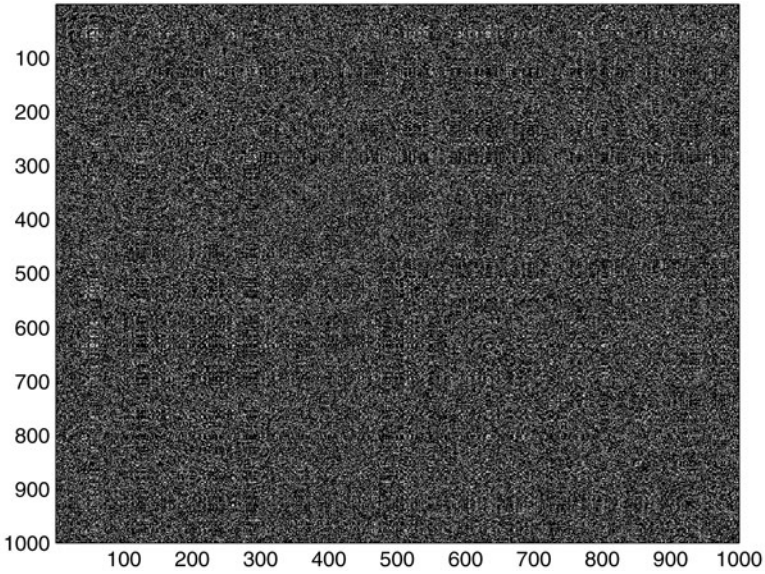


Fig. 1 The adjacency matrix of the 1,000 node example. The nodes are randomly permutated and one cannot distinguish some clear cluster structure. Black means that there is an edge and white that is missing

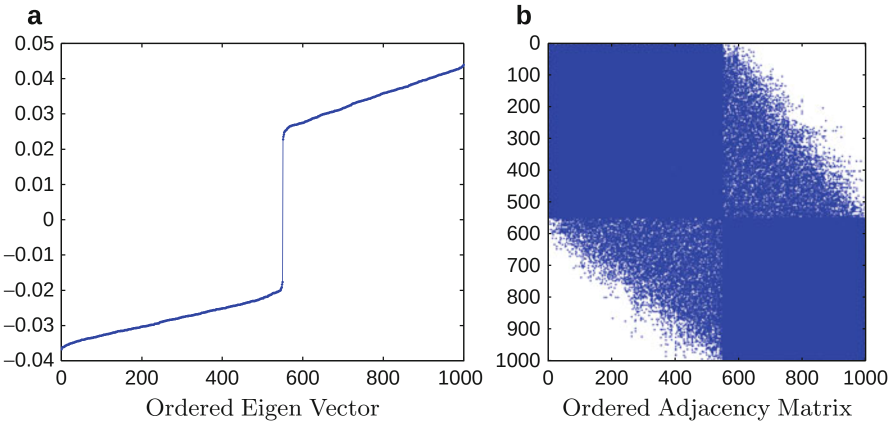


Fig. 2 This corresponds to the eigenvector of second lowest eigenvalue of the Laplacian matrix. It is easy to see the two cluster structure of the dataset. By simple thresholding we estimate the two classes

The most common approach in the image segmentation literature is to relax the y variable over the set of real numbers and then solve the corresponding eigenvalue problem. If $y \in R$ then the optimal solution is given from the solution of the following problem.

$$(D - L)y = \lambda D y \tag{12}$$

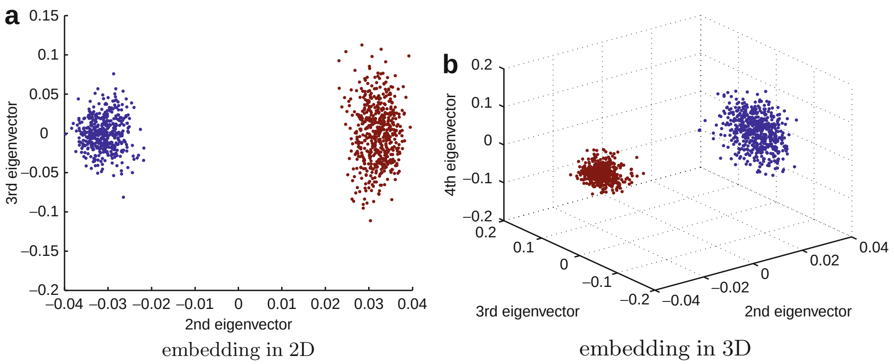


Fig. 3 Continuing from the previous example we compute, instead of just the second smallest eigenvector, the 3rd and the 4th also. In the eigenspace the cluster structure is very clear. Using clustering on the eigenspace we can cluster the data in k clusters (instead of just 2) assuming that it is meaningful for our problem

Although the second eigenvector is the most important one for the clustering, it is also common to compute the k first eigenvectors and then cluster the data in the eigenspace defined by the first k eigenvectors. Clustering in the eigenspace is used when we want to cluster the data in more than two clusters. The embedding of data in the eigenspace is visually demonstrated in Fig. 3.

In [9] Ng, Jordan, and Weiss alternatively use the cheeger constant as the minimization criterion for producing balanced cuts. Cheeger constant is similar to the Ncut formulation and defined by

$$\Phi(A,B) = \min_{A,B} \frac{\mathrm{cut}(A,B)}{min(\mathrm{vol}(A),\mathrm{vol}(B))} \tag{13}$$

Computing cheeger constant on a graph is again equivalent to minimizing, a slightly different, eigendecomposition problem subject to binary constraints. More specifically we are interested in computing the largest eigenvectors of

$$A = \sqrt{D} \cdot W \cdot \sqrt{D} \tag{14}$$

In [9], a clustering methodology was proposed using k-means on the eigenspace spanned by the first k eigenvectors of A.

An alternative approach for solving the eigenvalue problem subject to binary constraints is through semidefinite programing [10–12]. Semidefinite programing provides approximations of the exact solution with known bounds, but owing to their practical limitations they were not actively pursued by researchers and a majority of clustering algorithms were based on eigendecomposition.

3 Cross-correlation as a Bivariate Synchronization Measure

Data mining in time series data has always been a very challenging problem, where one tries to determine the similarities among the time series. There are extensive literatures on how to define these similarities. It has a great impact on a wide variety of field from financial time series to earthquake data synchronization. In this chapter, we will consider networks produced from electrophysiological recordings recorded from the human scalp (EEG).

Many metrics have been proposed over years in order to define and capture the notion of time series synchronization in EEG. These metrics range from simple statistical correlation, cross-correlation, spectral coherence [13,14], wavelet coherence, phase synchronization [15] to complex non-linear measures such as approximate entropy [16], non-linear Interdependence [17]. A comprehensive survey discussing a broad number of synchronization measures applied to EEG signals was provided in [18]. Synchronization between time series recorded between different electrode sites might be very important in evaluating effects of drugs prescribed for brain conditions or investigation of connections between the brain sites could prove useful in analyzing and predicting the brain activities. It is very difficult to see the global picture owing to the enormous amount of similarities between the edges that increases exponentially with the number of nodes despite the useful information we obtain from the bivariate measure between the two time series.

In this chapter, we suggest graph partitioning based on normalized cut optimization criteria to visualize and seek meaningful patterns in clusters. The dimensions of the graphs that arises in EEG graph partitioning is not significant from a computational perspective (instances usually are of the order 30 nodes), but it provides a useful tool for information visualization in the field of computational neuroscience.

Cross-correlation is a well-studied index used in capturing linear synchronization between time series. Cross-correlation C_{xy} between two time series $X = \{x_i\}_{i=1}^{N}$ and $Y = \{y_i\}_{i=1}^{N}$ is given by:

$$C_{xy}(\tau) = E[x_{n+\tau} y_n^*] = E[x_n y_{n-\tau}^*], \tag{15}$$

where $E[\cdot]$ is the expected value of some random variable. A simple observation of the cross-correlation function reveals its symmetric property (that is $C_{xy} = C_{yx}$) and hence this construction results in a graph with undirected edges and symmetric adjacency matrix. For computational purposes, we use the following formula to estimate the cross-correlation function:

$$\hat{C}_{xy}(\tau) = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} \frac{(x_i - \overline{x})}{\sigma_x} \frac{(y_{i+\tau} - \overline{y})}{\sigma_y} \tag{16}$$

where $\overline{x}$ and $\overline{y}$ are the mean values of X and Y, respectively. In order to construct our network, we considered only the maximum value of cross-correlation for each pair of electrodes. The advantage of cross-correlation is that it is a time-invariant measure. In other words, it can capture correlation even if the time series are not fully aligned.

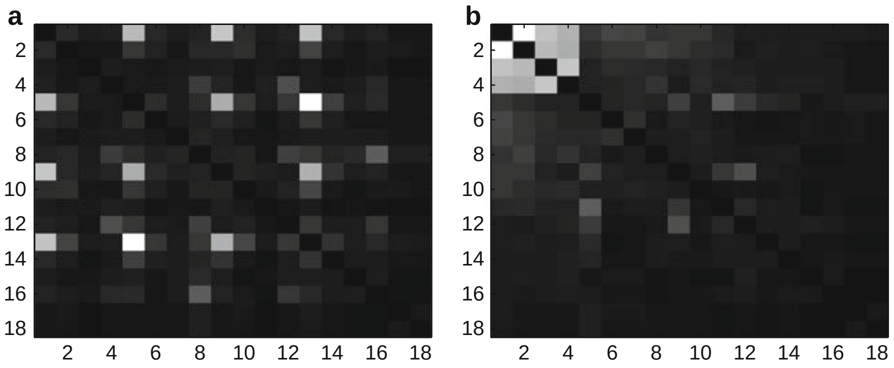
4 Computational Results

In order to demonstrate the spectral graph clustering in EEG recordings we used human EEG recordings from patients with absence epilepsy. Absence epilepsy is a form of non-convulsive epilepsy. Its main EEG characteristics are spike and wave discharges appearing in the frequency band of 3 Hz. The international 10–20 electrode placement system with 19 electrodes were used and the following 16 bipolar channels were chosen: 1:Fp1-F3, 2:F3-C3, 3:C3-P3, 4:P3-O1, 5:Fp2-F4, 6:F4-C4, 7:C4-P4, 8:P4-O2, 9:Fp1-F7, 10:F7-T3, 11:T3-T5, 12:T5-O1, 13:Fp2-F8, 14:F8-T4, 15:T4-T6, 16:T6- O2. Data points were collected at a sampling rate of 200 Hz for each channel.

We employed the algorithm proposed by Ng, Jordan, and Weiss [9] for clustering and we embedded the data in the two most important eigenvectors. Heatmap results can be seen in Fig. 4.

In the eigenspace one can have a clearer view of the clustered components. We used k-means with two classes to find the clusters (Fig. 5). The electrodes participating in the strongly connected cluster are 13,5,1,9.

These channels were plotted and we could observe a muscle artifact appearing in all four of them as shown in Fig. 6. Therefore as expected strong linear correlation among time series produces well-formed clusters. Detection of such clusters might



Cross correlation heatmap

Cross Correlation Clustered heatmap

Fig. 4 In (**a**) we can see the cross-correlation heatmap (autocross correlations were not computed). In (**b**) we can see the well-formed cluster between four first electrodes. Spectral clustering gives us some hint that there is something that we need to investigate in these electrode channels

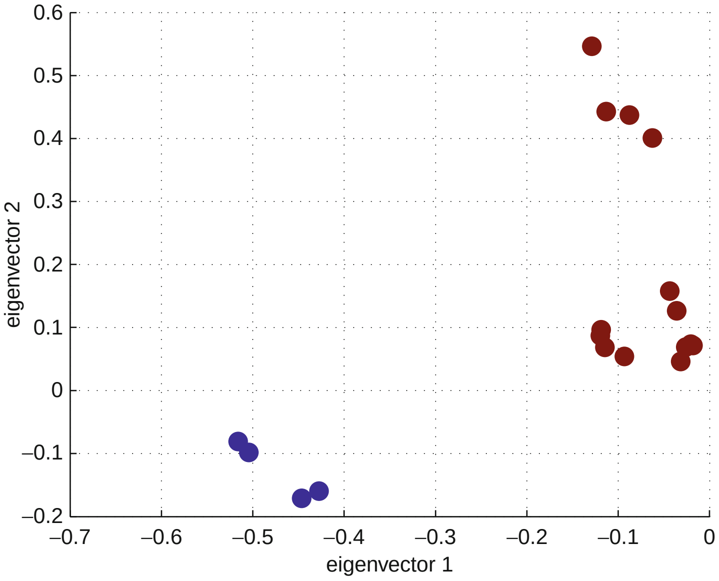


Fig. 5 If we embed the data into the space spanned by the two first eigenvectors it is easy to see the two formed clusters. Computationally it is easy to detect the two clusters by running a clustering algorithm such as k-mean

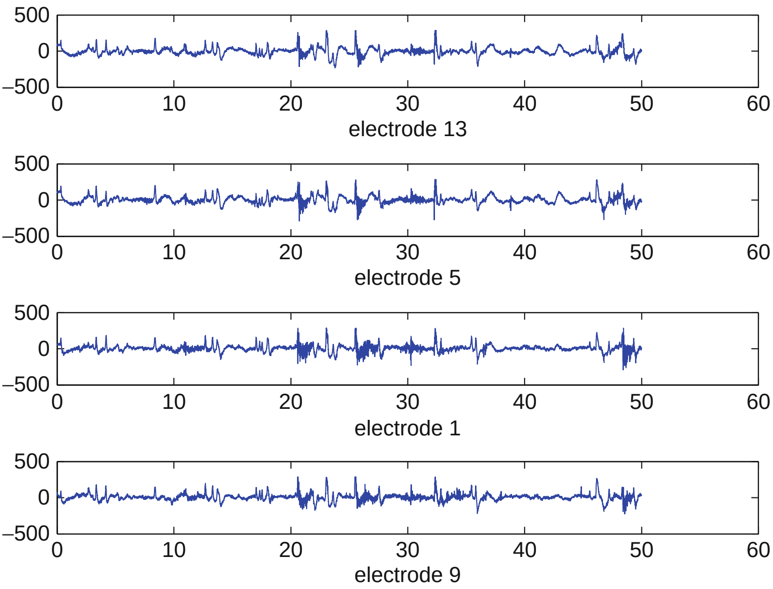


Fig. 6 By plotting the electrodes that form the four electrode cluster we find out that there is a strong muscle artifact appearing in all four channels

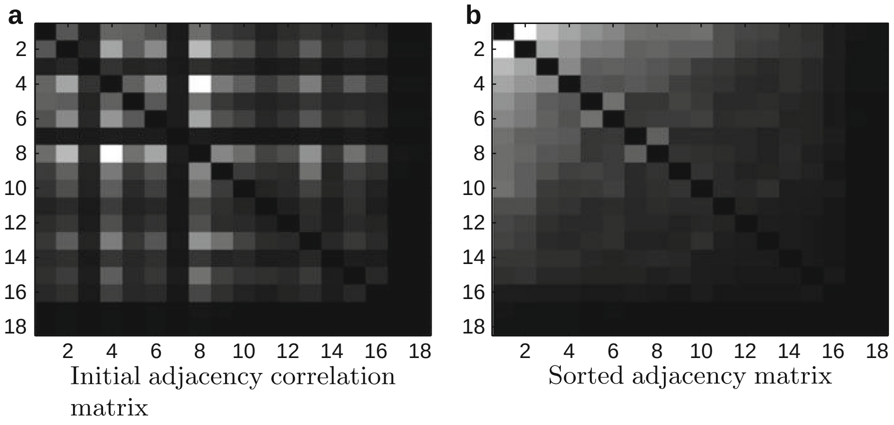


Fig. 7 Correlation adjacency matrix of EEG recordings with absence epilepsy. There is no clear cluster formation but there is clear connectivity degradation

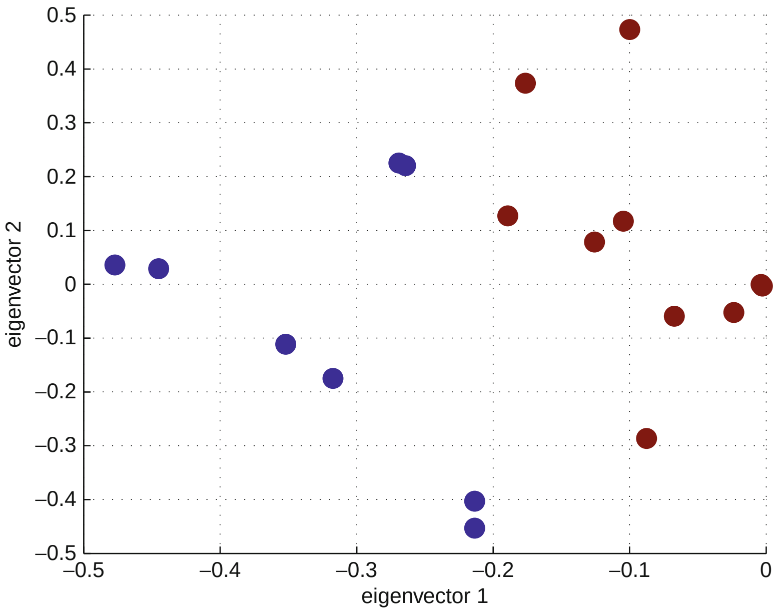


Fig. 8 Clustering of the EEG electrodes in the 2D eigenspace using k-means

be important in order to characterize the quality of EEG recording, e.g., we can use this clustering method in order to mark parts of EEG that carry strong artifacts and might be inappropriate for further computation analysis. In the second example, we illustrate the linear connectivity during an actual epileptic seizure. In absence epilepsy, a seizure is characterized by continuous 3 Hz spike and wave bursts that are generalized and can last up to ∼15 s (Fig. 7). The cross-correlation heatmap and the clustered by the algorithm heatmap can be shown in Fig. 8.

In these heatmaps, a strong cluster cannot be detected but we can notice an organization with respect to the linear connectivity that is not spent during normal EEG. In the eigenspace using k-means we can cluster these electrodes in two clusters (see Fig. 8).

5 Conclusion

In this chapter we used spectral clustering techniques to partition a sensor graph based on cross-correlation of EEG time series. From the examples illustrated in previous section we saw that spectral graph clustering can be useful to identity artifacts that appear in many channels of the recordings. It can also be useful in order to visualize the connectivity structure during some special neurological condition, such as an epileptic seizure.

Further investigation can be done in graphs defined by other bivariate similarity measures. Some of them, such as mutual information, mean square coherence, phase locking value, synchronization likelihood, non-linear interdependence measures, have been proven useful in characterizing synchronization effects in several pathological conditions (e.g., see [19, 20]).

In addition, some similarity measures, such as the one based on Lyapunov exponents (STLmax), are known for being able to predict non-state transition of the EEG and have been proposed as a basis on some seizure prediction algorithms [21, 22]. The proposed framework can be used in order to describe the global picture of the interactions in such a network and mine features based on individual clusters.

References

1. J. Abello, P.M. Pardalos, and M.G.C. Resende. On maximum clique problems in very large graphs. *American Mathematical Society*, 50:119–130, 1999.
2. V. Boginski, S. Butenko, and P.M. Pardalos. Mining market data: a network approach. *Comput. Oper. Res.*, 33(11):3171–3184, 2006.
3. R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Upper Saddle River/Un, 1993.
4. M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23:298–305, 1973.
5. L. Hagen and A.B. Kahng. New spectral methods for ratio cut partitioning and clustering. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions*, 11(9):1074–1085, 1992.
6. A. Pothen, H.D. Simon, and K.-P. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM. J. Matrix Anal. & Appl.*, 11:430–452, 1990.
7. M. Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, 25, pp. 619–633, 1975.

8. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.
9. A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, Boston, 2001.
10. J Falkner, F Rendl, and H Wolkowicz. A computational study of graph partitioning. *Math. Program.*, 66(2):211–239, 1994.
11. S.E. Karisch and F. Rendl. Semidefinite programming and graph equipartition. In *In Topics in Semidefinite and Interior-Point Methods*, pages 77–95. AMS, Providence Rhode island, 1998.
12. F. Rendl and H. Wolkowicz. A projection technique for partitioning the nodes of a graph. *Annals of Operations Research*, 58(3):155–179, May 1995.
13. S. Micheloyannis, V. Sakkalis, M. Vourkas, C.J. Stam, and P.G. Simos. Cortical networks involved in mathematical thinking: Evidence from linear and non-linear cortical synchronization of electrical activity. *Neuroscience Letters*, 373:212–217, 2005.
14. M. Ford, J. Goethe, and D. Dekker. Eeg coherence and power changes during a continuous movement task. *Int. J. Psychophysiol*, vol. 4:99110, 1986.
15. J.P. Lachaux, E. Rodriguez, J. Martinerie, and F. J. Varela. Measuring phase synchrony in brain signals. *Human Brain Mapping*, 8:194–208, 1999.
16. S.M. Pincus. Approximate entropy as a measure of system complexity. *Proc Natl Acad Sci*, 88:2297–2301, 1991.
17. Q. Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger. Performance of different synchronization measures in real data: a case study on electroencephalographic signals,. *Phys. Rev. E*, 65:041903, 2002.
18. E. Pereda, R. Q. Quiroga, and J. Bhattacharya. Nonlinear multivariate analysis of neurophysiological signals,. *Progress in Neurobiology*, 77:1–37, 2005.
19. J. Dauwels, F. Vialatte, and A. Cichocki. *Neural Information Processinge*, volume 4984 of *Lecture Notes in Computer Science*, chapter A Comparative Study of Synchrony Measures for the Early Detection of Alzheimers Disease Based on EEG. Springer, Berlin / Heidelberg, 2008.
20. V. Sakkalis, C.D. Giurcaneanu, P. Xanthopoulos, M. Zervakis, V. Tsiaras, Y. Yang, E. Karakonstantaki, and S. Michelyannis. Assessment of linear and nonlinear synchroniztion measures for analyzing eeg in a mild epileptic paradigm. *IEEE transaction on Information technology in biomedicine*, 2008. DOI:10.1109/TITB.2008.923141. 13(4):433-441, Jul 2009.
21. L.D. Iasemidis, D.-S. Shiau, P.M. Pardalos, W. Chaovalitwongse, K. Narayanana, A. Prasad, K. Tsakalis, P.R. Carney, and J.C. Sackellares. Long-term prospective on-line real-time seizure prediction. *Clinical Neurophysiology*, 116(3):532–544, 2005.
22. L.D. Iasemidis, D.-S. Shiau, W. Chaovalitwongse J.C. Sackellares, P.M. Pardalos, J.C. Principe, P.R. Carney, A. Prasad, B. Veeramani, and K. Tsakalis. Adaptive epileptic seizure prediction system. *IEEE Transactions on Biomedical Engineering*, 50(5):616–627, 2003.