AN APPROXIMATION SCHEME FOR A GEOMETRICAL NP-HARD PROBLEM

Joonmo Kim

ABSTRACT. In some wireless sensor networks, the sensor nodes are required to be located sparsely at designated positions over a wide area, introducing the problem of adding minimum number of relay nodes to interconnect the sensor nodes. The problem finds its abstract form in literature: the Minimum number of Steiner Points. Since it is known to be NP-hard, this paper proposes an approximation scheme to estimate the minimum number of relay nodes through the properties of the abstract form. Note that by reducing the number of nodes in a sensor network, the amount of data exchange over the network will be far decreased.

1. Introduction

Sensor networks are wireless networks in an ad hoc fashion. There are many sensor nodes in a sensor network, and a sensor node is autonomous device equipped with sensing, processing, and communication capabilities. Sensor nodes are spread over an area to gather information there and send it to a data sink, the user of the information. There are many different purposes of using sensor networks, and accordingly many ways of placing the sensor nodes over the area [1, 2]. In a military battle field, the sensor nodes would be dropped from an airplane and then they will form a self-configured network. However, deployment of redundant nodes may cause much more message exchanges than enough, resulting in as much resource consumptions including powers. Another kind of placing sensor nodes is the one for the temperature surveillance over a wide area for ecological purposes [3]. In this case, one may need to figure out the locations of some sensor nodes based on the geological information. This paper deals with the case that the sensor nodes are likely to be located sparsely over a wide area, and some sensors may have reasons to be located at specific locations.

AMS Subject Classification: 68R10, 90C35, 94C15
Key Words and Phrases: Sensor networks, interconnection, deployment, optimizations.
So, in the general case of measuring the temperatures over an ecological area, the sensor nodes are assumed to be located sparsely over a wide area and each of them can be put at any point. To make such a sparse placement of sensor nodes over a wide area, we should introduce relay nodes so that the sensor nodes should be interconnected when some sensor nodes are located beyond the transmission radius \(r\) of their near-by neighbors. Interconnection here means to have close enough distance for each node to more than one of its neighbors to be able to communicate with each other in the wireless network. We may regard that the relay nodes have the same hardware to the sensor nodes, but their function of sensing may be turned off. Assuming that the locations of the sensor nodes are predetermined by given conditions, the interconnection turns out to be the problem of getting the locations of the relay nodes among the sparsely scattered sensor nodes so that the minimum number of relay nodes should be deployed over the area. This interconnection problem finds its abstract form in literature, which is named as the Minimum number of Steiner Points (denoted STP-MSP). Since STP-MSP is NP-hard [4], challenges have been made to get the best approximation ratio to the optimal solutions. This paper proposes an approximation scheme that shows: one may build a polynomial-time algorithm that gives the approximation ratio 2 to the optimal number of relay nodes. The algorithm is not presented explicitly, instead, the scheme involves enough descriptions on how it can be implemented. With the adaptations to the practical conditions that may take place in the real-world sensor networks, the output from the proposed scheme will make the best ever layout toward the optimal deployment.

2. Definitions

The problem definition of Minimum number of Steiner Points from [6] is: Given \(n\) terminals (mathematical points) on the Euclidean plane \(\mathbb{R}^2\) and a positive constant \(r (\in \mathbb{R})\), find a Steiner Tree [5] interconnecting all the terminals with the minimum number of Steiner points [5] such that the Euclidean length of each edge is \(\leq r\). One may see that STP-MSP is analogous to the interconnection of sensor networks: the terminals are the sensor nodes, the Steiner points are the relay nodes, and \(r\) is the transmission radius. For the approximation, we are to build a scheme for the loose version of the problem. The condition of finding the ‘minimum number of Steiner points’ is taken off from STP-MSP and so the loose version is: Given \(n\) terminals on the Euclidean plane \(\mathbb{R}^2\) and a positive constant \(r\), find a Steiner tree that interconnects all terminals with Steiner points such that the Euclidean length of each edge is \(\leq r\). Then, we are to show that we may build a scheme for the loose version so that the algorithm from the scheme should produce feasible solutions such that the number of the Steiner points of every feasible solution is \(\leq\) twice the optimal number of Steiner points. For a problem in the loose version (i.e., a set of terminals on \(\mathbb{R}^2\)), one may find a set of Steiner points as a solution, and draw circles so that each Steiner point should be the center of each circle; let a Steiner-Cover be the resulting set of circles. There exist many Steiner-Covers for a problem instance. When the Steiner points of a Steiner-Cover are optimal, the Steiner-Cover
s called a Steiner-Cover. All circles in this paper have the constant radius \( r \). Let the bounding-box of a set of terminals on \( r^2 \) be the smallest rectangle enclosing the set. A rectangle is an axis-aligned one that is a partition of the bounding-box. The size of the rectangle is the length of its longer edge. A line-separator of a rectangle is a straight line segment which is parallel to the shorter edge of the rectangle: the line-separator partitions the rectangle into two of at least \( 1/3 \)rd of the area each.

**Definition 1.** (Tiling: dividing the given area) A tiling of a rectangle \( R \) is a binary tree (a hierarchy) of sub-rectangles of \( R \). The rectangle \( R \) is at the root. If the size of \( R \) is \( \leq 1 \) (unit distance), the hierarchy contains nothing else. Otherwise, the root contains a line-separator for \( R \), and has two sub-trees that are tilings of the two rectangles into which the line-separator divides \( R \).

![Sensor node](image_url)  
**Figure 1** Dividing given area

Note that rectangles at depth \( d \) in the tiling form a partition of the root rectangle. The set of all rectangles at depth \( d + 1 \) is a refinement of the partition obtained by putting a line-separator through each depth \( d \) rectangle of size \( > 1 \).

**Definition 2.** (portals) A portal in a tiling is any point that lies on the perimeters of rectangles in the tiling.

A set of portals \( P \) is called \( m \)-regular for the tiling when there are \( m \) (an integer) equidistant portals on the line-separator. Likewise, \( p \) (an integer) points that lie equidistantly on the perimeter of a circle are named indexed points.

**Definition 3.** (\( m \)-light Steiner-Cover) Let \( m \in \mathbb{Z}_+ \), \( S \) be a tiling of the bounding-box, and \( P \) be an \( m \)-regular set of portals on this tiling. Then, a Steiner-Cover in which each circle crosses with at least one portal in \( P \) at an indexed point is \( m \)-light with respect to \( S \).
3. Shorter radius

Given a set of terminals on $\mathbb{R}^2$, there are infinitely many candidate solutions of the problem. To form a feasible scheme we are to set up a frame (partitions with portals) over the set of terminals so that there may exist only polynomially many feasible solutions over the frame. Note that a solution is a Steiner-Cover for the given terminals. We are to show that there is a subset of the polynomially many feasible solutions such that each solution in it is a Steiner-$r$-Cover that has twice as many circles as a Steiner-Cover$_{opt}$ does; let $m$-light Steiner-Cover$_{opt}$ be a solution in the subset. A polynomial time Dynamic Programming (DP) can be designed to identify all the polynomially many feasible solutions over the frame, and the one with the minimum number of circles will be chosen as the desired solution. Since all the $m$-light Steiner-Cover$_{opt}$'s must also be checked by the DP, # (circles in the desired solution) \leq # (circles in $m$-light Steiner-Cover$_{opt}$), resulting in the ratio 2 to the optimal solutions. Proof is given on the following properties. There exist optimal solution trees of the STP-MSP problem with the properties from [6]: • No two edges cross each other. • Two edges meeting at a vertex form an angle of at least $60^\circ$. • If two edges form an angle $60^\circ$, then they have the same length.

![Diagram](image_url)

Figure 2 Shorter Radius by two Sensor nodes

Left side of Figure 2 Shorter Radius by two Sensor nodes shows the radius, the distance to nearby sensor nodes, of the central sensor node, where each of $r_a$, $r_b$, and $r_c$ are $\leq r$. The right side of Figure 2 Shorter Radius by two Sensor nodes shows that two replaced sensor no
An Approximation Scheme for a Geometrical NP-hard Problem

des linked by $r_0$ may interconnect the same terminals with the property that each $r_1, r_2, r_3, r_4, r_5$ and $r_6$ are $\leq r$. Following lemma shows this shortening of radii.

**Lemma.** (shortening radii)

![Diagram showing shortening the radius](image-url)

Figure 3  Shortening the radius

For a Sensor node in a Sensor Network, let the transmission radius $r_a$ be given. Replacing the central sensor node by two ones within the distance $r_c = r/2$, where the line of $r_c$ passes through the center. Then the triangular inequality implies: There exists $r_b$ such that $r_b \leq r_a - x$, where $0 \leq x \leq r/2$

$\square$

**Theorem** (Structure Theorem) For a set of terminals on $\mathbb{R}^2$, there are infinitely many Steiner $r$-Cover$_{2*r_{opt}}$'s that cover a Steiner-Cover$_{opt}$. The set also has an associated tiling of the bounding-box such that some Steiner-Cover$_{2*r_{opt}}$'s are m-light on this tiling. (See Figure 4  Two new nodes in distance $r/2$)

![Diagram showing two new nodes](image-url)

Figure 4  Two new nodes in distance $r/2$ of $r_c$

**Proof:** By Lemma, each circle $c_{opt}$ from an assumed Steiner-Cover$_{opt}$ can be covered by two other new sensor nodes: for each circle $c_{opt}$, one may interconnect the sensor nodes inside. Fr
om each $c_{opt}$ covered by two other new sensor nodes we may build up a $Steiner-Cover_{kin}$ that covers $Steiner-Cover_{opt}$. Let $pair_{kin}$ be $c_{opt}$ covered by two other new sensor nodes. By definition, there are infinitely many $Steiner-Cover_{kin}$ and all of them are actually $Steiner-Cover_{2*opt}$. One may build up a tiling over the bounding-box on $\mathbb{R}^2$ so that a circle formed by new sensor nodes can be passed through by a line-separator at least once. The tiling can be acquired whenever more than one bottom-most rectangles can be put inside a circle of radius $r$. No wider, one may set the value $w$ as the width between the inside circle and the outer one, and $w$ is wider than the inter-portal distance $d$ so that there should lie more than one portals along the part of the line-separator. Since there are infinitely many $pair_{kin}$, some of them should turn out to be $m$-light $pair_{kin}$. Therefore, there exist some $m$-light $Steiner-Cover_{2*opt}$'s on the tiling.

\[\square\]

![Figure 5: Infinitely many pair$_{kin}$ around the central sensor node](image)

A DP finds minimum $m$-light Steiner-Cover over the tiling in a polynomial time, where $\#$ (the circles in the minimum $m$-light Steiner-Cover) $\leq \#$ (the circles in $Steiner-Cover_{2*opt}$). The size of the bottom most rectangle of the tiling should be chosen to be small so that brute-force search can be made to identify all the $m$-light Steiner-Covers that cover the terminals in the rectangle. According to the tiling, two neighboring bottom most rectangles meet along a line, the line-separator, forming an upper-level rectangle. We may union two minimum $m$-light Steiner-Covers from each of the two bottom-most rectangles, forming $2^m$ upper-level $m$-light Steiner-Cover. For the union, firstly enumerate all the $2^m$ combinations of portals (chosen portals) out of $m$ portals along the line-separator. For each case of the combinations (chosen portals), join two minimum $m$-light Steiner-Covers from each of the two bottom level rect
angles respectively, where the Steiner-Covers should pass through the chosen portals. Each case of the combinations may have many cases of the joins. The one with the minimum number of circles among the joins is kept as the m-light Steiner-Cover for the case of the combinations. When joining, if duplicating circles come up from the two rectangles for an upper-level portal, one of them should be deleted. This process goes on for all the line-separators at the bottom-most level. Again up along the tiling, the upper-level rectangles also have neighboring rectangles at the same level, and the same process repeats. The process goes up until the top level, where the final minimum m-light Steiner-Cover is acquired. As mentioned, DP scans all the m-light Steiner-Cover and finds the minimum one in a polynomial time, which is described as follows. The time complexity, analogously by [7], is bounded by the number of rectangles in the tiling and other combinatorial factors that can be chosen to be bounded.

Now, we need to show the number of entries of the lookup table for this DP is polynomial and the run time for each of the entries is poly time. An entry is mixed by the triple: (a) A rectangle, (b) A set of k_i (≤ 4m) portals along the perimeter of the rectangle, and (c) The choice s of k_i circle positions, i.e., the permutation of size k_i out of the p indexed-points, {0, 1, 2, ..., p-1}.

For (a), the number of distinct rectangles is at most \( \binom{n}{4} \). For (b), each rectangle has 4 sides which are the parts of the line-separators of some upper-level rectangles. The m portals on the line separator are evenly distanced, so they are completely determined once we know the line-separators. But the number of choices of a line-separator is at most the number of pairs of points, which is \( \binom{n}{2} \). This accounts for the factor \( O((n^2)^s) \). Furthermore, once we have identified the set of \( \leq 4m \) portals on the four sides, the number of ways choosing a set of k_i portals is \( \binom{4m}{k_i} \). So the choices in (b) is \( n^s \times \sum_{i=1}^{c} \binom{4m}{k_i} \), where \( c = 4 \cdot \frac{8L}{r} \), the maximum number of crossings. For (c), for each portal chosen above, there are p choices of a circle shape s. Hence we ca upper bound the size of the lookup table by \( n^s \times n^p \times \sum_{i=1}^{c} \binom{4m}{k_i} p^s = O(\sqrt{2p}^{m}) = n^{O(1)} \), where m should be chosen as \( m = O(\log n) \) to for m a polynomial expression, c is a constant as \( c = 4 \cdot \frac{8L}{r} \), p and r are mentioned before, and L is the size of the bounding-box.

4. Conclusion

The high-degree polynomial time of DP can be accommodated since the computation for a solution will be a batch process before the deployment of sensor nodes. On the other hand, to reduce the computation time, one may divide the computation into parallel ones or design a randomized algorithm. The computation by the proposed scheme may produce the layout o
ver which twice of the minimum number of relay nodes can be deployed at worst, far decrea
sing the number of latent messages that might have been carried over the sensor network wit
h many more relay nodes.

Acknowledgements

The present research was conducted by the research fund of Dankook University in 2005.

References

ance in Distributed Sensor Networks, Proc. IEEE Wireless Communications and Networkin
[4] G.-H. Lin and G.L. Xue, Steiner tree problem with minimum number of Steiner points a
ter trees with minimum number of Steiner points, Theoretical Computer Science, vol.262, pp

School of Electronics and Computer Engineering
Dankook University
q888@dankook.ac.kr
ARARO: Aggregate Router-Assisted Route Optimization for Mobile Network Support

Kyung-Taeg Rho and Soo-Mok Jung

ABSTRACT
Network Mobility basic support protocol (NEMO Basic) extends the operation of Mobile IPv6 to provide uninterrupted Internet connectivity to the communicating nodes of mobile networks. The protocol uses a mobile router (MR) in the mobile network to perform prefix scope binding updates with its home agent (HA) to establish a bi-directional tunnel between the HA and MR. This solution reduces location-update signaling by making network movements transparent to the mobile nodes (MNs) behind the MR. However, delays in data delivery and higher overheads are likely to occur because of sub-optimal routing and multiple encapsulation of data packets. To manage the mobility of the mobile network, it is important to minimize packet overhead, to optimize routing, and to reduce the volume of handoff signals over the nested mobile network. This paper proposes an aggregate routers-assisted route optimization (ARARO) scheme for nested mobile networks support which introduces a local anchor router in order to localize handoff and to optimize routing. With ARARO, a mobile network node (MNN) behind a MR performs route optimization with a correspondent node (CN) as the MR sends a binding update message (BU) to aggregate router (AGR) via root-MR on behalf of all active MNNs when the mobile network moves. This paper describes the new architecture and mechanisms and provides simulation results which indicate that our proposal reduces transmission delay, handoff latency and signaling overhead. To evaluate the scheme, we present the results of simulation.

Keywords: handoff, nested mobile network, route optimization, ARARO, transmission delay

1. Introduction

The vast address space of IPv6 will enable mobile objects such as cars, buses, trains, airplanes, or ships to carry an IPv6 network, in which many kinds of information devices can act as an IPv6 host having the IPv6 address. If at least one of the routers in the IPv6 network connects to a router on the Internet, any host in the IPv6 network can communicate with any host on the Internet.

The Network Mobility (NEMO) working group [1] of the Internet Engineering Task Force (IETF) has developed a conceptual architecture of a mobile network and a NEMO basic support protocol (NEMO Basic) [2] by extending the operation of the Mobile IPv6 (MIPv6) protocol [3].

A mobile network, which is composed of one or more IP subnets, moves as a single unit in the Internet topology [4]. It uses a MR as a gateway to provide Internet connectivity via an access router (AR) to the mobile network nodes (MNNs). The MNNs are categorized into three groups: local fixed nodes (LFNs),
local mobile nodes (LMNs), and visiting mobile nodes (VMNs). The home addresses (HoAs) of the LFNs and LMNs are associated with the mobile network’s IP subnet prefix, whereas the HoAs of the VMNs are associated with other networks. Therefore, a VMN arriving at a mobile network first configures a care-of address (CoA) from the mobile network’s IP prefix and then registers the CoA with its HA. A VMN may represent a single host or a network itself, such as a PAN (personal area network), resulting in nested mobility. In the case of nesting, MRs belonging to each mobile network form a hierarchy, with the (upper) parent-MR providing connectivity to (lower) sub-MRs.

In NEMO Basic, MRs associate their CoAs with their network prefix in the BUs sent to their HA. This provides connectivity of the MR and consequently to each node in its embedded network. A major drawback of NEMO is that all communications to and from the mobile network must go through the MR-HA tunnel. This results in extra overhead and high delays. Moreover, with nested mobile networks, the problem increases with each nested level. Outbound packets must go through the HAs of all MRs of higher levels before reaching their destination. This is known as the "pinball problem" and causes high delay. Another major drawback of these IP-in-IP encapsulations is related to overhead. Indeed, each nested level introduces additional overhead which in turn increases the network load and the risk of congestion.

To deploy this mobile IP service widely, the Hierarchical Mobile IPv6 (HMIPv6) technology [5] is also being studied in IETF. By adding a mobility anchor point (MAP) in a visited network to manage local mobility there, we can limit HAs to providing only global or inter-MAP mobility management. This technology lets us to avoid frequent locational registration of MNs with HAs, which may be a long way from the MNs, and to reduce the time required for handovers.

In response to overcome some of aforementioned drawbacks, our proposed scheme reduces the number of control messages and the handoff latency as it enables a aggregate router (AGR) to just update the binding information for a MR, using a BU from the MR instead of the MNs behind the MR when the MR moves locally within the AGR domain. This scheme also enables packets to be optimally routed to MNs in the mobile network via AGR. The remainder of this paper is organized as follows: Section 2 summarizes the related literature. Section 3 describes our proposed scheme on supporting network mobility. A performance evaluation of the proposed architecture and mechanisms is described in Section 4. Finally, in Section 5, we present some concluding remarks.

2. Related Work

Route optimization is a mechanism that not only shortens the data delivery path between a MN and CN, but also reduces the potential level of encapsulation. Nevertheless, route optimization requires some route update signaling and/or additional information in the IP headers of data packets to enable packets to follow the optimal path and reach their destination intact. The generic consideration in designing a route
optimization scheme is to use a minimum of signaling and/or additional information in the packet header. Some solutions to the problems of route optimization in NEMO have been published [6] [7]. Thubert et al. proposed the use of a new routing header, routing header (RH) type 4, also called a reverse routing header (RRH), for MNN-originated outbound packets, and a modified RH type 2 for inbound packets destined for MNN [6]. The RH type 4 collects the CoAs of all nested MRs, which are later included in the modified RH type 2 to reduce the number of nested encapsulations for inbound packets. This scheme, however, optimizes the path between the HA and the MR serving the MNN, not between the CN and MR. Moreover, it requires MRs to modify packet headers, which would increase computational overheads.

Perera et al. [7] proposed assigning the MR a prefix pertaining to the visited network and advertising the prefix in a router advertisement to the MNNs. MIPv6-enabled MNNs first configure their CoAs from the prefix using a stateless address auto-configuration, and then perform route optimization by sending BU to their HAs and CNs. As this scheme requires MNNs to configure their CoAs every time the network moves, it could possibly cause a binding explosion problem [8].

3. Proposed Route Optimization Scheme

We describe the operations of the ARARO scheme to provide the route optimization for a nested mobile network which has multiple levels of MRs as shown in Fig. 1. In this scheme, all MRs keep a binding cache, which we call MR-binding cache (MR-BC) for all the nested mobile routers behind them. Additionally, the AOR keeps another binding cache (AOR-BC) for all active MNNs that have ongoing communication sessions with CNs.

![Nested Mobile Network Architecture for proposed routing method](image-url)

The MR-BC is used to store binding between the local CoAs (LCoAs) of nested MRs and their mobile network prefixes (MNNs). As shown in Fig. 1, when the sub-MR (MR3) attached to a parent-MR (MR2), MR3 sends a routing update message (RU) to MR2 to register the MR3's LCoA (MR3-LCoA) and the
MR3's MNP (MNP3). MR2 use this information to forward a packet addressed to the MNN1-LCoA. Similarly, MR2 registers its LCoA and all mobile network prefixes (MNP2, MNP3) with MR1 by sending RU (MR2-LCoA, MNP2, MNP3). In other words, if a MR (MR3) detects the movement of its mobile network, it sends a RU to the parent-MR (MR2), containing the MR3's LCoA and MR3's MNP. Then, the parent-MR (MR2) updates its binding cache entry and resend the RU to its parent-MR (MR1) recursively. At this time, the RU from the parent-MR (MR2) contains MR2-LCoA and MR2-MNP (MNP2), and sub-MR's MNP (MR3-MNP), with MR2-LCoA being used as next hop address for MR1 to send packets destined to MR2 or MR3. In conclusion, if a RU reaches the root-MR, the routing update procedure is completed.

To localize handoff signals, reduce handoff latency time and support route optimization in this scheme, the AGR maintains binding information for all MNNs and MRs in its domain in its binding cache (AGR-BC). AGR-BC is composed of HoA, LCoA, MNP, MR/MNN, and parent-MR field as shown in Fig. 1. Our proposed scheme applies the hierarchical location management method to a mobile network and manages the location of the mobile network and MNNs within it hierarchically. AGR1 periodically sends its address to ARs, which are connected as subordinates of AGR1. When root-MR (MR1) connects to AR2, MR1 sends a router solicitation message (RS) containing a request for AGR address and MNP. Next, AR2 sends a router advertisement message (RA) to MR1, containing AGR1 address and AR2-prefix. Then, MR1 creates its LCoA using AR2-prefix, sets AGR1 address as its regional CoA (RCoA). MR1 then sends a BU containing MR1-HoA, MR1-LCoA, MR1-MNP (MNP1), 1 (meaning a MR) and its parent MR (AR2) to AGR. AGR1 register the information to bind the MR1-LCoA and MR1-HoA, MR1-MNP 1, and AR2. After this, MR1 sends a BU to its HA (MR1-HA) to register AGR1 address as its RCoA. Next, MR1 sends a RA to MNNs in its mobile network, containing AGR1 address and its MNP (MNP1).

Each MNN behind MR1 also detects new AGR address in the case where MR1 is already connected to AR2. The MNN (MNN2) creates MNN2-LCoA and its RCoA (AGR1 address) from the RA received from MR1. MNN2 sends a BU with MNN2-HoA, MNN2-LCoA, 0 (Not a MR) and its parent MR (MR1-HoA) to AGR1, and a BU containing the AGR1 address and MNN2-HoA to its HA (MNN2-HA). At this time, AGR1 caches the relation information to bind MNN2-HoA, MNN2-LCoA, 0 (Not a MR), and MR1-HoA. Other sub MRs and other MNNs in Fig. 1 could be applied the same procedures as above-mentioned method.

Fig. 2 shows the sequence of route optimization from CN1 to MNN1. MNN1 sends a BU to CN1 to register binding information that MNN1-RCoA is AGR1 address. On arrival of the BU, the CN1 creates a binding between the MNN1-HoA and AGR1 address mentioned in the alternate CoA option field. In this situation, CN1 can send packets destined to MNN1 via AGR1 using RH type2. RH type 2 is an extension header defined in IPv6. AGR1 checks the home address option (HAO) field of the RH type 2 header of an inbound packet to get the HoA of the MNN1 that the packet is addressed to. The MNN1's HoA
(MNNI-HoA) is used to search for the corresponding LCoA (MNNI-LCoA) in the AORI-BC. AORI encapsulates this packet with MNNI-LCoA after searching for it in its binding cache and transmits the packet to MNNI.

Fig. 2 Procedures – Optimization of the route from a CN in proposed ARARO scheme

Similarly, MNN-oriented outbound packets have the AOR address and the address of the CN in the source and destination address fields, respectively. These packets are tunneled to the AOR using the AGR address as the destination address and the MNN-LCoA as the source address in the outer IP header. The AGR decapsulates and forwards the packet normally to the CN.

Fig. 3 Procedures – Mobile Networks Handoff

Fig. 3 shows the procedure of mobile networks handoff within the same AGR1 domain where MR1 performs the handoff to a new link and configures a new LCoA. MR1 then sends a BU just to AGR1 containing its new LCoA and parent MR’s address which is AR3. Then, AGR1 updates the LCoA and parent MR’s address for MR1. The need to send BUs from each MNN within mobile networks managed by MR1 to AGR1 is eliminated like NEMO Basic. That means our scheme avoids increasing signaling.
volume due to handoff management. However, if the AGR address is changed, it is necessary to updates
the AGR addresses of MNNs and sub-MRs within MR1’s nested mobile network.

4. Performance Evaluation

![Network Model for Simulation](image)

Fig. 4 Network Model for Simulation

In this section, we present some comparative performance results with the NEMO Basic. We consider a
three-level nested mobile network as depicted in Fig. 4. HA1 to HA4 are respectively HAs of MR1 to
MR4. The remaining HAs is the HAs of MNNs. We set up the CN as a traffic source at a Constant Bit
Rate (CBR) over a User Datagram Protocol (UDP), producing fixed length packets of 1500 bytes every
10 ms. Then the MNN acts as a sink node receiving packets from CN. The setup link topology consists of
wired link and wireless link. The wired link is fixed and used at the connection of CN to AGR, CN to HA,
HA to AGR, and AGR to the AR. The wired link bandwidth is set to 100 Mbps. The wireless link
bandwidth is set to 11 Mbps with the wireless link latency set to 2 ms. The packet service rate was 100
packets/second corresponding to data rates of 1.2 Mbps. The handoff interval was set to 2 seconds. We
evaluated each scheme assuming 5, 10, 100 MNNs in the mobile network. The simulation assumes that
delay between HA and HA is 100 ms, delays between CN and HA, CN and AGR and HA and AGR are
the same, 50 ms and delay between MMR and AR is 50 ms. Furthermore, packet header size, BU size and
binding acknowledgement message (BACK) size are also predefined: 40 bytes, 112 bytes and 56 bytes
respectively.

4.1 End-to-End Packet Delay

One of the main advantages of using route optimization techniques lie in delay reduction. Packet
transmission delay measurements from a CN to a MN in the mobile network are depicted in Fig. 5. This is
related to the reduction of the number of nested tunnels. Indeed, the proposed solution requires only a
unique tunnel from AGR to MN regardless of the number of nested levels in the mobile network. The
packets, in NEMO Basic, must pass through multiple tunnels from the MN to MN-HA. The packet transmission delay saving time between ARARO and RRH method is 56.33 ms at level 0 and 56.58 ms at level 3. RRH method is superior to NEMO Basic but is inferior to ARARO.

Fig. 5 Packet Delivery Delay

4.2 Handoff Latency

Handoff latency is the mean time from handoff initiation to completion. Fig. 6 shows the handoff latency for each scheme when a mobile network is assumed to move locally at nesting level 2 within AGR domain. NEMO Basic requires that MR only performs a registration operation to its HA/CNs instead of MNNs connected to the MR. However, ARARO requires that MR also performs a registration operation to AGR instead of MNNs behind the MR.

Fig. 6 Handoff Latency

As we can see, the handoff delay with the proposed approach is decreased to 551 ms regardless of the number of MNNs in the mobile network, compared to NEMO Basic. On the other hand, the difference of ARARO and RRH depends on the number of MNNs, e.g., 659 ms with 5 MNNs, 13903 ms with 100 MNNs since RRH require each MN in the mobile network to send BUs to its CNs/HA when handoff occurs. As the number of MNs in the mobile network is increased, the handoff latency of RRH increases enormously. Therefore, when compared to other schemes, ARARO is the most efficient.
4.3 Handoff Signals

The handoff signals are - RS, RA, BU and BACK [3]. Fig. 7 shows the number of handoff signals required in each scheme whenever handoff occurs. With regard to signaling overheads, both NEMO Basic and ARARO offer low and constant values. On the other hand, in RRH, increasing the numbers of MNNs increases the number of handoff signals. If the number of MNNs is 100, ARARO provides about the same level of performance as NEMO Basic, while it requires about 500 fewer handoff signals than RRH.

[Diagram showing the number of handoff signals against the number of MNNs]

Fig. 7 Number of Handoff Signal

5. Conclusions

The NEMO Basic provides advantages by reducing location update overheads. However, it has the side effect of increasing packet delivery overheads due to pinball routing and multi-layer encapsulation of data packets. To solve this problem, this paper has a new mobility management mechanism for optimizing the end-to-end route for MNNs/MRs within a nested mobile network environment. By developing AGR to provide a mobility anchor point (MAP) function in HMIPv6 and slightly changing in the implementation of the NEMO Basic in the local components of a mobile network such as MRs and MNNs, the proposed approach could provide more effective route optimization that would reduce the burden of location registration for handoffs. The ARARO enables a CN to forward packets directly to the mobile network without any tunneling, which reduces packet delays and encapsulation overhead in the core network. It also reduces handoff latency and the volume of handoff signals. Our future subjects of study include investigating security issues between MRs and distributing the processing load by locating multiple MRs in a mobile network while retaining most of the predicted benefits.
References

APPLICATIONS OF GRAPH THEORY

S. Pirzada and Ashay Dharwadker

Abstract

Graph theory is becoming increasingly significant as it is applied to other areas of mathematics, science and technology. It is being actively used in fields as varied as biochemistry (genomics), electrical engineering (communication networks and coding theory), computer science (algorithms and computation) and operations research (scheduling). The powerful combinatorial methods found in graph theory have also been used to prove fundamental results in other areas of pure mathematics. This paper, besides giving a general outlook of these facts, includes new graph theoretical proofs of Fermat’s Little Theorem and the Nielson-Schreier Theorem. New applications to DNA sequencing (the SNP assembly problem) and computer network security (worm propagation) using minimum vertex covers in graphs are discussed. We also show how to apply edge coloring and matching in graphs for scheduling (the timetabling problem) and vertex coloring in graphs for map coloring and the assignment of frequencies in GSM mobile phone networks. Finally, we revisit the classical problem of finding re-entrant knight’s tours on a chessboard using Hamiltonian circuits in graphs.

INTRODUCTION

Graph theory is rapidly moving into the mainstream of mathematics mainly because of its applications in diverse fields which include biochemistry, electrical engineering (communications networks and coding theory), computer science (algorithms and computations) and operations research (scheduling). The wide scope of these and other applications has been well-documented cf. [5] [19]. The powerful combinatorial methods found in graph theory have also been used to prove significant and well-known results in a variety of areas in mathematics itself. The best known of these methods are related to a part of graph theory called
matchings, and the results from this area are used to prove Dilworth's chain decomposition theorem for finite partially ordered sets. An application of matching in graph theory shows that there is a common set of left and right coset representatives of a subgroup in a finite group. This result played an important role in Dharwadker's 2000 proof of the four-color theorem [8] [18]. The existence of matchings in certain infinite bipartite graphs played an important role in Laczkovich's affirmative answer to Tarski's 1925 problem of whether a circle is piecewise congruent to a square. The proof of the existence of a subset of the real numbers \( \mathbb{R} \) that is non-measurable in the Lebesgue sense is due to Thomas [21].

Surprisingly, this theorem can be proved using only discrete mathematics (bipartite graphs). There are many such examples of applications of graph theory to other parts of mathematics, but they remain scattered in the literature [3] [16]. In this paper, we present a few selected applications of graph theory to other parts of mathematics and to various other fields in general.

1. **THE CANTOR-SCHRÖDER-BERNSTEIN THEOREM**

Here we discuss the graph theoretical proof of the classical result of Schröder and Bernstein. This theorem was presumed to be an obvious fact by Cantor (cf. remark 1.2) and later proved independently by Schröder (1896) and Bernstein (1905). The proof given here can be found in [14] and is attributed to König.

1.1. **Theorem (Cantor-Schröder-Bernstein).** For the sets \( A \) and \( B \), if there is an injective mapping \( f: A \rightarrow B \) and an injective mapping \( g: B \rightarrow A \), then there is a bijection from \( A \) onto \( B \), that is, \( A \) and \( B \) have the same cardinality.

**Proof.** Without loss of generality, assume \( A \) and \( B \) to be disjoint. Define a bipartite graph \( G = (A, B, E) \), where \( xy \in E \) if and only if either \( f(x) = y \) or \( g(y) = x \), \( x \in A, y \in B \). By the hypothesis, \( 1 \leq d(v) \leq 2 \) for each \( v \) of \( G \). Therefore, each component
of $G$ is either a one-way infinite path (that is, a path of the form $x_0, x_1, \ldots, x_n, \ldots$), or a two-way infinite path (of the form $\ldots, x_{-n}, x_{-n+1}, \ldots, x_{-1}, x_0, x_1, \ldots, x_n, \ldots$), or a cycle of even length with more than two vertices, or an edge. Note that a finite path of length greater or equal to two cannot be a compound of $G$. Thus, in each component there is a set of edges such that each vertex in the component is incident with precisely one of these edges. Hence, in each component, the subset of vertices from $A$ is of the same cardinality as the subset of vertices from $B$. ■

1.2. Remark. Cantor inferred the result as a corollary of the well-ordering principle. The above argument shows that the result can be proved without using the axiom of choice.

2. Fermat's (Little) Theorem

There are many proofs of Fermat’s Little Theorem. The first known proof was communicated by Euler in his letter of March 6, 1742 to Goldbach. The idea of the graph theoretic proof given below can be found in [12] where this method, together with some number theoretic results, was used to prove Euler’s generalization to non-prime modulus.

2.1. Theorem (Fermat). Let $a$ be an integer and let $p$ be a prime such that $a$ is not divisible by $p$. Then, $a^p - a$ is divisible by $p$.

Proof. Consider the graph $G = (V, E)$, where the vertex set $V$ is the set of all sequences $(a_1, a_2, \ldots, a_p)$ of natural numbers between 1 and $a$ (inclusive), with $a_i \neq a_j$ for some $i \neq j$. Clearly, $V$ has $a^p - a$ elements. Let $u = (u_1, u_2, \ldots, u_p)$, $v = (u_{p+1}, u_1, \ldots, u_{p-1}) \in V$. Then, we say $uv \in E$. With this assumption, each vertex
of $G$ is of degree 2. So, each component of $G$ is a cycle of length $p$. Therefore, the number of components is $\frac{a^p - a}{p}$. That is, $p | a^p - a$. ■

3. The Nielsen-Schreier Theorem

Let $H$ be a group and $S$ be a set of generators of $H$. The product of generators and their inverses which equals identity (1) is called a trivial relation among the generators in $S$ if 1 can be obtained from that product by repeatedly replacing $xx^{-1}$ or $x^{-1}x$ by 1. Otherwise such a product is called a non-trivial relation. A group $H$ is free if $H$ has a set of generators such that all relations among the generators are trivial.

Babai [2] proved the Nielsen-Schreier Theorem for subgroups of free groups, as well as other results in diverse areas, from his Contraction Lemma. The particular case of this lemma when $G$ is a tree, and its use in proving the Nielsen-Schreier Theorem, was also observed by Serre [20].

3.1. Contraction Lemma. Let $H$ be a semi-regular subgroup of the automorphism group of a connected graph $G$. Then, $G$ is contractible onto some Cayley graph of $H$. The proof of this lemma is technical, although it only uses ideas from group theory and graph theory.

Let $H$ be a group and $h \in H$. Let $h_R$ be a permutation of $H$ obtained by multiplying all the elements of $H$ on the right by $h$. The collection $H_R = \{h_R, h \in H\}$ is a regular group of permutations (under composition) and is called the (right) regular permutation representation of $H$.

It can be seen [2] that $G$ is a Cayley graph of the group $H$ if and only if $G$ is connected and $H_R$ is a subgroup of the automorphism group of $G$. 
The automorphism group of a graph $G$ is the group of all permutations $p$ of the vertices of $G$ with the property that $p(x)p(y)$ is an edge of $G$ if and only if $xy$ is an edge of $G$.

A group $H$ of permutations acting on a set $V$ is called semi-regular if for each $x \in V$, the stabilizer $H_x = \{ h \in H : x^h = x \}$ consists of the identity only, where $x^h$ denotes the image of $x$ under $h$. If $H$ is transitive and semi-regular, then it is regular.

Let $(H, \circ)$ be a group and $S$ be a set of generators of $H$, not necessarily minimal.

The Cayley graph $G(H, S)$ of $(H, \circ)$ with respect to $S$, has vertices $x, y, \ldots \in H$, and $xy$ is an edge if and only if either $x = yoa$ or $y = xoa$, for some $a \in S$.

If $G$ is any graph and $e = xy$ an edge of $G$, then by contraction along $e$, we mean the graph $G'$ obtained by identifying the vertices $x$ and $y$.

We say that a graph $G_1$ is contractible onto a graph $G_2$ if there is a sequence of contractions along edges which transforms $G_1$ to $G_2$.

3.2. Corollary. If $J$ is a subgroup of a group $H$, then any $G(H, S)$ is contractible onto $G(J, T)$ for some set $T$ of generators of $J$.

3.3. Theorem (Nielsen-Schreier). Any subgroup of a free group is free.

Proof. We first show that in any group $H$ and for any set $S$ of generators of $H$, the Cayley graph $G(H, S)$ contains a cycle of length $> 2$ if and only if there is a nontrivial relation among the generators in $S$. To show this, suppose $x_0, x_1, \ldots, x_n = x_0$ is a cycle of $G(H, S)$. Then, there are $a_i \in S$, $1 \leq i \leq n$, such that $x_{i-1}a_i^{e_i} = x_i$, where $e_i \in \{1, -1\}$. Hence, $x_n = x_{n-1}a_n^{e_n} = x_{n-2}a_n^{e_n}a_{n-1}^{e_{n-1}} = \ldots = x_0a_1^{e_1}a_2^{e_2} \ldots a_n^{e_n}$, that is, the identity
1 = a_1^{e_1} a_2^{e_2} \ldots a_n^{e_n}. If this were a trivial relation, then there would exist an integer \( i \), \( 1 \leq i \leq n \), such that \( a_i = a_{i+1} \) and \( e_i = -e_{i+1} \). However, this implies that \( x_{i+1} = x_{i+1} \), a contradiction. Similarly, if \( a_1^{e_1} a_2^{e_2} \ldots a_n^{e_n} = 1 \) is a nontrivial relation, then \( x_0, x_1, \ldots, x_{n-1}, x_n \), where \( x_i = x_{i+1}^{e_i} \), \( 1 \leq i \leq n \), and \( x_0 = x_n \), is a closed trial in \( G(H, S) \), which must contain a cycle.

Suppose now that \( H \) is a free group, \( S \) a minimal set of generators of \( H \), and \( J \) a subgroup of \( H \). Since there is no nontrivial relation on the elements of \( S \), \( G(H, S) \) does not contain a cycle. Also, from the Corollary above, \( G(H, S) \) is contractible onto \( G(J, T) \) for some set \( T \) of generators of \( J \). Because any contraction of a cycle-free graph is again cycle free, \( G(J, T) \) must be cycle free, and, thus, there is no nontrivial relation on the elements of \( T \). Hence, \( J \) must be a free group, freely generated by \( T \). □

4. The SNP Assembly Problem

In computational biochemistry there are many situations where we wish to resolve conflicts between sequences in a sample by excluding some of the sequences. Of course, exactly what constitutes a conflict must be precisely defined in the biochemical context. We define a conflict graph where the vertices represent the sequences in the sample and there is an edge between two vertices if and only if there is a conflict between the corresponding sequences. The aim is to remove the fewest possible sequences that will eliminate all conflicts. Recall that given a simple graph \( G \), a vertex cover \( C \) is a subset of the vertices such that every edge has at least one end in \( C \). Thus, the aim is to find a minimum vertex cover in the conflict graph \( G \). (In general, this is known to be an NP-complete problem [13]). We look at a specific example of the SNP assembly problem given in [15] and show how to solve this problem using the vertex cover algorithm [6].

A Single Nucleotide Polymorphism (SNP, pronounced “snip”) [15] is a single base mutation in DNA. It is known that SNPs are the most common source of genetic
polymorphism in the human genome (about 90% of all human DNA polymorphisms). The SNP Assembly Problem [15] is defined as follows. A SNP assembly is a triple $(S, F, R)$ where $S = \{s_1, ..., s_n\}$ is a set of $n$ SNPs, $F = \{f_1, ..., f_m\}$ is a set of $m$ fragments and $R$ is a relation $R: S \times F \rightarrow \{0, A, B\}$ indicating whether a SNP $s_i \in S$ does not occur on a fragment $f_j \in F$ (marked by 0) or if occurring, the non-zero value of $s_i$ ($A$ or $B$). Two SNPs $s_i$ and $s_j$ are defined to be in conflict when there exist two fragments $f_i$ and $f_j$ such that exactly three of $R(s_i, f_i)$, $R(s_i, f_j)$, $R(s_j, f_i)$, $R(s_j, f_j)$ have the same non-zero value and exactly one has the opposing non-zero value. The problem is to remove the fewest possible SNPs that will eliminate all conflicts. The following example from [15] is shown in the table below. Note that the relation $R$ is only defined for a subset of $S \times F$ obtained from experimental values.
<table>
<thead>
<tr>
<th></th>
<th>( R )</th>
<th>( f_1 )</th>
<th>( f_2 )</th>
<th>( f_3 )</th>
<th>( f_4 )</th>
<th>( f_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1 )</td>
<td>( A )</td>
<td>( B )</td>
<td></td>
<td></td>
<td></td>
<td>( B )</td>
</tr>
<tr>
<td>( s_2 )</td>
<td>( B )</td>
<td>( A )</td>
<td>( A )</td>
<td>( A )</td>
<td>( 0 )</td>
<td></td>
</tr>
<tr>
<td>( s_3 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( B )</td>
<td>( B )</td>
<td>( A )</td>
<td></td>
</tr>
<tr>
<td>( s_4 )</td>
<td>( A )</td>
<td>( 0 )</td>
<td>( A )</td>
<td>( 0 )</td>
<td>( B )</td>
<td></td>
</tr>
<tr>
<td>( s_5 )</td>
<td>( A )</td>
<td>( B )</td>
<td>( B )</td>
<td>( B )</td>
<td>( A )</td>
<td></td>
</tr>
<tr>
<td>( s_6 )</td>
<td>( B )</td>
<td></td>
<td>( A )</td>
<td>( A )</td>
<td>( 0 )</td>
<td></td>
</tr>
</tbody>
</table>

Note, for instance, that \( s_1 \) and \( s_5 \) are in conflict because \( R(s_1, f_2) = B, R(s_1, f_3) = B, R(s_2, f_2) = B, R(s_5, f_3) = A \). Again, \( s_4 \) and \( s_6 \) are in conflict because \( R(s_4, f_1) = A, R(s_4, f_3) = A, R(s_6, f_1) = B, R(s_6, f_3) = A \). Similarly, all pairs of conflicting SNPs are easily determined from the table. The conflict graph \( G \) corresponding to this SNP assembly problem is shown below in figure 4.2.

![Conflict Graph G](image)

**Figure 4.2. The conflict graph \( G \)**

We now use the vertex cover algorithm [6] to find minimal vertex covers in the conflict graph \( G \). The input is the number of vertices 6, followed by the adjacency matrix of \( G \) shown below in figure 4.3. The entry in row \( i \) and column \( j \) of the adjacency matrix is 1 if the vertices \( s_i \) and \( s_j \) have an edge in the conflict graph and 0 otherwise.
The vertex cover program [6] finds two distinct minimum vertex covers, shown in figure 4.4.

\begin{figure}[h]
\centering
\begin{tikzpicture}
  \node (s1) at (0,0) [shape=circle, fill=black!20] {$s_1$};
  \node (s2) at (1,0) [shape=circle, fill=black!20] {$s_2$};
  \node (s3) at (2,0) [shape=circle, fill=black!20] {$s_3$};
  \node (s4) at (1,-1) [shape=circle, fill=black!20] {$s_4$};
  \node (s5) at (2,-1) [shape=circle, fill=black!20] {$s_5$};
  \node (s6) at (1,-2) [shape=circle, fill=black!20] {$s_6$};

  \draw (s1) -- (s2);
  \draw (s2) -- (s1);
  \draw (s1) -- (s4);
  \draw (s4) -- (s5);
  \draw (s5) -- (s6);
  \draw (s4) -- (s6);
  \end{tikzpicture}
\caption{Minimum Vertex Cover: $s_1$, $s_4$}
\end{figure}

\begin{figure}[h]
\centering
\begin{tikzpicture}
  \node (s1) at (0,0) [shape=circle, fill=black!20] {$s_1$};
  \node (s2) at (1,0) [shape=circle, fill=black!20] {$s_2$};
  \node (s3) at (2,0) [shape=circle, fill=black!20] {$s_3$};
  \node (s4) at (1,-1) [shape=circle, fill=black!20] {$s_4$};
  \node (s5) at (2,-1) [shape=circle, fill=black!20] {$s_5$};
  \node (s6) at (1,-2) [shape=circle, fill=black!20] {$s_6$};

  \draw (s1) -- (s2);
  \draw (s2) -- (s1);
  \draw (s1) -- (s4);
  \draw (s4) -- (s5);
  \draw (s5) -- (s6);
  \draw (s4) -- (s6);
  \end{tikzpicture}
\caption{Minimum Vertex Cover: $s_4$, $s_5$}
\end{figure}

Thus, either removing $s_1$, $s_4$ or removing $s_4$, $s_5$ solves the given SNP assembly problem. \hspace{1em} ∎
5. COMPUTER NETWORK SECURITY

A team of computer scientists led by Eric Filiol [11] at the Virology and Cryptology Lab, ESAT, and the French Navy, ESCANSIC, have recently used the vertex cover algorithm [6] to simulate the propagation of stealth worms on large computer networks and design optimal strategies for protecting the network against such virus attacks in real-time.

![Diagram of computer network](image_url)

**Figure 5.1.** The set {2, 4, 5} is a minimum vertex cover in this computer network.

The simulation was carried out on a large internet-like virtual network and showed that that the combinatorial topology of routing may have a huge impact on the worm propagation and thus some servers play a more essential and
significant role than others. The real-time capability to identify them is essential to greatly hinder worm propagation. The idea is to find a minimum vertex cover in the graph whose vertices are the routing servers and whose edges are the (possibly dynamic) connections between routing servers. This is an optimal solution for worm propagation and an optimal solution for designing the network defense strategy. Figure 5.1 above shows a simple computer network and a corresponding minimum vertex cover \{2, 4, 5\}.

6. The Timetabling Problem

In a college there are \(m\) professors \(x_1, x_2, \ldots, x_m\) and \(n\) subjects \(y_1, y_2, \ldots, y_n\) to be taught. Given that professor \(x_i\) is required (and able) to teach subject \(y_j\) for \(p_{ij}\) periods (\(P = [p_{ij}]\) is called the teaching requirement matrix), the college administration wishes to make a timetable using the minimum possible number of periods. This is known as the timetabling problem [4] and can be solved using the following strategy. Construct a bipartite multigraph \(G\) with vertices \(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n\) such that vertices \(x_i\) and \(y_j\) are connected by \(p_{ij}\) edges.

We presume that in any one period each professor can teach at most one subject and that each subject can be taught by at most one professor. Consider, first, a single period. The timetable for this single period corresponds to a matching in the graph and, conversely, each matching corresponds to a possible assignment of professors to subjects taught during this period. Thus, the solution to the timetabling problem consists of partitioning the edges of \(G\) into the minimum number of matchings. Equivalently, we must properly color the edges of \(G\) with the minimum number of colors. We shall show yet another way of solving the problem using the vertex coloring algorithm [7]. Recall that the line graph \(L(G)\) of \(G\) has as vertices the edges of \(G\) and two vertices in \(L(G)\) are connected by an edge if and only if the corresponding edges in \(G\) have a vertex in common. The line graph \(L(G)\) is a simple graph and a proper vertex coloring of \(L(G)\) yields a proper edge coloring of \(G\) using the same number of colors. Thus, to solve the
timetabling problem, it suffices to find a minimum proper vertex coloring of \( L(G) \) using [7]. We demonstrate the solution with a small example.

Suppose there are four professors \( x_1, x_2, x_3, x_4 \) and five subjects \( y_1, y_2, y_3, y_4, y_5 \) to be taught [4]. The teaching requirement matrix \( p = [p_{ij}] \) is given below in figure 6.1.

<table>
<thead>
<tr>
<th></th>
<th>( y_1 )</th>
<th>( y_2 )</th>
<th>( y_3 )</th>
<th>( y_4 )</th>
<th>( y_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

*Figure 6.1. The teaching requirement matrix*

*Figure 6.2. The bipartite multigraph \( G \)*
We first construct the bipartite multigraph $G$ shown above in figure 6.2. Next, we construct the line graph $L(G)$. The adjacency matrix of $L(G)$ is given below.

\[
\begin{array}{cccccccccccc}
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
\end{array}
\]

Now, we use the vertex coloring algorithm [7] to find a minimum proper 4-coloring of the vertices of $L(G)$.

*Figure 6.3. A minimum proper 4-coloring of the vertices of $L(G)$*
Vertex Coloring: (1, green) (2, red) (3, blue) (4, yellow) (5, yellow) (6, green) (7, green) (8, yellow) (9, red) (10, blue) (11, yellow). This, in turn, yields a minimum proper edge 4-coloring of the bipartite multigraph $G$.

![Graph of vertex coloring](image)

**Figure 6.4. A minimum proper 4-coloring of the edges of $G$**

Edge Coloring: ({$x_1,y_1$}, green) ({$x_1,y_1$}, red) ({$x_1,y_2$}, blue) ({$x_1,y_4$}, yellow) ({$x_2,y_2$}, yellow) ({$x_2,y_4$}, green) ({$x_3,y_2$}, green) ({$x_3,y_3$}, yellow) ({$x_3,y_4$}, red) ({$x_4,y_4$}, blue) ({$x_4,y_5$}, yellow). Interpret the colors green, red, blue, yellow as periods 1, 2, 3, 4 respectively. Then, from the edge coloring of $G$, we obtain a solution of the given timetabling problem as shown below in figure 6.5.
7. **Map Coloring and GSM Mobile Phone Networks**

Given a map drawn on the plane or the surface of a sphere, the famous four color theorem asserts that it is always possible to properly color the regions of the map such that no two adjacent regions are assigned the same color, using at most four distinct colors [8] [18] [1]. For any given map, we can construct its dual graph as follows. Put a vertex inside each region of the map and connect two distinct vertices by an edge if and only if their respective regions share a whole segment of their boundaries in common. Then, a proper vertex coloring of the dual graph yields a proper coloring of the regions of the original map.

![Map of India](image)

*Figure 7.1. The map of India*
Figure 7.2. The dual graph of the map of India

We use the vertex coloring algorithm [7] to find a proper coloring of the map of India with four colors, see figures 7.1 and 7.2 above.

The Groupe Special Mobile (GSM) was created in 1982 to provide a standard for a mobile telephone system. The first GSM network was launched in 1991 by Radiolinja in Finland with joint technical infrastructure maintenance from Ericsson. Today, GSM is the most popular standard for mobile phones in the world, used by over 2 billion people across more than 212 countries. GSM is a cellular network with its entire geographical range divided into hexagonal cells. Each cell has a communication tower which connects with mobile phones within the cell. All mobile phones connect to the GSM network by searching for cells in the immediate vicinity. GSM networks operate in only four different frequency ranges. The reason why only four different frequencies suffice is clear: the map of the cellular regions can be properly colored by using only four different colors! So, the vertex coloring algorithm may be used for assigning at most four different frequencies for any GSM mobile phone network, see figure 7.2 below.
8. Knight's Tours

In 840 A.D., al-Adli [17], a renowned shatranj (chess) player of Baghdad is said to have discovered the first re-entrant knight's tour, a sequence of moves that takes the knight to each square on an 8×8 chessboard exactly once, returning to the original square. Many other re-entrant knight's tours were subsequently discovered but Euler [10] was the first mathematician to do a systematic analysis in 1766, not only for the 8×8 chessboard, but for re-entrant knight's tours on the general $n \times n$ chessboard. Given an $n \times n$ chessboard, define a knight's graph with a vertex corresponding to each square of the chessboard and an edge connecting vertex $i$ with vertex $j$ if and only if there is a legal knight's move from the square corresponding to vertex $i$ to the square corresponding to vertex $j$. Thus, a re-entrant knight's tour on the chessboard corresponds to a Hamiltonian circuit in the knight's graph. The Hamiltonian circuit algorithm [9] [13] has been used to find re-entrant knights tours on chessboards of various dimensions.
Figure 8.1. A re-entrant knight's tour on the 8×8 chessboard

REFERENCES


Elsevier Science Publishing Company Inc.


---

**S. Pirzada**

Department of Mathematics  
University of Kashmir  
Srinagar 190006  
India  
sdpirzada@yahoo.co.in  
http://www.geocities.com/dharwadker/pirzada

---

**Ashay Dharwadker**  

H-501 Palam Vihar  
District Gurgaon  
Haryana 122017  
India  
dharwadker@yahoo.com  
http://www.geocities.com/dharwadker
Solving a Matrix Polynomial by Conjugate Gradient Methods *

Hyun Ji Ko† Hyun-Min Kim†

Abstract

One of well known and much studied nonlinear matrix equations is the matrix polynomial which has the form \( G(X) = A_0 X^m + A_1 X^{m-1} + \cdots + A_m \), where \( A_0, A_1, \ldots, A_m \) and \( X \) are \( n \times n \) real matrices. We show how the minimization methods can be used to solve the matrix polynomial \( G(X) \) and give some numerical experiments. We also compare Polak and Ribière version and Fletcher and Reeves version of conjugate gradient method.

Keywords. matrix polynomial, solvent, gradient, Hessian, conjugate gradient method

AMS subject classifications. 65F30, 65H10

1 Introduction

Nonlinear matrix equations often occur in applications and modelling of scientific problems. In this work we specially consider one of the nonlinear matrix equations which is called the matrix polynomial

\[
G(X) = A_0 X^m + A_1 X^{m-1} + \cdots + A_m, \quad \text{where} \quad A_0, \ldots, A_m, X \in \mathbb{R}^{n \times n}.
\]

A matrix \( S \) satisfying the equation \( G(S) = 0 \) is called a solvent, more precisely, a right solvent of \( G(X) \) to distinguish it from a left solvent, which is a solution of the related matrix equation

\[
X^m A_0 + X^{m-1} A_1 + \cdots + A_m = 0.
\]

For solving the quadratic matrix equation

\[
AX^2 + BX + C = 0, \quad \text{where} \quad A, B, C, X \in \mathbb{R}^{n \times n},
\]

Davis [1], [2] considered Newton's method and Higham and Kim [4], [5] incorporated exact line searches into Newton's method and gave the generalized Schur decomposition approach and Berroulli's method. Also, the conjugate gradient method was suggested by Kim [8].

In matrix polynomials Newton's method with and without line searches and Berroulli's method were considered [6], [7], [9], [11]. Our work is to extend the conjugate gradient method for solving the matrix polynomial (1.1). Defining the object function and finding the gradient of the object function we can apply both conjugate gradient methods which are suggested by Fletcher and Reeves [3] and Polak and Ribière [10] for minimizing the nonlinear equations.

*This work was supported for two years by Pusan National University Research Grant.
†Department of Mathematics, Pusan National University, Busan, Republic of Korea.
‡Department of Mathematics, Pusan National University, Busan, Republic of Korea. Corresponding author(hyunmin@pusan.ac.kr).
2 Conjugate Gradient Methods for Solving a Matrix Polynomial

For solving matrix polynomials by adapting nonlinear conjugate gradient methods, first we define the gradient and Hessian of the objective function

\[ f(X) = \frac{1}{2} \| G(X) \|_F^2 \]

where \( f : \mathbb{R}^{n \times n} \to \mathbb{R} \). To derive the gradient of \( f \)

\[ \nabla f(X) = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{n1}} & \cdots & \frac{\partial f}{\partial x_{nn}} \end{bmatrix} \in \mathbb{R}^{n^2} \]

and the Hessian of \( f \)

\[ \nabla^2 f(X) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_{11}^2} & \cdots & \frac{\partial^2 f}{\partial x_{11} \partial x_{1n}} & \cdots & \frac{\partial^2 f}{\partial x_{1n} \partial x_{1n}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f}{\partial x_{n1} \partial x_{n1}} & \cdots & \frac{\partial^2 f}{\partial x_{n1} \partial x_{n1}} & \cdots & \frac{\partial^2 f}{\partial x_{n1} \partial x_{nn}} \\ \frac{\partial^2 f}{\partial x_{n1} \partial x_{nn}} & \cdots & \cdots & \cdots & \frac{\partial^2 f}{\partial x_{n1} \partial x_{nn}} \end{bmatrix} \in \mathbb{R}^{n^2 \times n^2} \]

suppose that the nonlinear matrix equation \( G(X) \) is twice continuously differentiable. By expanding \( G(X + E) \) we obtain

\[ G(X + E) = G(X) + G'_X(E) + O_2(E), \]

where \( G'_X(E) : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n} \) is the Fréchet derivative of \( G \) at \( X \) in the direction \( E \) and \( O_2(E) = O(\|E\|^2) \). Using vec operator

\[
\text{vec}(G(X + E)) = \text{vec}(G(X) + G'_X(E) + O_2(E)) = \text{vec}(G(X)) + \text{vec}(G'_X(E)) + \text{vec}(O_2(E)).
\]

Now, the function \( f(X + E) \) can be written

\[
f(X + E) = \frac{1}{2} \| G(X + E) \|_F^2 = \frac{1}{2} \left[ \text{vec}(G(X)) + \text{vec}(G'_X(E)) + \text{vec}(O_2(E)) \right]^T \left[ \text{vec}(G(X)) + \text{vec}(O_2(E)) \right] = \frac{1}{2} \left[ \text{vec}(G(X))^T \text{vec}(G(X)) + 2 \text{vec}(G(X))^T \text{vec}(G'_X(E)) + O(\|E\|^2) \right],
\]

where

\[
O(\|E\|^2) = \text{vec}(G(X))^T \text{vec}(O_2(E)) + \text{vec}(G'_X(E))^T \left[ \text{vec}(G(X)) + \text{vec}(O_2(E)) \right] + \text{vec}(O_2(E))^T \left[ \text{vec}(G(X)) + \text{vec}(O_2(E)) \right].
\]

By Taylor series of \( f(X + E) \) we have

\[
f(X + E) = f(X) + \text{vec}(\nabla f(X))^T \text{vec}(E) + \text{vec}(E)^T \nabla^2 f(X) \text{vec}(E) + O(\|E\|^2) = f(X) + \text{trace}(\nabla f(X)^T E) + \text{vec}(E)^T \nabla^2 f(X) \text{vec}(E) + O(\|E\|^2).
\]

From the equations (2.1) and (2.2),

\[
\text{vec}(G(X))^T \text{vec}(G'_X(E)) = \text{trace}(\nabla f(X)^T E)
\]

(2.3)
and writing $\text{quad}(y)$ for the quadratic part of $y$ in the variable $E$,
\[
\text{vec}(E)^T \nabla^2 f(X) \text{vec}(E) = \frac{1}{2} \left[ \text{vec}(G_X'(E))^T \text{vec}(G_X'(E)) + 2 \text{vec}(G(X)^T \text{quad}(\text{vec}(N_X(E)))) \right]
= \frac{1}{2} \text{trace}(G_X'(E)^T G_X'(E)) + \langle G(X)^T \text{quad}(N_X(E)) \rangle.
\]

By the way, we can transform these representations into different forms. In the left side of the equation (2.3) we have
\[
\text{vec}(G(X)^T \text{vec}(G_X'(E))) = \text{trace} \left[ G(X)^T G_X'(E) \right]
\]
and from the right side of the equation (2.3) we obtain
\[
\text{trace}(\nabla f(X)^T E) = \text{trace}(E^T \nabla f(X)).
\]
So, by setting $E = e_i e_j^T$ and using $\text{trace}(AB) = \text{trace}(BA)$, we can get the gradient of the function $f$ which is
\[
(\nabla f)_{ij} = \text{trace}(e_i e_j^T \nabla f) = \text{trace} \left[ G(X)^T G_X'(E) \right].
\]

There is no necessity for expressing more. But it is important to know the positive definiteness of the Hessian of $f(X)$ at a solution. At a solution $S$, $G(S) = 0$, so
\[
\text{vec}(E)^T \nabla^2 f(X) \text{vec}(E) = \frac{1}{2} \text{trace}(G_X'(E)^T G_X'(E)).
\]

Hence, the Hessian of $f(X)$ is positive definite at $X$ if and only if $\text{trace}(G_X'(E)^T G_X'(E)) > 0$ for all nonzero $E$ if and only if $G'(X)$ is nonsingular. We can now apply minimization method for solving the equation $G(X) = 0$. Again, the gradient of $f(X)$ can be written
\[
(\nabla f)_{ij} = \text{trace} \left[ G(X)^T G_X'(e_i e_j^T) \right]
= \text{trace} \left[ G(X)^T \sum_{p=1}^{m-p} \left( \sum_{\psi=0}^{m-p} A_{pq} X^{m-\psi-p} e_i e_j^T X^{p-1} \right) \right]
= \text{trace} \left[ \sum_{p=1}^{m-p} \left( \sum_{\psi=0}^{m-p} G(X)^T A_{pq} X^{m-\psi-p} e_i e_j^T X^{p-1} \right) \right]
= \sum_{p=1}^{m-p} \sum_{\psi=0}^{m-p} e_j^T X^{p-1} G(X)^T A_{pq} X^{m-\psi-p} e_i
\]

hence
\[
\nabla f = \sum_{p=1}^{m-p} \sum_{\psi=0}^{m-p} \left( X^{p-1} G(X)^T A_{pq} X^{m-\psi-p} \right)^T
= \sum_{p=1}^{m-p} \sum_{\psi=0}^{m-p} \left( A_{pq} X^{m-\psi-p} \right)^T G(X)(X^{p-1})^T.
\]

Finally, by (2.4), the Hessian of $f(X)$ can be obtained by
\[
\text{vec}(E)^T \nabla^2 f(X) \text{vec}(E)
= \frac{1}{2} \text{trace} \left[ \sum_{l=1}^{t} \sum_{k=0}^{m-p} \sum_{p=1}^{m-p} \sum_{\psi=0}^{m-p} (X^{p-1})^T E (X^{m-\psi-p}) A^T \right] AX^{m-l-k} EX^{l-1}.
\]
3 Algorithms for Solving $G(X) = 0$ by Conjugate Gradient Methods

We now develop the algorithm of the conjugate gradient for solving a matrix polynomial. Let $X_0$ be given and $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$. To construct the approximations $X_i$ and the search directions $D_i$ for $i = 1, 2, \cdots$, we use the steepest decent direction $D_i = -\nabla f(X_i)$. The following algorithm attempts to minimize $f(X)$.

**Algorithm 3.1.**

Evaluate $f_0 = f(X_0)$, $\nabla f_0 = \nabla f(X_0)$

$i = 0$; $D_0 = -\nabla f(X_0)$

while $\nabla f_i \neq 0$

find $\alpha_i$ that minimize $\|G(X_i + \alpha_i D_i)\|_F^2$

$X_{i+1} = X_i + \alpha_i D_i$

$D_{i+1} = -\nabla f(X_{i+1}) + \beta_i D_i$

end

The two forms of $\beta_i$ are suggested by Fletcher and Reeves [3] and Polak and Ribiére [8],

$$\beta_i^{FR} = \frac{\|\nabla f_{i+1}\|_F^2}{\|\nabla f_i\|_F^2},$$

$$\beta_i^{PR} = \frac{\text{trace}((\nabla f_{i+1} - \nabla f_i)^T \nabla f_{i+1})}{\|\nabla f_i\|_F^2}.$$

We now call the conjugate gradient method with $\beta_i^{FR}$ the $CG_{FR}$ method and the conjugate gradient method with $\beta_i^{PR}$ the $CG_{PR}$ method. Algorithm 3.1 can be considered with exact line searches for a step length $\alpha_i$. First, we apply exact line searches for the quadratic matrix equation

$$Q(X) = AX^2 + BX + C$$

where $A, B, C, X$ are $n \times n$ matrices [8]. From $Q(X + \alpha D) = Q(X) + \alpha D^T Q + \alpha^2 D^T A D$ we have a quartic polynomial

$$p_{CG}(t) = \|Q(X + \alpha D)\|_F^2$$

$$= a_4 \alpha^4 + a_3 \alpha^3 + a_2 \alpha^2 + a_1 \alpha + a_0,$$

where

$$a_4 = \|AD^2\|_F^2,$$

$$a_3 = \text{trace}(DX(D)^T AD^2 + (AD^2)^T DX(D)),$$

$$a_2 = \text{trace}(Q^T AD + (AD^2)^T Q) + \|DX(D)\|_F^2,$$

$$a_1 = \text{trace}(Q^T DX(D) + DX(D)^T Q),$$

$$a_0 = \|Q\|_F^2.$$

Since $p_{CG}(t)$ is quartic and the coefficient $a_4$ is positive it has a minimization.

Here we will generalize the exact line searches for the quadratic matrix equation to the matrix polynomial. Let matrix polynomial $G(X)$ in (1.1) be given. For implementation of $p_{CG}(t) = \|G(X + \alpha D)\|_F^2$, we must find exact expansion of $G(X + \alpha D)$. So we introduce some notion of repeated permutation [11]. A function $\Phi_{X,Y}[n, m - n]$ is the sum of the products of
all repeated permutation of which the number of matrices $X$ is $n$ and the number of matrices $Y$ is $m - n$. For example,

$$
\Phi_{X,Y}^{[2,3]} = \Phi_{X,Y}^{[2,5-2]}
= XYYYY + XYYXY + XXYXY + XYXYX + YXYXY
+ YXYXY + YYXYY + YYXYX + YYYXY + YYYXX.
$$

Also,

$$
\Phi_{X,Y}^{[0,0]} = I, \quad \Phi_{X,Y}^{[m,0]} = X^m, \quad \Phi_{X,Y}^{[0,m]} = Y^m \text{ for all nonnegative integer } m.
$$

And we can easily verify

$$
(\Phi_{X,Y}^{[n,m-n]} Y)^T = \Phi_{Y,X}^{[n,m-n]} Y.
$$

By using this notation, we can describe the expansion of $G(X + \alpha D)$ which has the form

$$
G(X + \alpha D) = A_0(X + \alpha D)^m + A_1(X + \alpha D)^{m-1} + \cdots + A_m
= A_0 \sum_{i=0}^m \alpha^i \Phi_{X,Y}^{[m-i,i]} + A_1 \sum_{i=0}^{m-1} \alpha^i \Phi_{X,Y}^{[m-1-i,i]} + \cdots + A_m
= \sum_{j=0}^m \sum_{i=0}^j \alpha^j A_{m-j} \Phi_{X,Y}^{[j-i,i]}.
$$

Finally, we obtain

$$
\|G(X + \alpha D)\|_F^2 = \text{trace}(G(X + \alpha D))^T G(X + \alpha D)
= \text{trace} \left( G^T(X) + \alpha D_X(D)^T + \sum_{j=2}^m \sum_{j=2}^j \alpha^j \Phi_{X,Y}^{T,DF}[j-i,h] A_{m-j} \right)
= \text{trace} \left( G(X) + \alpha D_X(D)^T + \sum_{j=2}^m \sum_{j=2}^j \alpha^j A_{m-j} \Phi_{X,Y}^{T,DF}[j-i,h] \right)
= \|G(X)\|_F^2 + \cdots + \alpha^{2m} \|A_0 D_m\|_F^2.
$$

Note that exact line searches always satisfy the equation.

$$
\text{trace}(\nabla f_{i+1})^T D_i = \text{vec}(\nabla f_{i+1})^T \text{vec}(D_i) = 0. \quad (3.1)
$$

By applying the vec operation to $D_{i+1} = -\nabla f(X_{i+1}) + \beta D_i$ and premultiplying by $\text{vec}(\nabla f_{i+1})^T$, we have

$$
\text{vec}(\nabla f_{i+1})^T \text{vec}(D_{i+1}) = -\|\nabla f_{i+1}\|_F^2 + \beta \text{vec}(\nabla f_{i+1})^T \text{vec}(D_i).
$$

Therefore, by (3.1), we have $\text{vec}(\nabla f_{i+1})^T \text{vec}(D_{i+1}) < 0$, which means that $D_{i+1}$ is descent direction.
4 Numerical Experiment and Conclusion

In this section we show and compare some experimental results $\text{CG}_{\text{FR}}$ method and $\text{CG}_{\text{PR}}$ method. Our experiment were done in MATLAB 7.1, which has the unit roundoff $u = 2^{-53} \approx 1.1 \times 10^{-16}$. Iterations for two methods are terminated when the relative residual $\rho(X)$ satisfies

$$\rho(X) = \frac{\|H(G(X))\|_F}{\|A_0\|_F \|X_0\|_2 + \cdots + \|A_m\|_F} \leq nu.$$

First, we consider two examples

$$G_1(X) = X^2 + X + \begin{bmatrix} -6 & -5 \\ 0 & -8 \end{bmatrix} = 0,$$

$$G_2(X) = X^2 + \begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix} X + \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = 0 \quad (4.1)$$

from [1], [8]. Figure 4.1 and Table 4.1 show that we have exactly same results in [8] and our default matrix is, as in [1],

$$\text{Default}_X = \left( \frac{\|B\|_F + \sqrt{\|B\|_F^2 + 4 \|A\|_F \|C\|_F}}{2\|A\|_F} \right) I.$$

Figure 4.1: Convergence for problem $G_1(X)$ in (4.1) with $\text{CG}_{\text{FR}}$ and $\text{CG}_{\text{PR}}$ methods.

Table 4.1: Number of iterations for convergence for problem $G_2(X)$ in (4.2) with conjugate gradient methods.

<table>
<thead>
<tr>
<th>$X_0$</th>
<th>$\text{CG}_{\text{FR}}$</th>
<th>$\text{CG}_{\text{PR}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default $X_0$</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>$10^I$</td>
<td>83</td>
<td>8</td>
</tr>
<tr>
<td>$10^5I$</td>
<td>34</td>
<td>8</td>
</tr>
<tr>
<td>$10^{10}I$</td>
<td>39</td>
<td>10</td>
</tr>
</tbody>
</table>
We now consider cubic matrix equations. Two examples of matrix polynomial with degree 3 are

\[ G_3(X) = X^3 + X^2 + X \begin{bmatrix} -6 & -5 \\ 0 & -6 \end{bmatrix} = 0 \]  \hspace{1cm} (4.3)

and

\[ G_4(X) = X^3 + \begin{bmatrix} 0 & -1 \\ -1 & 1 \end{bmatrix} X^2 + X \begin{bmatrix} -10 & -7 \\ 4 & 0 \end{bmatrix} = 0. \] \hspace{1cm} (4.4)

Starting matrices are both \( I_3 \) and we can see the convergence results in Figures 4.2 and 4.3. Note that using \( CFTP \) method we could not find the solvent of the equation (4.4).

![Figure 4.2: Convergence for problem (4.3) with conjugate gradient methods.](image1)

![Figure 4.3: Convergence for problem (4.4) with conjugate gradient methods.](image2)

Figure 4.1, 4.2 and 4.3 show that the \( CFTP \) method gives better results than the \( CFTP \) method.

Finally, we give a summary of our results and compare experimental results. For solving matrix polynomials Newton’s method and functional iterations were introduced. Also the conjugate gradient method for solving the quadratic matrix equations was considered. So we
generalized the conjugate gradient method to the matrix polynomial. We also experimented with some examples and compare the CGFR method and CGPR method. Although not all of them converge, we find some examples to converge. The examples in this paper, CGFR method is more efficient for converging.

References


SOME COMPUTATIONS AND EXTREMAL PROPERTIES OF OPERATORS

Kyung Young Moon and Sun Hyun Park

Abstract

In [6] some computation of spectral measures induced by normal operators $T^nT^m$ was introduced. In this note we improve some computations by using spectral measures, which are related to extremal vectors. Also, we discuss the extremal value properties and apply our spectral measure equations to moment sequences which are induced by weighted shifts.


Keywords and phrases: extremal vectors, spectral measures, weighted shifts.

1. Introduction

Let $\mathcal{H}$ be a separable, infinite dimensional, complex Hilbert space, and denote by $\mathcal{L}(\mathcal{H})$ the algebra of all bounded linear operators on $\mathcal{H}$. And also the study of invariant subspaces is very important to study of structure of Hilbert space operators. In these sequel studies, Per Enflo introduced a new technique involving some “extremal vectors” for producing invariant subspaces for certain quasinilpotent operators in $\mathcal{L}(\mathcal{H})$ (cf. [1]). One modified those techniques and expanded in [4] to produce better invariant subspace theorems for some quasinilpotent operators, and continued to explore the limits of Enflo’s technique for producing invariant and hyperinvariant subspaces in [5]. In [7], the spectral techniques were contributed to solve the hyperinvariant subspace problem of subnormal operators.

In this paper we study improved some computations in the techniques introduced in [6]. In Section 2, we discuss some spectral equations which
are related to extremal vectors. In Section 3, we study the extremal value properties. Finally, in Section 4 we apply our spectral measure equations to moment sequences which are induced by weighted shifts.

Throughout this paper, \( \mathbb{N} \) is the set of natural numbers.

2. Some constructions

Let \( N \) be a normal operator in \( \mathcal{L}(\mathcal{H}) \). Then there exists a corresponding spectral measure \( E \) on \( \sigma(N) \) such that for each Borel function \( f \) on \( \sigma(N) \), \( f(N) = \int f(\lambda) \, dE \) exists, in particular, if \( f = \chi_{\Delta} \) is a characteristic function on a Borel subset \( \Delta \subset \sigma(N) \), then \( f(N) = E(\Delta) \) and for \( f \in C(\sigma(N)) \) there exist polynomials \( p_n(\lambda) \) with \( \|f - p_n\|_{\infty} \to 0 \) such that \( f(N) = \lim \int p_n(\lambda) \, dE \), where \( C(\sigma(N)) \) is the Banach space of all continuous functions on \( \sigma(N) \).

By the spectral mapping theorem (cf. [2]), there exists a corresponding spectral measure \( E^{(n,p)} \) to \( (T^{n}T^*)^{p} \) such that

\[
(T^{n}T^*)^{p} = \int_{[0,\|T^n\|\|T^*\|]} \lambda dE^{(n,p)} \quad (n \in \mathbb{N})
\]

and

\[
T^{n}T^* = \int_{[0,\|T^n\|\|T^*\|]} \lambda dE^{(n)} \quad (n \in \mathbb{N}).
\]

For brevity we write \( E^{(n,p)}_\lambda = E^{(n,p)}([0,\lambda]) \) and \( E^{(n,p)}_{\lambda^-} = E^{(n,p)}([0,\lambda]) \) (see [5] for this notation). First we discuss some fundamental properties of these spectral measures.

**Proposition 2.1.** Under the above notation we have the following statements:

(i) \( E^{(n,p)}_\lambda = E^{(n)}_{\lambda p} \) for all \( \lambda \in [0,\infty) \),

(ii) \( E^{(n,p)}_{\lambda p} = E^{(n)}_\lambda \) for all \( \lambda \in [0,\infty) \).
Proof. (i) First we claim that $E^{(n,p)}([0, \lambda]) = E^{(n)}([0, \lambda^{1/p}])$. Observe that

$$
\int_{[0,\|T^n\|^{2p}]} \lambda dE^{(n,p)} = \left( \int_{[0,\|T^n\|^{2p}]} T^{inT^n} \right)^p = \int_{[0,\|T^n\|^{2p}]} \lambda^p dE^{(n)}
$$
$$= \int_{[0,\|T^n\|^{2p}]} \varphi(\lambda) dE^{(n)} \quad \text{(where } \varphi(\lambda) = \lambda^p)$$
$$= \int_{[0,\|T^n\|^{2p}]} \lambda dE^{(n)} \circ \varphi^{-1}$$
$$= \int_{[0,\|T^n\|^{2p}]} \lambda dE^{(n)} \circ \varphi^{-1}.
$$

Hence

$$
\int_{[0,\|T^n\|^{2p}]} \lambda dE^{(n,p)} = \int_{[0,\|T^n\|^{2p}]} \lambda dE^{(n)} \circ \varphi^{-1}.
$$

Continuing this process, we have that

$$
\int_{[0,\|T^n\|^{2p}]} \lambda^m dE^{(n,p)} = \int_{[0,\|T^n\|^{2p}]} \lambda^m dE^{(n)} \circ \varphi^{-1} \text{ for all } m \in \mathbb{N}.
$$

Then obviously we have

$$
\int_{[0,\|T^n\|^{2p}]} p(\lambda) dE^{(n,p)} = \int_{[0,\|T^n\|^{2p}]} p(\lambda) dE^{(n)} \circ \varphi^{-1} \text{ for any polynomial } p,
$$

which implies that $E^{(n,p)} = E^{(n)} \circ \varphi^{-1}$. Also we have that

$$
E^{(n,p)}_{\lambda} = E^{(n,p)}([0, \lambda]) = E^{(n)} \circ \varphi^{-1}([0, \lambda]) = E^{(n)}([0, \lambda^{1/p}]) = E^{(n)}_{\lambda^{1/p}}.
$$

(ii) This follows obviously from (i). \qed

The following lemma is a well-known property.

**Lemma 2.2** [2, p.266]. Let $N$ be a normal operator in $L(H)$ and let $E$ be the corresponding spectral measure to $N$. Then $\lambda \in \sigma_p(N)$ if and only if $E(\{\lambda\}) \neq 0$.

**Proposition 2.3.** Let $x_0 \in H$ be a unit vector. Under the above notation we have the following statements:

(i) The functions $\lambda \rightarrow E^{(n,p)}_{\lambda}$ and $\lambda \rightarrow E^{(n,p)}_{\lambda}$ are monotone increasing,

(ii) $\lambda \rightarrow \langle E^{(n,p)}_{\lambda} x_0, x_0 \rangle$ is continuous from the right,
(iii) \( \lambda \rightarrow \langle E^{(n,p)}_{\lambda} x_0, x_0 \rangle \) is continuous from the left. In particular, the functions in (ii) and (iii) are continuous at \( \lambda_0 \) if \( \lambda_0 \notin \sigma_p((T^{*n}T^n)^p) \).

Proof. (i) If \( \lambda_1 < \lambda_2 \), then
\[
E^{(n,p)}_{\lambda_2} = E^{(n,p)}([0, \lambda_2]) = E^{(n,p)}([0, \lambda_1] \cup (\lambda_1, \lambda_2)) = E^{(n,p)}([0, \lambda_1]) + E^{(n,p)}((\lambda_1, \lambda_2]) \geq E^{(n,p)}([0, \lambda_1]) = E^{(n,p)}_{\lambda_1}.
\]

(ii) Let \( \lambda_k \to \lambda_0^+ \ (k \to \infty) \). Then
\[
\langle E^{(n,p)}_{\lambda_k} x_0, x_0 \rangle - \langle E^{(n,p)}_{\lambda_0} x_0, x_0 \rangle = \langle (E^{(n,p)}([0, \lambda_k]) - E^{(n,p)}([0, \lambda_0])) x_0, x_0 \rangle \leq \|E^{(n,p)}((\lambda_0, \lambda_k]] x_0, x_0 \rangle \langle E^{(n,p)}((\lambda_0, \lambda_k]) \|.
\]

Since
\[
\lim_{k \to \infty} \|E^{(n,p)}((\lambda_0, \lambda_k]) \| = \|E^{(n,p)}(\cap_{k=1}^\infty (\lambda_0, \lambda_k]) \| = \|E^{(n,p)}(\emptyset) \| = 0,
\]
we have that
\[
\langle E^{(n,p)}_{\lambda_k} x_0, x_0 \rangle \to \langle E^{(n,p)}_{\lambda_0} x_0, x_0 \rangle.
\]

We will show the left continuity at \( \lambda_0 \notin \sigma_p((T^{*n}T^n)^p) \). To do so, let \( \lambda_k \to \lambda_0^-(k \to \infty) \). By the similar method, we have that
\[
\lim_{k \to \infty} \|E^{(n,p)}((\lambda_k, \lambda_0]) \| = \|E^{(n,p)}(\cap_{k=1}^\infty (\lambda_k, \lambda_0]) \| = \|E^{(n,p)}(\{\lambda_0\}) \|.
\]

By Lemma 2.2, \( \lim_{k \to \infty} \|E^{(n,p)}((\lambda_k, \lambda_0]) \| = 0 \). Hence
\[
\langle E^{(n,p)}_{\lambda_k} x_0, x_0 \rangle \to \langle E^{(n,p)}_{\lambda_0} x_0, x_0 \rangle.
\]

(iii) Similar to (ii). \( \blacksquare \)

3. Extremal value property

In this section we discuss extremal values of \( T \in \mathcal{L}(\mathcal{H}) \).

Definition 3.1. For \( \theta \) with \( 0 < \theta < 1 \) and \( x_0 \in \mathcal{H} \) with \( \|x_0\| = 1 \), we define
\[
\lambda_n(\theta, x_0, p) = \inf\{\lambda \in [0, \|[T^n]^{2p}\|] : \|E^{(n,p)}_{\lambda} x_0\| \geq \theta\}.
\]
In this case \( \lambda_{n,p} := \lambda_n(\theta, x_0, p) \) is called an extremal value of \( T \) with respect to \((\theta, x_0, p)\).

Let \( A \) be a positive operator and let \( Ax = 0 \). Some computation shows that \( A^p x = 0 \) for any \( p > 0 \).

And for \( p > 0 \), obviously we have that \( T^m T^n \) is one to one if and only if \((T^m T^n)^p\) is one to one.

**Proposition 3.2 (Boundedness Property).** Let \( T \in \mathcal{L}(\mathcal{H}) \). Then
\[
\lambda_n(\theta, x_0, p) \leq |\sigma(T)|^{2m} \quad \text{for all} \quad n \in \mathbb{N},
\]
where \(|\sigma(T)|\) is the spectral radius of \( T \).

**Proof.** Since \( \lambda_{n,p} \leq ||T^n||^{2p} \), we have \((\lambda_{n,p})^{\frac{1}{n}} \leq ||T^n||^{\frac{2p}{n}}\). Hence
\[
(\lambda_{n,p})^{\frac{1}{n}} \leq \limsup_{n \to \infty} (\lambda_{n,p})^{\frac{1}{n}} \leq \left( \limsup_{n \to \infty} ||T^n|| \right)^{\frac{1}{n}} = |\sigma(T)|^{2p}.
\]

Hence the proof is complete. \( \blacksquare \)

The following is the main theorem of this note.

**Theorem 3.3 (Extremal and injective property).** Let \( T \in \mathcal{L}(\mathcal{H}) \). Then \( T \) is one to one if and only if \( \lambda_{n}(\theta, x_0, p) > 0 \) for any \( x_0 \) in \( \mathcal{H} \) with \( ||x_0|| = 1 \), any \((n, p) \in \mathbb{N} \times (0, \infty) \) and any \( \theta \in (0, 1) \).

**Proof.** (\( \Rightarrow \)) Assume that \( T \) is one to one. To the contrary, we suppose that there exist \( x_0 \) in \( \mathcal{H} \) with \( ||x_0|| = 1 \), \((n, p) \in \mathbb{N} \times (0, \infty) \) and \( \theta \in (0, 1) \) such that \( \lambda_{n,p} = 0 \). Then by the definition of infimum of \( \lambda_{n}(\theta, x_0, p) \), there exists a sequence \( \{\lambda_{n,p}^{(k)}\}_{k=1}^{\infty} \) such that \( \lambda_{n,p}^{(1)} \geq \lambda_{n,p}^{(2)} \geq \cdots \to \lambda_{n,p} (= 0) \). Then we have
\[
0 < \theta \leq \lim_{k \to \infty} ||E^{(n,p)}([0, \lambda_{n,p}^{(k)})]x_0|| \leq \lim_{k \to \infty} ||E^{(n,p)}([0, \lambda_{n,p}^{(k)})]|| ||x_0||
= \lim_{k \to \infty} ||E^{(n,p)}([0, \lambda_{n,p}^{(k)})]|| = ||E^{(n,p)}(\cap_{k=1}^{\infty}([0, \lambda_{n,p}^{(k)})]|)
= ||E^{(n,p)}(\{0\})||.
\]

To show \( E^{(n,p)}(\{0\}) = 0 \), we will prove that \( 0 \notin \sigma_p((T^m T^n)^p) \), i.e. \((T^m T^n)^p\) is one to one). Since \( T \) is one to one, \( T^m \) is one to one. Suppose \( T^m T^n x = 0 \). Then \( \langle T^m T^n x, x \rangle = 0 \). So \( ||T^n x|| = 0 \) and so \( T^n x = 0 \). Hence \( x = 0 \). Thus \( T^m T^n \) is one to one. By the above remark, so \((T^m T^n)^p\) is one to one.
Therefore by Lemma 2.2 \( E^{(n,p)}(\{0\}) = 0 \). This contradiction proves this implication.

\((\Leftarrow)\) Suppose that \( T \) is not one to one. Then obviously, \((T^* T)^n P \) is not one to one by the above remark. Hence \( 0 \in \sigma_p ((T^* T)^n P) \). So \( E^{(n,p)}(\{0\}) \neq 0 \) for all natural number \( n \) and all \( p > 0 \). There exists a vector \( x_0 \in \mathcal{H} \) with \( \|x_0\| = 1 \) such that \( \|E^{(n,p)}(\{0\}) x_0\| > \theta_0 > 0 \). By the definition of extremal value, \( \lambda_n(\{0\}, x_0, p) = 0 \). This contradiction proves this implication. \( \blacksquare \)

4. Examples

Let \( \alpha : \alpha_0, \alpha_1, \ldots \) be a weight sequence of positive real numbers. Let \( W_\alpha \) be a hyponormal weighted shift with a weight sequence \( \alpha \). Let \( \{e_i\}_{i=0}^\infty \) be an orthonormal basis for \( \mathcal{H} \). We may consider \( \mathcal{H} \) as \( l^2(\mathbb{Z}_+) \), where \( l^2(\mathbb{Z}_+) \) is the set of square summable sequences in \( \mathbb{C} \). Let \( W_\alpha \) be a weighted shift with a weight sequence \( \alpha \) and let \( E^{(n,p)} \) be the spectral measure corresponding to \((W^*_\alpha W_\alpha)^n P \). Let \( E^{(n,p)}_{\epsilon_i, \epsilon_j} \) be the scalar valued spectral measure on \([0, \|W_\alpha^n\|^{2p}] \) which is defined by \( E^{(n,p)}_{\epsilon_i, \epsilon_j}(\Delta) = \langle E^{(n,p)}(\Delta) e_i, e_j \rangle \) for every Borel set \( \Delta \in \mathcal{B}(\omega, \|W_\alpha^n\|^{2p}) \). Then, since \( \langle \int f(\lambda) dE(\lambda) x, y \rangle = \int f(\lambda) dE_{x,y} \),

\[
\int_{[0,\|W_\alpha^n\|^{2p}]} \lambda^m dE^{(n,p)}_{\epsilon_i, \epsilon_j} = \langle (W^*_\alpha W_\alpha)^{mp} e_i, e_j \rangle.
\]

**Example 4.1.** Let \( W_\alpha \) be a weighted shift with a weight sequence \( \alpha \). Since 

\[
(W^*_\alpha W_\alpha)^{mp} = \text{Diag}\{\alpha_0^{2mp}, \ldots, \alpha_{n-1}^{2mp}, \alpha_1^{2mp}, \ldots, \alpha_n^{2mp}, \ldots, \alpha_i^{2mp}, \ldots, \alpha_{n+i-1}^{2mp}, \ldots\},
\]

we have that for each \( m \in N \),

\[
\int_{[0,\|W_\alpha^n\|^{2p}]} \lambda^m dE^{(n,p)}_{\epsilon_i, \epsilon_j} = \langle (W^*_\alpha W_\alpha)^{mp} e_i, e_j \rangle = \begin{cases} \alpha_i^{2mp} \ldots \alpha_{n+i-1}^{2mp} & (i = j) \\ 0 & (i \neq j) \end{cases}.
\]

(4.1)

**Example 4.2.** Let \( W_\alpha \) be a Bergmann shift (i.e. \( \alpha : \sqrt{\frac{1}{2}}, \sqrt{\frac{2}{3}}, \sqrt{\frac{3}{4}}, \ldots \)). By Example 4.1., we have

\[
\int_0^1 \lambda^m dE^{(n,p)}_{\epsilon_i, \epsilon_j}(\lambda) = \frac{(i + 1) \ldots (i + n + 1)}{(i + 2) \ldots (i + n + 1 + m - 1)}.
\]
Hence for each \( m \in \mathbb{N} \),
\[
\int_0^1 \lambda^m dE^{(n,p)}_{\alpha_i,\alpha_j}(\lambda) = \begin{cases} 
\frac{(i+1)^{mp}}{n+i+1} & (i = j) \\
0 & (i \neq j) 
\end{cases}.
\]

By Example 4.1., we have
\[
\int_0^1 \lambda^m dE^{(n,p)}_{\alpha_i,\alpha_j}(\lambda) = \frac{2^{mp} \alpha_i \alpha_{i+1} \cdots \alpha_{n+i-1}}{(i+1)(i+2)(i+3) \cdots (n+i+1)} \frac{1}{n+i+1}.
\]

**Example 4.3.** Let \( W_\alpha \) be a unilateral shift of multiplicity one. Since \( \|W_\alpha\|^{2p} = 1 \), by Example 4.1 we have
\[
\int_{[0,\|W_\alpha\|^{2p}]} \lambda^m dE^{(n,p)}_{\alpha_i,\alpha_j}(\lambda) = \int_0^1 \lambda^m dE^{(n,p)}_{\alpha_i,\alpha_j}(\lambda) = \delta_{ij},
\]
where \( \delta_{ij} \) is the usual notation.

**References**


Department of Mathematics, College of National Sciences, Kyungpook National University, Daegu 702-701, Korea.
ANALYSIS OF FIRST-ORDER SYSTEM LEAST-SQUARES FOR THE OPTIMAL CONTROL PROBLEMS FOR THE NAVIER-STOKES EQUATIONS

YOUNGMI CHOI*, SANG DONG KIM*, HYUNG-CHUN LEE†, AND EYEONG-CHUN SHIN‡

ABSTRACT. First-order least-squares method of a distributed optimal control problem for the incompressible Navier-Stokes equations is considered. An optimality system for the optimal solution are reformulated to the equivalent first-order system by introducing velocity-flux variables and then the least-squares functional corresponding to the system is defined in terms of the sum of the squared $L^2$ norm of the residual equations of the system. The optimal error estimates for least-squares finite element approximations are obtained.

1. INTRODUCTION

In [3], Bochev, Cai, Manteuffel, and MacCormick developed first-order system least-squares functionals for formulation of the incompressible Navier-Stokes equations. They recast the Navier-Stokes equations as a first-order system by introducing a velocity-flux variable. A least-squares principle based on $L^2$-norm applied to this first-order system and optimal discretization error estimates are obtained.

The goal of this paper is to extend this methodology to the optimal control problem for the Navier-Stokes equations in two and three dimensions. We first obtain a coupled optimality system related to two Navier-Stokes type equations associated with state variables and adjoint variables. The optimality system may be written as first-order system of partial differential equations by introducing velocity-flux variables. The Euler-Lagrange equations for the corresponding least-squares principle are then recast in the canonical form. This allows us to apply conventional abstract theory and our results to obtain optimal error estimates for least-squares finite element method.

The optimal control problem we consider is to minimize the functional

\[(1.1) \quad J(u, p, f) = \frac{1}{2} \| u - u_d \|^2 + \frac{\beta}{2} \| f \|^2,\]

This work was supported by Korea Research Foundation Grant (KRF-2002-070-C0014).

*Department of Mathematics, Kyungpook National University, Daegu 702-701, Korea.
†Department of Mathematics, Ajou University, Suwon 443-749, Korea.
‡Department of Mathematics, Chonnam National University, Gwangju 500-757, Korea.
subject to the incompressible Navier-Stokes equations

\begin{align}
(1.2) \quad -\Delta \mathbf{u} + \frac{1}{\nu} (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p &= \mathbf{f} \quad \text{in } \Omega, \\
(1.3) \quad \nabla \cdot \mathbf{u} &= 0 \quad \text{in } \Omega, \\
(1.4) \quad \mathbf{u} &= 0 \quad \text{on } \partial \Omega,
\end{align}

where \( \mathbf{u}_d \) is a given desired function. Here, \( \Omega \subset \mathbb{R}^n (n = 2 \text{ or } 3) \) be an open, connected, and bounded domain with Lipschitz boundary \( \partial \Omega = \Gamma \) and \( \mathbf{u} \) a candidate velocity field, \( p \) the pressure, \( \mathbf{f} \) a prescribed forcing term and \( \nu \) the viscous constant. Assume that \( p \) satisfies the zero mean constraint, \( \int_{\Omega} p \, dx = 0. \)

The objective of this optimal control problem is to seek a state variables \( \mathbf{u} \) and \( p \), and the control \( \mathbf{f} \) which minimize the \( L^2 \)-norm distances between \( \mathbf{u} \) and \( \mathbf{u}_d \) and satisfy (1.2)–(1.4). The second term in (1.1) is added as a limiting the cost of control and the positive penalty parameter \( \delta \) can be used to change the relative importance of the two terms appearing in the definition of the functional.

The plan of the paper is as follows. In the next section, we give a precise statement of the optimization problem. Then we reformulate the optimality systems to the first-order system and define the \( L^2 \)-norm least squares functional. In §3, we obtain the optimal error estimates for least-squares finite element method for the optimality system.

1.1. Notations. The standard Sobolev spaces \( H^m(\Omega) \) and \( H^m_0(\Omega) \) will be used with the associated standard inner products \( (\cdot, \cdot)_m \) and their respective norms \( \| \cdot \|_m \). In particular, for \( m = 0 \) we replace \( H^m(\Omega) \) by \( L^2(\Omega) \) with the norm \( \| \cdot \| \) and inner product \( (\cdot, \cdot) \), and denote \( L^2_0(\Omega) \) as the subspace of square integrable functions with zero mean. For positive values of \( m \) the space \( H^{-m}(\Omega) \) is defined as the dual space of \( H^m_0(\Omega) \) equipped with the norm \( \| \cdot \|_{-m} = \sup_{\| \phi \|_m \leq 1} \langle \phi, \cdot \rangle \) where \( \langle \cdot, \cdot \rangle \) is the duality pairing between \( H^{-m}(\Omega) \) and \( H^m_0(\Omega) \). Define the product spaces \( H^m_n(\Omega)^d = \prod_{i=1}^d H^m_0(\Omega) \) and \( H^{-m}(\Omega)^d = \prod_{i=1}^d H^{-m}(\Omega) \) with standard product norms. All subspaces are equipped with the norms inherited from the corresponding underlying spaces. Throughout the paper, we use boldface lower case font to denote vectors and underline boldface upper case font to denote matrices.

2. The optimal control problem

2.1. The optimization problem. Let \( \mathbf{u} \in H^1_0(\Omega)^n \) and \( p \in L^2(\Omega) \) denote the state variables, and let \( \mathbf{f} \in H^{-1}(\Omega)^n \) denote the distributed control. The state and control variables are also constrained to satisfy the system (1.2)–(1.4), which recast into the weak form:

\begin{align}
(2.5) \quad a(\mathbf{u}, \mathbf{w}) + \frac{1}{\nu} c(\mathbf{u}, \mathbf{u}, \mathbf{w}) - b(\mathbf{w}, p) &= \langle \mathbf{f}, \mathbf{w} \rangle \quad \forall \mathbf{w} \in H^1(\Omega)^n, \\
(2.6) \quad b(\mathbf{u}, \mathbf{q}) &= 0 \quad \forall \mathbf{q} \in L^2(\Omega),
\end{align}
where
\[
\begin{align*}
    a(u, w) &= \int_{\Omega} \nabla u : \nabla w \, dx = \frac{1}{2} \int_{\Omega} (\nabla u + \nabla u^T) : (\nabla w + \nabla w^T) \, dx, \\
    b(w, p) &= \int_{\Omega} p \nabla \cdot w \, dx, \\
    c(u, v, w) &= \int_{\Omega} (u \cdot \nabla) v \cdot w \, dx.
\end{align*}
\]

With \( J(\cdot) \) given by (1.1), the admissibility set \( U_{ad} \) is defined by
\[
(2.7) \quad U_{ad} = \{(u, p, f) \in H^1_0(\Omega)^n \times L^2_0(\Omega)^n \times H^{-1}(\Omega)^n : J(u, p, f) < \infty \text{ and } (u, p, f) \text{ satisfies (2.5) and (2.6)}\}.
\]

Then \((\hat{u}, \hat{p}, \hat{f}) \in U_{ad}\) is called an optimal solution if there exists \( \epsilon > 0 \) such that
\[
J(\hat{u}, \hat{p}, \hat{f}) \leq J(u, p, f) \quad \forall (u, p, f) \in U_{ad}
\]

satisfying
\[
\|\hat{u} - u\|_1 + \|\hat{p} - p\| + \|\hat{f} - f\| < \epsilon.
\]

The optimal control problem can now be formulated as a constrained minimization in a Hilbert space
\[
(2.8) \quad \min_{(u, p, f) \in U_{ad}} J(u, p, f)
\]

2.2. An optimality system. From the Lagrangian
\[
L(u, p, f, v, q : u_d) = J(u, p, f) - (\Delta u - \frac{1}{\nu}(u \cdot \nabla)u - \nabla p + f, v) - (\nabla \cdot u, q)
\]

where \( J(\cdot, \cdot, \cdot) \) is defined by (1.1), one may derive an optimality system of equations for the solution of (2.8). The constrained problem (2.8) can now be recast as the unconstrained problem of finding stationary points of \( L(\cdot) \). We now apply the necessary conditions for the latter problem. Clearly, setting to zero the first variations with respect to \( u, p, f, v \) and \( q \) yields the optimality system
\[
(2.9) \quad \begin{cases}
    -\frac{1}{\nu}(u \cdot \nabla)u - \nabla p + f, \nabla \hat{v} = 0, & \forall \hat{v} \in H^1_0(\Omega)^n, \\
    -\nabla \cdot \hat{u} = 0, & \forall \hat{q} \in L^2_0(\Omega), \\
    u = 0, & \text{on } \Gamma, \\
    \hat{u} = 0, & \text{in } \Omega, \\
    (u - u_d, \hat{u}) - \frac{1}{\nu}(\hat{u} \cdot \nabla)u - \frac{1}{\nu}(u \cdot \nabla)\hat{u}, v = 0, & \hat{u} \in H^1_0(\Omega)^n, \\
    -\nabla \cdot \hat{v} = 0, & \hat{p} \in L^2_0(\Omega), \\
    v = 0, & \text{on } \Gamma, \\
    (\hat{f} - v, \hat{f}) = 0, & \hat{f} \in H^{-1}(\Omega)^n.
\end{cases}
\]
The strong form of the optimality system is as follows.

\[
\begin{align*}
\begin{cases}
-\Delta u + \frac{1}{\nu}(u \cdot \nabla)u + \nabla p &= f \quad \text{in } \Omega, \\
\nabla \cdot u &= 0 \quad \text{in } \Omega, \\
u \cdot u &= 0 \quad \text{on } \Gamma,
\end{cases}
\end{align*}
\]

\[
(u - u_d) - \Delta v + \frac{1}{\nu}(\nabla u)^T v - \frac{1}{\nu}(u \cdot \nabla) v + \nabla q = 0 \quad \text{in } \Omega,
\]

\[
\nabla \cdot v = 0 \quad \text{in } \Omega, \\
v &= 0 \quad \text{on } \Gamma, \\
\beta f &= v \quad \text{in } \Omega.
\]

Note that this system is coupled, i.e., the constraint equations for the state variables depend on the unknown controls, the adjoint equations for the Lagrange multipliers depend on the state, and optimality conditions for the controls depend on the Lagrange multipliers.

2.3. **First-order system.** To formulate the least-squares method, system (2.10) will be transformed into an equivalent first-order system. Introduce the velocity-flux variable

\[U = \nabla u^T\]

which is a matrix with entries \(U_{ij} = \partial u_j / \partial x_i, 1 \leq i, j \leq n\). Then

\[(\nabla^T U)^T = \Delta u\]

and it is easy to see that the new variable satisfies the identities

\[\text{tr} U = 0, \quad \nabla \times U = 0 \quad \text{in } \Omega\]

and

\[n \times U = 0 \quad \text{on } \Gamma\]

where \(\text{tr} U = \sum_{i=1}^{n} U_{ii}\) and \(n\) is the outward unit normal on \(\Gamma\).
The optimality condition (the last equation in (2.10)) can be substituted into the state equations and thus, we have the first-order optimality system

\begin{align*}
-(\nabla'U)^t + \frac{1}{\nu'} U' u + \nabla p &= \frac{v}{\beta} & \text{in } \Omega, \\
\nabla' u &= 0 & \text{in } \Omega, \\
U - \nabla u^* &= 0 & \text{in } \Omega, \\
\nabla (\text{tr} U) &= 0 & \text{in } \Omega, \\
\nabla \times U &= 0 & \text{in } \Omega, \\
u &= 0 & \text{on } \Gamma, \\
\int p \, dx &= 0 & \text{in } \Omega, \\
n \times U &= 0 & \text{on } \Gamma, \\
(\mathbf{u} - \mathbf{u}_d) - (\nabla' \mathbf{V})^t + \frac{1}{\nu'} U' \mathbf{V} - \frac{1}{\nu'} \nabla q &= 0 & \text{in } \Omega, \\
\nabla' \mathbf{V} &= 0 & \text{in } \Omega, \\
\mathbf{V} - \nabla \mathbf{v}^* &= 0 & \text{in } \Omega, \\
\nabla (\text{tr} \mathbf{V}) &= 0 & \text{in } \Omega, \\
\nabla \times \mathbf{V} &= 0 & \text{in } \Omega, \\
\mathbf{v} &= 0 & \text{on } \Gamma, \\
\int q \, dx &= 0 & \text{in } \Omega, \\
n \times \mathbf{V} &= 0 & \text{on } \Gamma.
\end{align*}

\textbf{(2.11)}

3. Least-Squares finite element method

3.1. Least-Squares. The $L^2$ least-squares functional for first-order system (2.11) is defined as follows:

\begin{align*}
\mathcal{J}_1(\mathbf{U}, u, p, \mathbf{V}, v, q : \mathbf{u}_d) &= \| - (\nabla' \mathbf{U})^t + \frac{1}{\nu'} U' \mathbf{u} + \nabla p - \frac{v}{\beta} \|^2 + \| \nabla' \mathbf{u} \|^2 + \| U - \nabla u^* \|^2 + \| \nabla (\text{tr} U) \|^2 \\
&+ \| \nabla \times \mathbf{U} \|^2 + \| - (\nabla' \mathbf{V})^t + \frac{1}{\nu'} U' \mathbf{V} - \frac{1}{\nu'} \nabla q + \mathbf{u} - \mathbf{u}_d \|^2 + \| \nabla' \mathbf{V} \|^2 \\
&+ \| \mathbf{V} - \nabla \mathbf{v}^* \|^2 + \| \nabla (\text{tr} \mathbf{V}) \|^2 + \| \nabla \times \mathbf{V} \|^2.
\end{align*}

(3.12)

To define the least-squares method, we need a suitable minimization problem.
Let
\[ X := \left[ H^1(\Omega)^n \times H^1(\Omega)^n \times [H^1(\Omega) \cap L_0^2(\Omega)] \right]^2 \]
and let \( X_0 \) be a subspace of \( X \):

\[
X_0 := \{ (U, u, p, V, v, q) \in X \mid u = 0, n \times U = 0, v = 0, n \times V = 0, \text{ on } \Gamma \}. 
\]

Then the least-squares principle is to find \((U, u, p, V, v, q) \in X_0\) such that
\[
J_1(U, u, p, V, v, q : u_\delta) = \inf_{(\tau, w, r, \psi, x, \xi : x \in X_0)} J_1(\tau, w, r, \psi, x, \xi : u_\delta). 
\]

It is easy to see that the Euler-Lagrange equation for this minimization problem is given by the variational problem:

find \((U, u, p, V, v, q) \in X_0\) such that

\[
B\left((U, u, p, V, v, q), (\bar{U}, \bar{u}, \bar{p}, \bar{V}, \bar{v}, \bar{q})\right) = F\left(\bar{U}, \bar{u}, \bar{p}, \bar{V}, \bar{v}, \bar{q}\right) 
\]

where

\[
B\left((U, u, p, V, v, q), (\bar{U}, \bar{u}, \bar{p}, \bar{V}, \bar{v}, \bar{q})\right) 
= \left(- (\nabla^2 U)^t + \frac{1}{\nu} U^t u + \nabla_p - \frac{v}{\beta}, -(\nabla^2 \bar{U})^t + \frac{1}{\nu} \bar{U}^t \bar{u} + \nabla_{\bar{p}} - \frac{\bar{v}}{\bar{\beta}} \right) 
+ \left( \nabla^2 u, \nabla^2 \bar{u} \right) + \left( U - \nabla u, \bar{U} - \nabla \bar{u} \right) + \left( \nabla(\text{tr} U), \nabla(\text{tr} \bar{U}) \right) + \left( \nabla \times U, \nabla \times \bar{U} \right) 
+ \left( - (\nabla^2 V)^t + \frac{1}{\nu} U v - \frac{1}{\nu} V^t u + \nabla_q + u, -(\nabla^2 \bar{V})^t + \frac{1}{\nu} \bar{U} \bar{v} - \frac{1}{\nu} \bar{V}^t \bar{u} - \frac{1}{\nu} \bar{V}^t \bar{u} + \nabla q + \bar{u} \right) 
+ \left( \nabla \bar{V}, \nabla \bar{V} \right) + \left( V - \nabla \bar{V}, \bar{V} - \nabla \bar{V} \right) + \left( \nabla(\text{tr} V), \nabla(\text{tr} \bar{V}) \right) + \left( \nabla \times V, \nabla \times \bar{V} \right) 
\]

and

\[
F\left(\bar{U}, \bar{u}, \bar{p}, \bar{V}, \bar{v}, \bar{q}\right) = \left(u_\delta, -(\nabla^2 \bar{V})^t + \frac{1}{\nu} \bar{U} \bar{v} - \frac{1}{\nu} \bar{V}^t \bar{u} - \frac{1}{\nu} \bar{V}^t \bar{u} + \nabla q + \bar{u} \right). 
\]

for all \((\bar{U}, \bar{u}, \bar{p}, \bar{V}, \bar{v}, \bar{q}) \in V\).

Let \( X_h \) denote a finite-dimensional subspace of \( X_0 \). Then the least-squares discretization method of the optimal control problem for the Navier-Stokes equations is defined by the following discrete variational problem:

find \((U^h, u^h, p^h, V^h, v^h, q^h) \in X_h\) such that
(3.15) \[ B \left( (U^h, \bar{u}^h, p^h, \nabla^h, \bar{V}^h, \bar{V}^h, \bar{V}^h, \bar{q}^h) \right) = F \left( \bar{U}^h, \bar{u}^h, p^h, \bar{V}^h, \bar{V}^h, \bar{q}^h, \bar{q}^h \right) \]
for all \((\bar{U}^h, \bar{u}^h, p^h, \bar{V}^h, \bar{V}^h, \bar{q}^h, \bar{q}^h) \in X_h\)

It is easy to see that the discrete variational problem (3.15) corresponds to the necessary condition for the following discrete least-squares principle for (3.12):

find \((U^h, u^h, p^h, V^h, q^h) \in X_h\) such that

(3.16) \[ F_1(U^h, u^h, p^h, V^h, q^h : u_d) \leq F_1(\bar{U}^h, \bar{u}^h, \bar{p}^h, \bar{V}^h, \bar{q}^h : u_d) \]
for all \((\bar{U}^h, \bar{u}^h, \bar{p}^h, \bar{V}^h, \bar{V}^h, \bar{q}^h) \in X_h\)

For space \(X_h\), we assume the following approximation property: there exists an integer \(d \geq 1\) such that, for all \(U \in H^{d+1}(\Omega)^n\), \(u \in H^{d+1}(\Omega)^n\), \(p \in H^{d+1}(\Omega)\), \(V \in H^{d+1}(\Omega)^n\), \(v \in H^{d+1}(\Omega)^n\) and \(q \in H^{d+1}(\Omega)\), one can find \((\bar{U}^h, \bar{u}^h, \bar{p}^h, \bar{V}^h, \bar{v}^h, \bar{q}^h) \in X_h\) such that

(3.17) \[ \| U - U^h \|_\mu + \| u - \bar{u}^h \|_\mu + \| p - \bar{p}^h \|_\mu + \| V - \bar{V}^h \|_\mu + \| v - \bar{v}^h \|_\mu + \| q - \bar{q}^h \|_\mu \leq Ch^{d+1-\mu} \left( \| U \|_{d+1} + \| u \|_{d+1} + \| p \|_{d+1} + \| V \|_{d+1} + \| v \|_{d+1} + \| q \|_{d+1} \right) , \]

\(\mu = 0, 1\). Note, for example, that (3.17) can be satisfied with \(d = 1\) by choosing continuous piecewise linear functions for all variables.

3.2. Discretization error estimates. The main goal of this section is to derive error estimates for least-squares method (3.15). For this purpose, we show how to cast nonlinear problems (3.14) and (3.15) in the respective canonical forms

(3.18) \[ F(\lambda, \mu) \equiv \mu + T \cdot G(\lambda, \mu) = 0 \]
and

(3.19) \[ F^h(\lambda, \mu) \equiv \mu + T_h \cdot G(\lambda, \mu) = 0. \]

The following function spaces will be needed below (with \(m\) representing some nonnegative integer):

(3.20) \[ X_m^h = \left[ H^{m+1}(\Omega)^n \times H^{m+1}(\Omega)^n \times H^{m+1}(\Omega) \right]^2 \cap X_0, \]

(3.21) \[ Y = X_0^h, \]

(3.22) \[ Z = \left[ L^{3/2}(\Omega)^n \times L^{3/2}(\Omega)^n \times L^{3/2}(\Omega) \right]^2, \]

where \(X_0^h\) denotes the dual of \(X_0\) with respect to the \(L^2\) inner product.

We make identifications \(\mu = (U, u, p, V, q), \mu_0 = (U^h, \bar{u}^h, p^h, \bar{V}^h, \bar{q}^h)\), \(V = (\bar{U}, \bar{u}, \bar{p}, \bar{V}, \bar{V}, \bar{q}^h)\) and \(\lambda = \frac{1}{\nu}\), and we assume that \(\lambda \in \Lambda\), where \(\Lambda\) is a compact subset of \(\mathbb{R}^+\). We then introduce the following:
Define $T : Y \mapsto X_0$ through $u = Tg$ for $g \in Y$ if and only if
\begin{align}
B_S(u, v) &= \left( - (\nabla^4 U)^t - \nabla p - \frac{v}{\beta} - \left( \nabla^4 U \right)^t + \nabla \bar{p} - \frac{v}{\beta} \right) + \left( \nabla^4 u, \nabla^4 \bar{u} \right) \\
&+ \left( U - \nabla u^t, \bar{U} - \nabla \bar{u}^t \right) + \left( \nabla (\text{tr} U), \nabla (\text{tr} \bar{U}) \right) + \left( \nabla \times U, \nabla \times \bar{U} \right) \\
&\quad + \left( - (\nabla^t V)^t + \nabla q + \left( u - u_0 \right), -(\nabla^t \bar{V})^t + \nabla \bar{q} + \bar{u} \right) + \left( \nabla V, \nabla \bar{V} \right) \\
&\quad + \left( - (\nabla^t \bar{V})^t + \nabla \bar{q} + \left( u - u_0 \right), -(\nabla^t V)^t + \nabla q + \bar{u} \right) + \left( \nabla \bar{V}, \nabla V \right)
\end{align}
(3.23)
for all $(\bar{U}, \bar{u}, \bar{p}, \bar{V}, \bar{v}, \bar{q}) \in X_0$. Define $T_h : Y \mapsto X_h$ through $u^h = Tg$ for $g \in Y$ if and only if
\begin{align}
B_S(u^h, v^h) &= (g_1, \bar{U}^h) + (g_2, \bar{u}^h) + (g_3, \bar{p}^h) + (g_4, \bar{V}^h) + (g_5, \bar{v}^h) + (g_6, \bar{q}^h)
\end{align}
(3.24)
for all $(\bar{U}^h, \bar{u}^h, \bar{p}^h, \bar{V}^h, \bar{v}^h, \bar{q}^h) \in X_h$.

We also define $G : \Lambda \times X_0 \to Y$ through $\mathbf{g} = G(\lambda, \mathbf{u})$ for $\mathbf{u} \in X_0$ if and only if
\begin{align}
(g_1, \bar{U}) + (g_2, \bar{u}) + (g_3, \bar{p}) + (g_4, \bar{V}) + (g_5, \bar{v}) + (g_6, \bar{q})
\end{align}
\begin{align}
&= \left( - (\nabla^4 U)^t + \nabla p - \frac{v}{\beta} \right) + \left( - (\nabla^4 U)^t + \nabla \bar{p} - \frac{v}{\beta} \right) + \left( - (\nabla^t V)^t + \nabla q + \left( u - u_0 \right), \nabla \bar{V} - \frac{v}{\beta} \right) \\
&\quad + \left( - (\nabla^t \bar{V})^t + \nabla \bar{q} + \left( u - u_0 \right), \nabla V - \frac{v}{\beta} \right) \\
&\quad + \left( - (\nabla^t \bar{V})^t + \nabla \bar{q} + \left( u - u_0 \right), \nabla V - \frac{v}{\beta} \right) \\
&\quad + \left( - (\nabla^4 U)^t + \nabla p - \frac{v}{\beta} \right) + \left( - (\nabla^4 U)^t + \nabla \bar{p} - \frac{v}{\beta} \right)
\end{align}
for all $(\bar{U}, \bar{u}, \bar{p}, \bar{V}, \bar{v}, \bar{q}) \in X_0$.

**Lemma 3.1.** Assume that $T$, $T_h$, and $G$ are defined by (3.23), (3.24), and (3.25), respectively. Then nonlinear problem (3.14) is equivalent to (3.18) and discrete nonlinear problem (3.15) is equivalent to (3.19).

**Proof.** Assume that $\mathbf{u} = (U, u, p, V, v, q)$ solves problem (3.18) with $T$ and $G$ given by (3.23) and (3.25), respectively. Then $\mathbf{u} = - Tg$ if and only if
\begin{align}
B_S(u, v) = (g, v)
\end{align}
for all $v \in X_0$
and $\mathbf{g} = G(\lambda, \mathbf{u})$ if and only if (3.25) holds. It follows that $\mathbf{u}$ also solves variational problem (3.14). Conversely, if $\mathbf{u}$ solves (3.14), let $\mathbf{g}$ be defined by (3.25). Then $B_S(u, v) = (g, v)$ for all $v \in X_0$, i.e., $\mathbf{u} = - Tg$. Thus, (3.14) and (3.18) are equivalent. Proof of the equivalence of (3.15) and (3.19) is identical. \(\square\)
Error estimates for least-squares method (3.15) will now be derived from the abstract approximation theory of [8]. Below we state the main result of this theory for general $T$ and $T_h$, but otherwise specialized to our needs. Here we let $D_\lambda G(\lambda, U)$ and $D_\lambda F(\lambda, U)$ denote the Fréchet derivative of $G$ and $F$ with respect to $U$. We refer to $\{(\lambda, U(\lambda))|\lambda \in \Lambda\}$ as a regular branch of solutions of (3.18) if $U = U(\lambda)$ is a weak solution of (3.18) for each $\lambda \in \Lambda$, $\lambda \mapsto U(\lambda)$ is a continuous map $\Lambda \mapsto X_b$, and $D_\lambda F(\lambda, U)$ is an isomorphism of $X_b$.

**Theorem 3.1.** Let $F(\lambda, U) = 0$ denote abstract form (3.18) and assume that $\{(\lambda, U(\lambda))|\lambda \in \Lambda\}$ is a branch of regular solutions of (3.18). Furthermore, assume that $T \in L(Y, X_0)$, that $G$ is a $C^2$ map $\Lambda \times X_0 \mapsto Y$ such that all second derivatives of $G$ are bounded on bounded subsets of $\Lambda \times X_0$, and that there exists a space $Z \subset Y$, with continuous imbedding, such that $D_\lambda G(\lambda, U) \in L(X_0, Z)$ for all $\lambda \in \Lambda$ and $U \in X_0$. If approximate problem (3.19) is such that

$$\lim_{h \to 0} \| (T - T_h) g \|_{X_0} = 0$$

for all $g \in Y$ and

$$\lim_{h \to 0} \| (T - T_h) h \|_{L(Z, X_0)} = 0.$$

Then:

1. there exists a neighborhood $O$ of the origin in $X_0$ and, for $h$ sufficiently small, a unique $C^2$ function $\lambda \mapsto U^h(\lambda) \in X_b$ such that $\{(\lambda, U^h(\lambda))|\lambda \in \Lambda\}$ is a branch of regular solutions of discrete problem (3.19) and $U(\lambda) - U^h(\lambda) \in O$ for all $\lambda \in \Lambda$;

2. for all $\lambda \in \Lambda$ we have

$$\| U^h(\lambda) - U(\lambda) \|_{X_0} \leq C \| (T - T_h) G(\lambda, U(\lambda)) \|_{X_0};$$

3. if the regular branch is such that $U(\lambda) \in X_0^n$ for some integer $m \geq 1$ and $d = \min\{d, m\}$, where $d$ is the largest integer satisfying (3.17), then

$$\| U(\lambda) - U^h(\lambda) \|_1 + \| u(\lambda) - u^h(\lambda) \|_1 + \| p(\lambda) - p^h(\lambda) \|_1$$

$$+ \| v(\lambda) - v^h(\lambda) \|_1 + \| q(\lambda) - q^h(\lambda) \|_1$$

$$\leq C h^d (\| U(\lambda) \|_{d+1} + \| u(\lambda) \|_{d+1} + \| p(\lambda) \|_{d+1} + \| v(\lambda) \|_{d+1} + \| q(\lambda) \|_{d+1}).$$

In the next few lemmas, we verify the hypotheses of Theorem 3.1 for our least-squares formulation. We begin by establishing essential properties of operators $T$ and $T_h$, which we assume, for this and the next section, are defined by (3.23) and (3.24), respectively.

**Lemma 3.2.** $T \in L(Y, X_0)$ and $T_h \in L(Y, X_0)$.

**Proof.** From $B_S(\cdot, \cdot)$ is continuous and coercive on $X_0 \times X_0$ (see [9], $B_S(\cdot, \cdot)$ is equivalent to functional $G_2$ of [9]) and, by virtue of the inclusion $X_b \subset X_0$, it is also continuous and coercive on $X_b \times X_b$. 


Furthermore, for each $g \in Y$, $(g, v)$ defines a continuous functional on $X_0$. Thus, the Lax-Milgram theorem implies that, for all $g \in Y$, variational problems (3.23) and (3.24) have unique respective solutions $u \in X_0$ and $u_h \in X_h$, i.e., $T : Y \mapsto X_0$ and $T_h : Y \mapsto X_h$ are well-defined linear operators. From

$$C\|u\|_{X_0}^2 \leq B_s(u, u) = (g, u) \leq \|g\|_Y \|u\|_{X_0},$$

it follows that

$$\|Tg\|_{X_0} = \|u\|_{X_0} \leq C\|g\|_Y,$$

i.e., $T$ is in $L(Y, X_0)$. The proof that $T_h \in L(Y, X_h)$ is similar. \hfill \Box

Before continuing with the approximation properties of $T_h$, consider the choice of $Y$ and $Z$ in (3.21) and (3.22). When $Z \subset Y$ with compact imbedding, the proof of (3.26) in Theorem 3.1 can be simplified. Since $L^{n/2}(\Omega)$ is compactly imbedded the duals of $H^1_0(\Omega)$, $H^1(\Omega) = \{v \in H^1(\Omega) | n \times v = 0 \text{ on } \Gamma\}$, and $H^1(\Omega)$, the imbedding $Z \subset Y$ is compact. (see[3])

**Lemma 3.3.** Convergence properties (3.25) and (3.26) hold. If, in addition, $g \in Y$ is such that $Tg \in X_0^m$ for some $m \geq 1$ and $d = \min(d, m)$, where $d$ is the largest integer satisfying (3.17), then

$$\|(T - T_h)g\|_{X_0} \leq Ch^d\|Tg\|_{X_0^{d+1}}.$$  \hfill (3.29)

**Proof.** It is similar to Lemma 3 in [3]. \hfill \Box

The only hypotheses of Theorem 3.1 that remain to be verified are the assumptions concerning the nonlinear operator $G$. For this purpose, we need the weak and strong forms of the first Fréchet derivative $D_1 G(\lambda, u)$ and second Fréchet derivative $D_2^3 G(\lambda, u)$. To determine the weak form of $D_1 G(\lambda, u)$, let $\tilde{u} \in X_0$, substitute $u + \tilde{u}$ into (3.25), and expand about $u$. This yields the following weak representation of $D_1 G(\lambda, u)$:
Analysis of First-Order System Least-Squares for the Optimal Control Problems for the Navier-Stokes Equations

\[ D_\varepsilon G(\lambda, \mathcal{U}) : \Lambda \times X_0 \to Y \text{ defined by } g = D_\varepsilon G(\lambda, \mathcal{U}) \mathcal{U} \text{ for } \mathcal{U} \in X_0 \text{ if and only if} \]

\[
(g_1, \vec{U}) + (g_2, \vec{u}) + (g_3, \vec{p}) + (g_4, \vec{V}) + (g_5, \vec{v}) + (g_6, \vec{q})
= \left( - (\nabla \vec{V})^s + \nabla \vec{q} + \vec{u} - u_a \right)_\nu + \left( \vec{U}^s u + \frac{1}{\nu} \vec{U}^s \vec{u} \right)_\nu + \left( \vec{U}^s u - \vec{V}^s u + \frac{1}{\nu} \vec{V}^s \vec{u} \right)_\nu + \left( \vec{U}^s u - \vec{V}^s u + \nabla \vec{q} + \vec{u} \right)
\]

\[(3.30)\]

for all \((\vec{U}, \vec{u}, \vec{p}, \vec{V}, \vec{v}, \vec{q}) \in X_0.\)

The strong form of \(D_\varepsilon G(\lambda, \mathcal{U}) \mathcal{U}\) can be found from (3.30) using standard integration by parts:

\[
g_1 = \frac{1}{\nu} \left( \vec{U}^s u + \nabla \vec{q} + \vec{u} - u_a \right)_\nu + \left( \vec{U}^s u - \vec{V}^s u + \frac{1}{\nu} \vec{V}^s \vec{u} \right)_\nu + \left( \vec{U}^s u - \vec{V}^s u + \nabla \vec{q} + \vec{u} \right)
\]

\[(3.31)\]

\[
g_2 = \frac{1}{\nu} \left( \vec{U}^s u + \nabla \vec{q} + \vec{u} - u_a \right)_\nu + \left( \vec{U}^s u - \vec{V}^s u + \frac{1}{\nu} \vec{V}^s \vec{u} \right)_\nu + \left( \vec{U}^s u - \vec{V}^s u + \nabla \vec{q} + \vec{u} \right)
\]

\[(3.32)\]

\[
g_3 = - \nabla \left( \vec{U}^s u + \vec{V}^s u \right)
\]

\[(3.33)\]
\[ g_4 = \frac{1}{\nu} u \left( - (\nabla V)^t + \nabla \hat{u} + (u - u_d) + \frac{1}{\nu} \nabla u - \frac{1}{\nu} V^t u \right)^t \]
\[ g_5 = - \frac{1}{\nu} \left( \frac{1}{\nu} U^t u + \frac{1}{\nu} U^t \hat{u} + \frac{1}{\nu} U \hat{v} + \frac{1}{\nu} \hat{v} \right) \left( - (\nabla V)^t + \nabla \hat{u} + (u - u_d) + \frac{1}{\nu} \nabla u - \frac{1}{\nu} V^t u \right) \]
\[ g_6 = - \nabla \left( \frac{1}{\nu} U^t u + \frac{1}{\nu} \hat{u} - \frac{1}{\nu} V^t u - \frac{1}{\nu} \hat{v} u \right) \]
for all \((\hat{U}, \hat{u}, \hat{\hat{u}}, \hat{V}, \hat{v}, \hat{\hat{v}}) \in X_0\).

Finally, the weak form of the second Fréchet derivative is
\[ D^2 \hat{g}(\lambda, \hat{u}) : \Lambda \times [X_0 \times X_0] \to Y \text{ defined by } g = D^2 \hat{g}(\lambda, \hat{u})[\hat{u}, \hat{\hat{u}}] \text{ for } \hat{u} \in X_0 \text{ if and only if} \]
\[ (g_1, \hat{U}) + (g_2, \hat{u}) + (g_3, \hat{\hat{u}}) + (g_4, \hat{V}) + (g_5, \hat{v}) + (g_6, \hat{\hat{v}}) \]
\[ = \left( - (\nabla^t \hat{U})^t + \nabla \hat{\hat{u}} + \frac{1}{\nu} U^t \hat{u} + \frac{1}{\nu} U^t \hat{\hat{u}} - \frac{1}{\nu} U^t \hat{\hat{u}} \right) \]
\[ \left( - (\nabla^t \hat{V})^t + \nabla \hat{\hat{v}} + \frac{1}{\nu} U^t \hat{v} + \frac{1}{\nu} U^t \hat{\hat{v}} - \frac{1}{\nu} U^t \hat{\hat{v}} \right) \]
\[ \left( - (\nabla^t \hat{V})^t + \nabla \hat{\hat{v}} + \frac{1}{\nu} U^t \hat{v} + \frac{1}{\nu} U^t \hat{\hat{v}} - \frac{1}{\nu} U^t \hat{\hat{v}} \right) \]
for all \((\hat{U}, \hat{u}, \hat{\hat{u}}, \hat{V}, \hat{v}, \hat{\hat{v}}) \in X_0\).

The next lemma summarizes technical results that we use below.

**Lemma 3.4.** Let \( D_t \) denote the derivative with respect to the \( i \)-th coordinate variable in \( \mathbb{R}^n \), \( 1 \leq i \leq n \), and assume that \( u, v, w, \) and \( z \) are in \( H^1(\Omega) \). Then
\[ \int_{\Omega} D_t u v w d\Omega \leq C \|u\|_1 \|v\|_1 \|w\|_1, \]
\[ 1 \leq i \leq n, \text{ and} \]
\[
\left| \int_{\Omega} u v w z \, d\Omega \right| \leq C \| u \|_1 \| v \|_1 \| w \|_1 \| z \|_1.
\]

Moreover, \((u, v) \mapsto uv\) is a continuous bilinear mapping from \(L^2(\Omega) \times H^1(\Omega)\) into \(L^{3/2}(\Omega)\) and \((u, v, w) \mapsto uvw\) is a continuous trilinear mapping from \(H^1(\Omega) \times H^1(\Omega) \times H^1(\Omega)\) into \(L^{3/2}(\Omega)\); i.e.,
\[
\| uv \|_{0,3/2} \leq C \| u \|_{0,2} \| v \|_{1,2} \quad \text{for all } u \in L^2(\Omega) \text{ and } v \in H^1(\Omega),
\]
\[
\| uvw \|_{0,3/2} \leq C \| u \|_{1,2} \| v \|_{1,2} \| w \|_{1,2} \quad \forall u, v, w \in H^1(\Omega).
\]

**Proof.** It is similar to Lemma 4 in [3]. The first part of the lemma follows easily from the imbedding \(H^1(\Omega) \subset L^4(\Omega)\) in two and three dimensions and the Hölder inequality. The second part follows directly from a result in [8] (see Corollary 1.1, p. 5). \(\square\)

In the next lemma, we establish properties of \(G\) that are required for the validity of the approximation result in Theorem 3.1.

**Lemma 3.5.** Assume that mapping \(G\) is defined by (3.25). For \(X_0, Y, \text{ and } Z\) given by (3.13), (3.21) and (3.22), respectively, the following are true.
1. For all \(U \in X_0, D\mu G(\lambda, U) \in L(X_0, Z).\)
2. The second Fréchet derivative \(D^2 G(\lambda, U)\) is bounded on bounded subsets of \(\Lambda \times X_0.\)

**Proof.** To prove 1, consider strong form (3.31)–(3.36) of \(D\mu G(\lambda, U).\) By assumption, \(U \in X_0;\) i.e., \(U \in H^1(\Omega)^n, u \in H^1(\Omega)^n, p \in H^1(\Omega), V \in H^1(\Omega)^n, v \in H^1(\Omega)^n,\) and \(q \in H^1(\Omega).\) Now each equation (3.31)–(3.35) and (3.36) consists of terms of the form \(D\mu uv\) and \(uvw,\) where \(u, v,\) and \(w\) belong to \(H^1(\Omega),\) so the second part of Lemma 3.4 implies that \((g_1, g_2, g_3, g_4, g_5, g_6) \in Z.\) Using (3.40) and (3.41), it also follows that
\[
\| D\mu G(\lambda, U) U \|_Z \leq C \| U \|_{X_0},
\]
\(\text{i.e., that } D\mu G(\lambda, U) \in L(X_0, Z).\)

To prove 2, consider weak form (3.37) of the second Fréchet derivative. Assume that \((\lambda, U)\) belongs to a bounded subset of \(\Lambda \times X_0\) and let \(\hat{U}, \hat{U} \in X_0\) be arbitrary. Then it is not difficult to see that weak form (3.37) involves only terms of the form \(D\mu uvw\) and \(uvw,\) where \(u, v, w,\) and \(z\) belong to
Thus, each term can be estimated using (3.38) or (3.39):

\begin{align*}
\|(g_1, \tilde{U})\| & \leq C_1(\lambda, \mathcal{U}, \mathcal{U}_0)(\|\tilde{U}\|_{X_0} + \|\tilde{U}\|_{X_0})\|\tilde{U}\|_1, \\
\|(g_2, \tilde{u})\| & \leq C_2(\lambda, \mathcal{U}, \mathcal{U}_0)(\|\tilde{u}\|_{X_0} + \|\tilde{u}\|_{X_0})\|\tilde{u}\|_1, \\
\|(g_3, \tilde{p})\| & \leq C_3(\lambda, \mathcal{U}, \mathcal{U}_0)(\|\tilde{p}\|_{X_0} + \|\tilde{p}\|_{X_0})\|\tilde{p}\|_1, \\
\|(g_4, \tilde{V})\| & \leq C_4(\lambda, \mathcal{U}, \mathcal{U}_0)(\|\tilde{V}\|_{X_0} + \|\tilde{V}\|_{X_0})\|\tilde{V}\|_1, \\
\|(g_5, \tilde{v})\| & \leq C_5(\lambda, \mathcal{U}, \mathcal{U}_0)(\|\tilde{v}\|_{X_0} + \|\tilde{v}\|_{X_0})\|\tilde{v}\|_1, \\
\|(g_6, \tilde{q})\| & \leq C_6(\lambda, \mathcal{U}, \mathcal{U}_0)(\|\tilde{q}\|_{X_0} + \|\tilde{q}\|_{X_0})\|\tilde{q}\|_1,
\end{align*}

where \(C_i\) is polynomial function of \(\lambda\), \(\|\tilde{u}\|_{X_0}\), and \(\|\mathcal{U}_0\|_{X_0}\). In combination with the fact that \(\lambda\) and \(\|\tilde{u}\|_{X_0}\) are in bounded subsets of \(\Lambda \times X_0\), and that \(\|\mathcal{U}_0\|_{\mathcal{V}}\) is fixed, it follows that \(D^2_{\mathcal{U}}G(\lambda, \mathcal{U})\) is bounded in the norm of \(L(X_0, L(X_0, \mathcal{Y}))\).

This completes verification of all assumptions of Theorem 3.1. As a result, we can conclude that error estimates (3.27) and (3.28) hold for the least-squares finite element approximation as long as problem (3.14) has a regular branch of solutions with sufficient regularity.

References

New large-update primal-dual interior point algorithms based on kernel functions for LCPs

Min-Kyung Kim\textsuperscript{*} Gyeong-Mi Cho\textsuperscript{†}

Abstract

In this paper we propose new large-update primal-dual interior point algorithms for \( P_\kappa(\kappa) \) linear complementarity problems (LCPs). New search directions and proximity measures are proposed based on a specific class of kernel functions, \( \psi(t) = \frac{\log(1+t)}{p+1} + \frac{t^{q-1}}{q} \), \( q > 0, p \in [0,1] \), which are the generalized form of the ones in [3] and [12]. It is the first to use this class of kernel functions in the complexity analysis of interior point method (IPM) for \( P_\kappa(\kappa) \) LCPs. We showed that if a strictly feasible starting point is available, then new large-update primal-dual interior point algorithms for \( P_\kappa(\kappa) \) LCPs have the best known complexity \( O((1+2\kappa)^{3/2}(\log 2\kappa)(\log \frac{\kappa}{\varepsilon})) \) when \( p = 1 \) and \( q = \frac{1}{2}(\log 2\kappa) - 1 \).

1 Introduction

In this paper we consider the following linear complementarity problem (LCP) as follows:

\[
\begin{cases}
  s = Mx + q, \\
x s = 0, \\
x \geq 0, \quad s \geq 0,
\end{cases}
\]

where \( M \in \mathbb{R}^{n \times n} \) is a \( P_\kappa(\kappa) \) matrix and \( q \in \mathbb{R}^n \).

\textsuperscript{*}Department of Mathematics, Pusan National University, Busan 609-735, Korea.

\textsuperscript{†}Corresponding author. E-mail: gcho@dongseo.ac.kr. Tel.: 82 51 320 1728. Fax: 82 51 327 8955. Department of Multimedia Engineering, Dongseo University, Busan 617-716, Korea. This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD)(KRF-2005-204-C00009).
LCPs have many applications in mathematical programming and equilibrium problems. Indeed, it is known that by exploiting the first-order optimality conditions of the optimization problem, any differentiable convex quadratic program can be formulated into a monotone linear complementarity problem, i.e. $P_+(0)$ LCP, and vice versa([15]). And variational inequality problems are widely used in the study of equilibrium in, e.g., economics, transportation planning and game theory. And variational inequality problems have a close connection to the LCPs. The reader can refer to [4] for the basic theory, algorithms and applications.

The primal-dual IPM for linear optimization (LO) problem was first introduced in [6] and [10]. Kojima et al.(6) first proved the polynomial computational complexity of the algorithm for LO problem, and since then many other algorithms have been developed based on the primal-dual strategy. Kojima, Mizuno and Yoshise ([8]) proposed a polynomial time algorithm for monotone linear complementarity problems, i.e. $P_+(0)$ LCPs. They also proposed an $O(\sqrt{n}L)$ potential reduction algorithm ([9]). The existence of a central path is very important for IPMs. Kojima et al.([7]) proved the existence of the central path for any $P_+(\kappa)$ LCP and generalized previously known results to the wider class of so called $P_+(\kappa)$ LCPs and unified interior point methods (IPMs) for LCPs. Since then an interior point algorithm’s quality is measured by the fact whether it can be generalized to $P_+(\kappa)$ LCPs or not([5]). Miao ([11]) extended the Mizuno-Todd-Ye predictor-corrector method to $P_+(\kappa)$ LCPs. His algorithm uses the $l_2$ neighborhood of the central path and has $O((1 + \kappa)\sqrt{n})$ iteration complexity. Recently, Illés and Nagy([5]) give a version of Mizuno-Todd-Ye predictor-corrector interior point algorithm for the $P_+(\kappa)$ LCP and show the iteration complexity $O((1 + \kappa)^{\frac{3}{2}}\sqrt{n})$. They choose $\tau$ and $\tau'$ neighborhood parameters in such a way that a predictor step following by one corrector step at each iteration. For larger value of $\kappa$ the values of $\tau$ and $\tau'$ are fast decreasing, therefore the constant in the complexity result is increasing.

Most of polynomial-time interior point algorithms for LO are based on the use of the logarithmic barrier function, e.g. see [14]. Peng et al.[13]) introduced self-regular barrier functions for primal-dual IPMs for LO and obtained the best complexity result for large-update primal-dual IPMs for LO with some specific self regular barrier function. Recently, Bai et al.[2]) proposed a new class of kernel functions which are not log-barrier and not self regular and they obtained the polynomial complexity for LO and greatly simplified the analysis of IPMs.

In this paper we propose new large-update primal-dual interior point algorithms for $P_+$ LCPs and get the iteration complexity $O((1 + 2\kappa)(1 + p)(1 + q)(\frac{1 + 2 + 2 + p + q + n}{q(1 + p)3 + (1 + p)n}) log \frac{N}{\varepsilon})$. We use a specific class of kernel functions which is the generalized form of the ones in [3] and [12]. Cho et al.([3]) extended Bai et al.’s primal-dual interior point algorithm for LO to
$P_\kappa(\kappa)$ LCPs based on a simple kernel function which is a special case of our kernel functions. When $p = 0$, we obtain the iteration complexity $O((1 + 2\kappa)^{\left(\frac{q+1}{q}2(2q+1)\right)} n \log \frac{n}{\varepsilon})$ which is similar to the one in [3]. Also, for $p = 1$, the kernel function is self-regular which is studied in [12] for LO. If $p = 1$, we get the iteration complexity $O((1 + 2\kappa)(1 + q)^\frac{2+3q}{2q} n^{\frac{2+3q}{2q} \log \frac{n}{\varepsilon}})$ which is similar to the one in [12]. Consider the special case where $p = 1$ and $q = \frac{1}{2} (\log 2n) - 1$, the iteration bound is $O((1 + 2\kappa)^{\sqrt{2n}(\log 2n) \log \frac{n}{\varepsilon}})$ which is the best complexity result for large-update IPMs for $P_\kappa(\kappa)$ LCPs. Since we define a neighborhood and use a search direction based on a specific class of kernel functions which are not logarithmic barrier, the analysis is different from the ones in [5], [7], [8], [9], and [11].

This paper is organized as follows. In Section 2 we recall basic concepts and the notion of the central path. In Section 3 we describe the kernel function and its properties. In Section 4 we compute the feasible step size and derive the amount of decrease of the barrier function during an inner iteration. Finally, in Section 5 we obtain the bounds for the total number of iterations of the Algorithm.

We use the following notations throughout the paper: $R^n$ denotes the set of $n$ dimensional nonnegative vectors and $R^n_+$, the set of $n$ dimensional positive vectors. For $x = (x_1, x_2, \ldots, x_n)^T \in R^n$, $x_{\min} = \min\{x_1, x_2, \ldots, x_n\}$, i.e. the minimal component of $x$, $\|x\|$ is the 2-norm of $x$, and $X$ is the diagonal matrix from vector $x$, i.e. $X = \text{diag}(x)$. $x \otimes s$ denotes the componentwise product (Hadamard product) of vectors $x$ and $s$. $x^T s$ is the scalar product of the vectors $x$ and $s$. $e$ is the $n$-dimensional vector of ones and $I$ is the $n$-dimensional identity matrix. $J$ is the index set, i.e. $J = \{1, 2, \ldots, n\}$.

## 2 Preliminaries

$P_\kappa(\kappa)$ matrix is introduced by Kojima et al.([7]) and we first give some definitions about $P_\kappa(\kappa)$ matrix which is the generalization of positive semi-definite matrices, i.e. $P_\kappa(0)$ matrix.

**Definition 2.1** Let $\kappa \geq 0$ be a nonnegative number. A matrix $M \in R^{n \times n}$ is called a $P_\kappa(\kappa)$ matrix if

$$
(1 + 4\kappa) \sum_{i \in J_+(x)} x_i (Mx)_i + \sum_{i \in J_-(x)} x_i (Mx)_i \geq 0,
$$

for all $x \in R^n$, where

$$
J_+(x) = \{ i \in J : x_i (Mx)_i \geq 0 \} \quad \text{and} \quad J_-(x) = \{ i \in J : x_i (Mx)_i < 0 \}.
$$

**Definition 2.2** A matrix $M \in R^{n \times n}$ is called a $P_\kappa$ matrix if it is a $P_\kappa(\kappa)$ matrix for some $\kappa \geq 0$, i.e. $P_\kappa = \bigcup_{\kappa \geq 0} P_\kappa(\kappa)$. 
Note that the class $P_*$ contains the class PSD of positive semi-definite matrices, i.e. matrices $M$ satisfying $x^T M x \geq 0$ for all $x \in \mathbb{R}^n$, and the class $P$ of matrices with all the principal minors positive. In the following we give some definitions about convexity concepts which is essential in our analysis.

**Definition 2.3** A function $f : D(\subset \mathbb{R}) \rightarrow \mathbb{R}$ is exponentially convex if and only if $f(\sqrt{x_1 x_2}) \leq \frac{1}{2}(f(x_1) + f(x_2))$ for all $x_1, x_2 \in D$.

And now we cite some well-known results. For proofs and details see the book of Kojima et al.[(7)].

**Proposition 2.4** If $M \in \mathbb{R}^{n \times n}$ is a $P_*(\kappa)$ matrix, then

$$M' = \begin{pmatrix} -M & I \\ S & X \end{pmatrix}$$

is a nonsingular matrix for any positive diagonal matrices $X, S \in \mathbb{R}^{n \times n}$.

We use the following corollary to prove that the modified Newton-system has a unique solution.

**Corollary 2.5** Let $M \in \mathbb{R}^{n \times n}$ be a $P_*(\kappa)$ matrix and $x, s \in \mathbb{R}_+^n$. Then for all $a \in \mathbb{R}^n$ the system

$$\begin{cases} -M \Delta x + \Delta s = 0, \\ S \Delta x + X \Delta s = a \end{cases}$$

has a unique solution $(\Delta x, \Delta s)$.

To find an approximate solution for (1) we relax the complementarity condition, i.e. the second equation in (1), and we get the following parameterized system:

$$\begin{cases} s = Mx + q, \\ xs = \mu e, \\ x > 0, s > 0, \end{cases}$$

where $\mu > 0$. Without loss of generality, we assume that (1) is strictly feasible, i.e. there exists $(x^0, s^0)$ such that $s^0 = Mx^0 + q$, $x^0 > 0$, $s^0 > 0$, and moreover, we have an initial strictly feasible point with $\Psi(x^0, s^0, \mu^0) \leq \tau$ for some $\mu^0 > 0$. Indeed, we may not have an available strictly feasible point $(x^0, s^0)$. In order to solve this difficulty, we embed (1) to an artificial LCP which has a strictly feasible point([7]). For this given strictly feasible point
$(x^0, s^0)$ we can always find a $\mu^0 > 0$ such that $\Psi(x^0, s^0, \mu^0) \leq \tau$. Since $M$ is a $P_+ (\kappa)$ matrix and (1) is strictly feasible, (2) has a unique solution for any $\mu > 0$. We denote the solution of (2) as $(x(\mu), s(\mu))$ for given $\mu > 0$. We also call it $\mu-$center for given $\mu$ and the solution set $\{(x(\mu), s(\mu)) \mid \mu > 0\}$ the central path of system (2). As $\mu \to 0$, the sequence $(x(\mu), s(\mu))$ approaches to the solution $(x, s)$ of the system (1) ([7]). We define the following notations:

$$d = \sqrt{\frac{x}{s}}, \quad v = \sqrt{\frac{x s}{\mu}}, \quad dx = \frac{v \Delta x}{x}, \quad ds = \frac{v \Delta s}{s}. \quad (3)$$

Then we have the scaled Newton-system as follows:

$$\begin{cases}
-M dx + ds &= 0, \\
 dx + ds &= v^{-1} - v, \end{cases} \quad (4)$$

where $\bar{M} = DMD$ and $D = \text{diag}(d)$.

We consider a strictly convex function $\Psi(v)$ which is minimal at $v = c$ and $\Psi(c) = 0$. And we will see the function $\Psi(u)$ in chapter 3. Since $v^{-1} - v$ in (4) is exactly the negative gradient of the logarithmic barrier function $\Psi(u) = \sum_{i=1}^{n} (v_i^2 - 1)/2 - \log v_i$, we replace the scaled centering equation, the second equation in (4), with

$$dx + ds = -\nabla \Psi(v). \quad (5)$$

So we get the following modified Newton system:

$$\begin{cases}
-M \Delta x + \Delta s &= 0, \\
S \Delta x + X \Delta s &= -\mu v \nabla \Psi(v). \end{cases} \quad (6)$$

This system uniquely defines a search direction $(\Delta x, \Delta s)$ by Corollary 2.5, since $M$ is a $P_+ (\kappa)$ matrix and (1) is strictly feasible. Throughout the paper, we assume that a proximity parameter $\tau$ and a barrier update parameter $\theta$ are given and $\tau = O(n)$ and $0 < \theta < 1$, fixed. The algorithm works as follows. We assume that strictly feasible point $(x, s)$ is given which is in a $\tau-$neighborhood of the given $\mu-$center. Then after decreasing $\mu$ to $\mu_{+} = (1-\theta)\mu$, for some fixed $\theta \in (0, 1)$, we solve the modified Newton system (6) to obtain the unique search direction. The positivity condition of a new iterate is ensured with the right choice of the step size $\alpha$ which is defined by some line search rule. This procedure is repeated until we find a new iterate $(x_{+}, s_{+})$ that is in a $\tau-$neighborhood of the $\mu_{+}-$center and then we let $\mu := \mu_{+}$ and $(x, s) := (x_{+}, s_{+})$. Then $\mu$ is again reduced by the factor $1-\theta$ and we solve the modified Newton system targeting at the new $\mu_{+}-$center, and so on. This process is repeated
until \( \mu \) is small enough, e.g. \( n\mu \leq \varepsilon \). Throughout the paper, we use the proximity function (or the barrier function) \( \Psi(v) \) to find a search direction and to measure the proximity between the current iterates and the \( \mu \)-center. Then we get the following algorithm.

\[
\text{Algorithm}
\]

Input:
A threshold parameter \( \tau > 1 \);
an accuracy parameter \( \varepsilon > 0 \);
a fixed barrier update parameter \( \theta, \quad 0 < \theta < 1 \);
starting point \((x^0, s^0)\) and \( \mu^0 > 0 \) such that \( \Psi(x^0, s^0, \mu^0) \leq \tau \);

\[
\begin{align*}
x &:= x^0; \quad s := s^0; \quad \mu := \mu^0; \\
\text{while } n\mu \geq \varepsilon \text{ do} & \\
& \begin{align*}
\mu &:= (1 - \theta)\mu; \\
\text{while } \Psi(v) > \tau \text{ do} & \\
& \begin{align*}
& \text{solve the modified Newton-system (6) for } \Delta x \text{ and } \Delta s; \\
& \text{determine a step size } \alpha; \\
& x := x + \alpha \Delta x; \\
& s := s + \alpha \Delta s;
\end{align*}
& \end{align*}
& \end{align*}
\end{align*}
\]

One distinguishes IPMs as large-update methods when \( \theta = O(1) \) and small-update methods when \( \theta = O\left(\frac{1}{\sqrt{n}}\right) \). The small-update methods have the best known iteration complexity, but in practice large-update methods are more efficient than small-update. In this paper we define large-update IPMs and give the worst case of complexity results.
3 The kernel function and its growth behavior

In this section we define a barrier function which is not a logarithmic barrier. We consider a univariate function \( \psi(t) : D \to \mathbb{R}_+ \) with \( \mathbb{R}_+ \subseteq D \) as follows:

\[
\psi(t) = \frac{t^{p+1} - 1}{p+1} + \frac{t^{-q-1}}{q}, \quad q > 0 \quad \text{and} \quad p \in [0,1].
\] (7)

To simplify the analysis we will restrict ourselves to the case where the proximity function \( \Psi(v) \) is separable with identical coordinate functions. Thus, letting \( \psi \) denote the function on the coordinates, we have \( \Psi(v) = \sum_{i=1}^n \psi(v_i) \). We call the univariate function \( \psi(t) \) the kernel function of the proximity function \( \Psi(v) \). For \( \psi(t) \) we have

\[
\psi'(t) = tp^{p-1} - \frac{1}{t^{q+1}}, \quad \psi''(t) = ptp^{p-1} + \frac{q+1}{t^{q+2}}, \quad \psi'''(t) = p(p-1)tp^{p-2} - \frac{(q+1)(q+2)}{t^{q+3}}.
\] (8)

Since \( \psi''(t) > 0 \), \( \psi(t) \) is strictly convex. Note that

\[
\psi(1) = \psi'(1) = 0, \quad \psi''(t) < 0, \quad t > 0.
\]

And due to \( \psi(1) = \psi'(1) = 0 \), \( \psi(t) \) is determined by the second derivative:

\[
\psi(t) = \int_1^t \int_1^s \psi''(\xi) d\xi d\xi.
\]

We define the norm-based proximity measure \( \delta(v) \) as follows:

\[
\delta(v) = \frac{1}{2} \| \nabla \Psi(v) \| = \frac{1}{2} \| d_x + ds \|.
\] (9)

Note that since \( \Psi(v) \) is strictly convex and minimal at \( v = c \), we have

\[
\Psi(v) = 0 \iff \delta(v) = 0 \iff v = c.
\]

In the following lemma we give key properties which are important in the analysis of the Algorithm.

**Lemma 3.1** Kernel function \( \psi(t) \) in (7) satisfies the following properties.

(i) \( t\psi''(t) + \psi'(t) > 0, \quad t > 0 \).

(ii) \( \psi''(t)\psi'(\beta t) - \beta \psi'(t)\psi'(\beta t) > 0, \quad t > 1, \quad \beta > 1 \).

(iii) \( \psi(t) \leq \frac{t^{p+1}}{p+1}, \quad t \geq 1 \).
(iv) \[ 2\psi''(t)^2 - \psi'(t)\psi'''(t) > 0, \quad 0 < t \leq 1. \]

**Proof**: (i): From (8), \( t\psi''(t) + \psi'(t) = (pt^{p-1} + \frac{q}{p+1}) + (t^{p-1} - \frac{1}{p+1}) = (p+1)t^p - \frac{q}{p+1} > 0, \) for \( t > 0. \) (ii): From (8), \( \psi''(t)\psi'(t)\psi''(t) = \frac{p^q - 1}{p+1} + \frac{p^q}{p+1} = \frac{t^{p-1}}{p+1}(\beta^2 - \frac{1}{p+1}) > 0, \) for \( p \in [0,1], \) \( q > 0, \) \( t > 1 \) and \( \beta > 1. \) (iii): Since \( t^{q-1} \leq 0 \) for \( t \geq 1, \) \( \psi(t) = \frac{t^{p+1}-1}{p+1} - \frac{t^{q-1}}{q} \leq \frac{t^{p+1}}{p+1} \), \( t \geq 1. \) (iv): From (8), \( 2\psi''(t) - \psi'(t)\psi'''(t) = 2(pt^{p-1} + \frac{q}{p+1})^2 - (p - 1)(p(p-1)t^{p-2} - \frac{q}{t^{p+1}} - \frac{q}{t^{p+1}}) = q(q+1)(t^{p-1} + \frac{1}{t^{p+1}}) + p(p+1)t^{p-2} + (4pq + p^2 + 3p + (q+1)(q+2))t^{p-q-3} \geq 0, \) for \( 0 < t \leq 1. \)

By Lemma 3.1 (i) and Lemma 1 in [13], we get the following result.

**Corollary 3.2** Kernel function \( \psi(t) \) is exponentially convex.

Since \( \psi(t) \) is strictly convex, we can obtain the following Lemma. For the proof and details the reader can refer to Lemma 3.1 in [2].

**Lemma 3.3** Suppose that \( \psi(t_1) = \psi(t_2), \) with \( t_1 \leq t_2 \) and \( \beta \geq 1. \) Then we have
\[ \psi(\beta t_1) \leq \psi(\beta t_2). \]

Equality holds if and only if \( \beta = 1 \) or \( t_1 = t_2 = 1. \)

Let \( g : [0,\infty) \to [1,\infty) \) be the inverse function of \( \psi(t) \) for \( t \geq 1 \) and \( \rho : [0,\infty) \to (0,1] \) the inverse function of \( -\frac{1}{2}\psi'(t) \) for \( t \in (0,1). \) Then we have the following Lemma. The reader can refer to Lemma 3.4 and Lemma 3.8 in [1] for the proof.

**Lemma 3.4** (Modification of Lemma 3.4 and Lemma 3.8 in [1]) For each \( q > 0, \) we have

(i) \( ((p+1)s+1)^{1+q} \leq g(s) \leq (p+1)s + \frac{2+q}{q}, \) \( s \geq 0. \)

(ii) \( \rho(s) \geq \frac{1}{(2s+1)^{1+q}}, \) \( s > 0. \)

Note that at the start of outer iteration of the algorithm, just before the update of \( \mu \) with the factor \( 1 - \theta, \) we have \( \Psi(v) \leq \tau. \) Due to the update of \( \mu \) the vector \( \nu \) is divided by the factor \( \sqrt{1 - \theta}, \) with \( 0 < \theta < 1, \) which in general leads to an increase in the value of \( \Psi(v). \)

Then, during the subsequent inner iterations, \( \Psi(v) \) decreases until it passes the threshold \( \tau \) again. Hence, during the course of the algorithm the largest values of \( \Psi(v) \) occur just after the updates of \( \mu. \) The following lemma gives an estimate for the effect of a \( \mu \) update on the value of \( \Psi(v). \) By using Lemma 3.1 (ii) and Lemma 3.3, we can get the following lemma. The reader can refer to Theorem 3.2 in [2] for the proof.
Lemma 3.5 (Modification of Theorem 3.2 in [2]) Let \( g : [0, \infty) \to [1, \infty) \) be the inverse function of \( \psi(t) \) for \( t \geq 1 \). Then we have for any positive vector \( v \) and any \( \beta \geq 1 \),
\[
\Psi(\beta v) \leq n \psi\left( \beta g\left( \frac{\Psi(v)}{n} \right) \right).
\]
\[\square\]

By Lemma 3.5, Lemma 3.1 (iii) for \( \beta \geq 1 \) and \( g\left( \frac{\Psi(v)}{n} \right) \geq 1 \) and Lemma 3.4 (i), we obtain
\[
\Psi(\beta v) \leq \frac{n}{p+1} \leq \frac{\beta^{p+1} (p+1) \frac{\Psi(v)}{n} + \frac{p+1}{q}}{p+1}.
\]

We denote the value of \( \Psi(v) \) after the \( \mu \)-update as \( \Psi_0 \) and \( \beta = \frac{1}{\sqrt{1-\theta}} \). If \( \Psi(v) \leq \tau \), then
\[
\Psi_0 : = \Psi\left( \frac{v}{\sqrt{1-\theta}} \right) \leq n \psi\left( \frac{g\left( \frac{\Psi(v)}{n} \right)}{\sqrt{1-\theta}} \right) \leq \frac{\tau(p+1)}{(p+1)(1-\theta)^{\frac{p+1}{2}}}. \quad (10)
\]

Since \( \tau = O(n) \) and \( \theta = O(1) \), \( \Psi_0 = O(n) \). In the following theorem we provide a lower bound for \( \delta(v) \) in terms of the proximity function \( \Psi(v) \). By Theorem 9 in [2] and \( \psi''(t) < 0 \), we obtain the following theorem.

Theorem 3.6 (Modification of Theorem 9 in [2]) Let \( \delta(v) \) be the norm-based proximity measure as defined in (9). Then we have
\[
\delta(v) \geq (1/2) \psi'(g(\Psi(v))).
\]
\[\square\]

By Lemma 3.6 and Lemma 3.4 (i), we get
\[
\delta(v) \geq \frac{1}{2} \left[ \psi(\Psi(v))^p - \frac{1}{g(\Psi(v))^{p+1}} \right] \\
\geq \frac{1}{2} \left[ \left( (p+1) \Psi(v) + 1 \right)^{p+1} - \frac{1}{((p+1) \Psi(v) + 1)^{p+1}} \right].
\]

According to the Algorithm, at the start of each inner iteration we have \( \Psi(v) > \tau \) and \( \tau \geq 1 \). And since \( q > 0 \) and \( p \in [0, 1] \), we obtain
\[
\delta(v) \geq \frac{1}{2} \left[ \frac{\sqrt{(p+1)\Psi(v) + 1}}{\sqrt{(p+1)\Psi(v) + 1}} - \frac{1}{\sqrt{(p+1)\Psi(v) + 1}} \right]
= \frac{1}{2} \left[ \frac{\sqrt{(p+1)\Psi(v) + 1}}{\sqrt{(p+1)\Psi(v) + 1}} - \frac{1}{\sqrt{(p+1)\Psi(v) + 1}} \right] = \frac{1}{2} \frac{(p+1)\Psi(v)}{\sqrt{(p+1)\Psi(v) + 1}} \geq \frac{1}{2} \frac{\Psi(v)}{(3\Psi(v))^{1/3}} \geq \frac{1}{6} \Psi(v)^{2/3}.
\]

4 Computation of the step size and the decrease

In this section we compute the feasible step size \(\alpha\) such that the proximity function is decreasing and the bound for the decrease. Since \(P_\alpha\) LCPs are generalization of LO problems, we loose the orthogonality of vectors \(dx\) and \(ds\). So the analysis is different from LO case. After a damped step for fixed \(\mu\) we have

\[
x_+ = x + \alpha \Delta x, \quad s_+ = s + \alpha \Delta s.
\]

Then by (3), we have

\[
x_+ = x \left( e + \alpha \frac{\Delta x}{x} \right) = x \left( e + \alpha \frac{dx}{v} \right) = x \left( e + \alpha \frac{dx}{v+\alpha ds} \right) = s \left( e + \alpha \frac{ds}{s} \right) = s \left( e + \alpha \frac{ds}{v+\alpha ds} \right) = \frac{s}{v} (v + \alpha ds).
\]

Then we get

\[
e^2_+ = \frac{x_+ s_+}{\mu} = (v + \alpha dx)(v + \alpha ds).
\]

Throughout the paper we assume that the step size \(\alpha\) is such that the coordinates of the vectors \(v + \alpha dx\) and \(v + \alpha ds\) are positive. Hence by Corollary 3.2, we have

\[
\Psi(v_+) = \Psi(\sqrt{(v + \alpha dx)(v + \alpha ds)}) \leq \frac{1}{2} (\Psi(v + \alpha dx) + \Psi(v + \alpha ds)).
\]

For given \(\mu > 0\) by letting \(f(\alpha)\) be the difference of the new and old proximity measures, i.e.

\[
f(\alpha) = \Psi(v_+) - \Psi(v),
\]

we have \(f(\alpha) \leq f_1(\alpha)\), where

\[
f_1(\alpha) := \frac{1}{2} (\Psi(v + \alpha dx) + \Psi(v + \alpha ds)) - \Psi(v).
\]
Note that
\[ f(0) = f_1(0) = 0. \]

By taking the derivative of \( f_1(\alpha) \) with respect to \( \alpha \), we have
\[
f'_1(\alpha) = \frac{1}{2} \sum_{i=1}^{n} (\psi'(v_i + \alpha dx_i)dx_i + \psi'(v_i + \alpha ds_i)ds_i),
\]
where \( dx_i \) and \( ds_i \) denote the \( i \)-th component of the vector \( dx \) and \( ds \), respectively. From (5) and the definition of \( \delta \),
\[
f'_1(0) = \frac{1}{2} \nabla \Psi(v)^T (dx + ds) = -\frac{1}{2} \nabla \Psi(v)^T \nabla \Psi(v) = -2\delta(v)^2. \tag{12}
\]

By taking the derivative of \( f_1'(\alpha) \) with respect to \( \alpha \), we have
\[
f''_1(\alpha) = \frac{1}{2} \sum_{i=1}^{n} (\psi''(v_i + \alpha dx_i)dx_i^2 + \psi''(v_i + \alpha ds_i)ds_i^2). \tag{13}
\]

Since \( M \) is a \( P_+(\kappa) \) matrix and \( M \Delta x = \Delta s \) from (NS), for \( \Delta x \in \mathbb{R}^n \) we have
\[
(1 + 4\kappa) \sum_{i \in J_+} \Delta x_i \Delta s_i + \sum_{i \in J_-} \Delta x_i \Delta s_i \geq 0,
\]
where \( J_+ = \{ i \in J : \Delta x_i \Delta s_i \geq 0 \} \), \( J_- = J - J_+ \). Since \( dx ds = \frac{v^T \Delta x \Delta s}{\mu} = \frac{\Delta x \Delta s}{\mu} \) and \( \mu > 0 \),
\[
(1 + 4\kappa) \sum_{i \in J_+} dx_i ds_i + \sum_{i \in J_-} dx_i ds_i \geq 0. \tag{14}
\]

For notational convenience we define
\[
\delta := \delta(v), \quad \sigma_+ = \sum_{i \in J_+} dx_i ds_i, \quad \sigma_- = -\sum_{i \in J_-} dx_i ds_i.
\]

In the following we compute the bound of \( \|dx\| \) and \( \|ds\| \). To compute this, we need the following technical lemma. For the proofs the reader refers to Lemma 4.1 in [3].

**Lemma 4.1** (Lemma 4.1 in [3]) \( \sigma_+ \leq \delta^2 \) and \( \sigma_- \leq (1 + 4\kappa)\delta^2 \).

In the following lemma we compute the bound for \( \|dx\| \) and \( \|ds\| \). For the proof and details, the reader can refer to Lemma 4.2 in [3].
Lemma 4.2 (Lemma 4.2 in [3]): \[ \sum_{i=1}^{n} (dx_i^2 + ds_i^2) \leq 4(1+2\kappa)\delta^2, \quad \|dx\| \leq 2\sqrt{1+2\kappa}\delta, \quad \text{and} \quad \|ds\| \leq 2\sqrt{1+2\kappa}\delta. \]

To compute the upper bound for the difference of the new and old proximity measures, we need the following technical lemmas.

Lemma 4.3 \[ f_1''(\alpha) \leq 2(1+2\kappa)\delta^2\psi''(v_{\min} - 2\alpha\sqrt{1+2\kappa}\delta). \]

Proof: By Lemma 4.2, we have

\[ v_i + \alpha dx_i \geq v_{\min} - 2\alpha\sqrt{1+2\kappa}\delta, \quad v_i + \alpha ds_i \geq v_{\min} - 2\alpha\sqrt{1+2\kappa}\delta. \]

From (13), \( \psi''(t) < 0 \), and Lemma 4.2,

\[ f_1''(\alpha) = \frac{1}{2} \sum_{i=1}^{n} (\psi''(v_i + \alpha dx_i) dx_i^2 + \psi''(v_i + \alpha ds_i) ds_i^2) \]

\[ \leq \frac{1}{2} \sum_{i=1}^{n} (\psi''(v_{\min} - 2\alpha\sqrt{1+2\kappa}\delta) dx_i^2 + \psi''(v_{\min} - 2\alpha\sqrt{1+2\kappa}\delta) ds_i^2) \]

\[ = \frac{1}{2} \psi''(v_{\min} - 2\alpha\sqrt{1+2\kappa}\delta) \sum_{i=1}^{n} (dx_i^2 + ds_i^2) \]

\[ \leq \frac{1}{2} \psi''(v_{\min} - 2\alpha\sqrt{1+2\kappa}\delta) 4(1+2\kappa)\delta^2 \]

\[ = 2(1+2\kappa)\delta^2 \psi''(v_{\min} - 2\alpha\sqrt{1+2\kappa}\delta). \]

Lemma 4.4 \[ f_1'(\alpha) \leq 0 \quad \text{if} \quad \alpha \quad \text{is satisfying} \]

\[ -\psi'(v_{\min} - 2\alpha\delta\sqrt{1+2\kappa}) + \psi'(v_{\min}) \leq \frac{2\delta}{\sqrt{1+2\kappa}}. \] (15)

Proof: Using (12), Lemma 4.3, and (15),

\[ f_1'(\alpha) = f_1'(0) + \int_{0}^{\alpha} f_1''(\xi) d\xi \]

\[ \leq -2\delta^2 + 2(1+2\kappa)\delta^2 \int_{0}^{\alpha} \psi''(v_{\min} - 2\xi\sqrt{1+2\kappa}\delta) d\xi \]

\[ = -2\delta^2 - 2\delta - \int_{0}^{\alpha} \psi''(v_{\min} - 2\xi\sqrt{1+2\kappa}\delta) d(v_{\min} - 2\xi\sqrt{1+2\kappa}\delta) \]

\[ = -2\delta^2 - 2\delta - (\psi'(v_{\min} - 2\alpha\delta\sqrt{1+2\kappa}) - \psi'(v_{\min})) \]

\[ \leq -2\delta^2 + 2(1+2\kappa)\delta \frac{2\delta}{\sqrt{1+2\kappa}} \]

\[ = 0. \]
So we get the desired result. 

In the following lemma, we compute the feasible step size $\alpha$ such that the proximity measure is decreasing when we take a new iterate for fixed $\mu$.

**Lemma 4.5** Let $\rho : [0, \infty) \to (0, 1]$ denote the inverse function of the restriction of $-\frac{1}{2} \psi'(t)$ to the interval $(0, 1]$. Then the largest step size $\alpha$ that satisfies (15) is given by

$$
\alpha := \frac{1}{2\delta \sqrt{1 + 2\kappa}} \left( \rho(\delta) - \rho \left( \left( 1 + \frac{1}{\sqrt{1 + 2\kappa}} \right) \delta \right) \right).
$$

**Proof:** We want to compute the step size $\alpha$ such that (15) holds with $\alpha$ as large as possible. Since $\psi'' > 0$, the derivative of the left hand side in (15) with respect to $\alpha$ is $2\delta \sqrt{1 + 2\kappa} \psi'(v_{\text{min}} - 2\alpha \delta \sqrt{1 + 2\kappa}) > 0$. Hence the left hand side in (15) is monotone increasing in $\alpha$. So the largest possible value of $\alpha$ satisfying (15) occurs when

$$
-\psi'(v_{\text{min}} - 2\alpha \delta \sqrt{1 + 2\kappa}) + \psi'(v_{\text{min}}) = \frac{2\delta}{\sqrt{1 + 2\kappa}}.
$$

The derivative of the left hand side in (17) with respect to $v_{\text{min}}$ is $-\psi''(v_{\text{min}} - 2\alpha \delta \sqrt{1 + 2\kappa}) + \psi''(v_{\text{min}}) < 0$. Since $\psi'' < 0$, the left hand side in (17) is decreasing in $v_{\text{min}}$. This implies that for fixed $\delta$, $\alpha$ gets smaller if $v_{\text{min}}$ gets smaller. Note that by the definition of $\delta$ and $\Psi(v)$,

$$
\delta = \frac{1}{2} \| \nabla \Psi(v) \| = \frac{1}{2} \sqrt{\sum_{i=1}^{a} (\psi'(v_i))^2} \geq \frac{1}{2} | \psi'(v_{\text{min}}) | \geq -\frac{1}{2} \psi'(v_{\text{min}}).
$$

Equality holds if and only if $v_{\text{min}}$ is the only coordinate in $v$ which is different from 1 and $v_{\text{min}} \leq 1$, i.e. $\psi'(v_{\text{min}}) \leq 0$. Hence when $v_{\text{min}}$ satisfies

$$
-\frac{1}{2} \psi'(v_{\text{min}}) = \delta,
$$

the smallest step size $\alpha$ occurs. In this case by (18) and the definition of $\rho$,

$$
v_{\text{min}} = \rho(\delta).
$$

From (17) and (18),

$$
-\frac{1}{2} \psi'(v_{\text{min}} - 2\alpha \delta \sqrt{1 + 2\kappa}) = \delta \left( 1 + \frac{1}{\sqrt{1 + 2\kappa}} \right).
$$
Then by (20) and the definition of \( \rho \), \( v_{\text{min}} - 2a\bar{\delta}\sqrt{1 + 2\kappa} = \rho \left( \left( 1 + \frac{1}{\sqrt{1 + 2\kappa}} \right) \bar{\delta} \right) \). Thus by (19), the largest step size \( \alpha \) is given as follows: \( \alpha = \frac{1}{2a\bar{\delta}\sqrt{1 + 2\kappa}} \left( \rho(\bar{\delta}) - \rho \left( \left( 1 + \frac{1}{\sqrt{1 + 2\kappa}} \right) \bar{\delta} \right) \right) \). \( \square \)

In the following lemma we compute the lower bound for \( \bar{\alpha} \) in Lemma 4.5.

**Lemma 4.6** Let \( \rho \) and \( \bar{\alpha} \) be as defined in Lemma 4.5. Then we have

\[
\bar{\alpha} \geq \frac{1}{1 + 2\kappa} \frac{1}{\psi''(\rho((1 + \frac{1}{\sqrt{1 + 2\kappa}})\bar{\delta}))}.
\]

**Proof:** By the definition of \( \rho \), \( -\psi'(\rho(\bar{\delta})) = 2\bar{\delta} \). By taking the derivative with respect to \( \delta \), we get \( -\psi''(\rho(\bar{\delta}))\rho'(\bar{\delta}) = 2 \). So we have \( \rho'(\bar{\delta}) = -\frac{2}{\psi''(\rho(\bar{\delta}))} < 0 \) since \( \psi'' > 0 \). Hence \( \rho \) is monotonically decreasing. By (16) and the fundamental theorem of calculus, we have

\[
\bar{\alpha} = \frac{1}{2\bar{\delta}\sqrt{1 + 2\kappa}} \left( \rho(\bar{\delta}) - \rho \left( \left( 1 + \frac{1}{\sqrt{1 + 2\kappa}} \right) \bar{\delta} \right) \right)
= \frac{1}{2\bar{\delta}\sqrt{1 + 2\kappa}} \int_{\bar{\delta}}^{\bar{\delta}(1 + \frac{1}{\sqrt{1 + 2\kappa}})\bar{\delta}} \rho'(\xi) d\xi = \frac{1}{\delta\sqrt{1 + 2\kappa}} \int_{\bar{\delta}}^{\bar{\delta}(1 + \frac{1}{\sqrt{1 + 2\kappa}})\bar{\delta}} \psi''(\rho(\xi)) d\xi
\]

Since \( \delta \leq \xi \leq (1 + \frac{1}{\sqrt{1 + 2\kappa}})\bar{\delta} \) and \( \rho \) is monotonically decreasing, \( \rho(\xi) \geq \rho \left( \left( 1 + \frac{1}{\sqrt{1 + 2\kappa}} \right) \bar{\delta} \right) \). Since \( \psi'' \) is monotonically decreasing, \( \psi''(\rho(\xi)) \leq \psi''(\rho \left( \left( 1 + \frac{1}{\sqrt{1 + 2\kappa}} \right) \bar{\delta} \right)) \). Hence \( \frac{1}{\psi''(\rho(\bar{\delta}))} \geq \frac{1}{\psi''(\rho \left( \left( 1 + \frac{1}{\sqrt{1 + 2\kappa}} \right) \bar{\delta} \right))} \). Therefore we have

\[
\bar{\alpha} \geq \frac{1}{\delta\sqrt{1 + 2\kappa}} \frac{1}{\psi''(\rho \left( \left( 1 + \frac{1}{\sqrt{1 + 2\kappa}} \right) \bar{\delta} \right))} \int_{\bar{\delta}}^{\bar{\delta}(1 + \frac{1}{\sqrt{1 + 2\kappa}})\bar{\delta}} d\xi = \frac{1}{1 + 2\kappa} \frac{1}{\psi''(\rho((1 + \frac{1}{\sqrt{1 + 2\kappa}})\bar{\delta}))}.
\]

Define

\[
\tilde{\alpha} = \frac{1}{1 + 2\kappa} \frac{1}{\psi''(\rho((1 + \frac{1}{\sqrt{1 + 2\kappa}})\bar{\delta}))}.
\]

and we will use \( \tilde{\alpha} \) as the default step size in our Algorithm. By Lemma 4.6, we have \( \bar{\alpha} \geq \tilde{\alpha} \). In the following, we want to evaluate the decrease of the proximity function value. We cite the following result in [13] without proof.

**Lemma 4.7** Let \( h(t) \) be a twice differentiable convex function with \( h(0) = 0 \), \( h'(0) < 0 \) and let \( h(t) \) attains its (global) minimum at \( t^* > 0 \). If \( h''(t) \) is increasing for \( t \in [0, t^*] \), then

\[
h(t) \leq \frac{h'(0)}{2}, \quad 0 \leq t \leq t^*.
\]
Lemma 4.8 If the step size $\alpha$ is such that $\alpha \leq \bar{\alpha}$, then $f(\alpha) \leq -\alpha \delta^2$.

Proof: Define the univariate function $h$ as follows:

$$h(0) = f_1(0) = 0, \quad h'(0) = f_1'(0) = -2\delta^2, \quad h''(\alpha) = 2(1 + 2\kappa)\delta^2\psi'(v_{\text{min}} - 2\alpha \sqrt{1 + 2\kappa})$$

By Lemma 4.3, $f_1''(\alpha) \leq h''(\alpha)$. So we have $f_1'(\alpha) \leq h'(\alpha)$ and $f_1(\alpha) \leq h(\alpha)$. By the definition of $h(\alpha)$ and $\psi'(t) > 1$, $h''(\alpha) \geq 2(1 + \kappa)\delta^2$. This implies that $h'(\alpha)$ is strongly convex and hence $h(\alpha)$ attains its global minimum for some $\alpha^* > 0$. By taking $\alpha \leq \bar{\alpha}$, with $\bar{\alpha}$ as defined in Lemma 4.5, using the fundamental theorem of calculus, and Lemma 4.4, we have

$$h'(\alpha) = h'(0) + \int_0^\alpha h''(\xi)d\xi$$

$$= -2\delta^2 + 2(1 + 2\kappa)\delta^2 \int_0^\alpha \psi'(v_{\text{min}} - 2\kappa \sqrt{1 + 2\kappa})d\xi$$

$$= -2\delta^2 - \sqrt{1 + 2\kappa} \psi'(v_{\text{min}} - 2\alpha \sqrt{1 + 2\kappa}) - \psi'(v_{\text{min}})$$

$$\leq -2\delta^2 + \frac{2\delta}{\sqrt{1 + 2\kappa}}$$

$$= 0.$$

Since $h''(\alpha) = -4(1 + 2\kappa)^{\frac{3}{2}} \delta^3 \psi'(v_{\text{min}} - 2\alpha \sqrt{1 + 2\kappa})$ and $\psi''' < 0$, $h''(\alpha)$ is increasing in $\alpha$. By Lemma 4.7, we have $f_1(\alpha) \leq h(\alpha) \leq \frac{1}{2} \alpha h'(0) = -\alpha \delta^2$. Since $f(\alpha) \leq f_1(\alpha)$, the proof is completed. \hfill \Box

In the following theorem we obtain the upper bound for the difference $f(\alpha)$ between the new and old proximity measures in the Algorithm.

**Theorem 4.9** Let $\bar{\alpha}$ be a step size as defined in (21). Then we have

$$f(\bar{\alpha}) \leq -\frac{1}{1 + 2\kappa} \frac{\delta^2}{\psi'^{(1 + \frac{1}{\sqrt{1 + 2\kappa}})}}$$

(22)

Proof: By Lemma 4.6 and (21), $\bar{\alpha} \leq \bar{\alpha}$. By Lemma 4.8, we get the desired result. \hfill \Box

**Lemma 4.10** The right hand side in (22) is monotonically decreasing in $\delta$. 

Proof: Let \( t = \rho(a \delta) \) where \( a = 1 + \frac{1}{\sqrt{1 + 2 \kappa}} \). Then \( 0 < t \leq 1 \) and \( -\psi'(\rho(a \delta)) = 2a \delta \), i.e. \( \frac{1}{2} \psi'(t) = -\frac{1}{2} \psi'(\rho(a \delta)) = a \delta \). Then

\[
\frac{1}{1 + 2 \kappa} \psi''(\rho((1 + \frac{1}{\sqrt{1 + 2 \kappa}}) \delta)) = \frac{1}{4a^2} \left(1 + 2 \kappa\right) \psi'(t)^2.
\]

Define

\[
g(t) = \frac{1}{4a^2} \left(1 + 2 \kappa\right) \psi'(t)^2.
\]

Since \( \rho \) is monotonically decreasing, \( t \) is monotonically decreasing if \( \delta \) increases. Hence the right hand in (22) is monotonically decreasing in \( \delta \) if and only if the function \( g(t) \) is monotonically decreasing for \( 0 < t \leq 1 \). Note that \( g(1) = 0 \) and

\[
g'(t) = \frac{1}{4a^2} \left(1 + 2 \kappa\right) \frac{\psi'(t) \left(2 \psi''(t)^2 - \psi'(t) \psi'''(t)\right)}{\psi'(t)^2}.
\]

Since \( \psi'(1) = 0 \) and \( \psi'' > 0, \psi'(t) \leq 0 \) for \( 0 < t \leq 1 \). By Lemma 3.1 (ii), \( g(t) \) is monotonically decreasing for \( 0 < t \leq 1 \). Hence the lemma is proved. \( \Box \)

5 Complexity analysis

In this section we compute the total number of iterations of the Algorithm. We need the following technical lemma to obtain iteration bounds. For the proof the reader can refer to [13].

Lemma 5.1 ([13]) Let \( t_0, t_1, \ldots, t_K \) be a sequence of positive numbers such that

\[
t_{k+1} \leq t_k - \kappa t_k 1^{-\gamma}, \quad k = 0, 1, \ldots, K - 1,
\]

where \( \kappa > 0 \) and \( 0 < \gamma \leq 1 \). Then \( K \leq \left\lfloor \frac{t_0}{\kappa \gamma} \right\rfloor \).

We define the value of \( \Psi(v) \) after the \( \mu \)-update as \( \Psi_0 \) and the subsequent values in the same outer iteration are denoted as \( \Psi_k, k = 1, 2, \ldots \). Let \( K \) denote the total number of inner iterations in the outer iteration. Then by the definition of \( K \), we have

\[
\Psi_{K-1} > \tau, \quad 0 \leq \Psi_K \leq \tau.
\]

In the following lemma, we compute the upper bound for the total number of inner iterations which we needed to return to the \( \tau \)-neighborhood again. For notational convenience we denote \( \Psi(v) \) by \( \Psi \) and \( a = 1 + \frac{1}{\sqrt{1 + 2 \kappa}} \).
Lemma 5.2 Let $K$ be the total number of inner iterations in an outer iteration. Then we have

$$K \leq 36 \left(\frac{5}{3}\right)^{\frac{1}{1+\kappa}} (1 + 2\kappa)(1 + p)(1 + q)\Psi_0^{\frac{1 + p + q}{1 + p + q + 2}},$$

where $\Psi_0$ denotes the value of $\Psi(v)$ after the $\mu$-update.

Proof: Since $\psi''(t) < 0$, $\psi'(t)$ is decreasing function. So using (22) and Lemma 3.4 (ii), we have

$$f(\bar{\alpha}) \leq -\frac{1}{1 + 2\kappa} \frac{\delta^2}{\psi'(\rho(\delta\alpha))} \leq -\frac{1}{1 + 2\kappa} \frac{\delta^2}{\psi'\left((1 + 2a\delta)^{-\frac{1}{1+\tau}}\right)}.$$

By Lemma 4.10 and (11), we get

$$f(\bar{\alpha}) \leq -\frac{1}{1 + 2\kappa} \frac{1}{36\Psi^{\frac{2p}{1+p}} \psi' \left((1 + \frac{a}{3} \Psi^{\frac{p}{1+p}})^{-\frac{1}{1+\tau}}\right)} \leq -\frac{1}{36(1 + 2\kappa)} \frac{1}{\left(p \left(1 + \frac{a}{3} \Psi^{\frac{p}{1+p}}\right)^{\frac{1-p}{1+\tau}} + (q + 1) \left(1 + \frac{a}{3} \Psi^{\frac{p}{1+p}}\right)^{\frac{q+1}{1+\tau}}\right)}.$$

Also since $p \in [0,1]$ and $q > 0$, we have

$$f(\bar{\alpha}) \leq -\frac{\Psi^{\frac{2p}{1+p}}}{36(1 + 2\kappa)} \frac{1}{\left(1 + p + q\right) \left(1 + \frac{a}{3} \Psi^{\frac{p}{1+p}}\right)^{\frac{p+q}{1+\tau}}}. \quad (23)$$

By assuming $\Psi_0 \geq \Psi \geq \tau \geq 1$ and using $a = 1 + \frac{1}{\sqrt{4+2a}} \leq 2$, we have

$$1 + \frac{a}{3} \Psi^{\frac{p}{1+p}} \leq 1 + \frac{2}{3} \Psi^{\frac{p}{1+p}} \leq \frac{5}{3} \Psi^{\frac{p}{1+p}}.$$

From (23), we have

$$f(\bar{\alpha}) \leq -\frac{\Psi^{\frac{2p}{1+p}}}{36(1 + 2\kappa)} \frac{1}{\left(1 + p + q\right) \left(1 + \frac{a}{3} \Psi^{\frac{p}{1+p}}\right)^{\frac{p+q}{1+\tau}}} \leq -\frac{1}{36(1 + 2\kappa)(1 + p + q)} \Psi^{\frac{2p}{1+p}} \left(\frac{5}{3}\right)^{\frac{p+q}{1+\tau}}.$$
This implies that
\[ \Psi_{k+1} \leq \Psi_k - \beta \Psi_k^{1-\gamma}, \quad k = 0, 1, 2, \ldots, K - 1, \]
where
\[ \beta = \left( \frac{3}{5} \right)^{\frac{3+\alpha}{3+\alpha}} \frac{1}{36(1+2\kappa)(1+p+q)}, \quad \gamma = \frac{q+p+1}{(p+1)(q+1)}. \]
Hence by Lemma 5.1, \( K \) is bounded above by
\[ K \leq \frac{\Psi_0}{\beta \gamma} = 36 \left( \frac{5}{3} \right)^{\frac{3+\alpha}{3+\alpha}} (1+2\kappa)(1+p)(1+q)^{1+\frac{p+q}{1+\frac{p+q}{(1+\theta)(1+\tau)}}}. \tag{24} \]
This completes the proof.

From (10), we have
\[ \Psi_0 \leq \frac{q(p+1)\tau + n(p + q + 1)}{q(p+1)(1 - \theta)^{\frac{p+q}{2}}} = \Theta. \]
From (24), we have
\[ K \leq 36 \left( \frac{5}{3} \right)^{\frac{3+\alpha}{3+\alpha}} (1+2\kappa)(1+p)(1+q)^{1+\frac{p+q}{1+\frac{p+q}{(1+\theta)(1+\tau)}}}. \]
The upper bound for the total number of iterations is obtained by multiplying the number \( K \) by the number of central path parameter updates. If the central path parameter \( \mu \) has the initial value \( \mu^0 \) and is updated by multiplying \( 1 - \theta \), with \( 0 < \theta < 1 \), then after at most
\[ \left\lceil \frac{1}{\theta} \log \frac{n\mu^0}{\epsilon} \right\rceil \]
iterations we have \( n\mu \leq \epsilon \). Thus the total number of iterations is bounded above by
\[ 36 \left( \frac{5}{3} \right)^{\frac{3+\alpha}{3+\alpha}} (1+2\kappa)(1+p)(1+q)^{1+\frac{p+q}{1+\frac{p+q}{(1+\theta)(1+\tau)}}} \frac{1}{\theta} \log \frac{n\mu^0}{\epsilon}. \]
So we obtain the main result.

**Theorem 5.3** Let a linear complementarity problem for any \( P_t(\kappa) \) matrix \( M \) be given, where \( \kappa \geq 0 \). Assume that a strictly feasible starting point \((x^0, s^0)\) is available with \( \Psi(x^0, s^0, \mu^0) \leq \tau \) for some \( \mu^0 > 0 \). Then the total number of iterations for our Algorithm is bounded above by
\[ \left\lceil 36 \left( \frac{5}{3} \right)^{\frac{3+\alpha}{3+\alpha}} (1+2\kappa)(1+p)(1+q)^{1+\frac{p+q}{1+\frac{p+q}{(1+\theta)(1+\tau)}}} \frac{1}{\theta} \log \frac{n\mu^0}{\epsilon} \right\rceil. \]
Remark 5.4  For our Algorithm since $\tau = O(n)$ and $\theta = O(1)$, we obtain $O\left((1 + 2\kappa)(1 + p)(1 + q)\left(\frac{1 + (q + 1)(2n)}{q} \log \frac{n}{\varepsilon} \right)^{\frac{2}{q + 1}}\log \frac{n}{\varepsilon}\right)$ which is the worst case iteration complexity. Note that if $p = 0$, then we get $O\left((1 + 2\kappa)\frac{(q + 1)(2n + 1)}{q} \log \frac{n}{\varepsilon}\right)$ complexity and if $p = 1$, then we have $O\left((1 + 2\kappa)(1 + q)\left(\frac{2q + 1}{2q} \log \frac{n}{\varepsilon}\right)^{\frac{q + 1}{q}}\right)$ complexity. Also note that for small value of $q$ the bound is bad. Now consider the special case where $q \geq 2$, $\tau = \frac{q - 2q}{2q} n$ and $p = 1$. By omitting $\theta = O(1)$, the bound becomes $(1 + 2\kappa)(q + 1)(2n)\frac{\log n}{\varepsilon}$. Then by taking $q = \frac{1}{2}(\log 2n) - 1$, we get

\[
(1 + 2\kappa)(q + 1)(2n)\frac{\log n}{\varepsilon} = \frac{1}{2}(1 + 2\kappa)(\log 2n)(2n)\frac{\frac{1}{2}(\log 2n) - 1}{\log \frac{n}{\varepsilon}} = \frac{1}{2}(1 + 2\kappa)\exp(1)\sqrt{2n}(\log 2n)\log \frac{n}{\varepsilon}.
\]

Thus for $p = 1$ and $q = \frac{1}{2}(\log 2n) - 1$, we have the polynomial complexity

$O\left((1 + 2\kappa)\sqrt{2n}(\log 2n)\log \frac{n}{\varepsilon}\right),$

which is the best worst case complexity so far.

References


Upper and lower solutions for a singular p-Laplacian system

Chan-Gyun Kim and Eun Kyoung Lee
Department of Mathematics, Pusan National University, Pusan 609-735, Korea
E-mail: cgkim75@pusan.ac.kr (C.G. Kim), lek915@pusan.ac.kr (E.K. Lee)

Abstract

In this paper, we define the upper and lower solutions for a p-Laplacian system with singular nonlinearity at the boundaries. And we prove the theorem for the upper and lower solutions method.

Keywords: singular p-Laplacian system; upper solution; lower solution
MSC: 34B08, 34B16, 35J25

1 Introduction

In this paper, we consider the upper and lower solution method for the following p-Laplacian system,

\[
\begin{cases}
\varphi_p((u'(t))') + F(t, u(t), v(t)) = 0, & t \in (0, 1), \\
\varphi_p((v'(t))') + G(t, u(t), v(t)) = 0, & t \in (0, 1), \\
u(0) = A, & u(1) = B, \\
\varphi_p(v'(0)) = G(0, u(0), 0), & \varphi_p(v'(1)) = G(1, u(1), 0), \\
v(0) = C, & v(1) = D,
\end{cases}
\]

\( (H) \)

where \( F, G : (0, 1) \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} \) are continuous. And \( F \) and \( G \) may be singular at \( t = 0 \) or \( 1 \).

Upper and lower solution method is very useful to show the existence of the solutions for nonlinear boundary value problems and under suitable monotonicity conditions on \( F \) and \( G \), it is well known for continuous case (i.e. \( F \) and \( G \) are continuous on \([0, 1] \times \mathbb{R} \times \mathbb{R}\) ) that existence of a lower solution \( \alpha \) and an upper solution \( \beta \) with \( \alpha \leq \beta \) implies the existence of solution \((u,v)\) of problem \((H)\) with \( \alpha \leq (u,v) \leq \beta \). One may refer to De Coster and Habets([2]) for recent development up to \( L^1\)-Carathéodory case.

Recently, there are vast amount of literatures ([1],[3],[4],[5],[6],[7]) about nonlinear boundary value problems having an indefinite weight which is, in general,
not in $L^1(0,1)$. In particular, when $p = 2$, the theorem of upper and lower solution method for system (H) was proved by Lee([5]). On the other hand, when $p \neq 2$, the theorem for scalar equation was proved by Lü and O'Regan([6]).

The theorem for system (H) is not extended by obvious manner from neither the linear case ($p = 2$) nor scalar equation case and, as far as the authors know, the proof is not given in any previous studies. Thus the aim of this paper is to prove the theorem of upper and lower solution method for $p$-Laplacian system (H).

2 Main Result

Definition 2.1 For a function $F : [0,1] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we say $F(t,u,v)$ is quasi-monotone nondecreasing with respect to $v$ (or $u$) if for fixed $t$,

\[ F(t,u,v_1) \leq F(t,u,v_2) \quad \text{whenever } v_1 \leq v_2, \]

( or $F(t,u_1,v) \leq F(t,u_2,v)$ whenever $u_1 \leq u_2$.)

Definition 2.2 We say that $(\alpha_u, \alpha_v)$ is a lower solution of (H) if $(\alpha_u, \alpha_v) \in C^1(0,1) \times C^1(0,1)$, $(\varphi_p(\alpha'_u(t)), \varphi_p(\alpha'_v(t))) \in C^1(0,1) \times C^1(0,1)$ and

\[
\begin{aligned}
\varphi_p(\alpha'_u(t))' + F(t,\alpha_u(t),\alpha_v(t)) &\geq 0, \\
\varphi_p(\alpha'_v(t))' + G(t,\alpha_u(t),\alpha_v(t)) &\geq 0, \\
\alpha_u(0) &\leq A, \quad \alpha_v(1) \leq B, \\
\alpha_v(0) &\leq C, \quad \alpha_v(1) \leq D.
\end{aligned}
\]

We also say that $(\beta_u, \beta_v)$ is an upper solution of the problem (H) if $(\beta_u, \beta_v) \in C^1(0,1) \times C^1(0,1)$, $(\varphi_p(\beta'_u(t)), \varphi_p(\beta'_v(t))) \in C^1(0,1) \times C^1(0,1)$ and it satisfies the reverse of the above inequalities.

Let $D^\alpha = \{(t,u,v) | t \in [0,1] \text{ and } (\alpha_u(t), \alpha_v(t)) \leq (u,v) \leq (\beta_u(t), \beta_v(t))\}$. As the main result of this paper, the fundamental theorem on upper and lower solutions for the singular problem (H) is given as follows.

Theorem 2.3 Let $(\alpha_u, \alpha_v)$ and $(\beta_u, \beta_v)$ be, respectively, a lower solution and an upper solution of the problem (H) such that

\[ (a_1) \quad (\alpha_u(t), \alpha_v(t)) \leq (\beta_u(t), \beta_v(t)) \text{ for all } t \in [0,1]. \]

Assume also that there exist functions $h_F, h_G \in C([0,1], \mathbb{R}^+) \text{ such that}$

\[ (a_2) \quad |F(t,u,v)| \leq h_F(t), \quad |G(t,u,v)| \leq h_G(t), \text{ for all } (t,u,v) \in D^\alpha. \]

\[ (a_3) \quad \int_0^1 \varphi_p^{-1} \left( \int_s^t h_F(\tau) \, d\tau \right) \, ds + \int_0^1 \varphi_p^{-1} \left( \int_s^t h_G(\tau) \, d\tau \right) \, ds < \infty. \]

\[ \int_0^1 \varphi_p^{-1} \left( \int_s^t h_G(\tau) \, d\tau \right) \, ds + \int_0^1 \varphi_p^{-1} \left( \int_s^t h_G(\tau) \, d\tau \right) \, ds < \infty. \]
(a4) \( F(t, u, v) \) and \( G(t, u, v) \) are quasi-monotone nondecreasing w.r.t. \( v \) and \( u \), respectively.

Then the problem (H) has at least one solution \((u, v)\) such that
\[
(\alpha_u(t), \alpha_v(t)) \leq (u(t), v(t)) \leq (\beta_u(t), \beta_v(t)) \quad \text{for all } t \in [0, 1].
\]

To prove the above theorem, we need the following Lemmas.

**Lemma 2.4** Assume that there exist \( h_F, h_G \in L^1(0, 1) \) such that
\[
|F(t, u, v)| \leq h_F(t), \quad |G(t, u, v)| \leq h_G(t), \quad \text{for all } (t, u, v) \in (0, 1) \times \mathbb{R}^2.
\]

Then (H) has a solution.

**Proof.** For fixed \( l \in C[0, 1] \), define \( G_l : \mathbb{R} \to \mathbb{R} \) by
\[
G_l(a) = \int_0^1 \varphi_p^{-1}(a + l(t)) \, dt.
\]

First, we will show that the equation \( G_l(a) = B - A \) has a unique solution. For fixed \( l \in C[0, 1] \), since \( a + l(t) \to \pm \infty \) uniformly on \([0, 1]\), as \( a \to \pm \infty \), we can say that \( G_l(a) \to \pm \infty \), as \( a \to \pm \infty \). By intermediate value theorem, there is \( \bar{a}(l) \in \mathbb{R} \) such that \( G_l(\bar{a}(l)) = B - A \). Since \( \varphi_p^{-1} \) is monotone increasing, \( G_l \) is also monotone increasing and thus \( \bar{a}(l) \) is unique. Second, we will show that \( \bar{a} : C[0, 1] \to \mathbb{R} \) defined by \( G_l(\bar{a}(l)) = B - A \), sends bounded sets into bounded sets and is continuous. Let \( \Sigma \) be a bounded set in \( C[0, 1] \). There exists \( M > 0 \) such that \( \|l\|_{\infty} \leq M \), for all \( l \in \Sigma \). From
\[
\int_0^1 \varphi_p^{-1}(\bar{a}(l) + l(t)) \, dt = B - A,
\]
using Mean Value Theorem, there exists \( c \in (0, 1) \) such that
\[
\varphi_p^{-1}(\bar{a}(l) + l(c)) \, dt = B - A.
\]

Then,
\[
|\bar{a}(l)| = |\varphi_p(B - A) - l(c)| \leq |\varphi_p(B - A)| + \|l\|_{\infty} \leq |\varphi_p(B - A)| + M.
\]

Thus \( \bar{a}(\Sigma) \) is bounded. To show the continuity of \( \bar{a} \), let \( l_n \to l \) in \( C[0, 1] \) as \( n \to \infty \). Since \( [\bar{a}(l_n)] \) is bounded sequence in \( \mathbb{R} \), there is a convergent subsequence, say \( \{\bar{a}(l_{n_j})\} \) for convenience, such that \( \bar{a}(l_{n_j}) \to \bar{a} \) as \( n \to \infty \). By the boundedness of \( \{\bar{a}(l_{n_j})\} \) and \( \{l_n\} \), using Lebesgue Dominated Convergence Theorem, if \( n \to \infty \) in
\[
\int_0^1 \varphi_p^{-1}(\bar{a}(l_{n_j}) + l_{n_j}(t)) \, dt = B - A,
\]
then we find that
\[
\int_0^1 \varphi_p^{-1}(\bar{a} + l(t)) \, dt = B - A.
\]
Thus $\tilde{a}(t) = \tilde{a}$, i.e. $\tilde{a}(t) \to a(t)$, which shows the continuity of $\tilde{a}$. Define $\tilde{G} : L^1(0,1) \to C[0,1]$ by
\[
\tilde{G}(h)(t) = A + \int_0^t \varphi_p^{-1} \left( a(h) + \int_0^s h(\tau)d\tau \right) ds,
\]
where $a : L^1(0,1) \to \mathbb{R}$ is satisfying
\[
\int_0^1 \varphi_p^{-1} \left( a(h) + \int_0^s h(\tau)d\tau \right) ds = B - A.
\]
Then $a$ is continuous which sends bounded sets in $L^1(0,1)$ into bounded set in $\mathbb{R}$. Thus we can say that $\tilde{G}$ is continuous and maps bounded set of $L^1(0,1)$ into relatively compact sets of $C[0,1]$. Now, Define $H_F$, $H_G : C[0,1] \times C[0,1] \to L^1(0,1)$ by
\[
H_F(u,v)(t) = F(t,u(t),v(t)), \\
H_G(u,v)(t) = G(t,u(t),v(t)).
\]
By hypothesis, $H_F$ and $H_G$ maps $C[0,1] \times C[0,1]$ into bounded set. Thus $\tilde{G}(H_F(C[0,1] \times C[0,1]))$ and $\tilde{G}(H_G(C[0,1] \times C[0,1]))$ are bounded and $\tilde{G} \circ H_F$ and $\tilde{G} \circ H_G$ are completely continuous. Define $T : C[0,1] \times C[0,1] \to C[0,1] \times C[0,1]$ by
\[
T(u,v) = \left( \tilde{G}(H_F(u,v)), \tilde{G}(H_G(u,v)) \right).
\]
Then $T$ is completely continuous and $T(C[0,1] \times C[0,1])$ is bounded in $C[0,1] \times C[0,1]$. By Schauder's fixed point theorem, $T$ has a fixed point. We can easily check that the fixed point of $T$ is a solution of $(H)$ and the proof is done. \qed

**Lemma 2.5** Assume that there exist $h_F, h_G \in C((0,1), \mathbb{R}^+)$ such that
\[
|F(t,u,v)| \leq h_F(t), \quad |G(t,u,v)| \leq h_G(t), \quad \text{for all } (t,u,v) \in (0,1) \times \mathbb{R}^2
\]
and satisfying (a2). Then $(H)$ has a solution.

**Proof.** We will use the truncation technique for this proof. Consider
\[
\begin{cases}
\varphi_p(u'(t))' + \eta_n(t)F(t,u(t),v(t)) = 0, & t \in (0,1), \\
\varphi_p(v'(t))' + \eta_n(t)G(t,u(t),v(t)) = 0, & t \in (0,1), \\
u(0) = A, u(1) = B, v(0) = C, v(1) = D,
\end{cases}
\]
where $n \geq 4$ and $\eta_n \in C[0,1]$ defined by
\[
\eta_n(t) = \begin{cases}
0, & 0 \leq t \leq \frac{1}{2n}, \\
1, & \frac{1}{n} \leq t \leq 1 - \frac{1}{n}, \\
0, & 1 - \frac{1}{2n} \leq t \leq 1,
\end{cases}
\]
Since \(|\eta_n(t)F(t, u, v)| \leq \eta_n(t)h_F(t), \ |\eta_n(t)G(t, u, v)| \leq \eta_n(t)h_G(t)| \in C([0, 1])\) and \(\eta_n(t)h_G(t) \in C([0, 1])\), by Lemma 2.4, we know that \((H^n)\) has a solution \((u_n, v_n)\). Define \(Z_F\) and \(Z_G\) by

\[
Z_F(t) = A + \int_0^t \varphi_p^{-1} \left( \varphi_p(B - A) + \int_s^t h_F(r)dr \right) ds \\
- B - \int_0^1 \varphi_p^{-1} \left( \varphi_p(A - B) + \int_s^t h_F(r)dr \right) ds,
\]

\[
Z_G(t) = C + \int_0^t \varphi_p^{-1} \left( \varphi_p(D - C) + \int_s^t h_G(r)dr \right) ds \\
- D - \int_0^1 \varphi_p^{-1} \left( \varphi_p(C - D) + \int_s^t h_G(r)dr \right) ds.
\]

We can check that \(Z_F(t)\) and \(Z_G(t)\) are strictly increasing in \((0, 1)\). \(Z_F(0^+) < 0 < Z_F(1^-)\) and \(Z_G(0^+) < 0 < Z_G(1^-)\). Thus, \(Z_F(t)\) and \(Z_G(t)\) have the unique zeros in \((0, 1)\). Let \(\alpha_F\) and \(\alpha_G\) be the unique zeros of \(Z_F(t)\) and \(Z_G(t)\), respectively. Let

\[
X(t) = \begin{cases} 
A + \int_0^t \varphi_p^{-1} \left( \varphi_p(B - A) + \int_s^t h_F(r)dr \right) ds, & 0 \leq t \leq \alpha_F, \\
B + \int_0^t \varphi_p^{-1} \left( \varphi_p(A - B) + \int_s^t h_F(r)dr \right) ds, & \alpha_F \leq t \leq 1,
\end{cases}
\]

\[
Y(t) = \begin{cases} 
C + \int_0^t \varphi_p^{-1} \left( \varphi_p(D - C) + \int_s^t h_G(r)dr \right) ds, & 0 \leq t \leq \alpha_G, \\
D + \int_0^t \varphi_p^{-1} \left( \varphi_p(C - D) + \int_s^t h_G(r)dr \right) ds, & \alpha_G \leq t \leq 1.
\end{cases}
\]

Then \((X, Y)\) is solution of

\[
\begin{align*}
\varphi_p(u'(t))' + h_F(t) = 0, & \quad t \in (0, 1), \\
\varphi_p(v'(t))' + h_G(t) = 0, & \quad t \in (0, 1), \\
u(0) = A, \ u(1) = B, \ v(0) = C, \ v(1) = D.
\end{align*}
\]

And let

\[
x(t) = \begin{cases} 
A + \int_0^t \varphi_p^{-1} \left( \varphi_p(B - A) - \int_s^t h_F(r)dr \right) ds, & 0 \leq t \leq \beta_F, \\
B + \int_0^t \varphi_p^{-1} \left( \varphi_p(A - B) - \int_s^t h_F(r)dr \right) ds, & \beta_F \leq t \leq 1,
\end{cases}
\]

\[
y(t) = \begin{cases} 
C + \int_0^t \varphi_p^{-1} \left( \varphi_p(D - C) - \int_s^t h_G(r)dr \right) ds, & 0 \leq t \leq \beta_G, \\
D + \int_0^t \varphi_p^{-1} \left( \varphi_p(C - D) - \int_s^t h_G(r)dr \right) ds, & \beta_G \leq t \leq 1.
\end{cases}
\]

\(\beta_F\) and \(\beta_G\) are the unique zeros of \(x_F(t)\) and \(z_G(t)\), respectively, where
\[ z(t) = A + \int_0^t \varphi_p^{-1} \left( \varphi_p(B - A) - \int_s^t h_F(r) dr \right) ds \\
- B - \int_t^1 \varphi_p^{-1} \left( \varphi_p(A - B) - \int_s^t h_F(r) dr \right) ds, \]

\[ z_G(t) = C + \int_0^t \varphi_p^{-1} \left( \varphi_p(D - C) - \int_s^t h_G(r) dr \right) ds \\
- D - \int_t^1 \varphi_p^{-1} \left( \varphi_p(C - D) - \int_s^t h_G(r) dr \right) ds. \]

Then \((x, y)\) is solution of

\[
\begin{align*}
\varphi_p(u(t))' - h_F(t) &= 0, & t \in (0, 1), \\
\varphi_p(v(t))' - h_G(t) &= 0, & t \in (0, 1), \\
u(0) = A, & u(1) = B, & v(0) = C, & v(1) = D.
\end{align*}
\]

We now claim that for all \(n\),

\[(x(t), y(t)) \leq (u_n(t), v_n(t)) \leq (X(t), Y(t)), \quad t \in [0, 1].\]

First, we will show that \((x(t), y(t)) \leq (u_n(t), v_n(t))\) for all \(t \in [0, 1]\) and \(n\). For the case \((u_n(t), v_n(t)) \leq (X(t), Y(t))\), we can check similarly. Suppose not, then \(x(t) \not\leq u_n(t)\) or \(y(t) \not\leq v_n(t)\). Assume \(x(t) \not\leq u_n(t)\). Since \(u_n(0) = x(0) = A\) and \(u_n(1) = x(1) = B\), there exists \(t_0 \in (0, 1)\) with \(u_n(t_0) < x(t_0)\). By continuity of \(x - u_n\), there exist \(a_n\) and \(b_n\) such that \(x(t) - u_n(t) > 0\) on \((a_n, b_n)\) and \(x(a_n) - u_n(a_n) = x(b_n) - u_n(b_n) = 0\). If we say that \(x(t) - u_n(t)\) has a positive maximum at \(t \in (a_n, b_n)\), then

\[ x'(t) = u_n'(t). \quad (2.1) \]

For \(a_n < s < b_n\),

\[ x'(s) = \varphi_p^{-1} \left( \varphi_p(u'(s)) - \int_s^t h_F(r) dr \right), \]

\[ u_n'(s) = \varphi_p^{-1} \left( \varphi_p(u_n'(s)) - \int_s^t \eta_n(r) F(r, u(r), v(r)) dr \right). \]

Integrate from \(a_n\) to \(b_n\), we get

\[ x(t) - x(a_n) = \int_{a_n}^t \varphi_p^{-1} \left( \varphi_p(x'(s)) - \int_s^t h_F(r) dr \right) ds, \]

\[ u_n(t) - u_n(a_n) = \int_{a_n}^t \varphi_p^{-1} \left( \varphi_p(u_n'(s)) - \int_s^t \eta_n(r) F(r, u(r), v(r)) dr \right) ds. \]
By monotonicity of $\varphi_p^{-1}$ and (2.1), we get the following contradiction

$$0 < x(\hat{t}) - u_n(\hat{t}) \leq 0.$$  

Similarly, we can get a same contradiction for the case $y(t) \not\leq v_n(t)$. Next, we show the equicontinuity of $\{(u_n, v_n)\}_{n=1}^{\infty}$ on $[0, 1]$. For any $\varepsilon > 0$, from the continuity of $x, y, X, Y$ on $[0, 1]$, it follows that there is $\delta_1 \in (0, \frac{1}{4})$ such that

$$A - \varepsilon \leq x(t) \leq X(t) < A + \varepsilon,$$

$$C - \varepsilon \leq y(t) \leq Y(t) < C + \varepsilon,$$

for $t \in [0, 2\delta_1]$. And

$$B - \varepsilon \leq x(t) \leq X(t) < B + \varepsilon,$$

$$D - \varepsilon \leq y(t) \leq Y(t) < D + \varepsilon,$$

for $t \in [1 - 2\delta_1, 1]$. Thus

$$|(u_n(t_1) - u_n(t_2)) - (v_n(t_1) - v_n(t_2))| \leq |u_n(t_1) - u_n(t_2)| + |v_n(t_1) - v_n(t_2)| < 4\varepsilon,$$

for $t_1, t_2 \in [0, 2\delta_1] \cup [1 - 2\delta_1, 1]$. Now, we consider $t_1, t_2 \in [\delta_1, 1 - \delta_1]$. We notice that

$$\sup_{t \in [\delta_1, 1 - \delta_1]} \{|\eta_n(t)F(t, u, v)| : (u, v) \in \mathbb{R}^2\} \leq \sup_{t \in [\delta_1, 1 - \delta_1]} h_F(t) < \infty, \quad (2.2)$$

$$\sup_{t \in [\delta_1, 1 - \delta_1]} \{|\eta_n(t)G(t, u, v)| : (u, v) \in \mathbb{R}^2\} \leq \sup_{t \in [\delta_1, 1 - \delta_1]} h_G(t) < \infty.$$

Also notice that for $t \in [\delta_1, 1 - \delta_1]$,

$$u_n(t) = u_n(\delta_1) + \int_{\delta_1}^{t} \varphi_p^{-1}\left(\tau_n + \int_{\tau_n}^{1-\delta_1} \eta_n(t)F(s, u_n(s), v_n(s))ds\right)dr,$$

where $\tau_n$ is a solution of the equation

$$\int_{\delta_1}^{1-\delta_1} \varphi_p^{-1}\left(\tau_n + \int_{\tau_n}^{1-\delta_1} \eta_n(t)F(s, u_n(s), v_n(s))ds\right)dr = u_n(1 - \delta_1) - u_n(\delta_1).$$

By Mean-Value Theorem, there is $\xi_n \in (\delta_1, 1 - \delta_1)$ with

$$\varphi_p^{-1}\left(\tau_n + \int_{\xi_n}^{1-\delta_1} \eta_n(t)F(s, u_n(s), v_n(s))ds\right) = \frac{u_n(1 - \delta_1) - u_n(\delta_1)}{(1 - \delta_1) - \delta_1}.$$

That is

$$\tau_n = -\int_{\xi_n}^{1-\delta_1} \eta_n(t)F(s, u_n(s), v_n(s))ds + \varphi_p\left(\frac{u_n(1 - \delta_1) - u_n(\delta_1)}{(1 - \delta_1) - \delta_1}\right).$$
By the boundedness of \( \{(u_n, v_n)\} \) and (2.2), there is \( C > 0 \) such that
\[
|\tau_n| \leq \int_{-\delta_1}^{1-\delta_1} h_F(s)ds + \varphi_p \left( \frac{|u_n(1-\delta_1) - u_n(\delta_1)|}{(1-\delta_1) - \delta_1} \right) < C.
\]

Then for \( t_1, t_2 \in [\delta_1, 1 - \delta_1] \),
\[
|u_n(t_1) - u_n(t_2)| = \left| \int_{t_1}^{t_2} \varphi_p^{-1}(r_n + \int_{s}^{1-\delta_1} \eta_n(t)F(s, u_n(s), v_n(s))ds) dr \right| \\
\leq L|t_1 - t_2|,
\]
where \( L = \varphi_p^{-1} \left( C + \int_{-\delta_1}^{1-\delta_1} h_F(s)ds \right) \). Put \( \delta_2 = \frac{\delta}{2} \), then for \( t_1, t_2 \in [\delta_1, 1 - \delta_1] \), \( |t_2 - t_1| < \delta_2 \) implies
\[
|u_n(t_1) - u_n(t_2)| \leq L|t_2 - t_1| < L\delta_2 = \varepsilon.
\]

Similarly, we can take \( \delta_3 \) such that \( |t_2 - t_1| < \delta_3 \) implies \( |v_n(t_1) - v_n(t_2)| \leq \varepsilon \). Take \( \delta = \min\{\delta_1, \delta_2, \delta_3\} \), then for \( t_1, t_2 \in [0, 1] \), \( |t_2 - t_1| < \delta \) implies \( |(u_n, v_n)(t_1) - (u_n, v_n)(t_2)| \leq 6\varepsilon \). Thus by Arzelà-Ascoli theorem, there is uniformly convergent subsequence of \( \{(u_n, v_n)\}_{n=4}^{\infty} \), say \( \{(u_n, v_n)\} \) which converges uniformly to \((u, v)\) on \([0, 1] \). It is clear that \( u(0) = A, v(0) = B, u(1) = C, v(1) = D \). Now, let \( \Gamma = [a, b] \subset (0, 1) \) be arbitrary compact interval, then there exists \( n^* = n^*(\Gamma) \) such that \( \Gamma \subset \left[ \frac{a}{n_n}, 1 - \frac{b}{n_n} \right] \). For \( n > n^* \) and \( t \in [a, b] \),
\[
0 = \varphi_p(u_n'(t))' + \eta_n(t)F(t, u_n(t), v_n(t)) \\
= \varphi_p(u_n'(t))' + F(t, u_n(t), v_n(t)) \\
0 = \varphi_p(v_n'(t))' + \eta_n(t)G(t, u_n(t), v_n(t)) \\
= \varphi_p(v_n'(t))' + G(t, u_n(t), v_n(t)).
\]
Thus \((u_n, v_n)\) is solution of \((H)\) for \( n > n^* \) and \( t \in [a, b] \). We notice that
\[
\sup \{|F(t, u, v)| : t \in [a, b], \ (u, v) \in \mathbb{R}^2 \} \leq \sup_{t \in [a, b]} h_F(t) < \infty,
\]
\[
\sup \{|G(t, u, v)| : t \in [a, b], \ (u, v) \in \mathbb{R}^2 \} \leq \sup_{t \in [a, b]} h_G(t) < \infty.
\]
Also notice that for \( t \in [a, b] \),
\[
u_n(t) = \int_{a}^{t} \varphi_p^{-1} \left( \tau_n + \int_{s}^{t} F(s, u_n(s), v_n(s))ds \right) dr,
\]
where \( \tau_n \) is a solution of the equation
\[
\int_{a}^{b} \varphi_p^{-1} \left( \tau_n + \int_{s}^{b} F(s, u_n(s), v_n(s))ds \right) dr = u_n(b) - u_n(a).
\]
By the similar argument, we can check that \( \{\tau_n\}_{n=1}^{\infty} \) is also bounded. Thus it has a convergent subsequence, say \( \{\tau_n\} \), which converges to \( \tau_F \). In addition we know that \( \{u_n\}_{n=1}^{\infty} \) has a convergent subsequence, say \( \{u_n\} \), which converges to \( u \) on \([a, b]\). Using Lebesgue Dominated Convergence theorem, we can get for \( t \in [a, b] \),

\[
    u(t) = u(a) + \int_{a}^{t} \varphi_p^{-1} \left( \tau_F + \int_{r}^{b} F(s, u(s), v(s)) \, ds \right) \, dr.
\]

Similarly, we can get for \( t \in [a, b] \),

\[
    v(t) = v(a) + \int_{a}^{t} \varphi_p^{-1} \left( \tau_G + \int_{r}^{b} G(s, u(s), v(s)) \, ds \right) \, dr.
\]

And \( \tau_F \) and \( \tau_G \) are the solutions of the equations

\[
    \int_{a}^{b} \varphi_p^{-1} \left( \tau_F + \int_{r}^{b} F(s, u(s), v(s)) \, ds \right) \, dr = u(b) - u(a)
\]

\[
    \int_{a}^{b} \varphi_p^{-1} \left( \tau_G + \int_{r}^{b} G(s, u(s), v(s)) \, ds \right) \, dr = v(b) - v(a).
\]

Thus \((u, v)\) is a solution of \((H)\) on the interval \( \Gamma = [a, b] \). Since \( \Gamma \) is arbitrary, \((u, v) \in C^1((0, 1), \mathbb{R}) \) and satisfying

\[
    \begin{cases}
        \varphi_p(u'(t))' + F(t, u(t), v(t)) = 0, & t \in (0, 1), \\
        \varphi_p(v'(t))' + G(t, u(t), v(t)) = 0, & t \in (0, 1), \\
        u(0) = A, & u(1) = B, \quad v(0) = C, & v(1) = D.
    \end{cases}
\]

Thus \((u, v)\) is a solution of \((H)\). \(\square\)

**Proof of Theorem 2.3.**

Define a modified function of \( F \) as follows

\[
    F_*(t, u, v) = \begin{cases}
        F(t, \beta_u(t), v) - h_F(t) \frac{u - \beta_u(t)}{1 + u - \beta_u(t)} & \text{if } u > \beta_u(t), \\
        F(t, u, v) & \text{if } \alpha_u(t) \leq u \leq \beta_u(t), \\
        F(t, \alpha_u(t), v) - h_F(t) \frac{u - \alpha_u(t)}{1 - u + \alpha_u(t)} & \text{if } u < \alpha_u(t),
    \end{cases}
\]

\[
    F^*(t, u, v) = \begin{cases}
        F(t, u, \beta_v(t)) & \text{if } v > \beta_v(t), \\
        F(t, u, v) & \text{if } \alpha_v(t) \leq v \leq \beta_v(t), \\
        F(t, u, \alpha_v(t)) & \text{if } v < \alpha_v(t),
    \end{cases}
\]

\[
    G_*(t, u, v) = \begin{cases}
        G(t, \beta_u(t), v) & \text{if } u > \beta_u(t), \\
        G(t, u, v) & \text{if } \alpha_u(t) \leq u \leq \beta_u(t), \\
        G(t, \alpha_u(t), v) & \text{if } u < \alpha_u(t),
    \end{cases}
\]

\[
    G^*(t, u, v) = \begin{cases}
        G(t, u, \beta_v(t)) & \text{if } v > \beta_v(t), \\
        G(t, u, v) & \text{if } \alpha_v(t) \leq v \leq \beta_v(t), \\
        G(t, u, \alpha_v(t)) & \text{if } v < \alpha_v(t),
    \end{cases}
\]
and
\[ G^*(t, u, v) = \begin{cases} 
G_s(t, u, \beta_v(t),) - h_G(t) \frac{v - \beta_v(t)}{1 + v - \beta_v(t)} & \text{if } v > \beta_v(t), \\
G_s(t, u, v) & \text{if } \alpha_v(t) \leq v \leq \beta_v(t), \\
G_s(t, u, \alpha_v(t)) - h_G(t) \frac{v - \alpha_v(t)}{1 - v + \alpha_v(t)} & \text{if } v < \alpha_v(t).
\end{cases} \]

Then \( F^*, G^* : (0, 1) \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \) are continuous and
\[
[ F^*(t, u, v) ] \leq 2 h_F(t), \\
[ G^*(t, u, v) ] \leq 2 h_G(t)
\]
for all \((t, u, v) \in (0, 1) \times \mathbb{R} \times \mathbb{R}\). By Lemma 2.5, the problem
\[
\begin{cases}
\varphi_p(u'(t))' + F^*(t, u(t), v(t)) = 0, & t \in (0, 1), \\
\varphi_p(v'(t))' + G^*(t, u(t), v(t)) = 0, & t \in (0, 1), \\
u(0) = A, & u(1) = B, \\
v(0) = C, & v(1) = D
\end{cases}
\tag{H^*}

has a solution. We claim that any solution \((u, v)\) of the problem \((H^*)\) satisfies
\[
(\alpha_u(t), \alpha_v(t)) \leq (u(t), v(t)) \leq (\beta_u(t), \beta_v(t)), \quad \text{for all } t \in [0, 1].
\]
Suppose, by contradiction, that \((\alpha_u, \alpha_v) \not\leq (u, v), i.e., \alpha_u \not\leq u \text{ or } \alpha_v \not\leq v).\) Assume that \(\alpha_u \not\leq u\). There exists \(t_0 \in (0, 1)\) such that \((u - \alpha_u)(t_0) = \min_{t \in [0, 1]} (u - \alpha_u)(t) < 0\). Then we know that \(u'(t_0) - \alpha_u'(t_0) = 0\) and \(v'(t) - \alpha_v'(t) \geq 0\) on \(t \in [t_0, t_0 + \delta]\), for some \(\delta > 0\). We consider two cases that \(\alpha_u(t_0) \leq v(t_0)\) or \(\alpha_v(t_0) > v(t_0)\). First, if \(\alpha_v(t_0) \leq v(t_0)\), then by \(\alpha_u(t_0) > u(t_0)\) and quasi-monotonicity of \(F\),
\[
F^*(t_0, u(t_0), v(t_0)) = F(t_0, \alpha_u(t_0), v(t_0)) - h_F(t_0) \frac{u(t_0) - \alpha_u(t_0)}{1 - u(t_0) + \alpha_u(t_0)} \\
> F(t_0, \alpha_u(t_0), v(t_0)) \\
\geq F(t_0, \alpha_u(t_0), \alpha_v(t_0)).
\]

Next, if \(\alpha_v(t_0) > v(t_0)\),
\[
F^*(t_0, u(t_0), v(t_0)) = F(t_0, \alpha_u(t_0), \alpha_v(t_0)) - h_F(t_0) \frac{u(t_0) - \alpha_u(t_0)}{1 - u(t_0) + \alpha_u(t_0)} \\
> F(t_0, \alpha_u(t_0), \alpha_v(t_0)).
\]

By continuity of \(F^*\), there is \(\varepsilon_0 > 0\) such that for \(t \in [t_0, t_0 + \varepsilon_0],\)
\[
F^*(t, u(t), v(t)) > F(t, \alpha_u(t), \alpha_v(t)).
\]

For \(t \in [t_0, t_0 + \varepsilon_0],\)
\[
\varphi_p(u'(t)) = \varphi_p(u'(t_0)) - \int_{t_0}^{t} F^*(s, u(s), v(s))ds \\
\leq \varphi_p(u'(t_0)) - \int_{t_0}^{t} \frac{u(s) - \alpha_u(s)}{1 - u(s) + \alpha_u(s)} \varphi_p(u'(s))ds \\
\leq \varphi_p(u'(t_0)) - \int_{t_0}^{t} (\varphi_p'(\alpha_u'(s)))'ds = \varphi_p(\alpha_u'(t)).
\]
By monotonicity of $\varphi_p$, we obtain a contradiction that $u'(t) < \alpha'_u(t)$ for $t \in [t_0, t_0 + \varepsilon_0)$. And we can say the same contradiction for the case that $\alpha_u \not\leq v$. Thus

$$(\alpha_u(t), \alpha_v(t)) \leq (u(t), v(t)).$$

Similarly, we can show that

$$(u(t), v(t)) \leq (\beta_u(t), \beta_v(t)).$$

\[\square\]

References


