DISTRIBUTED ALGORITHMS FOR NETWORKED MULTI-AGENT SYSTEMS: OPTIMIZATION AND COMPETITION

BY

JAYASH KOSHAL

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Industrial Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Associate Professor Angelia Nedić, Chair
Associate Professor Vinayak V. Shanbhag, Contingent Chair
Professor Jong-Shi Pang
Professor Bruce Hajek

Abstract

This thesis pertains to the development of distributed algorithms in the context of networked multi-agent systems. Such engineered systems may be tasked with a variety of goals, ranging from the solution of optimization problems to addressing the solution of variational inequality problems. Two key complicating characteristics of multi-agent systems are the following: (i) the lack of availability of system-wide information at any given location; and (ii) the absence of any central coordinator. These intricacies make it infeasible to collect all the information at a location and preclude the use of centralized algorithms. Consequently, a fundamental question in the design of such systems is the need for developing algorithms that can support their functioning. Accordingly, our goal lies in developing distributed algorithms that can be implemented at a local level while guaranteeing a global system-level requirement. In such techniques, each agent uses locally available information, including that accessible from its immediate neighbors, to update its decisions, rather than availing of the decisions of all agents. This thesis focuses on multi-agent systems tasked with the solution of three sets of problems: (i) convex optimization problems; (ii) Cartesian variational inequality problems; and (iii) a sub-class of Nash games.

In the first part of this thesis, we consider a multiuser convex optimization problem. Traditionally, a multiuser problem is a constrained optimization problem characterized by a set of users (or agents). Such problems are characterized by an objective given by a sum of user-specific utility functions, and a collection of separable constraints that couple user decisions. We assume that user-specific utility information is private while users may communicate values of their decision variables. The multiuser problem is to maximize the sum of the users-specific cost functions subject to the coupling constraints, while abiding by the informational requirements of each user. In this part of the thesis, we focus on generalizations of convex multiuser optimization problems where the objective and constraints are not separable by user and instead consider instances where user decisions are coupled, both in the objective and through nonlinear coupling constraints. To solve this problem, we consider the application of distributed gradient-based algorithms on an approximation of the multiuser problem. Such an approximation is obtained through a regularization

and is equipped with bounds of the difference between the optimal function values of the original problem and its regularized counterpart. In the algorithmic development, we consider constant stepsize primal-dual and dual schemes in which the iterate computations are distributed naturally across the users, i.e., each user updates its own decision only. We observe that a generalization of this result is also available when users choose their stepsize and regularization parameters independently from a prescribed range.

The second part of this thesis is devoted to the solution of a Cartesian variational inequality (VI) problem. A Cartesian VI provides a unifying framework for studying multi-agent systems including regimes in which agents either cooperate or compete in a Nash game. Under suitable convexity assumptions, sufficiency conditions of such problems can be cast as a Cartesian VI. We consider a monotone stochastic Cartesian variational inequality problem that naturally arise from convex optimization problems or a subclass of Nash games over continuous strategy sets. Almost sure convergence of standard implementations of stochastic approximation rely on strong monotonicity of the mappings arising in such variational inequality problems. Our interest lies in weakening this requirement and this motivates the development of distributed iterative stochastic approximation algorithms. We introduce two classes of stochastic approximation methods, each of which requires exactly one projection step at every iteration, and provide convergence analysis for them. Of these, the first is a stochastic iterative Tikhonov regularization method which necessitates the update of regularization parameter after every iteration. The second method is a stochastic iterative proximal-point method, where the centering term is updated after every iteration. Conditions are provided for recovering global convergence in limited coordination extensions of such schemes where agents are allowed to choose their stepsize sequences, regularization and centering parameters independently, while meeting a suitable coordination requirement. We apply the proposed class of techniques and their limited coordination versions to a stochastic networked rate allocation problem.

The focus of the third part of the thesis is on a class of games, termed as aggregative games, being played over a networked system. In an aggregative game, an agent's objective function is coupled across agents through a function of the aggregate of all agents decisions. Every agent maintains an estimate of the aggregate and agents exchange this information over a connected network. We study two classes of distributed algorithm for information exchange and computation of equilibrium. The first method, a diffusion-based algorithm, operates in a synchronous setting which can contend with time-varying connectivity of the underlying network graph model. The second method, a gossip-based distributed algorithm, is inherently asynchronous and is applicable when the network is static. Our primary emphasis is on proving the convergence of these algorithms under an assumption of a diminishing (agent-specific) stepsize sequence. Under standard conditions, we establish the almost-sure convergence of these algorithms to an equilibrium point. Moreover, we also develop and analyze the associated error bounds when a constant stepsize (user-specific) is employed in the gossip-based method. Finally, we present numerical results to assess the performance of the diffusion and the gossip algorithm for a class of aggregative games for various network models and sizes.

To my parents, brother and friends.

Acknowledgements

No words can possibly capture my deep gratitude towards my advisers, Professor Angelia Nedić and Professor Uday V. Shanbhag. Over the past four years, they have sustained me with their guidance and patience which has been instrumental in shaping this work. Not only they have been a great source of knowledge and inspiration, their great efforts and sound advice, constant support, unfailing encouragement and unflinching trust in my academic abilities has always fuelled my motivation to excel in all spheres of life. The brisk completion of my PhD is an outcome of their foresight and meticulous planning coupled with impeccable execution. I shall be indebted to all their efforts throughout my life.

I am grateful to Professor Jong-Shi Pang and Professor Bruce Hajek for accepting to be a member of my doctoral committee members. Their helpful stylistic criticisms and characteristically stylish suggestions made significant improvement in the quality of this work. I would also like to thank Professor Jong-Shi Pang for teaching me complementarity theory. It was truly an honour to learn from him.

I take this opportunity to sincerely acknowledge the ISE Department at University of Illinois at Urbana-Champaign (UIUC) and National Science Foundation for providing financial assistance in the form of Research/Teaching Assistantship and CMMI 09-48905 EAGER ARRA award, respectively, which buttressed me to perform my work comfortably.

I consider myself extremely fortunate to have an excellent set of friends at UIUC. I have been helped in various ways by many people and the following are the one who are primarily responsible to making life at UIUC wonderfully enjoyable.

- My officemate from TB 406, Ankur, Aswin, Behrouz, Fuyuan, Farzad, Rasoul and Uma helped in making an environment conducive for fun-filled learning. Rasoul has been a great motorcycle companion over the last year, an activity which I am going to miss dearly. The presence of Fuyuan has significantly lightened the mood and the glorious time spent with her is forever etched in my memory.

- My others friends from ISE namely, Alok, Dane, Deniz, Limeng, Vijitashwa and Yerkin had an enormous impact on the time spent in Champaign. I specially thank Ankur, Alok, John and Vijitashwa for being a great friend cum mentor and for always willing to help me in every possible way.

- My classmates from IIT Bombay, Deep and Pravesh, along with Avinash, Hitesh and Siddharta shared wonderful initial years. Their friendship and laughter certainly made the transition into Champaign- Urbana less gruelling.
- I would specially like to thank Alok, Fuyuan, John and Yuhan for taking an initiative of forming a study group which helped me honing my skills and also learn new ones. The discussions with them on various topics were insightful yet entertaining. I am indebted to John in helping me envision a good career goal and also pave the path to attain one.

I am grateful to the proficient staff at the ISE Department: Amy, Debi, Donna, Holly, Kevin, Lynnell, Randy and Regina who have taken onus of innumerable arduous tasks making the life at ISE Department smooth. Special thanks to Amy, for making the coordination with Prof. Pang extremely smooth, and Holly, for her help in the entire process even during her vacation.

Finally, I dedicate this thesis to my parents and my bother. It is a small token to honour the immense sacrifice and hardship of my parents over the years. Their love, support and trust has enabled to me take that extra step and be infinitesimally better.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

a.s.	Almost sure
ITR	Iterative Tikhonov Regularization
PITR	Partially coordinated Iterative Tikhonov Regularization
IPP	Iterative Proximal-Point
PIPP	Partially coordinated Iterative Proximal-Point

List of Symbols

$\mathbb{R}$	Set of all real numbers
$\mathbb{R}_+$	Set of non-negative real numbers
$\|\cdot\|$	Euclidean norm of a vector
Π_X	Euclidean projection on set X
$\mathrm{VI}(K,F)$	Variational inequality problem for mapping F over set K
$\mathrm{SOL}(K,F)$	Set containing solution(s) of $\mathrm{VI}(K,F)$
$\mathbf{F}^{\mathrm{nat}}$	Natural map of $\mathrm{VI}(K,F)$
N	Number of agents in the network
i	Index for ith agent
x_i	Agent i decision variable
K_i	Set constraint on agent i decision variable
x_{-i}	The tuple $(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_N)$
x	The tuple $(x_i;x_{-i})$ with x_i as variable and x_{-i} as parameter
$\acute{x}$	A Slater vector
$\bar{x}$	The aggregate of the component of x, i.e. $\bar{x} = x_1 + \ldots + x_N$
$\bar{K}$	Minkowski's sum of set K_i, i.e. $\bar{K} = K_1 + \ldots + K_N$
x^*	Optimal solution or equilibrium point
K^*	Set consisting of all optimal solution or equilibrium points
$f_i(x)$	Objective function for agent i with x_{-i} viewed as parameter
x_i^k	Iterate of agent i at the kth update

α_i	Constant steplength for agent i
$\alpha_{k,i}$	Steplength for agent i for the kth iteration
$D(a)$	Diagonal matrix with diagonal entries being the components of the vector a
$[x]_l$	The lth coordinate of the vector x
$[A]_{ij}$	The jth entry of the i row of the matrix A.
$\mathbb{I}$	Identity matrix
e_i	A vector with ith entry 1
λ	Dual multiplier
λ^k	Dual multiplier iterate at kth update
$\mathcal{L}(x,\lambda)$	Lagrangian function
$\nabla_x \mathcal{L}(x,\lambda)$	Gradient w.r.t. to x of the Lagrangian
$\nabla_\lambda \mathcal{L}(x,\lambda)$	Gradient w.r.t. to λ of the Lagrangian
ν	Regularization parameter in primal space (space of x)
ν_i	Regularization parameter of agent i
V	Diagonal matrix with diagonal entries being agent dependent regularization parameters
ε	Regularization in dual space (space of λ)
Λ^*	Set consisting of optimal dual multipliers
$\mathcal{D}$	Set containing the set of optimal dual multipliers
w^k	Stochastic error in evaluating the gradient
$\mathcal{F}_k$	$\sigma-$field generated by stochastic errors
$\mathbb{E}[X]$	Expectation of random vector X
$\mathbb{E}[X\|\mathcal{F}_k]$	Conditional expectation of random variable X conditioned on the σ-algebra $\mathcal{F}_k$

Chapter 1

Introduction

The theory of complex networked systems continues to galvanize a wide range of scientific communities including those in mathematics, engineering, computer science, biology, sociology, among others. Complex networks arise in a multitude of application regimes and this ubiquity has inspired a range of questions in the development and analysis of algorithms. While early inquiries focused on gaining *consensus* on a particular parameter under possibly evolving connectivity graphs, subsequent efforts have concentrated on the optimization of multi-agent systems. The field of distributed algorithms continues to grow at a brisk pace and presents a plethora of challenging problems that will have profound impact on the ability to optimize and control networked systems as well as understand their associated emergent behavior.

A complex networked system may be modelled as a graph where the nodes represent agents and the underlying connectivity graph provides the foundation for information exchange amongst these agents[1]. The resulting multi-agent system consists of agents communicating and exchanging information over a complex topological network. Following are some of the important characteristics of a multi-agent system:

1. *Limited communication*: In truly large networks, agents may communicate with only a subset of other agents, often referred to as its neighbors;

2. *Lack of a central coordinator*: There is no designated central coordinator tasked with the collection and broadcast of system-wide information to make it accessible to every agent;

3. *Local information*: Each agent has access to only its local information and may control only its own decision.

Multi-agent systems, as mentioned earlier, may be tasked with a wide array of functionalities, ranging from consensus to optimization to equilibrium seeking and may operate in uncertain and dynamic environments. The precise implications of these characteristics are context sensitive and

[1] In competitive regimes, the words agent and player will be used interchangeably through this monograph.

their discussion is delayed until the next section. Examples of multi-agent systems include, but are not limited to: (i) Computer networks where computers are the agents and communication channels (wired or wireless) form the edges; (ii) Transportation networks where the routes (edges) connect various locations (nodes); (iii) Social networks where individual or organizations are the agents and the dyadic ties between them form the connection.

Naturally, the type of network significantly influences the problem at hand. To address the challenges posed by networked systems, two broad classes of design approaches may be employed. The first of these is a *top-down* approach and entails decomposing a large system into smaller subsystems that are relatively easy to analyse. A hierarchical chain of decompose and analyse is executed, and so until the final subsystem falls into a class of problems that is well understood. In contrast, the second approach, referred to as the *bottom-up* approach, takes an altogether different path. Here, almost all the effort is devoted to the atomic level problem. Once the entities at the atomic level are made self-efficient, all that is left to be done is an assembly, which itself may be automated.

To highlight the subtle differences across the approaches, we consider the following question:

> *Suppose we want to understand the pattern formation of a large population of migratory birds through a mathematical simulation? What would be the right approach for developing such a computer simulation?*

Any experienced programmer would ascertain that almost all the effort must be devoted to modelling the behaviour of the single entity, the bird. The above example, though naïve, serves the purpose of highlighting the importance of a bottom-up approach in truly large complex networked systems. Bottom-up approaches start with a rigorously predefined set of rules for agent behaviour and interactions, and then a desirable system-level behaviour emerges through such interactions.

With the advent of computers and other advanced technologies that came along with them, the size of network systems increased from large to colossal. This has led to the development and refinement of *distributed algorithms* for multi-agent system theoretic problems, a topic of central importance of this research work.

1.1 Distributed algorithms in multi-agent systems

This thesis focuses on the development of distributed algorithms in multi-agent systems. The local information structure of the multi-agent systems when coupled with the absence of a central coordinator does impose a host of restrictions on choices of algorithms. Thus in this work, we focus on developing distributed algorithms that are in compliance with the distributed information structure of the problem:

(i) Limited informational requirements: Any given user does not have access to the utility functions or the constraints of other users;

(ii) Single iteration schemes: Two nested iterative schemes require coordination across outer and inner iteration level, a challenging prospect in networked setting. Our goal lies in developing *single iteration* schemes that require less coordination;

(iii) Limited coordination of algorithm parameters: In truly large-scale networks, enforcing consistency across algorithm parameters is often challenging and ideally, one would like minimal coordination across users in specifying algorithm parameters;

(iv) Uncertainty: In many practical problems some elements may involve uncertain data, for instance, communication over the wireless medium.

The aforementioned criteria will persistently feature in all of the algorithms that are developed and studied in this thesis. More generally, the theme of algorithms fall in the bottom-up design approach. In effect, our algorithms endow each agent in the network with a mandatory protocol, and we then establish the emergence of a desirable behaviour.

The nature of agents in the sense of cooperation gives rise to different problem classes within multi-agent systems. For instance, if the agents are cooperative then the multi-agent system is essentially a multiuser optimization problem where all agents are collectively working towards a common goal. A popular choice of these problems is one where agents are collectively optimizing the sum of local individual objective[2] or a min-max problem where the worst case scenario loss is optimized. A problem of this kind may arise in network resource allocation such as rate allocation among multiple users, where the coupling cost may be due to congestion or delay, while the coupling constraints may be due to the network link capacities [1, 2, 3].

While multiuser optimization problems present a way to capture cooperative behaviour of the multi-agent system, Nash games may be natural models for capturing the strategic behaviour in a multi-agent system. A competitive counterpart of multiuser optimization problems is a Nash game. More precisely, agents are only interested in optimizing their own individual goals which are also affected by decisions of other agents in the network. Game-theoretic models find application in a range of settings ranging from wired and wireless communication networks [4, 5, 6], bandwidth

[2]The total sum of individual goals is considered to be a good measure of fair allocation.

allocation [7, 8, 5], spectrum allocation in radio networks [9], cognitive radio networks [10], Nash-Cournot games [11, 12] and optical networks [13, 14]. While aforementioned work lies in continuous strategy games, [15, 16] have examined a range of questions, under varying informational assumptions, in the regime of finite-strategy games.

The connection between multiuser optimization and Nash games is interesting from two important standpoints. The first of these is the *efficiency*[3]. The efficiency is a measure used to assess the quality of a Nash equilibrium with respect to the centralized welfare problem. More precisely, a Nash equilibrium is regarded as efficient if no single player can be made better-off without making at least one other player worse off. Interestingly, an efficient Nash equilibrium for a restricted class of Nash games is characterized as an optimal solution of a centralized optimization problem[4]. Notably, Nash games are known to arise in engineered multi-agent systems, as motivated by the development of decentralized control schemes. For instance, in massive networks, it may be impossible to exercise direct control over the users and a possible approach lies in allowing users to behave selfishly and compete for resources. If the aggregate cost of the Nash equilibria associated with the Nash counterpart is identical to its centralized counterpart, then the equilibria are deemed as efficient. In effect, the game achieves the same outcome as the centralized optimization problem.

We now discuss the second connection standpoint between multiuser optimization problems and Nash games, the *computation*. For the design of algorithms for computation of optimal or equilibrium point, the theory of *variational inequalities* has been instrumental. Variational inequalities (VI) provide a unifying framework for studying and analyzing both multiuser optimization problems and convex Nash games. In fact, under convexity assumptions, the (sufficient) optimality conditions of a multiuser optimization problem or the equilibrium condition of a Nash game give rise to a Cartesian variational inequality. We have a responsibility of explaining the main advantages of taking the VI path, and we take this towards the end of this section. We pause here to briefly introduce the theory of VI which is of importance to this work (see xiii for an overview of notation). To begin with, the problem $\mathrm{VI}(K,F)$ requires finding a vector $x \in K \subseteq \mathbb{R}^n$ such that

$$(y-x)^T F(x) \geq 0 \quad \text{for all } y \in K, \qquad \mathrm{VI}(K,F)$$

where $F : K \to \mathbb{R}^n$. Usually, K is a closed convex set and F is a continuous map. Additionally, when the set K is a Cartesian set, the $\mathrm{VI}(K,F)$ is called a Cartesian variational inequality.

[3]To be more precise, the notion of efficiency considered here is Pareto-efficiency.
[4]The generality of this statement and its discussion is beyond the scope of this thesis.

In this thesis, we also consider a stochastic Cartesian variational inequality problem SVI(K,F) for a mapping F over the Cartesian set K, which can be viewed as a natural extension of deterministic Cartesian variational inequality problem VI(K,F) with the mapping F given as an expectation of some uncertainty. Due to lack of analytical expression of F, approaches of deterministic VI cannot be directly applied. Our interest is in developing *distributed* stochastic approximation schemes for such a problem when the mapping F is monotone[5] over the Cartesian set K.

We resume with our discussion of problems arising in multi-agent systems. So far we have eluded the discussion of two important aspects of of multi-agent systems, specifically the information and communication. In fact, throughout our preceding discussion on multi-agent systems we have been implicitly assuming that agents have the access to the desired information either directly or through observation. In context of multiuser optimization problems or Nash games, this essentially requires every agent to have access to decision of all the other agents in the network; however, utility functions and constraint sets are not public knowledge. Such an assumption may be difficult to satisfy in many application, especially in truly large networks. Inspired by the need to address this restriction, we next consider a game-theoretic networked setting where agents have limited access to the decisions of other agents.

Towards this end, we consider a special class of Nash games called aggregative games, being played over a network of agents. An aggregative games is a game with the specific property that each player's cost function is represented as a function which depends only upon its decision and an aggregate of the decisions of all agents in the network. Aggregative games enjoy the luxury of belonging to a class of Nash games for which existence and uniqueness results are well-established. Problems arising in networked Nash-Cournot games [17, 18, 19, 20], rate allocation in communication networks including public goods games [21], common resource games [22], cost sharing games [23] usually belong to the the class of aggregative games. In all the preceding work related to the computation of Nash equilibrium, it is intrinsically assumed that each player is aware of the aggregate of the network. However, in the absence of a central coordinator, computation of the aggregate and its dissemination is a challenging proposition. The problem aggravates further if the players are precluded from sharing their decision with any other player but are allowed to share their belief of the aggregate information. To motivate the situation arising in the preceding discussion, consider the following question:

> *A group of people (at least 3) are in a meeting and they are all interested in knowing the average salary (or any other personal characteristics, age for instance). How-*

[5] A mapping $F : K \to \mathbb{R}^n$ is said to be monotone over a set $K \subseteq \mathbb{R}^n$ if $(F(x)-F(y))^T(x-y) \geq 0$ for all $x, y \in K$. It is said to be strictly monotone over K if $(F(x)-F(y))^T(x-y) > 0$ for all $x, y \in K$ with $x \neq y$. In addition, it is said to be strongly monotone if there exists a positive scalar η such that $(F(x)-F(y))^T(x-y) \geq \eta\|x-y\|^2$ for all $x, y \in K$.

ever, being conscious and also humble, they abstain to disclose their own information. Is there a way to compute the average abiding the informational restriction?[6]

Under such a restrained scenario, a distributed algorithm can possibly aid agents to to build an estimate of true information by constructively using partial information obtained through communication with their immediate neighbours (possibly time varying). Our interest is lies in exploring this avenue and establishing the emergence of a Nash equilibria following a distributed protocol.

More generally, the distributed methods are developed and studied in this thesis with the goal of computation of optimal decision in case of optimization problem or equilibrium decision in game-theoretic regimes. However, we follow a very similar approach in all the problems that we address: we attempt to cast the original problem into a Cartesian variational inequality. The solution to the Cartesian variational inequality originating through such a transformation is also a solution to the original problem. This approach provides us with two advantages: first, the ability to discuss the existence/uniqueness of a solution; and second, an avenue for constructing convergent methods to obtain these solutions. In particular, a solution to Cartesian VI also solves a fixed point equation of the related natural map which serves as the impetus behind all our algorithmic endeavours.

Our primary interest is in first-order methods, as these methods have relatively small overhead per iteration. They also exhibit stable behaviour in the presence of various sources of noise in the computations, as well as in the information exchange due to possibly noisy links in the underlying communication network over which the users communicate. Another important and desirable property of distributed algorithms is scalability: the computational effort to solve the problem should grow slowly (ideally linearly) with the size of the network[7].

A final note is in order regarding certain terms that we use in this work. The term "error analysis" pertains to the development of bounds on the difference between a given solution or function value and its optimal counterpart. The term "coordination" assumes relevance in distributed schemes where certain algorithmic parameters may need to satisfy a prescribed requirement across all users. Finally, it is worth accentuating why our work assumes relevance in implementing distributed algorithms in practical settings. In large-scale networks, the success of standard distributed implementations is often contingent on a series of factors. For instance, convergence often requires that steplengths match across users, exact/inexact solutions are available in bounded time intervals and finally, users have access to recent updates by the other network participants. In practice, algorithms may not subscribe to these restrictions and one may be unable to specify the choice of algorithm parameters, such as steplengths and regularization parameters, across users. Accord-

[6]This is adapted from a famous interview question.

[7]The number of agents in the network is often a good proxy for the size of the network.

ingly, we extend standard fixed-steplength gradient methods to allow for heterogeneous steplengths and diversity in other algorithmic parameters.

The remainder of the chapter is organised as follows. We provide a brief summary of distributed algorithms in the context of multi-agent system and formally set out our goals. We consider three diverse problems arising in a multi-agent system, the treatment of which spans the next three chapters. We describe each of them briefly in Sections 1.2, 1.3 and 1.4.

1.2 Multiuser optimization problem

In this thesis we consider the generic forms of multiuser problems. A multiuser problem is a constrained optimization problem associated with a finite set of N users (or players). Each user i has a convex cost function $f_i(x_i)$ that depends only on its decision vector x_i. The decision vectors x_i, $i = 1, \ldots, N$ are typically subject to a finite system of linear inequalities $\sum_{i=1}^{N} a_{ji}^T x_i \leq b_j$ for $j = 1, \ldots, m$, which couple the user decision variables. The traditional multiuser problem is formulated as a convex minimization of the form

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} f_i(x_i) \\ \text{subject to} \quad & \sum_{i=1}^{N} a_{ji}^T x_i \leq b_j, \quad j = 1, \ldots, m \\ & x_i \in K_i, \qquad\qquad i = 1, \ldots, N, \end{aligned} \tag{1.1}$$

where K_i is the set constraint on user i decision x_i (often K_i is a box constraint). In many applications, users are characterized by their payoff functions rather than cost functions, in which case the multiuser problem is a concave maximization problem. Often, the informational restrictions dictate that the ith user only has access to his objective f_i and constraint set K_i. Furthermore, user i can modify only its own decision x_i but may observe the decisions $(x_j)_{j \neq i}$ of the other users. In effect, every user can see the entire vector x. Consequently, distributed schemes that abide by these requirements are of relevance.

The prior work on multiuser optimization problem arising in network resource allocation [1, 2, 3, 24, 25, 26] is dealing with users with *separable* objectives, but coupled *polyhedral* constraints. Methods discussed therein are typically in a continuous-time setting (with exception for [3] where discrete-time schemes are investigated). Discrete-time (approximate) schemes, combined with simple averaging, have been studied in [27, 28, 29] for a general convex constrained formulation. However, all of the aforementioned work establishes the convergence properties of therein proposed algorithms under the assumption that the users coordinate their steplengths, i.e., the steplength values are equal across all users.

This thesis work generalizes the standard multiuser optimization problem, defined in (1.1), in two distinct ways: (i) The user objectives are coupled by a congestion metric (as opposed to being separable). Specifically, the objective in (1.1) is replaced by a system cost given by $\sum_{i=1}^{N} f_i(x_i) + c(x_1,\ldots,x_N)$, with a convex coupling cost $c(x_1,\ldots,x_N)$; and (ii) The linear inequalities in (1.1) are replaced with general convex inequalities. In effect, the constraints are nonlinear and not necessarily separable by user decisions.

To this end, consider a generalization to the canonical multiuser optimization problem of the following form:

$$\begin{aligned} \text{minimize} \quad & f(x) \triangleq \sum_{i=1}^{N} f_i(x_i) + c(x) \\ \text{subject to} \quad & d_j(x) \leq 0 \qquad \text{for all } j = 1,\ldots,m, \\ & x_i \in K_i \qquad \text{for all } i = 1,\ldots,N, \end{aligned} \tag{1.2}$$

where N is the number of users, $f_i(x_i)$ is user i cost function depending on a decision vector x_i and K_i is the constraint set for user i. The function $c(x)$ is a joint cost that depends on the user decisions, i.e., $x = (x_1,\ldots,x_N)$. The functions f_i, c and d_j are *convex and continuously differentiable*.

Before proceeding on our approach to solve this problem, we motivate the problem of interest via an example drawn from communication networks [1, 2], which can capture a host of other problems (such as in traffic or transportation networks).

Example 1. *Consider a network (see Fig 1.1) with a set of J link constraints and b_j being the finite capacity of link j, for $j \in J$. Let R be a set of user-specific routes, and let A be the associated link-route incidence matrix, i.e., $A_{jr} = 1$ if $j \in r$ implying that link j is traversed on route r, and $A_{jr} = 0$ otherwise.*

Suppose, the rth user has an associated route r and a rate allocation (flow) denoted by x_r. The corresponding utility of such a rate is given by $U_r(x_r)$. Assume further that utilities are additive implying that total utility is merely given by $\sum_{r\in R} U_r(x_r)$. Further, let $c(x)$ represent the congestion cost arising from using the same linkages in a route. Under this model the system optimal rates solve the following problem.

$$\begin{aligned} \textit{maximize} \quad & \sum_{r\in R} U_r(x_r) - c(x) \\ \textit{subject to} \quad & Ax \leq b, \quad x \geq 0. \end{aligned} \tag{1.3}$$

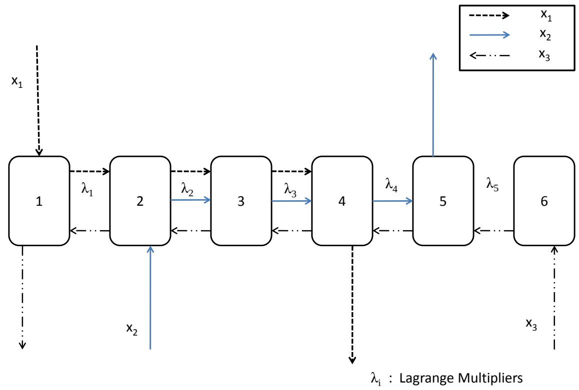


Figure 1.1: A network with 3 users and 5 links.

To handle these generalizations of the multiuser problem, we propose approximating the problems with their regularized counterparts and, then, solving the regularized problems in a distributed fashion in compliance with the user specific information (user functions and decision variables). We provide an error estimate for the difference between the optimal function values of the original and the regularized problems. For solving the regularized problems, we consider distributed primal-dual and dual approaches, including those requiring inexact solutions of Lagrangian subproblems. We investigate the convergence properties and provide error bounds for these algorithms using two different assumptions on the stepsizes, namely that the stepsizes are the same across all users and the stepsizes differ across different users. These results are extended to regimes where the users may select their regularization parameters from a broadcasted range. Hence, these algorithms satisfy the requirement we set out in the beginning of this chapter, namely, limited informational requirement, single level iteration, limited coordination of algorithm parameters [30, 31].

1.3 Stochastic Cartesian variational inequality problems

We consider a Cartesian monotone stochastic variational inequality problem, which is a problem that requires finding a vector $x = (x_1, \ldots, x_N)$ satisfying

$$\begin{aligned} (x_1 - y_1)^T F_1(x) \geq 0, &\quad \forall y_1 \in K_1, \\ &\vdots \\ (x_N - y_N)^T F_N(x) \geq 0, &\quad \forall y_N \in K_N, \end{aligned} \tag{1.4}$$

where $F_1(x) = \mathbb{E}[\tilde{f}_1(x,\xi)]$ and $\mathbb{E}[]$ denotes the expectation with respect to some uncertainty ξ. Our interest is in developing *distributed* stochastic approximation schemes for such a problem when the mapping

$$F(x) = (F_1(x), \ldots, F_N(x))$$

is monotone over the Cartesian product of the sets K_i, i.e., $(F(x)-F(y))^T(x-y) \geq 0$ for all $x,y \in K_1 \times \cdots \times K_N$. Cartesian stochastic variational inequalities arise from both stochastic optimization and game-theoretic problem as illustrated next. Consider a Nash game in which the ith player solves the following problem:

$$\begin{aligned} \text{minimize} \quad & f_i(x) \triangleq \mathbb{E}[f_i(x_i, x_{-i}, \xi_i)] \\ \text{subject to} \quad & x_i \in K_i. \end{aligned} \tag{1.5}$$

When $f_i(x)$ is convex over K_i for all $i = 1, \ldots, N$, then the equilibrium (sufficient) conditions of this problem are given by (1.4) where $F_i(x) = \nabla_{x_i}\mathbb{E}[f_i(x,\xi)]$ for $i = 1, \ldots, N$, and the resulting mapping F is monotone over $K_1 \times \cdots \times K_N$. This can be rewritten as a stochastic variational inequality SVI(K,F) where $F = (F_1^T, \ldots, F_N^T)^T$ and $K = K_1 \times \cdots \times K_N$.

For the Cartesian stochastic variational inequalities, it seems natural to exploit the presence of decoupled constraint sets and develop distributed schemes. When considering multiuser optimization problems, distributed optimization approaches have been natural candidates (cf. [2]). Yet, there appears to have been markedly little on stochastic variational inequalities that naturally extend multiuser optimization. Our work in this thesis intends to fill the lacuna through this framework. In a deterministic setting, distributed schemes for computing equilibria arising from monotone Nash games have received significant attention recently [5, 7, 32, 33, 14, 6, 34, 12]. Of particular relevance is the work in [14, 33], the latter employing an extragradient scheme [35] capable of accommodating deterministic monotone Nash games. Finally, Scutari et al. [10] examine an array of monotone Nash games and consider proximal-point based distributed schemes in a basic prox-setting where a sub-problem is solved at each iteration.

To the best of our knowledge, [36] appears to be the only existing work considering stochastic approximation methods for variational inequalities. Moreover, the convergence therein is established under the strong monotonicity assumption of mapping F. The work in this thesis is motivated by the challenges associated with solving stochastic variational problems when the mappings lose strong monotonicity. In solving deterministic variational inequalities, such a departure is ably handled through techniques, such as Tikhonov regularization [37, 38] and proximal-point [39]. A direct implementation of regularization techniques leads to a two nested level methods which are difficult to implement in a distributed networked setting. And the presence of uncertainty aggravates the implementation further. In the present work we overcome this challenge by emphasizing on iterative regularization for stochastic variational inequalities with monotone maps. In Chapter 3, we present and analyze two stochastic iterative regularization schemes:

1. *Stochastic iterative Tikhonov regularization method;*

2. *Stochastic iterative proximal-point method.*

Each of these schemes requires exactly one projection step at every iteration with users having autonomous choice of algorithmic parameters. Under some restrictions on the deviations across the users choices, we establish convergence properties of these methods in almost sure sense [40, 41].

1.4 Networked aggregative Nash games

In this section we introduce an aggregative game of our interest. An aggregative game is a non-cooperative game in which each player's payoff is parametrized by its action and the aggregate of the actions taken by all players [42, 43, 44]. Such games have been shown to be closely related with subclasses of potential games [45, 46] where a potential game refers to a Nash game in which the payoff functions admit a potential function [47]. Nash-Cournot games represent one instance of such games; here firms make quantity bids that fetch a price based aggregate quantity sold, implying that the payoff of any player is a function of the aggregate quantity sold [17, 48].

We consider aggregative games wherein the agents compete over a network. The players in this game are assumed to have local interactions with each other over time, where these interactions are modelled by time-varying connectivity graphs. To this end, consider a set of N players (or agents) indexed by $1,\ldots,N$, and let $\mathcal{N}=\{1,\ldots,N\}$. The ith player is characterized by a strategy set $K_i \subseteq \mathbb{R}^n$ and a payoff function $f_i(x_i,\bar{x})$, which depends on player i decision x_i and the aggregate $\bar{x}=\sum_{i=1}^N x_i$ of all player decisions. To formalize the game, let $\bar{K}$ denote the Minkowski sum of the sets K_i:

$$\bar{K}=\sum_{i=1}^N K_i. \tag{1.6}$$

In a generic aggregative game, player i faces the following parametrized optimization problem:

$$\begin{aligned} \text{minimize} \quad & f_i(x_i,\bar{x}) \\ \text{subject to} \quad & x_i \in K_i, \end{aligned} \tag{1.7}$$

where $K_i \subseteq \mathbb{R}^n$ and $\bar{x}$ is the aggregate of the agent's decisions x_i, i.e.,

$$\bar{x} = \sum_{j=1}^{N} x_j, \qquad \bar{x} \in \bar{K}. \tag{1.8}$$

The set K_i and the function f_i are assumed to be known by agent i only.

Next, we motivate our work by providing an example of aggregative games, whose broad range emphasizes the potential scope of our work.

Example 2 (Networked Nash-Cournot game)**.** *A classical example of an aggregative game is a networked Nash-Cournot game [17, 18, 19, 20]. Suppose a set of N firms compete over $\mathcal{L}$ locations. Firm i's production and sales at location l are denoted by g_{il} and s_{il}, respectively, while its cost of production at location l is denoted by $c_{il}(g_{il})$. Consequently, goods sold by firm i at location l fetch a revenue $p_l(\bar{s}_l)s_{il}$ where $p_l(\bar{s}_l)$ denotes the sales price at location l and $\bar{s}_l = \sum_{i=1}^{N} s_{il}$ represents the aggregate sales at location l. Finally, firm i's production at location l is capacitated by cap_{il} and its optimization problem is given by the following[8]:*

$$\begin{aligned} \textit{minimize} \qquad & \sum_{l=1}^{\mathcal{L}} \left(c_{il}(g_{il}) - p_l(\bar{s}_l)s_{il}\right) \\ \textit{subject to} \qquad & \sum_{l=1}^{\mathcal{L}} g_{il} = \sum_{l=1}^{\mathcal{L}} s_{il}, \\ & g_{il}, s_{il} \geq 0, \quad g_{il} \leq \mathrm{cap}_{il}, \qquad l = 1, \ldots, \mathcal{L}. \end{aligned} \tag{1.9}$$

In effect, firm i's payoff function is parametrized by nodal aggregate sales, thus rendering an aggregative game. Note that, in this example we have two independent networks, the first being used to model the communication of the firms and the second being used to model the physical layout of the firms production unit and locations. We allow the communication network to be dynamic but the layout network is assumed to be static.

Distributed computation of equilibria in such games is complicated by two challenges. First, the connectivity graphs of the underlying network may evolve over time. Second, in many settings, agents do not have ready access to aggregate decisions, implying that agents cannot compute their exact payoffs (or their gradients). Consequently, standard gradient-based or best-response schemes cannot be directly implemented since agents do not have ready access to the aggregate. Accordingly, we propose two distributed algorithms which allow agents to build estimates of the

[8] Note that the transportation costs are assumed to be zero.

aggregate and compute an equilibrium of aggregative games. Of these, the first is a *synchronous distributed algorithm* while the second is an *asynchronous distributed (gossip-based) algorithm*.

Distributed gradient-based algorithms for computing equilibria in deterministic regimes have received significant attention recently, particularly in the context of wireless and wireline communication networks and distributed control engineering [5, 7, 32, 33, 14]. Much of this work assumes a somewhat more restrictive strong monotonicity property on the mapping corresponding to the associated variational problem. By combining a regularization technique, this requirement can be weakened [20] while extensions to stochastic regimes can also be incorporated by examining regularized counterparts of stochastic approximation [40]. However, all of these approaches are under the assumption that agents have access to the decisions of all their competitors.

The novelty of the work in this chapter is in extending the realm of distributed algorithms in [49] for computation of a Nash equilibrium point while the majority of preceding efforts have been applied towards the solution of feasibility and optimization problems. Succinctly, the main contributions of this work lie in studying a distributed synchronous and asynchronous algorithm for aggregative Nash games and proving that they produce sequences that converge almost-surely to the unique equilibrium.

Chapter 2

Distributed Algorithms For Convex Multiuser Optimization Problem

In this chapter we present distributed algorithms aimed at solving system optimization problem (1.2). These algorithms are distributed in the sense that each user executes computations only in the space of its own decision variables.

The work presented in this chapter is closely related to the distributed algorithms in [49, 50] and the more recent work on shared-constraint games [51, 6], where several classes of problems with the structures admitting decentralized computations are addressed. However, the algorithms in the aforementioned work hinge on equal stepsizes for all users and exact solutions for their success. In most networked settings, these requirements fail to hold, thus complicating the application of these schemes. Furthermore, due to the computational complexity of obtaining exact solutions for large scale problems, one is often more interested in a good approximate solution (with a provable error bound) rather than an exact solution.

Related is also the literature on centralized projection-based methods for optimization (see for example books [52, 53, 35]) and variational inequalities [54, 55, 56, 57, 35, 58]. Recently, efficient projection-based algorithms have been developed in [59, 60, 61, 62] for optimization, and in [63, 64] for variational inequalities. The algorithms therein are all well suited for distributed implementations subject to some minor restrictions such as choosing Bregman functions that are separable across users' decision variables. The aforementioned algorithms will preserve their efficiency as long as the stepsize values are the same for all users. When the users are allowed to select their stepsizes within a certain range, there may be some efficiency loss. By viewing the stepsize variations as a source of noise, the work in this chapter may be considered as an initial step into exploring the effects of "noisy" stepsizes on the performance of first-order algorithms, starting with simple first-order algorithms which are known to be stable under noisy data.

The layout of the chapter is as follows. In Section 2.1 we present the approach which paves the path to the development of algorithms presented later in the chapter. We also formally present various assumptions we make on the multiuser problem (2.1) and recap the related fixed-point problem. In Section 2.2, we propose a regularized primal-dual method to allow for more general coupling among the constraints. Our analysis is equipped with error bounds when step-sizes and regularization parameters differ across users. Dual schemes are discussed in Section 2.3, where error bounds are provided for the case when inexact primal solutions are used. The behavior of the proposed methods is examined for a multiuser traffic problem in Section 2.4. We finally provide some concluding remarks in Section 2.5.

2.1 Fixed point approach

Before proceeding on with our approach, we re-present the problem and related notation for completeness.

$$\begin{aligned} &\text{minimize} && f(x) \triangleq \sum_{i=1}^{N} f_i(x_i) + c(x) && \\ &\text{subject to} && d_j(x) \le 0 && \text{for all } j = 1,\dots,m, \\ & && x_i \in K_i && \text{for all } i = 1,\dots,N, \end{aligned} \tag{2.1}$$

where N is the number of users, $f_i(x_i)$ is user i cost function depending on a decision vector $x_i \in \mathbb{R}^{n_i}$ and $K_i \subseteq \mathbb{R}^{n_i}$ is the constraint set for user i. The function $c(x)$ is a joint cost that depends on the user decisions, i.e., $x = (x_1,\dots,x_N) \in \mathbb{R}^n$, where $n = \sum_{i=1}^{N} n_i$. The functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}$ are *convex and continuously differentiable*. Further, we assume that $d_j : \mathbb{R}^n \to \mathbb{R}$ is a *continuously differentiable convex function* for every j. Often, when convenient, we will write the inequality constraints $d_j(x) \le 0,\ j = 1,\dots,m$, compactly as $d(x) \le 0$ with $d(x) = (d_1(x),\dots,d_m(x))^T$. Similarly, we use $\nabla d(x)$ to denote the vector of gradients $\nabla d_j(x),\ j = 1,\dots,m$, i.e., $\nabla d(x) = (\nabla d_1(x),\dots,\nabla d_m(x))^T$. The user constraint sets K_i are assumed to be *nonempty, convex and closed*. We denote by f^* and K^*, respectively, the optimal value and the optimal solution set of this problem.

Our approach is based on casting the system optimization problem as a fixed point problem through the variational inequality framework. Toward this goal, letting $\lambda \in \mathbb{R}^m_+$ denote the Lagrange multipliers of problem (2.1), the Lagrangian is given as:

$$\mathcal{L}(x,\lambda) = f(x) + \lambda^T d(x), \qquad K = K_1 \times K_2 \times \cdots \times K_N.$$

Under suitable strong duality conditions, from the first-order optimality conditions and the decomposable structure of K it can be seen that $(x^*,\lambda^*) \in K \times \mathbb{R}^m_+$ is a solution to (2.1) if and only x_i^* solves the parameterized variational inequalities $\mathrm{VI}(K_i, \nabla_{x_i}\mathcal{L}(x_i; x_{-i}^*, \lambda^*)),\ i = 1,\dots,N$, and

λ^* solves $\text{VI}(\mathbb{R}_m^+, -\nabla_\lambda \mathcal{L}(x^*,\lambda))$. A vector (x^*,λ^*) solves $\text{VI}(K_i, \nabla_{x_i}\mathcal{L}(x_i;x_{-i}^*,\lambda^*))$, $i=1,\ldots,N$ and $\text{VI}(\mathbb{R}_+^m, -\nabla_\lambda \mathcal{L}(x^*,\lambda))$ if and only if each x_i^* is a zero of the parameterized natural map[1] $\mathbf{F}_{K_i}^{\text{nat}}(x_i;x_{-i}^*,\lambda^*) = 0$, for $i=1,\ldots,N$, and λ^* is a zero of the parameterized natural map $\mathbf{F}_{\mathbb{R}_+^m}^{\text{nat}}(\lambda;x^*) = 0$, i.e.,

$$\mathbf{F}_{K_i}^{\text{nat}}(x_i;x_{-i}^*,\lambda^*) \triangleq x_i - \Pi_{K_i}(x_i - \nabla_{x_i}\mathcal{L}(x_i;x_{-i}^*,\lambda^*)) = 0 \qquad \text{for } i=1,\ldots,N,$$
$$\mathbf{F}_{\mathbb{R}_m^+}^{\text{nat}}(\lambda;x^*) \triangleq \lambda - \Pi_{\mathbb{R}_m^+}(\lambda + \nabla_\lambda \mathcal{L}(x^*,\lambda)) = 0.$$

Equivalently, letting $x^* = (x_1^*,\ldots,x_N^*) \in K$, a solution to the original problem is given by a solution to the following system of nonsmooth equations:

$$\begin{aligned} x^* &= \Pi_K(x^* - \nabla_x \mathcal{L}(x^*,\lambda^*)), \\ \lambda^* &= \Pi_{\mathbb{R}_+^m}(\lambda^* + \nabla_\lambda \mathcal{L}(x^*,\lambda^*)). \end{aligned} \tag{2.2}$$

Thus, x^* solves problem (2.1) if and only if it is a solution to the system (2.2) for some $\lambda^* \geq 0$. This particular relation motivates our algorithmic development. We now discuss the conditions that we use in the subsequent development. Specifically, we assume that the Slater condition holds for problem (2.1).

Assumption 1. *(Slater Condition) There exists a Slater vector $\acute{x} \in K$ such that $d_j(\acute{x}) < 0$ for all $j=1,\ldots,m$.*

Under the Slater condition, the primal problem (2.1) and its dual have the same optimal value, and a dual optimal solution λ^* exists. When K is compact for example, the primal problem also has a solution x^*. A primal-dual optimal pair (x^*,λ^*) is also a solution to the coupled fixed-point problems in (2.2). For a more compact notation, we introduce the mapping $\Phi(x,\lambda)$ as

$$\Phi(x,\lambda) \triangleq (\nabla_x \mathcal{L}(x,\lambda), -\nabla_\lambda \mathcal{L}(x,\lambda)) = (\nabla_x \mathcal{L}(x,\lambda), -d(x)), \tag{2.3}$$

and we let $z = (x,\lambda)$. In this notation, the preceding coupled fixed-point problems are equivalent to a variational inequality requiring a vector $z^* = (x^*,\lambda^*) \in K \times \mathbb{R}_+^m$ such that

$$(z - z^*)^T \Phi(z^*) \geq 0 \quad \text{for all } z = (x,\lambda) \in K \times \mathbb{R}_+^m. \tag{2.4}$$

In the remainder of the chapter, in the product space $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N}$, we use $\|x\|$ and $x^T y$ to denote the Euclidean norm and the inner product that are induced, respectively, by the Euclidean

[1] See [35], volume 1, 1.5.8 Proposition, page 83.

norms and the inner products in the component spaces. Specifically, for $x = (x_1, \ldots, x_N)$ with $x_i \in \mathbb{R}^{n_i}$ for all i, we have

$$x^T y = \sum_{i=1}^{N} x_i^T y_i \qquad \text{and} \qquad \|x\| = \sqrt{\sum_{i=1}^{N} \|x_i\|^2}.$$

We now state our basic assumptions on the functions and the constraint sets in problem (2.1).

Assumption 2. *The set K is closed, convex, and bounded. The functions $f_i(x_i)$, $i = 1, \ldots, N$, and $c(x)$ are continuously differentiable and convex.*

Next, we define the gradient map

$$F(x) = \left(\nabla_{x_1}(f_1(x_1) + c(x))^T, \ldots, \nabla_{x_N}(f_N(x_N) + c(x))^T\right)^T,$$

for which we assume the following.

Assumption 3. *The gradient map $F(x)$ is Lipschitz continuous with a constant L over the set K, i.e.,*

$$\|F(x) - F(y)\| \leq L\|x - y\| \qquad \text{for all } x, y \in K.$$

2.2 Regularized primal-dual method

In this section, we present a distributed gradient-based method that employs a fixed regularization in the primal and dual space. We present the regularized problem and proceed to provide bounds on the error. We then claim the monotonicity and Lipschitzian properties of the regularized mapping and develop the main convergence result of this section. Notably, the theoretical convergence results prescribe a set from which users may independently select stepsizes with no impact on the overall convergence of the scheme. Finally, we further weaken the informational restrictions of the scheme by allowing users to select regularization parameters from a broadcasted range, and we extend the Lipschitzian bounds and convergence rates to this regime.

2.2.1 Regularization

For approximately solving the variational inequality (2.4), we consider its regularized counterpart obtained by regularizing the Lagrangian in both primal and dual space. In particular, for $\nu > 0$ and

$\varepsilon > 0$, we let $\mathcal{L}_{\nu,\varepsilon}$ denote the regularized Lagrangian, given by

$$\mathcal{L}_{\nu,\varepsilon}(x,\lambda) = f(x) + \frac{\nu}{2}\|x\|^2 + \lambda^T d(x) - \frac{\varepsilon}{2}\|\lambda\|^2. \tag{2.5}$$

The regularized variational inequality requires determining a vector $z^*_{\nu,\varepsilon} = (x^*_{\nu,\varepsilon}, \lambda^*_{\nu,\varepsilon}) \in K \times \mathbb{R}^m_+$ such that

$$(z - z^*_{\nu,\varepsilon})^T \Phi_{\nu,\varepsilon}(z^*_{\nu,\varepsilon}) \geq 0 \quad \text{for all } z = (x,\lambda) \in K \times \mathbb{R}^m_+, \tag{2.6}$$

where the regularized mapping $\Phi_{\nu,\varepsilon}(x,\lambda)$ is given by

$$\Phi_{\nu,\varepsilon}(x,\lambda) \triangleq (\nabla_x \mathcal{L}_{\nu,\varepsilon}(x,\lambda), -\nabla_\lambda \mathcal{L}_{\nu,\varepsilon}(x,\lambda)) = (\nabla_x \mathcal{L}(x,\lambda) + \nu x, -d(x) + \varepsilon\lambda). \tag{2.7}$$

The gradient map $\nabla_x \mathcal{L}_{\nu,\varepsilon}(x,\lambda)$ is given by

$$\nabla_x \mathcal{L}_{\nu,\varepsilon}(x,\lambda) \triangleq (\nabla_{x_1} \mathcal{L}_{\nu,\varepsilon}(x,\lambda), \ldots, \nabla_{x_N} \mathcal{L}_{\nu,\varepsilon}(x,\lambda))$$

where $\nabla_{x_i} \mathcal{L}_{\nu,\varepsilon}(x,\lambda) = \nabla_{x_i}\left(f(x) + \lambda^T d(x)\right) + \nu x_i$. It is known that, under some conditions, the unique solutions $z^*_{\nu,\varepsilon}$ of the variational inequality in (2.6) converge, as $\nu \to 0$ and $\varepsilon \to 0$, to the smallest norm solution of the original variational inequality in (2.4) (see [35], Section 12.2). We, however, want to investigate approximate solutions and estimate the errors resulting from solving a regularized problem instead of the original problem, while the regularization parameters are kept fixed at some values.

To solve the variational inequality (2.6), one option lies in considering projection schemes for monotone variational inequalities (see Chapter 12 in [35]). However, the lack of Lipschitz continuity of the mapping precludes a direct application of these schemes. In fact, the Lipschitz continuity of $\Phi_{\nu,\varepsilon}(z)$ cannot even be proved when the functions f and d_j have Lipschitz continuous gradients. In proving the Lipschitzian property, we observe that the boundedness of the multipliers cannot be assumed in general. However, the "bounding of multipliers λ" may be achieved under the Slater regularity condition. In particular, the Slater condition can be used to provide a compact convex region containing all the dual optimal solutions. Replacing $\mathbb{R}^m_+$ with such a compact convex set results in a variational inequality that is equivalent to (2.6),

Determining a compact set containing the dual optimal solutions can be accomplished by viewing the regularized Lagrangian $\mathcal{L}_{\nu,\varepsilon}$ as a result of two-step regularization: we first regularize the original primal problem (2.1), and then we regularize its Lagrangian function. Specifically, for

$\nu > 0$, the regularized problem (2.1) is given by

$$\begin{aligned} \text{minimize} \quad & f_\nu(x) \triangleq \sum_{i=1}^{N} \left(f_i(x_i) + \frac{\nu}{2} \|x_i\|^2 \right) + c(x) \\ \text{subject to} \quad & d_j(x) \le 0 \quad \text{for all } j = 1, \ldots, m, \\ & x_i \in K_i \quad \text{for all } i = 1, \ldots, N. \end{aligned} \tag{2.8}$$

Its Lagrangian function is

$$\mathcal{L}_\nu(x, \lambda) = f(x) + \frac{\nu}{2} \|x\|^2 + \lambda^T d(x) \qquad \text{for all } x \in K,\ \lambda \ge 0, \tag{2.9}$$

and its corresponding dual problem is

$$\begin{aligned} \text{maximize} \quad & v_\nu(\lambda) \triangleq \min_{x \in K} \mathcal{L}_\nu(x, \lambda) \\ \text{subject to} \quad & \lambda \ge 0. \end{aligned}$$

We use v_ν^* to denote the optimal value of the dual problem, i.e., $v_\nu^* = \max_{\lambda \ge 0} v_\nu(\lambda)$, and we use Λ_ν^* to denote the set of optimal dual solutions. For $\nu = 0$, the value v_0^* is the optimal dual value of the original problem (2.1) and Λ_0^* is the set of the optimal dual solutions of its dual problem.

Under the Slater condition, for every $\nu > 0$, the optimal dual multipliers λ^* exist and if we assume that solution x_ν^* to problem (2.8) exists then strong duality holds [65]. In particular, the optimal values of problem (2.8) and its dual are equal, i.e., $f(x_\nu^*) = v_\nu^*$, and the dual optimal set Λ_ν^* is nonempty and bounded [66]. Specifically, we have

$$\Lambda_\nu^* \subseteq \left\{ \lambda \in \mathbb{R}^m \,\middle|\, \sum_{j=1}^{m} \lambda_j \le \frac{f(\acute{x}) + \frac{\nu}{2} \|\acute{x}\|^2 - v_\nu^*}{\min_{1 \le j \le m} \{-d_j(\acute{x})\}},\ \lambda \ge 0 \right\} \qquad \text{for all } \nu > 0.$$

When the Slater condition holds and the optimal value f^* of the original problem (2.1) is finite, the strong duality holds for that problem as well, and therefore, the preceding relation also holds for $\nu = 0$, with v_0^* being the optimal value of the dual problem for (2.1). In this case, we have $f^* = v_0^*$, while for any $\nu > 0$, we have $v_\nu^* = f(x_\nu^*)$ for a solution x_ν^* of the regularized problem (2.8). Since $f(x_\nu^*) \ge f^*$, it follows that $v_\nu^* \ge v_0^*$ for all $\nu \ge 0$, and therefore,

$$\Lambda_\nu^* \subseteq \left\{ \lambda \in \mathbb{R}^m \,\middle|\, \sum_{j=1}^{m} \lambda_j \le \frac{f(\acute{x}) + \frac{\nu}{2} \|\acute{x}\|^2 - v_0^*}{\min_{1 \le j \le m} \{-d_j(\acute{x})\}},\ \lambda \ge 0 \right\} \qquad \text{for all } \nu \ge 0,$$

where the set Λ_0^* is the set of dual optimal solutions for the original problem (2.1). Noting that a larger set on the right hand side can be obtained by replacing v_0^* with any lower-bound estimate of v_0^* [i.e., $v(\hat{\lambda})$ for some $\hat{\lambda} \geq 0$], we can define a compact convex set $\mathcal{D}_\nu$ for every $\nu \geq 0$, as follows:

$$\mathcal{D}_\nu = \left\{ \lambda \in \mathbb{R}^m \,\Big|\, \sum_{j=1}^m \lambda_j \leq \frac{f(\acute{x}) + \frac{\nu}{2}\|\acute{x}\|^2 - v(\hat{\lambda})}{\min_{1\leq j\leq m}\{-d_j(\acute{x})\}},\ \lambda \geq 0 \right\} \qquad \text{for every } \nu \geq 0, \tag{2.10}$$

which satisfies

$$\Lambda_\nu^* \subset \mathcal{D}_\nu, \qquad \text{for every } \nu \geq 0. \tag{2.11}$$

Observe that $v_0^* \leq v_\nu^* \leq v_{\nu'}$ for $0 \leq \nu \leq \nu'$, implying that $\mathcal{D}_0 \subseteq \mathcal{D}_\nu \subseteq \mathcal{D}_{\nu'}$. Therefore, the compact sets $\mathcal{D}_\nu$ are nested, and their intersection is a nonempty compact set $\mathcal{D}$ which contains the optimal dual solutions Λ_0^* of the original problem.

For the remainder of this chapter, we will assume that the Slater condition holds and the set K is compact (Assumption 2), so that the construction of such nested compact sets is possible. Specifically, we will *assume that a family of nested compact convex sets* $\mathcal{D}_\nu \subset \mathbb{R}_+^m$, $\nu \geq 0$, *satisfying relation (2.11) has already been determined*. In this case, the variational inequality of determining $z_{\nu,\varepsilon} = (x_{\nu,\varepsilon}, \lambda_{\nu,\varepsilon}) \in K \times \mathcal{D}_\nu$ such that

$$(z - z_{\nu,\varepsilon})^T \Phi_{\nu,\varepsilon}(z_{\nu,\varepsilon}) \geq 0 \qquad \text{for all } z = (x, \lambda) \in K \times \mathcal{D}_\nu, \tag{2.12}$$

has the same solution set as the variational inequality in (2.6), where λ is constrained to lie in the nonnegative orthant.

2.2.2 Regularization error

We now provide an upper bound on the distances between $x_{\nu,\varepsilon}$ and x_ν^*. Here, $x_{\nu,\varepsilon}$ is the primal component of $z_{\nu,\varepsilon}$, the solution of the variational inequality in (2.12) and x_ν^* is the solution of the regularized problem in (2.8) for given positive parameters ν and ε.

Proposition 1. *Let Assumption 2 hold except for the boundedness of K. Also, let Assumption 1 hold. Then, for any $\nu > 0$ and $\varepsilon > 0$, for the solution $z_{\nu,\varepsilon} = (x_{\nu,\varepsilon}, \lambda_{\nu,\varepsilon})$ of variational inequality (2.12), we have*

$$\nu\|x_\nu^* - x_{\nu,\varepsilon}\|^2 + \frac{\varepsilon}{2}\|\lambda_{\nu,\varepsilon}\|^2 \leq \frac{\varepsilon}{2}\|\lambda_\nu^*\|^2 \qquad \textit{for all } \lambda_\nu^* \in \Lambda_\nu^*,$$

where x_ν^* is the optimal solution of the regularized problem (2.8) and Λ_ν^* is the set of optimal solutions of its corresponding dual problem.

Proof. The existence of a unique solution $x_\nu^* \in K$ of problem (2.8) follows from the continuity and strong convexity of f_ν. Also, by the Slater condition, the dual optimal set Λ_ν^* is nonempty. In what follows, let $\lambda_\nu^* \in \Lambda_\nu^*$ be an arbitrary but fixed dual optimal solution for problem (2.8). To make the notation simpler, we use ξ to denote the pair of regularization parameters (ν, ε), i.e., $\xi = (\nu, \varepsilon)$. When the interplay between the parameters is relevant, we will write them explicitly.

From the definition of the mapping Φ_ξ it follows that the solution $z_\xi = (x_\xi, \lambda_\xi) \in K \times \mathcal{D}_\nu$ is a saddle-point for the regularized Lagrangian function $\mathcal{L}_\xi(x,\lambda) = \mathcal{L}(x,\lambda) + \frac{\nu}{2}\|x\|^2 - \frac{\varepsilon}{2}\|\lambda\|^2$, i.e.,

$$\mathcal{L}_\xi(x_\xi, \lambda) \le \mathcal{L}_\xi(x_\xi, \lambda_\xi) \le \mathcal{L}_\xi(x, \lambda_\xi) \qquad \text{for all } x \in K \text{ and } \lambda \in \mathcal{D}_\nu. \tag{2.13}$$

Recalling that $\Lambda_\nu^* \subseteq \mathcal{D}_\nu$, and by letting $\lambda = \lambda_\nu^*$ in the first inequality of the preceding relation, we obtain

$$0 \le \mathcal{L}_\xi(x_\xi, \lambda_\xi) - \mathcal{L}_\xi(x_\xi, \lambda_\nu^*) = (\lambda_\xi - \lambda_\nu^*)^T d(x_\xi) - \frac{\varepsilon}{2}\|\lambda_\xi\|^2 + \frac{\varepsilon}{2}\|\lambda_\nu^*\|^2. \tag{2.14}$$

We now estimate the term $(\lambda_\xi - \lambda_\nu^*)^T d(x_\xi) = \sum_{j=1}^m (\lambda_{\xi,j} - \lambda_{\nu,j}^*) d_j(x_\xi)$ by considering the individual terms, where $\lambda_{\nu,j}^*$ is the j-th component of λ_ν^*. By convexity of each d_j, we have

$$d_j(x_\xi) \le d_j(x_\nu^*) + \nabla d_j(x_\xi)^T (x_\xi - x_\nu^*) \le \nabla d_j(x_\xi)^T (x_\xi - x_\nu^*),$$

where the last inequality follows from x_ν^* being a solution to the primal regularized problem (hence, $d_j(x_\nu^*) \le 0$ for all j). By multiplying the preceding inequality with $\lambda_{\xi,j}$ (which is nonnegative) and by adding over all j, we obtain

$$\sum_{j=1}^m \lambda_{\xi,j} d_j(x_\xi) \le \sum_{j=1}^m \lambda_{\xi,j} \nabla d_j(x_\xi)^T (x_\xi - x_\nu^*).$$

By the definition of the regularized Lagrangian $\mathcal{L}_\xi(x,\lambda)$, we have

$$\begin{aligned}\sum_{j=1}^m \lambda_{\xi,j} \nabla d_j(x_\xi)^T (x_\xi - x_\nu^*) &= \nabla_x \mathcal{L}_\xi(x_\xi, \lambda_\xi)^T (x_\xi - x_\nu^*) - \left(\nabla f(x_\xi) + \nu x_\xi\right)^T (x_\xi - x_\nu^*) \\ &\le -\left(\nabla f(x_\xi) + \nu x_\xi\right)^T (x_\xi - x_\nu^*),\end{aligned}$$

where the inequality follows from $\nabla_x \mathcal{L}_\xi(x_\xi, \lambda_\xi)^T (x_\xi - x_\nu^*) \le 0$, which holds in view of the second inequality in saddle-point relation (2.13) with $x = x_\nu^* \in K$. Therefore, by combining the preceding two relations, we obtain

$$\sum_{j=1}^{m} \lambda_{\xi,j} d_j(x_\xi) \leq -\left(\nabla f(x_\xi) + \nu x_\xi\right)^T (x_\xi - x_\nu^*). \tag{2.15}$$

By convexity of each d_j, we have $d_j(x_\xi) \geq d_j(x_\nu^*) + \nabla d_j(x_\nu^*)^T (x_\xi - x_\nu^*)$ By multiplying the preceding inequality with $-\lambda_{\nu,j}^*$ (which is non-positive) and by adding over all j, we obtain

$$\begin{aligned} -\sum_{j=1}^{m} \lambda_{\nu,j}^* d_j(x_\xi) &\leq -\sum_{j=1}^{m} \lambda_{\nu,j}^* d_j(x_\nu^*) - \sum_{j=1}^{m} \lambda_{\nu,j}^* \nabla d_j(x_\nu^*)^T (x_\xi - x_\nu^*) \\ &= \sum_{j=1}^{m} \lambda_{\nu,j}^* \nabla d_j(x_\nu^*)^T (x_\nu^* - x_\xi), \end{aligned}$$

where the equality follows from $(\lambda_\nu^*)^T d(x_\nu^*) = 0$, which holds by the complementarity slackness of the primal-dual pair (x_ν^*, λ_ν^*) of the regularized problem (2.8). Using the definition of the Lagrangian function $\mathcal{L}_\nu$ in (2.9) for the problem (2.8), we have

$$\begin{aligned} \sum_{j=1}^{m} \lambda_{\nu,j}^* \nabla d_j(x_\nu^*)^T (x_\nu^* - x_\xi) &= \nabla_x \mathcal{L}_\nu(x_\nu^*, \lambda_\nu^*)^T (x_\nu^* - x_\xi) - \left(\nabla f(x_\nu^*) + \nu x_\nu^*\right)^T (x_\nu^* - x_\xi) \\ &\leq -\left(\nabla f(x_\nu^*) + \nu x_\nu^*\right)^T (x_\nu^* - x_\xi), \end{aligned}$$

where the inequality follows from relation $\nabla_x \mathcal{L}(x_\nu^*, \lambda_\nu^*)^T (x_\nu^* - x_\xi) \leq 0$, which in turn holds since (x_ν^*, λ_ν^*) is a saddle-point of the Lagrangian function $\mathcal{L}_\nu(x, \lambda)$ over $K \times \mathcal{D}_\nu$ and $x_\xi \in K$. Combining the preceding two relations, we obtain

$$-\sum_{j=1}^{m} \lambda_{\nu,j}^* d_j(x_\xi) \leq -\left(\nabla f(x_\nu^*) + \nu x_\nu^*\right)^T (x_\nu^* - x_\xi) = \left(\nabla f(x_\nu^*) + \nu x_\nu^*\right)^T (x_\xi - x_\nu^*).$$

The preceding relation and inequality (2.15), yield

$$(\lambda_\xi - \lambda_\nu^*)^T d(x_\xi) = \sum_{j=1}^{m} (\lambda_{\xi,j} - \lambda_{\nu,j}^*) d_j(x_\xi) \leq (\nabla f(x_\nu^*) - \nabla f(x_\xi))^T (x_\xi - x_\nu^*) - \nu \|x_\xi - x_\nu^*\|^2.$$

From the monotonicity of ∇f, we have $(\nabla f(x_\nu^*) - \nabla f(x_\xi))^T (x_\xi - x_\nu^*) \leq 0$, thus implying $(\lambda_\xi - \lambda_\nu^*)^T d(x_\xi) \leq -\nu \|x_\xi - x_\nu^*\|^2$. Finally, by combining the preceding relation with (2.14), and recalling notation $\xi = (\nu, \varepsilon)$, we obtain for any solution x_ν^*,

$$\nu \|x_{\nu,\varepsilon} - x_\nu^*\|^2 + \frac{\varepsilon}{2} \|\lambda_{\nu,\varepsilon}\|^2 \leq \frac{\varepsilon}{2} \|\lambda_\nu^*\|^2 \qquad \text{for all } \lambda^* \in \Lambda_\nu^*, \tag{2.16}$$

thus showing the desired relation. ■

As an immediate consequence of Proposition 1, in view of $\Lambda_\nu^* \subset \mathcal{D}_\nu$, we have

$$\|x_{\nu,\varepsilon} - x_\nu^*\| \leq \sqrt{\frac{\varepsilon}{2\nu}} \max_{\lambda^* \in \mathcal{D}_\nu} \|\lambda^*\| \qquad \text{for all } \nu > 0 \text{ and } \varepsilon > 0. \tag{2.17}$$

This relation provides a bound on the distances of the solutions x_ν^* of problem (2.8) and the component $x_{\nu,\varepsilon}$ of the solution $z_{\nu,\varepsilon}$ of the regularized variational inequality in (2.12). The relation suggests that a ν larger than ε would yield a better error bound. Note, however, that increasing ν would correspond to the enlargement of the set $\mathcal{D}_\nu$, and therefore, increasing value for $\max_{\lambda^* \in \mathcal{D}_\nu} \|\lambda^*\|$. When the specific structure of the sets $\mathcal{D}_\nu$ is available, one may try to optimize the term $\sqrt{\frac{\varepsilon}{2\nu}} \max_{\lambda^* \in \mathcal{D}_\nu} \|\lambda^*\|$ with respect to ν, while ε is kept fixed. In fact, the following result provides a simple result when $\mathcal{D}_\nu$ is specified using the Slater point $\acute{x}$.

Lemma 1. *Under the assumptions of Proposition 1, for a fixed $\varepsilon > 0$, the tightest bound for $\|x_{\nu,\varepsilon} - x_\nu^*\|$ is given by*

$$\|x_{\nu,\varepsilon} - x_\nu^*\| \leq \left(\frac{\sqrt{\varepsilon \left(f(\acute{x}) - v(\acute{\lambda}) \right)}}{\min_{1 \leq j \leq m}\{-d_j(\acute{x})\}} \|\acute{x}\| \right).$$

Proof. Using $\|x\|_2 \leq \|x\|_1$, from relation (2.17) we have

$$\|x_{\nu,\varepsilon} - x_\nu^*\| \leq \sqrt{\frac{\varepsilon}{2\nu}} \left(\max_{\lambda \in \mathcal{D}_\nu} \|\lambda\|_2 \right) \leq \sqrt{\frac{\varepsilon}{2\nu}} \left(\max_{\lambda \in \mathcal{D}_\nu} \|\lambda\|_1 \right).$$

But by the structure of the set D_ν, we have that

$$\begin{aligned}
\sqrt{\frac{\varepsilon}{2\nu}} \left(\max_{\lambda \in \mathcal{D}_\nu} \|\lambda\|_1 \right) &= \sqrt{\frac{\varepsilon}{2\nu}} \left(\frac{f(\acute{x}) + \frac{\nu}{2}\|\acute{x}\|^2 - v(\acute{\lambda})}{\min_{1 \leq j \leq m}\{-d_j(\acute{x})\}} \right) \\
&= \sqrt{\frac{\varepsilon}{2\nu}} \left(\frac{f(\acute{x}) - v(\acute{\lambda})}{\min_{1 \leq j \leq m}\{-d_j(\acute{x})\}} \right) + \sqrt{\frac{\nu\varepsilon}{2}} \left(\frac{\frac{1}{2}\|\acute{x}\|^2}{\min_{1 \leq j \leq m}\{-d_j(\acute{x})\}} \right) \\
&= \frac{\sqrt{\varepsilon}}{\sqrt{2}\min_{1 \leq j \leq m}\{-d_j(\acute{x})\}} \left(\frac{a}{\sqrt{\nu}} + b\sqrt{\nu} \right),
\end{aligned}$$

where $a = f(\acute{x}) - v(\acute{\lambda})$ and $b = \frac{1}{2}\|\acute{x}\|^2$. It can be seen that the function $h(\nu) = a/\sqrt{\nu} + b\sqrt{\nu}$ has a unique minimum at $\nu^* = \frac{a}{b}$, with the minimum value $h(\nu^*) = 2\sqrt{ab}$. Thus, we have

$$\sqrt{\frac{\varepsilon}{2\nu}}\left(\max_{\lambda\in\mathcal{D}_\nu}\|\lambda\|_1\right) \leq \left(\frac{\sqrt{\varepsilon\left(f(\acute{x}) - v(\acute{\lambda})\right)}}{\min_{1\leq j\leq m}\{-d_j(\acute{x})\}}\right)\|\acute{x}\|,$$

implying the desired estimate. ∎

When the set K is bounded, as another consequence of Proposition 1, we may obtain the error bounds on the sub-optimality of the vector $x_{\nu,\varepsilon}$ by using the preceding error bound. Specifically, we can provide bounds on the violation of the primal inequality constraints $d_j(x) \leq 0$ at $x = x_{\nu,\varepsilon}$. Also, we can estimate the difference in the values $f(x_{\nu,\varepsilon})$ and the primal optimal value f^* of the original problem (2.1). This is done in the following lemma.

Lemma 2. *Let Assumptions 2 and 1 hold. For any $\nu, \varepsilon > 0$, we have*

$$\max\left\{0, d_j(x_{\nu,\varepsilon})\right\} \leq M_{d_j} M_\nu \sqrt{\frac{\varepsilon}{2\nu}} \qquad \text{for all } j = 1,\ldots,m,$$
$$|f(x_{\nu,\varepsilon}) - f(x^*)| \leq M_f M_\nu \sqrt{\frac{\varepsilon}{2\nu}} + \frac{\nu}{2}D^2,$$

with $M_{d_j} = \max_{x\in K}\|\nabla d_j(x)\|$ for each j, $M_f = \max_{x\in K}\|\nabla f(x)\|$, $M_\nu = \max_{\lambda\in\mathcal{D}_\nu}\|\lambda\|$ and $D = \max_{x\in K}\|x\|$.

Proof. Let $\nu > 0$ and $\varepsilon > 0$ be given, and let $j \in \{1,\ldots,m\}$ be arbitrary. Since d_j is convex, we have

$$d_j(x_{\nu,\varepsilon}) \leq d_j(x^*_\nu) + \nabla d_j(x^*_\nu)^T(x_{\nu,\varepsilon} - x^*_\nu) \leq \|\nabla d_j(x^*_\nu)\|\,\|x_{\nu,\varepsilon} - x^*_\nu\|,$$

where in the last inequality we use $d_j(x^*_\nu) \leq 0$, which holds since x^*_ν is the solution to the regularized primal problem (2.8). Since K is compact, the gradient norm $\|\nabla d_j(x)\|$ is bounded by some constant, say M_{d_j}. From this and the estimate

$$\|x_{\nu,\varepsilon} - x^*_\nu\| \leq \sqrt{\frac{\varepsilon}{2\nu}}\|\lambda^*_\nu\|, \tag{2.18}$$

which follows by Proposition 1, we obtain

$$d_j(x_{\nu,\varepsilon}) \leq M_{d_j}\sqrt{\frac{\varepsilon}{2\nu}}\,\|\lambda^*_\nu\|,$$

where λ_ν^* is a dual optimal solution of the regularized problem. Since the set of dual optimal solutions is contained in the compact set $\mathcal{D}_\nu$, the dual solutions are bounded. Thus, for the violation of the constraint $d_j(x) \leq 0$, we have

$$\max\{0, d_j(x_{\nu,\varepsilon})\} \leq M_{d_j} M_\nu \sqrt{\frac{\varepsilon}{2\nu}},$$

where $M_\nu = \max_{\lambda \in \mathcal{D}_\nu} \|\lambda\|$. Next, we estimate the difference $|f(x_{\nu,\varepsilon}) - f(x^*)|$. We can write

$$|f(x_{\nu,\varepsilon}) - f(x^*)| \leq |f(x_{\nu,\varepsilon}) - f(x_\nu^*)| + f(x_\nu) - f^*, \tag{2.19}$$

where we use $0 \leq f(x_\nu^*) - f^*$. By convexity of f, we have

$$\nabla f(x_\nu^*)^T (x_{\nu,\varepsilon} - x_\nu^*) \leq f(x_{\nu,\varepsilon}) - f(x_\nu^*) \leq \nabla f(x_{\nu,\varepsilon})^T (x_{\nu,\varepsilon} - x_\nu^*).$$

Since $x_{\nu,\varepsilon}, x^* \in K$ and K is compact, by the continuity of the gradient $\|\nabla f(x)\|$, the gradient norm is bounded over the set K, say by a scalar M_f, so that

$$|f(x_{\nu,\varepsilon}) - f(x_\nu^*)| \leq M_f \|x_{\nu,\varepsilon} - x_\nu^*\|.$$

Using the estimate (2.18) and the boundedness of the dual optimal multipliers, similar to the preceding analysis, we obtain the following bound

$$|f(x_{\nu,\varepsilon}) - f(x_\nu^*)| \leq M_f M_\nu \sqrt{\frac{\varepsilon}{2\nu}}.$$

By substituting the preceding relation in inequality (2.19), we obtain

$$|f(x_{\nu,\varepsilon}) - f(x^*)| \leq M_f M_\nu \sqrt{\frac{\varepsilon}{2\nu}} + f(x_\nu) - f^*.$$

Further, by using the estimate $f(x_\nu^*) - f^* \leq \frac{\nu}{2} \max_{x \in K} \|x\|^2 = \frac{\nu}{2} D^2$ of Lemma 23 (see appendix), we obtain the desired relation. ■

Next, we discuss how one may specify ν and ε. Given a threshold error δ on the deviation of the obtained function value from its optimal counterpart, we have that $|f(x_{\nu,\varepsilon}) - f(x^*)| < \delta$, if the following holds

$$M_f M_\nu \sqrt{\frac{\varepsilon}{2\nu}} + \frac{\nu}{2} D^2 < \delta.$$

But by the structure of the set D_ν, we have that

$$
\begin{aligned}
M_\nu &= \sqrt{\frac{\varepsilon}{2\nu}}\left(\max_{\lambda\in\mathcal{D}_\nu}\|\lambda\|_1\right)\\
&= \sqrt{\frac{\varepsilon}{2\nu}}\left(\frac{f(\acute{x})+\frac{\nu}{2}\|\acute{x}\|^2-v(\hat{\lambda})}{\min_{1\le j\le m}\{-d_j(\acute{x})\}}\right)\\
&= \sqrt{\frac{\varepsilon}{2\nu}}\left(\frac{f(\acute{x})-v(\hat{\lambda})}{\min_{1\le j\le m}\{-d_j(\acute{x})\}}\right)+\sqrt{\frac{\nu\varepsilon}{2}}\left(\frac{\frac{\|\acute{x}\|^2}{2}}{\min_{1\le j\le m}\{-d_j(\acute{x})\}}\right)\\
&= \frac{1}{\sqrt{2}\min_{1\le j\le m}\{-d_j(\acute{x})\}}\left(\frac{a\sqrt{\varepsilon}}{\sqrt{\nu}}+b\sqrt{\varepsilon\nu}\right),
\end{aligned}
$$

where $a=f(\acute{x})-v(\hat{\lambda})$ and $b=\frac{\|\acute{x}\|^2}{2}$. Thus, we have

$$
M_f M_\nu\sqrt{\frac{\varepsilon}{2\nu}}+\frac{\nu}{2}D^2\le\frac{M_f}{\sqrt{2}\min_{1\le j\le m}\{-d_j(\acute{x})\}}\left(\frac{a\sqrt{\varepsilon}}{\sqrt{\nu}}+b\sqrt{\varepsilon\nu}\right)+\frac{\nu}{2}D^2<\delta.
$$

Next, we may choose parameters ν and ε so that the above inequality is satisfied. The expression suggests that one must choose $\varepsilon<\nu$ (as M_f could be large). Thus setting $\varepsilon=\nu^3$, we will obtain a quadratic inequality in parameter ν which can subsequently allow for selecting ν and therefore ε.

Unfortunately, the preceding results do not provide a bound on $\|x_{\nu,\varepsilon}-x^*\|$ and indeed for the optimal ν^* minimizing $\|x_{\nu,\varepsilon}-x^*\|$, the error in $\|x_{\nu,\varepsilon}-x^*_\nu\|$ can be large (due to error in $\|x^*_\nu-x^*\|$). The challenge in obtaining a bound on $\|x^*_{\nu,\varepsilon}-x^*\|$ implicitly requires a bound on $\|x^*_\nu-x^*\|$ which we currently do not have access to. Note that by introducing a suitable growth property on the function, one may obtain a handle on $\|x^*_\nu-x^*\|$.

2.2.3 Properties of $\Phi_{\nu,\varepsilon}$

We now focus on characterizing the mapping $\Phi_{\nu,\varepsilon}$ under the following assumption on the constraint functions d_j for $j=1,\ldots,m$.

Assumption 4. *For each j, the gradient $\nabla d_j(x)$ is Lipschitz continuous over K with a constant $L_j>0$, i.e.,*

$$
\|\nabla d_j(x)-\nabla d_j(y)\|\le L_j\|x-y\|\qquad \text{for all } x,y\in K.
$$

Under this and the Slater assumption, we prove and the strong monotonicity and the Lipschitzian nature of $\Phi_{\nu,\varepsilon}(x,\lambda)$.

Lemma 3. *Let Assumptions 2–4 hold and let* $\nu, \varepsilon \geq 0$. *Then, the regularized mapping* $\Phi_{\nu,\varepsilon}$ *is strongly monotone over* $K \times \mathbb{R}^m_+$ *with constant* $\mu = \min\{\nu, \varepsilon\}$ *and Lipschitz over* $K \times \mathcal{D}_\nu$ *with constant* $L_\Phi(\nu, \varepsilon)$ *given by*

$$L_\Phi(\nu, \varepsilon) = \sqrt{(L + \nu + M_d + M_\nu L_d)^2 + (M_d + \varepsilon)^2}, \quad L_d = \sqrt{\sum_{j=1}^m L_j^2},$$

where L *is a Lipschitz constant for* $\nabla f(x)$ *over* K, L_j *is a Lipschitz constant for* $\nabla d_j(x)$ *over* K, $M_d = \max_{x \in K} \|\nabla d(x)\|$, *and* $M_\nu = \max_{\lambda \in \mathcal{D}_\nu} \|\lambda\|$.

Proof. We use $\lambda_{1,j}$ and $\lambda_{2,j}$ to denote the jth component of vectors λ_1 and λ_2. For any two vectors $z_1 = (x_1, \lambda_1),\ z_2 = (x_2, \lambda_2) \in K \times \mathbb{R}^m_+$, we have

$$\begin{aligned}
& (\Phi_{\nu,\varepsilon}(z_1) - \Phi_{\nu,\varepsilon}(z_2))^T (z_1 - z_2) \\
= & \begin{pmatrix} \nabla_x \mathcal{L}(x_1, \lambda_1) - \nabla_x \mathcal{L}(x_2, \lambda_2) + \nu(x_1 - x_2) \\ -d(x_1) + \varepsilon\lambda_1 + d(x_2) - \varepsilon\lambda_2 \end{pmatrix}^T \begin{pmatrix} x_1 - x_2 \\ \lambda_1 - \lambda_2 \end{pmatrix} \\
= & (\nabla f(x_1) - \nabla f(x_2))^T (x_1 - x_2) + \nu \|x_1 - x_2\|^2 \\
& + \sum_{j=1}^m (\lambda_{1,j} \nabla d_j(x_1) - \lambda_{2,j} \nabla d_j(x_2))^T (x_1 - x_2) \\
& - \sum_{j=1}^m (d_j(x_1) - d_j(x_2))(\lambda_{1,j} - \lambda_{2,j}) + \varepsilon \|\lambda_1 - \lambda_2\|^2.
\end{aligned}$$

By using the monotonicity of $\nabla f(x)$, and by grouping the terms with $\lambda_{1,j}$ and $\lambda_{2,j}$, separately, we obtain

$$\begin{aligned}
& (\Phi_{\nu,\varepsilon}(z_1) - \Phi_{\nu,\varepsilon}(z_2))^T (z_1 - z_2) \geq \nu \|x_1 - x_2\|^2 \\
& + \sum_{j=1}^m \lambda_{1,j} \left(d_j(x_2) - d_j(x_1) + \nabla d_j(x_1)^T (x_1 - x_2) \right) \\
& + \sum_{j=1}^m \lambda_{2,j} \left(d_j(x_1) - d_j(x_2) - \nabla d_j(x_2)^T (x_1 - x_2) \right) + \varepsilon \|\lambda_1 - \lambda_2\|^2.
\end{aligned}$$

Now, by non-negativity of $\lambda_{1,j}, \lambda_{2,j}$ and convexity of $d_j(x)$ for each j, we have

$$\begin{aligned}
\lambda_{1,j} \left(d_j(x_2) - d_j(x_1) + \nabla d_j(x_1)^T (x_1 - x_2) \right) \geq 0, \\
\lambda_{2,j} \left(d_j(x_1) - d(x_2) - \nabla d_j(x_2)^T (x_1 - x_2) \right) \geq 0.
\end{aligned}$$

Using the preceding relations, we get

$$(\Phi_{\nu,\varepsilon}(z_1)-\Phi_{\nu,\varepsilon}(z_2))^T(z_1-z_2)\geq \nu\|x_1-x_2\|^2+\varepsilon\|\lambda_1-\lambda_2\|^2\geq \min\{\nu,\varepsilon\}\,\|z_1-z_2\|^2,$$

showing that $\Phi_{\nu,\varepsilon}$ is strongly monotone with constant $\mu=\min\{\nu,\varepsilon\}$.

Next, we show that $\Phi_{\nu,\varepsilon}$ is Lipschitz over $K\times\mathcal{D}_\nu$. Thus, given $\nu,\varepsilon\geq 0$, and any two vectors $z_1=(x_1,\lambda_1),\ z_2=(x_2,\lambda_2)\in K\times\mathcal{D}_\nu$, we have

$$\begin{aligned}
&\|\Phi_{\nu,\varepsilon}(z_1)-\Phi_{\nu,\varepsilon}(z_2)\|\\
=\ &\left\|\begin{pmatrix}\nabla f(x_1)-\nabla f(x_2)+\nu(x_1-x_2)+\sum_{j=1}^m\left(\lambda_{1,j}\nabla d_j(x_1)-\lambda_{2,j}\nabla d_j(x_2)\right)\\ -d(x_1)+d(x_2)+\varepsilon(\lambda_1-\lambda_2)\end{pmatrix}\right\|\\
\leq\ &\|\nabla f(x_1)-\nabla f(x_2)\|+\nu\|x_1-x_2\|+\left\|\sum_{j=1}^m(\lambda_{1,j}\nabla d_j(x_1)-\lambda_{2,j}\nabla d_j(x_2))\right\|\\
&+\|d(x_1)-d(x_2)\|+\varepsilon\|\lambda_1-\lambda_2\|.
\end{aligned}\tag{2.20}$$

By the compactness of K (Assumption 2) and the continuity of $\nabla d_j(x)$ for each j, the boundedness of $\nabla d(x)=(\nabla d_1(x),\ldots,\nabla d_m(x))^T$ follows, i.e.,

$$\|\nabla d(x)\|\leq M_d\qquad \text{for all } x\in K \text{ and some } M_d>0.\tag{2.21}$$

Furthermore, by using the mean value theorem (see for example [52], page 682, Prop. A.22), we can see that $d(x)$ is Lipschitz continuous over the set K with the same constant M_d. Specifically, for all $x,y\in K$, there exists a $\theta\in[0,1]$ such that

$$\|d(x)-d(y)\|=\|\nabla d(x+\theta(y-x))(x-y)\|\leq M_d\|x-y\|.$$

By using the Lipschitz property of $\nabla f(x)$ and $d(x)$, and by adding and subtracting the term $\sum_{j=1}^m\lambda_{1,j}\nabla d_j(x_2)$, from relation (2.20) we have

$$\begin{aligned}
\|\Phi_{\nu,\varepsilon}(z_1)-\Phi_{\nu,\varepsilon}(z_2)\|\ \leq\ &L\|x_1-x_2\|+\nu\|x_1-x_2\|+\sum_{j=1}^m\lambda_{1,j}\left\|\nabla d_j(x_1)-\nabla d_j(x_2)\right\|\\
&+\sum_{j=1}^m|\lambda_{1,j}-\lambda_{2,j}|\left\|\nabla d_j(x_2)\right\|+M_d\|x_1-x_2\|+\varepsilon\|\lambda_1-\lambda_2\|,
\end{aligned}$$

where we also use $\lambda_{1,j}\geq 0$ for all j. By using Hölder's inequality and the boundedness of the dual variables $\lambda_1, \lambda_2 \in \mathcal{D}_\nu$, we get

$$\sum_{j=1}^{m} \lambda_{1,j} \left\|\nabla d_j(x_1) - \nabla d_j(x_2)\right\| \leq \|\lambda_1\| \sqrt{\sum_{j=1}^{m} \|\nabla d_j(x_1) - \nabla d_j(x_2)\|^2}$$
$$\leq M_\nu \sqrt{\sum_{j=1}^{m} L_j^2}\, \|x_1 - x_2\|,$$

where in the last inequality we also use the Lipschitz property of $\nabla d_j(x)$ for each j. Similarly, by Hölder's inequality and the boundedness of $\nabla d(x)$ [see (2.21)], we have

$$\sum_{j=1}^{m} |\lambda_{1,j} - \lambda_{2,j}| \left\|\nabla d_j(x_2)\right\| \leq M_d \|\lambda_1 - \lambda_2\|.$$

By combining the preceding three relations and letting $L_d = \sqrt{\sum_{j=1}^{m} L_j^2}$, we obtain

$$\|\Phi_{\nu,\varepsilon}(z_1) - \Phi_{\nu,\varepsilon}(z_2)\| \leq (L + \nu + M_d + M_\nu L_d)\|x_1 - x_2\| + (M_d + \varepsilon)\|\lambda_1 - \lambda_2\|.$$

Further, by Hölder's inequality, we have

$$\|\Phi_{\nu,\varepsilon}(z_1) - \Phi_{\nu,\varepsilon}(z_2)\| \leq \sqrt{(L + \nu + M_d + M_\nu L_d)^2 + (M_d + \varepsilon)^2} \sqrt{\|x_1 - x_2\|^2 + \|\lambda_1 - \lambda_2\|^2}$$
$$= L_\Phi(\nu, \varepsilon)\|z_1 - z_2\|,$$

thus showing the Lipschitz property of $\Phi_{\nu,\varepsilon}$. ■

2.2.4 Primal-dual method

The strong monotonicity and Lipschitzian nature of the regularized mapping $\Phi_{\nu,\varepsilon}$ for given $\nu > 0$ and $\varepsilon > 0$, imply that standard projection algorithms can be effectively applied. Our goal is to generalize these schemes to accommodate the requirements of *limited* coordination. While in theory, convergence of projection schemes relies on consistency of primal and dual step-lengths, in practice, this requirement is difficult to enforce. In this section, we allow for different step-lengths and show that such a scheme does indeed result in a contraction.

Now, we consider solving the variational inequality in (2.12) by using a primal-dual method in which the users can choose their primal stepsizes independently with possibly differing dual stepsizes. In particular, we consider the following algorithm:

$$\begin{aligned} x_i^{k+1} &= \Pi_{K_i}(x_i^k - \alpha_i \nabla_{x_i} \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)), \\ \lambda^{k+1} &= \Pi_{\mathcal{D}_\nu}(\lambda^k + \tau \nabla_\lambda \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)), \end{aligned} \tag{2.22}$$

where $\alpha_i > 0$ is the primal steplength for user i and $\tau > 0$ is the dual steplength. Next, we present our main convergence result for the sequence $\{z^k\}$ with $z^k = (x^k,\lambda^k)$ generated using (2.22).

Theorem 2. *Let Assumptions 2–4 hold. Let $\{z^k\}$ be a sequence generated by (2.22). Then, we have*

$$\|z^{k+1} - z_{\nu,\varepsilon}\| \leq \sqrt{q_{\nu,\varepsilon}}\|z^k - z_{\nu,\varepsilon}\| \qquad \text{for all } k \geq 0,$$

where $q_{\nu,\varepsilon}$ is given by

$$q_{\nu,\varepsilon} = \begin{cases} 1 + \alpha_{\max}^2 L_\Phi^2(\nu,\varepsilon) - 2\mu\tau + (\alpha_{\min} - \tau)\max\{1-2\nu, M_d^2\} & \\ +2(\alpha_{\max} - \alpha_{\min})L_\Phi(\nu,\varepsilon), & \text{for } \tau < \alpha_{\min} \leq \alpha_{\max}; \\ 1 + \alpha_{\max}^2 L_\Phi^2(\nu,\varepsilon) - 2\alpha_{\min}\mu & \\ +2(\alpha_{\max} - \alpha_{\min})L_\Phi(\nu,\varepsilon), & \text{for } \alpha_{\min} \leq \tau < \alpha_{\max}; \\ 1 + \tau^2 L_\Phi^2(\nu,\varepsilon) - 2\mu\alpha_{\min} + (\tau - \alpha_{\min})\max\{1-2\varepsilon, M_d^2\} & \\ +2(\alpha_{\max} - \alpha_{\min})L_\Phi(\nu,\varepsilon), & \text{for } \alpha_{\min} \leq \alpha_{\max} \leq \tau, \end{cases}$$

where $\alpha_{\min} = \min_{1\leq i\leq N}\{\alpha_i\}$, $\alpha_{\max} = \max_{1\leq i\leq N}\{\alpha_i\}$, $M_d = \max_{x\in K}\|\nabla d(x)\|$, $\mu = \min\{\nu,\varepsilon\}$ and $L_\Phi(\nu,\varepsilon)$ is as defined in Lemma 3.

Proof. Let $\{\alpha_i\}_{i=1}^N$ be the user dependent stepsizes of the primal iterations and let

$$\alpha_{\min} = \min_{1\leq i\leq N}\{\alpha_i\} \qquad \text{and} \qquad \alpha_{\max} = \max_{1\leq i\leq N}\{\alpha_i\}$$

denote the minimum and maximum of the user stepsizes. Using

$$x_{i,\nu,\varepsilon} = \Pi_{K_i}(x_{i,\nu,\varepsilon} - \alpha_i \nabla_{x_i} \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})),$$

non-expansive property of projection operator and Cauchy-Schwartz inequality, it can be verified that

$$\|x^{k+1} - x_{\nu,\varepsilon}\|^2 \leq \|x^k - x_{\nu,\varepsilon}\|^2 + \alpha_{\max}^2 \|\nabla_x \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k) - \nabla_x \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\|^2$$
$$- 2\alpha_{\min}(\nabla_x \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k) - \nabla_x \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon}))^T (x^k - x_{\nu,\varepsilon})$$
$$+ 2(\alpha_{\max} - \alpha_{\min})\|\nabla_x \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k) - \nabla_x \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\|\|x^k - x_{\nu,\varepsilon}\|,$$

and

$$\begin{aligned}\|\lambda^{k+1} - \lambda_{\nu,\varepsilon}\|^2 \quad \leq \quad & \|\lambda^k - \lambda_{\nu,\varepsilon}\|^2 + \tau^2\|(-d(x^k) + \varepsilon\lambda^k) - (-d(x_{\nu,\varepsilon}) + \varepsilon\lambda_{\nu,\varepsilon})\|^2 \\ & -2\tau(-d(x^k) + \varepsilon\lambda^k + d(x_{\nu,\varepsilon}) - \varepsilon\lambda_{\nu,\varepsilon})^T(\lambda^k - \lambda_{\nu,\varepsilon}).\end{aligned}$$

Summing the preceding two relations, we obtain

$$\begin{aligned}\|z^{k+1} - z_{\nu,\varepsilon}\|^2 \leq \|z^k - z_{\nu,\varepsilon}\|^2 &+ \max\{\alpha_{\max}^2, \tau^2\}\|\Phi(z^k) - \Phi(z_{\nu,\varepsilon})\|^2 \\ &- 2\alpha_{\min}(\nabla_x \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k) - \nabla_x \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon}))^T (x^k - x_{\nu,\varepsilon}) \\ &+ 2(\alpha_{\max} - \alpha_{\min})\|\nabla_x \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k) - \nabla_x \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\|\,\|x^k - x_{\nu,\varepsilon}\| \\ &- 2\tau(\nabla_\lambda \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k) - \nabla_\lambda \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon}))^T(\lambda^k - \lambda_{\nu,\varepsilon}). \qquad (2.23)\end{aligned}$$

We now consider three cases:

Case 1 ($\tau < \alpha_{\min} \leq \alpha_{\max}$): By adding and subtracting

$$2\tau(\nabla_x \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k) - \nabla_x \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon}))^T (x^k - x_{\nu,\varepsilon}),$$

we see that relation in (2.23) can be written as

$$\begin{aligned}\|z^{k+1} - z_{\nu,\varepsilon}\|^2 \leq \|z^k - z_{\nu,\varepsilon}\|^2 &+ \alpha_{\max}^2\|\Phi_{\nu,\varepsilon}(z^k) - \Phi_{\nu,\varepsilon}(z_{\nu,\varepsilon})\|^2 \\ &- 2\tau(\Phi_{\nu,\varepsilon}(z^k) - \Phi_{\nu,\varepsilon}(z_{\nu,\varepsilon}))^T(z^k - z_{\nu,\varepsilon}) \\ &- 2(\alpha_{\min} - \tau)\left(\nabla_x \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k) - \nabla_x \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\right)^T (x^k - x_{\nu,\varepsilon}) \\ &+ 2(\alpha_{\max} - \alpha_{\min})\|\nabla_x \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k) - \nabla_x \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\|\|x^k - x_{\nu,\varepsilon}\|.\end{aligned}$$

By Lemma 3, the mapping $\Phi_{\nu,\varepsilon}$ is strongly monotone and Lipschitz with constants $\mu = \min\{\nu,\varepsilon\}$

and $L_\Phi(\nu,\varepsilon)$, respectively. Hence, from the preceding relation we obtain

$$\begin{aligned}\|z^{k+1}-z_{\nu,\varepsilon}\|^2 \leq & (1+\alpha_{\max}^2 L_\Phi^2(\nu,\varepsilon)-2\tau\mu)\|z^k-z_{\nu,\varepsilon}\|^2 \\ & -2(\alpha_{\min}-\tau)\left(\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\right)^T(x^k-x_{\nu,\varepsilon}) \\ & +2(\alpha_{\max}-\alpha_{\min})\|\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\|\|x^k-x_{\nu,\varepsilon}\|.\end{aligned}$$

Now

$$\begin{aligned}\|\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\|\|x^k-x_{\nu,\varepsilon}\| &\leq \|\Phi(z^k)-\Phi(z_{\nu,\varepsilon})\|\|z^k-z_{\nu,\varepsilon}\| \\ &\leq L_\Phi(\nu,\varepsilon)\|z^k-z_{\nu,\varepsilon}\|^2\end{aligned}$$

and thus we get

$$\begin{aligned}\|z^{k+1}-z_{\nu,\varepsilon}\|^2 \leq & (1+\alpha_{\max}^2 L_\Phi^2(\nu,\varepsilon)-2\tau\mu)\|z^k-z_{\nu,\varepsilon}\|^2+2(\alpha_{\max}-\alpha_{\min})L_\Phi(\nu,\varepsilon)\|z^k-z_{\nu,\varepsilon}\|^2 \\ & -2(\alpha_{\min}-\tau)\left(\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\right)^T(x^k-x_{\nu,\varepsilon}).\end{aligned} \tag{2.24}$$

We next estimate the last term in the preceding relation. By adding and subtracting $\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda^k)$, we have

$$\begin{aligned}&\left(\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\right)^T(x^k-x_{\nu,\varepsilon}) \\ &=\left(\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda^k)\right)^T(x^k-x_{\nu,\varepsilon}) \\ &+\left(\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\right)^T(x^k-x_{\nu,\varepsilon}).\end{aligned}$$

Using the strong monotonicity of $\nabla_x\mathcal{L}_{\nu,\varepsilon}$, and writing the second term on the right hand side explicitly, we get

$$\begin{aligned}&\left(\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\right)^T(x^k-x_{\nu,\varepsilon}) \\ &\geq \nu\|x^k-x_{\nu,\varepsilon}\|^2+\sum_{j=1}^m\left(\nabla d_j(x_{\nu,\varepsilon})(\lambda_j^k-\lambda_{\nu,\varepsilon,j})\right)^T(x^k-x_{\nu,\varepsilon}) \\ &\geq \nu\|x^k-x_{\nu,\varepsilon}\|^2-\frac{1}{2}\left\|\sum_{j=1}^m\nabla d_j(x_{\nu,\varepsilon})(\lambda_j^k-\lambda_{\nu,\varepsilon,j})\right\|^2-\frac{1}{2}\|x^k-x_{\nu,\varepsilon}\|^2,\end{aligned}$$

where the last step follows by noting that $ab\geq-\frac{1}{2}(a^2+b^2)$. Using Cauchy-Schwartz and Hölder's inequality, we have

$$\begin{aligned}\left\|\sum_{j=1}^{m}\nabla d_j(x_{\nu,\varepsilon})(\lambda_j^k-\lambda_{\nu,\varepsilon,j})\right\|^2 &\leq \left(\sum_{j=1}^{m}\|\nabla d_j(x_{\nu,\varepsilon})\|\,|\lambda_j^k-\lambda_{\nu,\varepsilon,j}|\right)^2\\ &\leq \left(\sum_{j=1}^{m}\|\nabla d_j(x_{\nu,\varepsilon})\|^2\right)\|\lambda^k-\lambda_{\nu,\varepsilon}\|^2\\ &\leq M_d^2\|\lambda^k-\lambda_{\nu,\varepsilon}\|^2,\end{aligned}$$

where in the last step, the boundedness of $\nabla d(x)$ over K was employed ($\|\nabla d(x)\| \leq M_d$). By combining the preceding relations, we obtain

$$\begin{aligned}\Big(\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k) &- \nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\Big)^T(x^k-x_{\nu,\varepsilon})\\ &\geq -\frac{1}{2}\Big((1-2\nu)\|x^k-x_{\nu,\varepsilon}\|^2+M_d^2\|\lambda^k-\lambda_{\nu,\varepsilon}\|^2\Big)\\ &\geq -\frac{1}{2}\max\{1-2\nu,\, M_d^2\}\|z^k-z_{\nu,\varepsilon}\|^2.\end{aligned}$$

If the above estimate is substituted in (2.24), we obtain

$$\|z^{k+1}-z_{\nu,\varepsilon}\|^2 \quad \leq q_{\nu,\varepsilon}\|z^k-z_{\nu,\varepsilon}\|^2,$$

where $q_{\nu,\varepsilon}=1+\alpha_{\max}^2L_\Phi^2(\nu,\varepsilon)-2\mu\tau+(\alpha_{\min}-\tau)\max\{1-2\nu,M_d^2\}+2(\alpha_{\max}-\alpha_{\min})L_\Phi(\nu,\varepsilon)$, thus showing the desired relation.

Case 2 ($\alpha_{\min}\leq\tau<\alpha_{\max}$): By adding and subtracting

$$2\alpha_{\min}(\nabla_\lambda\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_\lambda\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon}))^T(\lambda^k-\lambda_{\nu,\varepsilon}),$$

for $\tau<\alpha_{\max}$ relation (2.23) reduces to

$$\begin{aligned}\|z^{k+1}-z_{\nu,\varepsilon}\|^2 \leq\ & \|z^k-z_{\nu,\varepsilon}\|^2+\alpha_{\max}^2\|\Phi(z^k)-\Phi(z_{\nu,\varepsilon})\|^2\\ &-2\alpha_{\min}(\Phi(z^k)-\Phi(z_{\nu,\varepsilon}))^T(z^k-z_{\nu,\varepsilon})\\ &-2(\tau-\alpha_{\min})(\nabla_\lambda\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_\lambda\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon}))^T(\lambda^k-\lambda_{\nu,\varepsilon})\\ &+2(\alpha_{\max}-\alpha_{\min})\|\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\|\|x^k-x_{\nu,\varepsilon}\|,\end{aligned}$$

which by Lipschitz continuity and strong monotonicity of Φ implies,

$$\begin{aligned}\|z^{k+1}-z_{\nu,\varepsilon}\|^2 &\le (1+\alpha_{\max}^2 L_\Phi^2(\nu,\varepsilon)-2\alpha_{\min}\mu)\|z^k-z_{\nu,\varepsilon}\|^2\\&\quad+2(\alpha_{\max}-\alpha_{\min})\|\nabla_\lambda\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_\lambda\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\|\|\lambda^k-\lambda_{\nu,\varepsilon}\|\\&\quad+2(\alpha_{\max}-\alpha_{\min})\|\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\|\|x^k-x_{\nu,\varepsilon}\|.\end{aligned}$$

Using Hölder's inequality, we get

$$\begin{aligned}\|z^{k+1}-z_{\nu,\varepsilon}\|^2 &\le (1+\alpha_{\max}^2 L_\Phi^2(\nu,\varepsilon)-2\alpha_{\min}\mu)\|z^k-z_{\nu,\varepsilon}\|^2\\&\quad+2(\alpha_{\max}-\alpha_{\min})\|\Phi(z^k)-\Phi(z_{\nu,\varepsilon})\|\|z^k-z_{\nu,\varepsilon}\|.\end{aligned}$$

Finally using Lipschitz continuity of Φ we get

$$\|z^{k+1}-z_{\nu,\varepsilon}\|^2 \le q\|z^k-z_{\nu,\varepsilon}\|^2,$$

where $q_{\nu,\varepsilon}=1+\alpha_{\max}^2 L_\Phi^2(\nu,\varepsilon)-2\alpha_{\min}\mu+2(\alpha_{\max}-\alpha_{\min})L_\Phi(\nu,\varepsilon)$.

Case 3 ($\alpha_{\min}\le\alpha_{\max}\le\tau$): Note that

$$\begin{aligned}&(\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon}))^T(x^k-x_{\nu,\varepsilon})\\&=\left(\Phi_{\nu,\varepsilon}(z^k)-\Phi_{\nu,\varepsilon}(z_{\nu,\varepsilon})\right)^T(z^k-z_{\nu,\varepsilon})\\&-(-d(x^k)+\varepsilon\lambda^k+d(x_{\nu,\varepsilon})-\varepsilon\lambda_{\nu,\varepsilon})^T(\lambda^k-\lambda_{\nu,\varepsilon}).\end{aligned}$$

Thus, from the preceding equality and relation (2.23), where $\alpha_{\max}<\tau$, we have

$$\begin{aligned}\|z^{k+1}-z_{\nu,\varepsilon}\|^2 &\le \|z^k-z_{\nu,\varepsilon}\|^2+\tau^2\|\Phi(z^k)-\Phi(z_{\nu,\varepsilon})\|^2\\&\quad-2\alpha_{\min}(\Phi(z^k)-\Phi(z_{\nu,\varepsilon}))^T(z^k-z_{\nu,\varepsilon})\\&\quad-2(\tau-\alpha_{\min})(\nabla_\lambda\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_\lambda\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon}))^T(\lambda^k-\lambda_{\nu,\varepsilon})\\&\quad+2(\alpha_{\max}-\alpha_{\min})\|\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x\mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\|\|x^k-x_{\nu,\varepsilon}\|.\end{aligned}$$

By Lemma 3, the mapping $\Phi_{\nu,\varepsilon}$ is strongly monotone and Lipschitz with constants $\mu=\min\{\nu,\varepsilon\}$

and $L_\Phi(\nu,\varepsilon)$, respectively. Hence, it follows

$$\begin{aligned}\|z^{k+1}-z_{\nu,\varepsilon}\|^2 \le (1+\tau^2 L_\Phi^2(\nu,\varepsilon)-2\alpha_{\min}\mu)\|z^k-z_{\nu,\varepsilon}\|^2 \\ -2(\tau-\alpha_{\min})(\nabla_\lambda \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_\lambda \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon}))^T(\lambda^k-\lambda_{\nu,\varepsilon}) \\ +2(\alpha_{\max}-\alpha_{\min})\|\nabla_x \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_x \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon})\|\|x^k-x_{\nu,\varepsilon}\|,\end{aligned}$$

which can be further estimated as

$$\begin{aligned}\|z^{k+1}-z_{\nu,\varepsilon}\|^2 \le (1+\tau^2 L_\Phi^2(\nu,\varepsilon)-2\alpha_{\min}\mu)\|z^k-z_{\nu,\varepsilon}\|^2+2(\alpha_{\max}-\alpha_{\min})L_\Phi(\nu,\varepsilon)\|z^k-z_{\nu,\varepsilon}\|^2 \\ -2(\tau-\alpha_{\min})(\nabla_\lambda \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)-\nabla_\lambda \mathcal{L}_{\nu,\varepsilon}(x_{\nu,\varepsilon},\lambda_{\nu,\varepsilon}))^T(\lambda^k-\lambda_{\nu,\varepsilon}). \qquad (2.25)\end{aligned}$$

Next, we estimate the last term on the right hand side of the preceding relation. Through the use of Cauchy-Schwartz inequality, we have

$$\left(d(x^k)-d(x_{\nu,\varepsilon})\right)^T(\lambda^k-\lambda_{\nu,\varepsilon}) \le \frac{1}{2}\left\|d(x^k)-d(x_{\nu,\varepsilon})\right\|^2+\frac{1}{2}\|\lambda^k-\lambda_{\nu,\varepsilon}\|^2.$$

By the continuity of the gradient mapping of $d(x)=(d_1(x),\ldots,d_m(x))^T$ and its boundedness ($\|\nabla d(x)\| \le M_d$), using the Mean-value Theorem we further have

$$\left\|d(x^k)-d(x_{\nu,\varepsilon})\right\|^2 \le M_d^2\|x^k-x_{\nu,\varepsilon}\|^2.$$

From the preceding two relations we have

$$\left(d(x^k)-d(x_{\nu,\varepsilon})\right)^T(\lambda^k-\lambda_{\nu,\varepsilon}) \le \frac{M_d^2}{2}\|x^k-x_{\nu,\varepsilon}\|^2+\frac{1}{2}\|\lambda^k-\lambda_{\nu,\varepsilon}\|^2,$$

which when substituted in inequality (2.25) yields

$$\begin{aligned}\|z^{k+1}-z_{\nu,\varepsilon}\|^2 \;\le\; & \left(1+\tau^2 L_\Phi^2(\nu,\varepsilon)-2\mu\alpha_{\min}+2(\alpha_{\max}-\alpha_{\min})L_\Phi(\nu,\varepsilon)\right)\|z^k-z_{\nu,\varepsilon}\|^2 \\ & +(\tau-\alpha_{\min})(1-2\varepsilon)\|\lambda^k-\lambda_{\nu,\varepsilon}\|^2+(\tau-\alpha_{\min})M_d^2\|x^k-x_{\nu,\varepsilon}\|^2.\end{aligned}$$

The desired relation follows by observing that

$$(1-2\varepsilon)\|\lambda^k-\lambda_{\nu,\varepsilon}\|^2+M_d^2\|x^k-x_{\nu,\varepsilon}\|^2 \le \max\{1-2\varepsilon, M_d^2\}\|z^k-z_{\nu,\varepsilon}\|^2.$$

■

An immediate corollary of Theorem 2 is obtained when all users have the same steplength. More precisely, we have the following algorithm:

$$\begin{aligned} x^{k+1} &= \Pi_K(x^k - \alpha \nabla_x \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)), \\ \lambda^{k+1} &= \Pi_{\mathcal{D}_\nu}(\lambda^k + \tau \nabla_\lambda \mathcal{L}_{\nu,\varepsilon}(x^k,\lambda^k)), \end{aligned} \tag{2.26}$$

where $\alpha > 0$ and $\tau > 0$ are, respectively, primal and dual stepsizes. We present the convergence of the sequence $\{z^k\}$ with $z^k = (x^k, \lambda^k)$ in the next corollary.

Corollary 1. *Let Assumptions 2–4 hold. Let $\{z^k\}$ be a sequence generated by (2.26) with the primal and dual step-sizes chosen independently. Then, we have*

$$\|z^{k+1} - z_{\nu,\varepsilon}\| \leq \sqrt{q_{\nu,\varepsilon}}\|z^k - z_{\nu,\varepsilon}\| \qquad \text{for all } k \geq 0,$$

where $q_{\nu,\varepsilon}$ is given by

$$q_{\nu,\varepsilon} = 1 - 2\mu \min\{\alpha,\tau\} + \max\{\alpha^2,\tau^2\} L_\Phi^2(\nu,\varepsilon) + \theta(\alpha,\tau),$$

$$\text{and } \theta(\alpha,\tau) \triangleq \begin{cases} (\alpha-\tau)\max\{1-2\nu, M_d^2\} & \text{for } \tau \leq \alpha, \\ (\tau-\alpha)\max\{1-2\varepsilon, M_d^2\} & \text{for } \alpha < \tau, \end{cases}$$

$\mu = \min\{\nu,\varepsilon\}$ *and $L_\Phi^2(\nu,\varepsilon)$ is as given in Lemma 3.*

Note that when $\alpha_{\min} = \alpha_{\max} = \tau$ and $\tau < 2\mu/L_\Phi^2(\nu,\varepsilon)$, Theorem 2 implies the standard contraction result for a strongly monotone and Lipschitz mapping. However, Theorem 2 does not guarantee the existence of a tuple $(\alpha_{\min}, \alpha_{\max}, \tau)$ resulting in a contraction in general, i.e., does not ensure that $q_{\nu,\varepsilon} \in (0,1)$. This is done in the following lemma.

Lemma 4. *Let $q_{\nu,\varepsilon}$ be as given in Theorem 1. Then, there exists a tuple $(\alpha_{\min}, \alpha_{\max}, \tau)$ such that $q_{\nu,\varepsilon} \in (0,1)$.*

Proof. It suffices to show that there exists a tuple $(\alpha_{\min}, \alpha_{\max}, \tau)$ such that

$$\begin{aligned} &0 < 1 + \alpha_{\max}^2 L_\Phi^2(\nu,\varepsilon) - 2\mu\tau + (\alpha_{\min} - \tau)\max\{1-2\nu, M_d^2\} \\ &+2(\alpha_{\max} - \alpha_{\min}) L_\Phi(\nu,\varepsilon) < 1 && \tau < \alpha_{\min} \leq \alpha_{\max} \\ &0 < 1 + \alpha_{\max}^2 L_\Phi^2(\nu,\varepsilon) - 2\alpha_{\min}\mu \\ &+2(\alpha_{\max} - \alpha_{\min}) L_\Phi(\nu,\varepsilon) < 1 && \alpha_{\min} \leq \tau < \alpha_{\max} \\ &0 < 1 + \tau^2 L_\Phi^2(\nu,\varepsilon) - 2\mu\alpha_{\min} + (\tau - \alpha_{\min})\max\{1-2\varepsilon, M_d^2\} \\ &+2(\alpha_{\max} - \alpha_{\min}) L_\Phi(\nu,\varepsilon) < 1 && \alpha_{\min} \leq \alpha_{\max} \leq \tau. \end{aligned}$$

Also, it suffices to prove only one of the cases since the other cases follow by interchanging the roles of τ and $\alpha_{\min}$ or τ and $\alpha_{\max}$. We consider the case where $\tau < \alpha_{\min} \leq \alpha_{\max}$. Here, if $\alpha_{\max} < 2\mu/L_{\Phi}^2(\nu,\varepsilon)$ then there is $\beta < 1$ such that setting $\tau = \beta\alpha_{\max}$ we have $q < 1$. To see this let $\alpha_{\min} = \beta_1\alpha_{\max}$ such that $\beta < \beta_1 \leq 1$ and $\max\{1-2\nu, M_d^2\} = M_d^2$. Consider

$$q_{\nu,\varepsilon} - 1 = -2\mu\tau + \alpha_{\max}^2 L_{\Phi}^2(\nu,\varepsilon) + (\alpha_{\min} - \tau)M_d^2 + 2(\alpha_{\max} - \alpha_{\min})L_{\Phi}(\nu,\varepsilon).$$

Setting $\tau = \beta\alpha_{\max}$, $\alpha_{\min} = \beta_1\alpha_{\max}$, the preceding relation reduces to

$$q_{\nu,\varepsilon} - 1 = -2\mu\beta\alpha_{\max} + \alpha_{\max}^2 L_{\Phi}^2(\nu,\varepsilon) + \alpha_{\max}(\beta_1 - \beta)M_d^2 + 2\alpha_{\max}(1-\beta_1)L_{\Phi}(\nu,\varepsilon).$$

Using $\beta < \beta_1 \leq 1$ we obtain

$$q_{\nu,\varepsilon} - 1 \leq -2\mu\beta\alpha_{\max} + \alpha_{\max}^2 L_{\Phi}^2(\nu,\varepsilon) + \alpha_{\max}(1-\beta)M_d^2 + 2\alpha_{\max}(1-\beta)L_{\Phi}(\nu,\varepsilon).$$

We are done if we show that the expression on the right hand side of the preceding relation is negative for some β i.e.,

$$-2\mu\beta\alpha_{\max} + \alpha_{\max}^2 L_{\Phi}^2(\nu,\varepsilon) + \alpha_{\max}(1-\beta)M_d^2 + 2\alpha_{\max}(1-\beta)L_{\Phi}(\nu,\varepsilon) < 0.$$

Following some rearrangement it can be verified that

$$\beta > \frac{\alpha_{\max}L_{\Phi}^2(\nu,\varepsilon) + M_d^2 + 2L_{\Phi}(\nu,\varepsilon)}{2\mu + M_d^2 + 2L_{\Phi}(\nu,\varepsilon)}.$$

Since we have $\alpha_{\max}L_{\Phi}^2(\nu,\varepsilon) < 2\mu$ it follows that the expression on right hand side of the preceding relation is strictly less than 1 and we have

$$\beta \in \left(\frac{\alpha_{\max}L_{\Phi}^2(\nu,\varepsilon) + M_d^2 + 2L_{\Phi}(\nu,\varepsilon)}{2\mu + M_d^2 + 2L_{\Phi}(\nu,\varepsilon)}, 1\right),$$

implying that we have $\beta \in (0,1)$. ■

The previous result is motivated by several issues arising in practical settings. First there may be errors due to noisy links in the communication network, which may cause inconsistencies across stepsizes. Often, it may be difficult to even enforce this consistency. As a consequence, we examine the extent to which the convergence theory is affected by a lack of consistency. A related question is whether one can, in a distributed setting, impose alternative requirements that weaken consistency. This can be achieved by setting bounds on the primal and dual stepsizes which are independent. For instance, if $\alpha_{\max} < \frac{2\mu}{L_{\Phi}^2(\nu,\varepsilon)}$, then it suffices to choose τ independently as $\tau \leq \beta\alpha_{\max} \leq \beta\frac{2\mu}{L_{\Phi}^2(\nu,\varepsilon)}$, where β is chosen independently. Importantly, Lemma 4 provides a characterization of the relationship between $\alpha_{\min}$, $\alpha_{\max}$ and τ using the values of problem parameters, to ensure convergence of the scheme. Expectedly, as the numerical results testify, the performance does deteriorate when there α_i's and τ do not match.

Finally, we remark briefly on the relevance of allowing for differing stepsizes. In distributed settings, communication of stepsizes may be corrupted via error due to noisy communication links. A majority of past work on such problems (cf. [3, 27]) requires that stepsizes be consistent across users. Furthermore, in constrained regimes, there is a necessity to introduce both primal (user) stepsizes and dual (link) stepsizes. We show that there may be limited diversity across all of these parameters while requiring that these parameters together satisfy some relationship. One may question if satisfying this requirement itself requires some coordination. In fact, we show that this constraint is implied by a set of private user-specific and dual requirements on their associated stepsizes, allowing for ease of implementation.

2.2.5 Extension to independently chosen regularization parameters

In this subsection, we extend the results of the preceding section to a regime where the ith user selects its own regularization parameter ν_i. Before proceeding, we provide a brief motivation of such a line of questioning. In networked settings specifying stepsizes and regularization parameters for the users at every instant is generally challenging. Enforcing consistent choices across these users is also difficult. An alternative lies in *broadcasting* a range of choices for stepsizes, as done in the previous subsection. In this section, we show that an analogous approach can be leveraged for specifying regularization parameters, with limited impacts on the final results. Importantly, the benefit of these results lies in the fact that enforcement of consistency of regularization parameters is no longer necessary. We start with definition of the regularized Lagrangian function with user specific regularization terms. In particular, we let

$$\mathcal{L}_{V,\varepsilon}(x,\lambda) = f(x) + \frac{1}{2}x^T V x + \lambda^T d(x) - \frac{\varepsilon}{2}\|\lambda\|^2 \tag{2.27}$$

where V is a diagonal matrix with diagonal entries $\nu_1, \ldots, \nu_N$. In this case, letting

$$\nu_{\max} \triangleq \max_{i\in\{1,\ldots,N\}}\{\nu_i\},$$

for some $\acute{x} \in K$ and $\hat{\lambda} \geq 0$, we define the set $\mathcal{D}_{\nu_{\max}}$ given by:

$$\mathcal{D}_{\nu_{\max}} = \left\{ \lambda \in \mathbb{R}^m \,\middle|\, \sum_{j=1}^{m} \lambda_j \leq \frac{f(\acute{x}) + \frac{\nu_{\max}}{2}\|\acute{x}\|^2 - v(\hat{\lambda})}{\min_{1 \leq j \leq m}\{-d_j(\acute{x})\}},\ \lambda \geq 0 \right\}. \tag{2.28}$$

We consider the regularized primal problem (2.1) with the regularization term $\frac{1}{2}x^T V x$. We let Λ_V^* be the set of dual optimal solutions of such regularized primal problem. Then, relation (2.10) holds for Λ_V^* and $\mathcal{D}_{\nu_{\max}}$, i.e., $\Lambda_V^* \subseteq \mathcal{D}_{\nu_{\max}}$ and, therefore, the development in the preceding two sections extends to this case as well. We let x_V^* and λ_V^* denote primal-dual of the regularized primal problem with the regularization term $\frac{1}{2}x^T V x$. Analogously, we let $x_{V,\varepsilon}^*$ and $\lambda_{V,\varepsilon}^*$ denote respectively the primal and the dual part of the saddle point solution for $\mathcal{L}_{V,\varepsilon}(x,\lambda)$ over $K \times \mathbb{R}_+^m$. We present the modified results in the form of remarks and omit the details of proofs.

The bound of Proposition 1 when user i uses its own regularization parameter ν_i will reduce to:

$$(x_V^* - x_{V,\varepsilon})^T V (x_V^* - x_{V,\varepsilon}) + \frac{\varepsilon}{2}\|\lambda_{V,\varepsilon}\|^2 \leq \frac{\varepsilon}{2}\|\lambda_{\mathbf{V}}^*\|^2 \qquad \text{for all } \lambda_V^* \in \Lambda_V^*,$$

and thus we have the following bound

$$\|x_V^* - x_{V,\varepsilon}\| \leq \sqrt{\frac{\varepsilon}{2\nu_{\min}}} \max_{\lambda^* \in \mathcal{D}_{\nu_{\max}}} \|\lambda^*\|,$$

where $\nu_{\min} \triangleq \min_{i \in \{1,\dots,N\}} \{\nu_i\}$.

The result in Lemma 2 is replaced by the following one.

Lemma 5. *Let Assumptions 1 and 2 hold. For any $\nu_i > 0$, $i = 1,\dots,N$, and $\varepsilon > 0$, we have*

$$\max\{0, d_j(x_{V,\varepsilon})\} \leq M_{d_j} M_{\nu_{\max}} \sqrt{\frac{\varepsilon}{2\nu_{\min}}} \qquad \text{for all } j = 1,\dots,m,$$
$$|f(x_{V,\varepsilon}) - f(x^*)| \leq M_f M_{\nu_{\max}} \sqrt{\frac{\varepsilon}{2\nu_{\min}}} + \frac{\nu_{\max}}{2} D^2,$$

with $M_{d_j} = \max_{x \in K}\|\nabla d_j(x)\|$ for each $j = 1,\dots,m$, $M_f = \max_{x \in K}\|\nabla f(x)\|$, $M_{\nu_{\max}} = \max_{\lambda^* \in \mathcal{D}_{\nu_{\max}}} \|\lambda^*\|$ and $D = \max_{x \in K}\|x\|$.

Our result following Lemma 2 where we describe how one may choose parameters ε and ν to get within a given threshold error on the deviation of the obtained function value from its optimal counterpart will have to be reconsidered using the appropriate parameters $\nu_{\min}$ and $\nu_{\max}$. More precisely, we will have $|f(x_{V,\varepsilon}) - f(x^*)| < \delta$ if we have

$$M_f M_{v_{\max}} \sqrt{\frac{\varepsilon}{v_{\min}}} + \frac{v_{\max}}{2} D^2 < \delta.$$

Following a similar analysis and using the structure of the set $D_{v_{\max}}$, we have

$$\begin{aligned}
M_{v_{\max}} \sqrt{\frac{\varepsilon}{2v_{\min}}} &= \sqrt{\frac{\varepsilon}{2v_{\min}}} \left(\max_{\lambda \in \mathcal{D}_{v_{\max}}} \|\lambda\|_1 \right) \\
&= \sqrt{\frac{\varepsilon}{2v_{\min}}} \left(\frac{f(\acute{x}) + \frac{v_{\max}}{2}\|\acute{x}\|^2 - v(\hat{\lambda})}{\min_{1\le j\le m}\{-d_j(\acute{x})\}} \right) \\
&= \sqrt{\frac{\varepsilon}{2v_{\min}}} \left(\frac{f(\acute{x}) - v(\hat{\lambda})}{\min_{1\le j\le m}\{-d_j(\acute{x})\}} \right) + \sqrt{\frac{\varepsilon}{2v_{\min}}} \left(\frac{\frac{v_{\max}}{2}\|\acute{x}\|^2}{\min_{1\le j\le m}\{-d_j(\acute{x})\}} \right),
\end{aligned}$$

Letting $a = \frac{f(\acute{x}) - v(\hat{\lambda})}{\sqrt{2}\min_{1\le j\le m}\{-d_j(\acute{x})\}}$ and $b = \frac{\|\acute{x}\|^2}{2\sqrt{2}\min_{1\le j\le m}\{-d_j(\acute{x})\}}$, we have

$$M_f M_{v_{\max}} \sqrt{\frac{\varepsilon}{v_{\min}}} + \frac{v_{\max}}{2} D^2 \le M_f \left(\frac{a\sqrt{\varepsilon}}{\sqrt{v_{\min}}} + b\sqrt{\frac{\varepsilon v_{\max}^2}{v_{\min}}} \right) + \frac{v_{\max}}{2} D^2 < \delta.$$

Next, we may choose parameters $v_{\min}$, $v_{\max}$ and ε so that the above inequality is satisfied. The expression suggests that one must choose $\varepsilon < v_{\min}$ (as M_f could be large). Thus, setting $\varepsilon = v_{\min} v_{\max}^2$, we will obtain a quadratic inequality in parameter $v_{\max}$ which can subsequently allow for selecting $v_{\max}$ and, therefore, selecting $v_{\min}$ and ε.

Analogous to the definition of the mapping $\Phi_{v,\varepsilon}(x,\lambda)$ in (2.7), we define the regularized mapping corresponding to the Lagrangian in (2.27). Specifically, we have the regularized mapping $\Phi_{V,\varepsilon}(x,\lambda)$ given by

$$\Phi_{V,\varepsilon}(x,\lambda) \triangleq (\nabla_x \mathcal{L}_{V,\varepsilon}(x,\lambda), -\nabla_\lambda \mathcal{L}_{V,\varepsilon}(x,\lambda)) = (\nabla_x \mathcal{L}(x,\lambda) + Vx, -d(x) + \varepsilon\lambda).$$

The properties of $\Phi_{V,\varepsilon}$, namely, strong monotonicity and Lipschitz continuity remain. Specifically, $\Phi_{V,\varepsilon}$ is strongly monotone with the same constant μ as before, i.e., $\mu = \min\{v_{\min}, \varepsilon\}$. However, Lipschitz constant is not the same. Letting $L_\Phi(V,\varepsilon)$ denote a Lipschitz constant for $\Phi_{V,\varepsilon}$, we have

$$L_\Phi(V,\varepsilon) = \sqrt{(L + v_{\max} + M_d + M_{v_{\max}} L_d)^2 + (M_d + \varepsilon)^2}. \tag{2.29}$$

The result of Theorem 2 can be expressed as in the following corollary.

Corollary 2. *Let Assumptions 1-4 hold. Let $\{z^k\}$ be a sequence generated by* (2.22) *with each user using ν_i as its regularization parameter instead of ν. Then, we have*

$$\|z^{k+1} - z_{V,\varepsilon}\| \leq \sqrt{q_{V,\varepsilon}}\|z^k - z_{V,\varepsilon},\|$$

with $q_{V,\varepsilon}$ as given in Theorem 2, where $L_\Phi(\nu,\varepsilon)$ is replaced by $L_\Phi(V,\varepsilon)$ from (2.29).

2.3 A regularized dual method

The focus in Section 2.2 has been on primal-dual method dealing with problems where a set of convex constraints couples the user decisions. A key property of our primal-dual method is that both schemes have the same time-scales. In many practical settings, the primal and dual updates are carried out by very different entities so that the time-scales may be vastly different. For instance, the dual updates of the Lagrange multipliers could be controlled by the network operator and might be on a slower time-scale than the primal updates that are made by the users. Dual methods have proved useful in multiuser optimization problems and their convergence to the *optimal primal* solution has been studied for the case when the user objectives are strongly convex [3, 2].

In this section, we consider regularization to deal with the lack of strong convexity of Lagrangian subproblems and to also accommodate inexact solutions of the Lagrangian subproblems. For the inexact solutions, we develop error bounds. Inexactness is essential in constructing distributed online schemes that require primal solutions within a fixed amount of time. In the standard dual framework, for each $\lambda \in \mathbb{R}^m_+$, a solution $x(\lambda) \in K$ of a Lagrangian subproblem is given by a solution to $\mathrm{VI}(K, \nabla_x \mathcal{L}(x,\lambda))$, which satisfies the following inequality:

$$(x - x(\lambda))^T \nabla_x \mathcal{L}(x(\lambda),\lambda) \geq 0 \qquad \text{for all } x \in K,$$

where $\nabla_\lambda \mathcal{L}(x(\lambda),\lambda) = \nabla_x f(x(\lambda)) + \sum_{j=1}^m \lambda_j \nabla d_j(x(\lambda))$. An optimal dual variable λ is a solution of $\mathrm{VI}(\mathbb{R}^m_+, -\nabla_\lambda \mathcal{L}(x(\lambda),\lambda))$ given by

$$(\hat{\lambda} - \lambda)^T (-\nabla_\lambda \mathcal{L}(x(\lambda),\lambda)) \geq 0 \qquad \text{for all } \hat{\lambda} \in \mathbb{R}^m_+,$$

where $\nabla_\lambda \mathcal{L}(x,\lambda) = d(x)$. We consider a regularization in both primal and dual space as discussed in Section 2.2. In Section 2.3.1, we discuss the exact dual method and provide the contraction results in the primal and dual space as well as bounds on the infeasibility. These results are extended to allow for inexact solutions of Lagrangian subproblems in Section 2.3.2.

2.3.1 Regularized exact dual method

We begin by considering an *exact* dual scheme for the regularized problem given by

$$x^t = \Pi_K(x^t - \alpha \nabla_x \mathcal{L}_{\nu,\varepsilon}(x^t, \lambda^t)), \tag{2.30}$$

$$\lambda^{t+1} = \Pi_{\mathcal{D}_\nu}(\lambda^t + \tau \nabla_\lambda \mathcal{L}_{\nu,\varepsilon}(x^t, \lambda^t)) \qquad \text{for } t \geq 0, \tag{2.31}$$

where the set $\mathcal{D}_\nu$ is as defined in (2.10). In the primal step (2.30), the vector x^t denotes the solution $x(\lambda^t)$ of the fixed-point equation corresponding to the current Lagrange multiplier λ^t.

We now focus on the conditions ensuring that the sequence $\{\lambda^t\}$ converges to the optimal dual solution $\lambda^*_{\nu,\varepsilon}$ and that the corresponding $\{x(\lambda^t)\}$ converges to the primal optimal $x^*_{\nu,\varepsilon}$ of the regularized problem. We note that Proposition 1 combined with Lemma 2 provide bounds on the constraint violations, and a bound on the difference in the function values $f(x^*_{\nu,\varepsilon})$ and the primal optimal value f^* of the original problem.

Lemma 6. *Under Assumption 2, the function $-d(x(\lambda))$ is co-coercive in λ with constant $\frac{\nu}{M_d^2}$, where $M_d = \max_{x \in K} \|\nabla d(x)\|$.*

Proof. Let λ_1 and $\lambda_2 \in \mathbb{R}^m_+$. Let x_1 and x_2 denote the solutions to $\mathrm{VI}(K, \nabla_x \mathcal{L}_{\nu,\varepsilon}(x, \lambda_1))$ and $\mathrm{VI}(K, \nabla_x \mathcal{L}_{\nu,\varepsilon}(x, \lambda_2))$, respectively. Then, we have the following inequalities:

$$(x_2 - x_1)^T(\nabla f_\nu(x_1) + \sum_{j=1}^m \lambda_{1,j} \nabla d_j(x_1)) \geq 0,$$
$$(x_1 - x_2)^T(\nabla f_\nu(x_2) + \sum_{j=1}^m \lambda_{2,j} \nabla d_j(x_2)) \geq 0,$$

where $\lambda_{1,j}$ and $\lambda_{2,j}$ denote the jth component of vectors λ_1 and λ_2, respectively. Summing these inequalities, we get

$$\sum_{j=1}^m \lambda_{1,j}(x_2 - x_1)^T \nabla d_j(x_1) + \sum_{j=1}^m \lambda_{2,j}(x_1 - x_2)^T \nabla d_j(x_2)$$
$$\geq (x_2 - x_1)^T(\nabla f_\nu(x_2) - \nabla f_\nu(x_1)) \geq \nu \|x_2 - x_1\|^2. \tag{2.32}$$

By using the convexity of the functions d_j and inequality (2.32), we obtain

$$
\begin{aligned}
(\lambda_2-\lambda_1)^T(-d(x_2)+d(x_1)) &= \sum_{j=1}^{m}\lambda_{1,j}(d_j(x_2)-d_j(x_1))+\sum_{j=1}^{m}\lambda_{2,j}(d_j(x_1)-d_j(x_2)) \\
&\geq \sum_{j=1}^{m}\lambda_{1,j}(x_2-x_1)^T\nabla d_j(x_1)+\sum_{j=1}^{m}\lambda_{2,j}(x_1-x_2)^T\nabla d_j(x_2) \\
&\geq \nu\|x_2-x_1\|^2. \qquad (2.33)
\end{aligned}
$$

Now, by using the Lipschitz continuity of $d(x)$, as implied by Assumption 2, we see that $\|x_2 - x_1\|^2 \geq \frac{\nu}{M_d^2}\|d(x_2)-d(x_1)\|^2$ with $M_d = \max_{x\in K}\nabla d(x)$, which when substituted in the preceding relation yields the result. ■

We now prove our convergence result for the dual method, relying on the exact solution of the corresponding Lagrangian subproblem.

Proposition 3. *Let Assumptions 1 and 2 hold, and let the step size τ be such that*

$$
\tau < \frac{2\nu}{M_d^2+2\varepsilon\nu} \qquad \textit{with} \qquad M_d = \max_{x\in K}\nabla d(x).
$$

Then, for the sequence $\{\lambda_t\}$ generated by the dual method in (2.31), we have

$$
\|\lambda^{t+1}-\lambda^*_{\nu,\varepsilon}\| \leq q\|\lambda^t-\lambda^*_{\nu,\varepsilon}\| \qquad \textit{where } q = 1-\tau\varepsilon.
$$

Proof. By using the definition of the dual method in (2.31) and the non-expansivity of the projection, we obtain the following set of inequalities:

$$
\begin{aligned}
\|\lambda^{t+1}-\lambda^*_{\nu,\varepsilon}\|^2 &\leq \|\lambda^t+\tau(d(x^t)-\varepsilon\lambda^t)-\left(\lambda^*_{\nu,\varepsilon}+\tau(d(x^*_{\nu,\varepsilon})-\varepsilon\lambda^*_{\nu,\varepsilon})\right)\|^2 \\
&= \|(1-\tau\varepsilon)(\lambda^t-\lambda^*_{\nu,\varepsilon})-\tau\left(d(x^*_{\nu,\varepsilon})-d(x^t)\right)\|^2 \\
&= (1-\tau\varepsilon)^2\|\lambda^t-\lambda^*_{\nu,\varepsilon}\|^2+\tau^2\|d(x^*_{\nu,\varepsilon})-d(x^t)\|^2 \\
&\quad -2\tau(1-\tau\varepsilon)(\lambda^t-\lambda^*_{\nu,\varepsilon})^T\left(d(x^*_{\nu,\varepsilon})-d(x^t)\right).
\end{aligned}
$$

By invoking the co-coercivity of $-d(x)$ from Lemma 6, we further obtain

$$
\|\lambda^{t+1}-\lambda^*_{\nu,\varepsilon}\|^2 \leq (1-\tau\varepsilon)^2\|\lambda^t-\lambda^*_{\nu,\varepsilon}\|^2+\left(\tau^2-2\tau(1-\tau\varepsilon)\frac{\nu}{M_d^2}\right)\|d(x^*_{\nu,\varepsilon})-d(x^t)\|^2.
$$

A contraction may be obtained by choosing τ such that $(\tau^2 - 2\tau(1-\tau\varepsilon)\frac{\nu}{M_d^2}) < 0$

$$\text{and } \tau < \frac{1}{\varepsilon} \text{ as given by } \tau < \frac{2\nu/M_d^2}{1+2\varepsilon\nu/M_d^2} < \frac{1}{\varepsilon}.$$

We therefore conclude that $\|\lambda^{t+1} - \lambda^*_{\nu,\varepsilon}\|^2 \leq (1-\tau\varepsilon)^2 \|\lambda^t - \lambda^*_{\nu,\varepsilon}\|^2$ for all $t \geq 0$. ■

Next, we examine two remaining concerns. First, can a bound on the norm $\|x^t - x^*_{\nu,\varepsilon}\|$ be obtained, where $x^t = x(\lambda^t)$? Second, can one make a rigorous statement regarding the infeasibility of x^t, similar to that provided in the context of the primal-dual method in Section 2.2?

Proposition 4. *Let Assumptions 1 and 2 hold. Then, for the sequence $\{x^t\}$, with $x^t = x(\lambda^t)$, generated by the dual method* (2.30) *using a step-size τ such that $\tau < \frac{2\nu}{M_d^2 + 2\varepsilon\nu}$, we have for all $t \geq 0$,*

$$\|x^t - x^*_{\nu,\varepsilon}\| \leq \frac{M_d}{\nu}\|\lambda^t - \lambda^*_{\nu,\varepsilon}\| \qquad \textit{and} \qquad \max\{0, d_j(x^t)\} \leq \frac{M_d^2}{\nu}\|\lambda^t - \lambda^*_{\nu,\varepsilon}\|.$$

Proof. From relation (2.33) in the proof of Lemma 6, the Cauchy-Schwartz inequality and the boundedness of $\nabla d_j(x)$ for all $j = 1, \ldots, m$, we have

$$\begin{aligned}
\|x^t - x^*_{\nu,\varepsilon}\|^2 &= \|x(\lambda^t) - x(\lambda^*_{\nu,\varepsilon})\|^2 \\
&\leq \frac{1}{\nu}\left(\lambda^t - \lambda^*_{\nu,\varepsilon}\right)^T \left(-d(x(\lambda^t)) + d(x(\lambda^*_{\nu,\varepsilon}))\right) \\
&\leq \frac{M_d}{\nu}\|\lambda^t - \lambda^*_{\nu,\varepsilon}\| \|x(\lambda^t) - x(\lambda^*_{\nu,\varepsilon})\|,
\end{aligned}$$

implying that $\|x^t - x^*_{\nu,\varepsilon}\| \leq \frac{M_d}{\nu}\|\lambda^t - \lambda^*_{\nu,\varepsilon}\|$. Furthermore, a bound on $\max\{0, d_j(x^t)\}$ can be obtained by invoking the convexity of each of the functions d_j and the boundedness of their gradients, as follows:

$$d_j(x^t) \leq d_j(x^*_{\nu,\varepsilon}) + \nabla d_j(x^*_{\nu,\varepsilon})^T (x^t - x^*_{\nu,\varepsilon}) \leq M_d \|x^t - x^*_{\nu,\varepsilon}\| \leq \frac{M_d^2}{\nu}\|\lambda^t - \lambda^*_{\nu,\varepsilon}\|,$$

where in the second inequality we use $d_j(x^*_{\nu,\varepsilon}) \leq 0$. Thus, a bound on the violation of constraints $d_j(x) \leq 0$ at $x = x^t$ is given by $\max\{0, d_j(x^t)\} \leq \frac{M_d^2}{\nu}\|\lambda^t - \lambda^*_{\nu,\varepsilon}\|$. ■

2.3.2 Regularized inexact dual method

The *exact* dual scheme requires solving the Lagrangian subproblem to optimality for a given value of the Lagrange multiplier. In practical settings, primal solutions are obtained via distributed iterative schemes and exact solutions are inordinately expensive from a computational standpoint. This motivates our study of the error properties resulting from solving the Lagrangian subproblem inexactly for every iteration in dual space. In particular, we consider a method executing a specified *fixed number of iterations*, say K, in the primal space for every iteration in the dual space. Our intent is to provide error bounds contingent on K. The inexact form of the dual method is given by the following:

$$x^{k+1}(\lambda^t) = \Pi_K(x^k(\lambda^t) - \alpha\nabla_x \mathcal{L}_{\nu,\varepsilon}(x^k(\lambda^t),\lambda^t)) \qquad k = 0,\ldots,K-1,\ t \geq 0, \tag{2.34}$$

$$\lambda^{t+1} = \Pi_{\mathcal{D}_\nu}\left(\lambda^t + \tau\nabla_\lambda \mathcal{L}_{\nu,\varepsilon}(x^K(\lambda^t),\lambda^t)\right) \qquad t \geq 0. \tag{2.35}$$

Throughout this section, we omit the explicit dependence of x on λ, by letting $x^k(t) \triangleq x^k(\lambda^t)$. We have the following result.

Lemma 7. *Let Assumptions 1–4 hold. Let* $\{x^k(t)\}, k = 1,\ldots,K, t \geq 0$ *be generated by (2.34) using a step-size* α*, with* $0 < \alpha < \frac{2}{L_f}$ *where* $L_f = L + \nu + M_\nu L_d$*,* L *and* L_d *are Lipschitz constants for the gradient maps* ∇f *and* ∇d *respectively, while* $M_\nu = \max_{\lambda\in\mathcal{D}_\nu}\|\lambda\|$*. Then, we have for all* t *and all* $k = 1,\ldots,K$*,*

$$\|x^k(t) - x(t)\| \leq q_p^{k/2}\|x^0(t) - x(t)\|,$$

where $x(t) := x(\lambda^t)$ *solves the Lagrangian subproblem corresponding to the multiplier* λ^t *and* $q_p = 1 - \alpha\nu(2 - \alpha L_f)$*.*

Proof. We observe that for each λ_t the mapping $\nabla_x\mathcal{L}_{\nu,\varepsilon}(x^k(\lambda^t),\lambda^t)$ of the Lagrangian subproblem is strongly monotone and Lipschitz continuous. The geometric convergence follows directly from [57], page 164, Theorem 13.1. ■

Our next proposition provides a relation for $\|\lambda^{t+1} - \lambda^*_{\nu,\varepsilon}\|$ in terms of $\|\lambda^t - \lambda^*_{\nu,\varepsilon}\|^2$ with an error bound depending on K and t.

Proposition 5. *Let Assumptions 2–4 hold. Let the sequence* $\{\lambda^t\}$ *be generated by (2.34)–(2.35) using a step-size* α *as in Lemma 7 and a step-size* τ *such that*

$$\tau < \min\left\{\frac{2\nu}{M_d^2 + 2\varepsilon\nu}, \frac{2\varepsilon}{1+\varepsilon^2}\right\} \quad \text{with} \quad M_d = \max_{x\in K}\|\nabla d\|.$$

We then have for all $t \geq 0$,

$$\|\lambda^{t+1} - \lambda^*_{\nu,\varepsilon}\|^2 \leq q_d^{t+1}\|\lambda^0 - \lambda^*_{\nu,\varepsilon}\|^2 + \frac{1-q_d^{t+1}}{1-q_d} M_d^2 \left(q_d q_p^K + 2\tau^2 M_x q_p^{K/2} \right),$$

where $q_p = 1 - \alpha\nu(2-\alpha L_f)$, $q_d = (1-\tau\varepsilon)^2 + \tau^2$, and $M_x = \max_{x,y\in K}\|x-y\|$.

Proof. In view of (2.35) and the non-expansive property of the projection, we have

$$\begin{aligned} \|\lambda^{t+1} - \lambda^*_{\nu,\varepsilon}\|^2 &\leq \left\|(1-\tau\varepsilon)(\lambda^t - \lambda^*_{\nu,\varepsilon}) - \tau\left(d(x^*_{\nu,\varepsilon}) - d(x^K(t))\right)\right\|^2 \\ &= (1-\tau\varepsilon)^2\|\lambda^t - \lambda^*_{\nu,\varepsilon}\|^2 + \tau^2 \underbrace{\|d(x^*_{\nu,\varepsilon}) - d(x^K(t))\|^2}_{\textbf{Term 1}} \\ &\quad -2\tau(1-\tau\varepsilon)\underbrace{(\lambda^t - \lambda^*_{\nu,\varepsilon})^T\left(d(x^*_{\nu,\varepsilon}) - d(x^K(t))\right)}_{\textbf{Term 2}}. \end{aligned} \tag{2.36}$$

Next, we provide bounds on terms 1 and 2. For term 1 by adding and subtracting $d(x(t))$, we obtain

$$\begin{aligned} &\|d(x^*_{\nu,\varepsilon}) - d(x^K(t))\|^2 = \|d(x^*_{\nu,\varepsilon}) - d(x(t)) + d(x(t)) - d(x^K(t))\|^2 \\ &\leq \|d(x^*_{\nu,\varepsilon}) - d(x(t))\|^2 + \|d(x(t)) - d(x^K(t))\|^2 + 2\|d(x^*_{\nu,\varepsilon}) - d(x(t))\|\|d(x(t)) - d(x^K(t))\|. \end{aligned}$$

By using the Lipschitz continuity of $d(x)$ for $x \in K$, we further have for all $t \geq 0$,

$$\begin{aligned} \|d(x^*_{\nu,\varepsilon}) - d(x^K(t))\|^2 &\leq \|d(x^*_{\nu,\varepsilon}) - d(x(t))\|^2 + \|d(x(t)) - d(x^K(t))\|^2 \\ &\quad +2M_d^2\|x^*_{\nu,\varepsilon} - x(t)\|\,\|x(t) - x^K(t)\|. \end{aligned} \tag{2.37}$$

Now, we consider term 2, for which by adding and subtracting $d(x(t))$, and by using the co-coercivity of $-d(x(\lambda))$ (see (2.33)), we obtain

$$\begin{aligned} &(\lambda^t - \lambda^*_{\nu,\varepsilon})^T\left(d(x^*_{\nu,\varepsilon}) - d(x^K(t))\right) \\ &= (\lambda^t - \lambda^*_{\nu,\varepsilon})^T\left(d(x^*_{\nu,\varepsilon}) - d(x(t))\right) + (\lambda^t - \lambda^*_{\nu,\varepsilon})^T\left(d(x(t)) - d(x^K(t))\right) \\ &\geq \frac{\nu}{M_d^2}\|d(x^*_{\nu,\varepsilon}) - d(x(t))\|^2 + (\lambda^t - \lambda^*_{\nu,\varepsilon})^T\left(d(x(t)) - d(x^K(t))\right). \end{aligned}$$

Thus, we have

$$\begin{aligned} &-2\tau(1-\tau\varepsilon)(\lambda^t - \lambda^*_{\nu,\varepsilon})^T\left(d(x^*_{\nu,\varepsilon}) - d(x^K(t))\right) \leq -2\tau(1-\tau\varepsilon)\frac{\nu}{M_d^2}\|d(x^*_{\nu,\varepsilon}) - d(x(t))\|^2 \\ &+\tau^2\|\lambda^t - \lambda^*_{\nu,\varepsilon}\|^2 + (1-\tau\varepsilon)^2\left\|d(x(t)) - d(x^K(t))\right\|^2. \end{aligned} \tag{2.38}$$

From relations (2.36), (2.37) and (2.38), by grouping the corresponding expressions accordingly, we obtain

$$\begin{aligned}\|\lambda^{t+1}-\lambda^*_{\nu,\varepsilon}\|^2 \;\le\; & \left((1-\tau\varepsilon)^2+\tau^2\right)\|\lambda^t-\lambda^*_{\nu,\varepsilon}\|^2 \\ & +\left(\tau^2-2\tau(1-\tau\varepsilon)\frac{\nu}{M_d^2}\right)\|d(x^*_{\nu,\varepsilon})-d(x(t))\|^2 \\ & +\left((1-\tau\varepsilon)^2+\tau^2\right)\|d(x(t))-d(x^K(t))\|^2 \\ & +2\tau^2M_d^2\|x^*_{\nu,\varepsilon}-x(t)\|\,\|x(t)-x^K(t)\|.\end{aligned}$$

By Lemma 7, we have $\|x^K(t)-x(t)\|\le q_p^{K/2}\|x^0(t)-x(t)\|$ with $q_p=1-\alpha\nu(2-\alpha L_f)$. By using this, the Lipschitz continuity of $d(x)$ over K, and $\|x^*_{\nu,\varepsilon}-x(t)\|\le M_x$ where $M_x=\max_{x,y\in K}\|x-y\|$, we obtain

$$\begin{aligned}\|\lambda^{t+1}-\lambda^*_{\nu,\varepsilon}\|^2 \;\le\; & \left((1-\tau\varepsilon)^2+\tau^2\right)\|\lambda^t-\lambda^*_{\nu,\varepsilon}\|^2+\left(\tau^2-2\tau(1-\tau\varepsilon)\frac{\nu}{M_d^2}\right)M_d^2M_x^2 \\ & +\left((1-\tau\varepsilon)^2+\tau^2\right)M_d^2q_p^K+2\tau^2M_d^2M_x\,q_p^{K/2}.\end{aligned}$$

By choosing τ such that

$$\tau<\min\left\{\frac{2\nu}{M_d^2+2\varepsilon\nu},\frac{2\varepsilon}{1+\varepsilon^2}\right\},$$

we ensure that $(1-\tau\varepsilon)^2+\tau^2<1$ and $\tau^2-2\tau(1-\tau\varepsilon)\frac{\nu}{M_d^2}<0$. Therefore for such a τ, by letting $q_d=(1-\tau\varepsilon)^2+\tau^2$, we have

$$\|\lambda^{t+1}-\lambda^*_{\nu,\varepsilon}\|^2 \;\le\; q_d\|\lambda^t-\lambda^*_{\nu,\varepsilon}\|^2+q_dM_d^2q_p^K+2\tau^2M_d^2M_x\,q_p^{K/2},$$

and by recursively using the preceding estimate, we obtain

$$\|\lambda^{t+1}-\lambda^*_{\nu,\varepsilon}\|^2 \;\le\; q_d^{t+1}\|\lambda^0-\lambda^*_{\nu,\varepsilon}\|^2+\frac{1-q_d^{t+1}}{1-q_d}M_d^2\left(q_dq_p^K+2\tau^2M_x\,q_p^{K/2}\right).$$

■

Note that by Proposition 5, we have $\lim_{K\to\infty}q_p^K=0$ since $q_p<1$ and, hence, the term

$$\frac{1-q_d^{t+1}}{1-q_d}M_d^2(q_dq_p^K+2\tau^2M_x\,q_p^{K/2})$$

converges to zero. This is precisely what we expect: as the Lagrangian problem is solved to a greater degree of exactness, the method approaches the exact regularized counterpart of section 2.3.1. Also, note that when K is fixed the following limiting error holds

$$\lim_{t\to\infty} \|\lambda^{t+1} - \lambda^*_{\nu,\varepsilon}\|^2 \leq \frac{1}{1-q_d} M_d^2 \left(q_d q_p^K + 2\tau^2 M_x q_p^{K/2} \right).$$

We now establish bounds on the norm $\|x^K(t) - x^*_{\nu,\varepsilon}\|$ and the constraint violation $d_j(x)$ at $x = x^K(t)$ for all j.

Proposition 6. *Under assumptions of Proposition 5, for the sequence $\{x^K(t)\}$ generated by (2.34)–(2.35) we have for all $t \geq 0$,*

$$\|x^K(t) - x^*_{\nu,\varepsilon}\| \leq q_p^{K/2} M_x + \frac{M_d}{\nu} \|\lambda^t - \lambda^*_{\nu,\varepsilon}\|,$$
$$\max\{0, d_j(x^K(t))\} \leq M_d \left(q_p^{K/2} M_x + \frac{M_d}{\nu} \|\lambda^t - \lambda^*_{\nu,\varepsilon}\| \right),$$

where q_p, M_x and M_d are as defined in Proposition 5.

Proof. Consider $\|x^K(t) - x^*_{\nu,\varepsilon}\|$. By Lemma 7 we have $\|x^K(t) - x(t)\| \leq q_p^{K/2} \|x^0(t) - x(t)\|$, while by co-coercivity of $-d(x)$, it can be seen that $\|x(t) - x^*_{\nu,\varepsilon}\| \leq \frac{M_d}{\nu} \|\lambda^t - \lambda^*_{\nu,\varepsilon}\|$. Hence,

$$\|x^K(t) - x^*_{\nu,\varepsilon}\| \leq \|x^K(t) - x(t)\| + \|x(t) - x^*_{\nu,\varepsilon}\| \leq q_p^{K/2} M_x + \frac{M_d}{\nu} \|\lambda^t - \lambda^*_{\nu,\varepsilon}\|,$$

where we also use $\|x^0(t) - x(t)\| \leq M_x$. For the constraint d_j, by convexity of d_j and using $d_j(x^*_{\nu,\varepsilon})$ we have for any $t \geq 0$,

$$\begin{aligned} d_j(x^K(t)) &\leq d_j(x^*_{\nu,\varepsilon}) + \nabla d(x^*_{\nu,\varepsilon})^T (x^K(t) - x^*_{\nu,\varepsilon}) \\ &\leq \|\nabla d(x^*_{\nu,\varepsilon})\| \, \|x^K(t) - x^*_{\nu,\varepsilon}\| \leq M_d \left(q_p^{K/2} M_x + \frac{M_d}{\nu} \|\lambda^t - \lambda^*_{\nu,\varepsilon}\| \right), \end{aligned}$$

where in the last inequality we use the preceding estimate for $\|x^K(t) - x^*_{\nu,\varepsilon}\|$. Thus, for the violation of $d_j(x)$ at $x = x^K(t)$ we have,

$$\max\{0, d_j(x^K(t))\} \leq M_d \left(q_p^{K/2} M_x + \frac{M_d}{\nu} \|\lambda^t - \lambda^*_{\nu,\varepsilon}\| \right).$$

■

One may combine the result of Proposition 6 with the estimate for $\|\lambda^t - \lambda_{\nu,\varepsilon}\|$ of Proposition 5 to bound the norm $\|x^K(t) - x(t)\|$ and the constraint violation $\max\{0, d_j(x^K(t))\}$ in terms of initial multiplier λ^0 and the optimal dual solution $\lambda^*_{\nu,\varepsilon}$.

An obvious challenge in implementing such schemes is that convergence relies on exact primal solutions. Often, there is a fixed amount of time available for obtaining primal updates, leading us to consider whether one could construct error bounds for dual schemes where an approximate primal solution is obtained through a fixed number of gradient steps.

Finally, we discuss an extension of the preceding results to the case of independently chosen regularization parameters. Analogous to Section 2.2.5, we extend the results of dual method to the case when user i selects a regularization parameter ν_i for its own Lagrangian subproblem. As in Section 2.2.5, the results follow straight-forwardly from the results developed so far in this section. We briefly discuss the modified results here for completeness.

As in Section 2.2.5, Lagrange multiplier λ belongs to set $\mathcal{D}_{\nu_{\max}}$ defined in (2.28). In this case, similar to the proof of Lemma 6, it can be seen that the function $-d(x(\lambda))$ is co-coercive in λ with constant $\frac{\nu_{\min}}{M_d^2}$. The result of Proposition 3 will require the dual steplength τ to satisfy the following relation:

$$\tau < \frac{2\nu_{\min}}{M_d^2 + 2\varepsilon\nu_{\min}}.$$

Similarly, the result of Proposition 4 will hold with $\nu_{\min}$ replacing the regularization parameter ν i.e., for τ such that $\tau < \frac{2\nu_{\min}}{M_d^2+2\varepsilon\nu_{\min}}$, we have for all $t \geq 0$,

$$\|x^t - x^*_{V,\varepsilon}\| \leq \frac{M_d}{\nu_{\min}} \|\lambda^t - \lambda^*_{V,\varepsilon}\| \qquad \text{and} \qquad \max\{0, d_j(x^t)\} \leq \frac{M_d^2}{\nu_{\min}} \|\lambda^t - \lambda^*_{V,\varepsilon}\|.$$

Finally, Lemma 7 will hold with L_f defined by $L_f = L + \nu_{\max} + M_\nu L_d$ and $q_p = 1 - \alpha\nu_{\min}(2 - \alpha L_f)$. Also, for the result of Proposition 5 to hold, the dual steplength τ should be required to satisfy

$$\tau < \min\left\{\frac{2\nu_{\min}}{M_d^2 + 2\varepsilon\nu_{\min}}, \frac{2\varepsilon}{1+\varepsilon^2}\right\}.$$

2.4 Case study

In this section, we report some experimental results for the algorithms developed in preceding sections. We use the `knitro` solver [67] on Matlab 7 to compute a solution of the problem and examine the performance of our proposed methods on a multiuser optimization problem involving a serial network with multiple links. The problem captures traffic and communication networks where users are characterized by utility/cost functions and are coupled through a congestion cost. This case manifests itself through delay arising from the link capacity constraints. In Section 2.4.1, we describe the underlying network structure and the user objectives and we present the numerical results for the primal-dual and dual methods, respectively. In each instance, an emphasis will be laid on determining the impact of the extensions, specifically independent primal and dual step-lengths and independent primal regularization (primal-dual), and inexact solutions of the Lagrangian subproblems (dual).

2.4.1 Network and user data

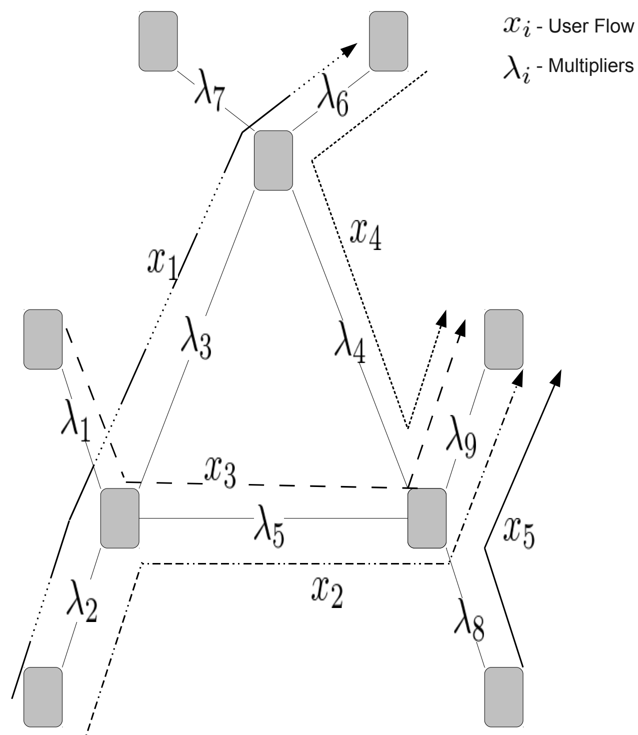


Figure 2.1: A network with 5 users and 9 links.

The network comprises of a set of N users sharing a set $\mathscr{L}$ of links (see Fig. 2.1 for an illustration). A user $i \in N$ has a cost function $f_i(x_i)$ of its traffic rate x_i given by

$$f_i(x_i) = -k_i \log(1 + x_i) \qquad \text{for } i = 1, \ldots, N. \tag{2.39}$$

Each user selects an origin-destination pair of nodes on this network and faces congestion based on the links traversed along the prescribed path connecting the selected origin-destination nodes.

We consider the congestion cost of the form:

$$c(x) = \sum_{i=1}^{N} \sum_{l \in \mathscr{L}} x_{li} \sum_{j=1}^{N} x_{lj}, \tag{2.40}$$

where, x_{lj} is the flow of user j on link l. The total cost of the network is given by

$$f(x) = \sum_{i=1}^{N} f_i(x) + c(x) = \sum_{i=1}^{N} -k_i \log(1+x_i) + \sum_{i=1}^{N} \sum_{l \in \mathscr{L}} x_{li} \sum_{j=1}^{N} x_{lj}.$$

Let A denote the adjacency matrix that specifies the set of links traversed by the traffic generated by the users. More precisely, $A_{li} = 1$ if traffic of user i goes through link l and 0 otherwise. It can be seen that $\nabla c(x) = 2A^T Ax$ and thus the Lipschitz constant of the gradient map $\nabla f(x)$ is given by $L = \sqrt{\sum_i k_i^2} + 2\|A^T A\|$. Throughout this section, we consider a network with 9 links and 5 users. Table 2.1 summarizes the traffic in the network as generated by the users and the parameters k_i of the user objective. The user traffic rates are coupled through the constraint of the

Table 2.1: Network and User Data

User(i)	Links traversed	k_i
1	L2, L3, L6	10
2	L2, L5, L9	0
3	L1, L5, L9	10
4	L6, L4, L9	10
5	L8, L9	10

form $\sum_{i=1}^{N} A_{li} x_i \leq C_l$ for all $l \in \mathscr{L}$, where C_l is the maximum aggregate traffic through link l. The constraint can be compactly written as $Ax \leq C$, where C is the link capacity vector and is given by $C = (10, 15, 20, 10, 15, 20, 20, 15, 25)$.

Regularized Primal-Dual Method. Figure 2.2 shows the number of iterations required to attain a desired error level for $\|z^k - z^*_{\nu,\varepsilon}\|$ with $\{z^k\}$ generated by primal-dual algorithm (2.26) for different values of the step-size ratio $\beta = \alpha/\tau$ between the primal step-size α and dual stepsize τ. Note that in this case each user has the same step-size and the regularization parameter. Relations in Lemma 4 are used to obtain the theoretical range for the ratio parameter β and the corresponding step-lengths. The regularization parameters ν and ε were both set at 0.1, such that $\mu = \min\{\nu, \varepsilon\} = 0.1$ and the algorithm was terminated when $\|z^k - z^*_{\nu,\varepsilon}\| \leq 10^{-3}$. It can be observed that the number of iterations required for convergence decreases as the step-size ratio of approaches the value 1.

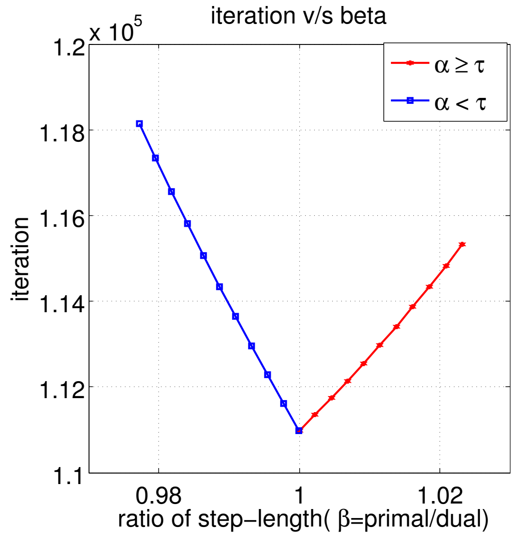


Figure 2.2: Performance of Primal-Dual Method for independent step-sizes in primal and dual space.

Figure 2.3 illustrates the performance of the primal-dual algorithm in terms of the number of iterations required to attain $\|z^k - z^*_{V,\varepsilon}\| < 10^{-3}$ as the steplength deviation in primal space $\alpha_{\max} - \alpha_{\min}$ increases. All users employ the same regularization parameter $v_i = v = 0.1$ and the dual regularization parameter ε is chosen to be 0.1. The plot demonstrates that, as the deviation between users' step-sizes increases, the number of iteration for a desired accuracy also increases.

Next, we let each user choose its own regularization parameter v_i with uniform distribution over interval $(v_{\min}, 0.1)$ for a given $v_{\min} \leq 0.1$. Figure 2.4 shows the performance of the primal-dual algorithm in terms of the number of iterations required to attain the error $\|z^k - z^*_{V,\varepsilon}\| < 10^{-3}$ as $v_{\min}$ is varied from 0.01 to 0.1. The dual steplength was set at $\tau = 1.9\mu/L^2_\Phi$, where $\mu = \min\{v_{\min}, \varepsilon\}$ with $\varepsilon = 0.1$. The primal stepsizes that users employ are the same across the users and are given by $\alpha_i = \alpha = \beta\tau$, where β is as given in Lemma 4. As expected, the number of iterations increases when $v_{\min}$ decreases.

Regularized Dual Method. Figure 2.5(a) compares dual iterations required to reach an accuracy level of $\|\lambda^k - \lambda^*_{v,\varepsilon}\| \leq 10^{-6}$ for each K where $\{\lambda^k\}$ is generated using dual method (2.35) and K is the number of iterations in the primal space for each λ^k. The regularization parameter ε is varied from 0.0005 to 0.0025, while v is fixed at 0.001. The primal step-size is set at $\alpha = 0.25/L_f$ and the dual step-size is taken as $\tau = 0.75v/M_d^2$ (see Section 2.3). Faster dual convergence was observed as K was increased for all ranges of parameters tested. For the case when $v = 0.001$ and $\varepsilon = 0.001$, Figure 2.5(b) shows the dependency of total number of iterations required (primal $\times$ dual) for $\|\lambda^k - \lambda^*_{v,\varepsilon}\| \leq 10^{-6}$ as the number K of primal iterations is varied. It can be observed

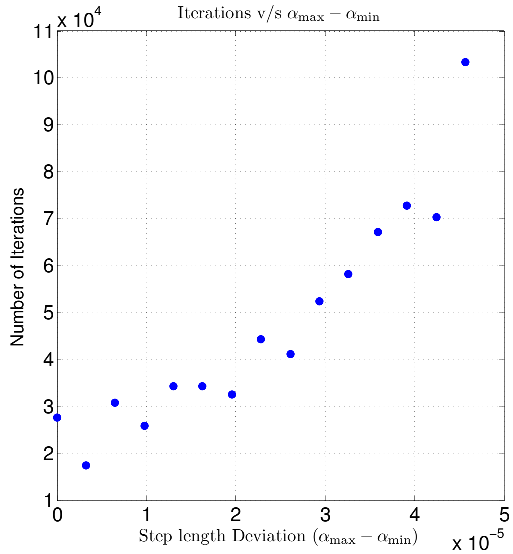


Figure 2.3: Performance of Primal-Dual method for deviation in user step-size.

that beyond a threshold level for K, the total number of iterations starts increasing. In effect, the extra effort in obtaining increasingly exact solutions to the Lagrangian subproblem is not met with faster convergence in the dual space.

2.5 Summary and conclusions

This chapter focuses on a class of multiuser optimization problems in which user interactions are seen in the user objectives (through congestion or delay functions) and in the coupling constraints (as arising from shared resources). Traditional algorithms rely on a high degree of separability and cannot be directly employed. They also rely on coordination in terms of uniform or equal step-sizes across users. The coordination requirements have been weakened to various degrees in this chapter, which considers primal-dual and dual gradient algorithms, derived from the fixed-point formulations of the regularized problem. These schemes are analyzed in an effort to make rigorous statements regarding convergence behavior as well as provide error bounds in regularized settings that limited coordination across step-length choices and inexact solutions. The main contributions are summarized next:

(1) Regularized primal-dual method: Under suitable convexity assumptions, we consider a regularized primal-dual projection scheme and provide error bounds for the regularized solution and optimal function value with respect to their optimal counterparts. In addition, we also

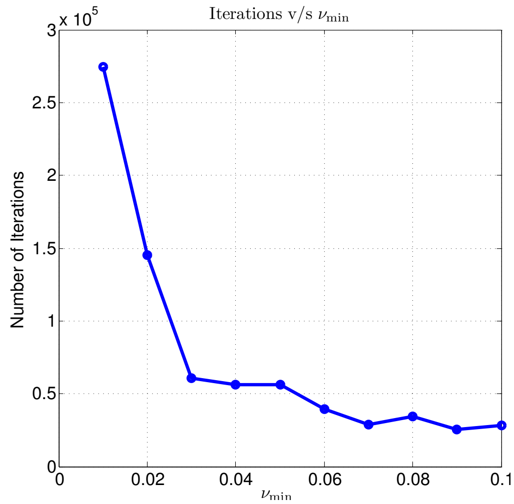


Figure 2.4: Performance of Primal-Dual method as user minimum regularization parameter $\nu_{\min}$ varies.

obtain a bound on the infeasibility for the regularized solution. We also show that, under some conditions, the method can be extended to allow not only for independent selection of primal and dual stepsizes as well as independently chosen stepsizes by every user but also when users choose their regularization parameter independently.

(2) Regularized dual method: In contrast with (1), applying dual schemes would require an optimal primal solution for every dual step. We show the contractive nature of a regularized dual scheme reliant on exact primal solutions. Furthermore, we develop asymptotic error bounds where for each dual iteration, the primal method for solving the Lagrangian subproblem terminates after a fixed number of steps. We also provide error bounds for the obtained solution and Lagrange multiplier as well as an upper bound on the infeasibility. Finally, we extend these results to the case when each user independently chooses its regularization parameter.

It is of future interest to consider the algorithms proposed in [63, 64] as applied to multiuser problem, whereby the users are allowed to implement step-sizes within a prescribed range of values. For this, at first, we would have to develop the error bounds for the algorithms in [63, 64] for the case when different users employ different stepsizes.

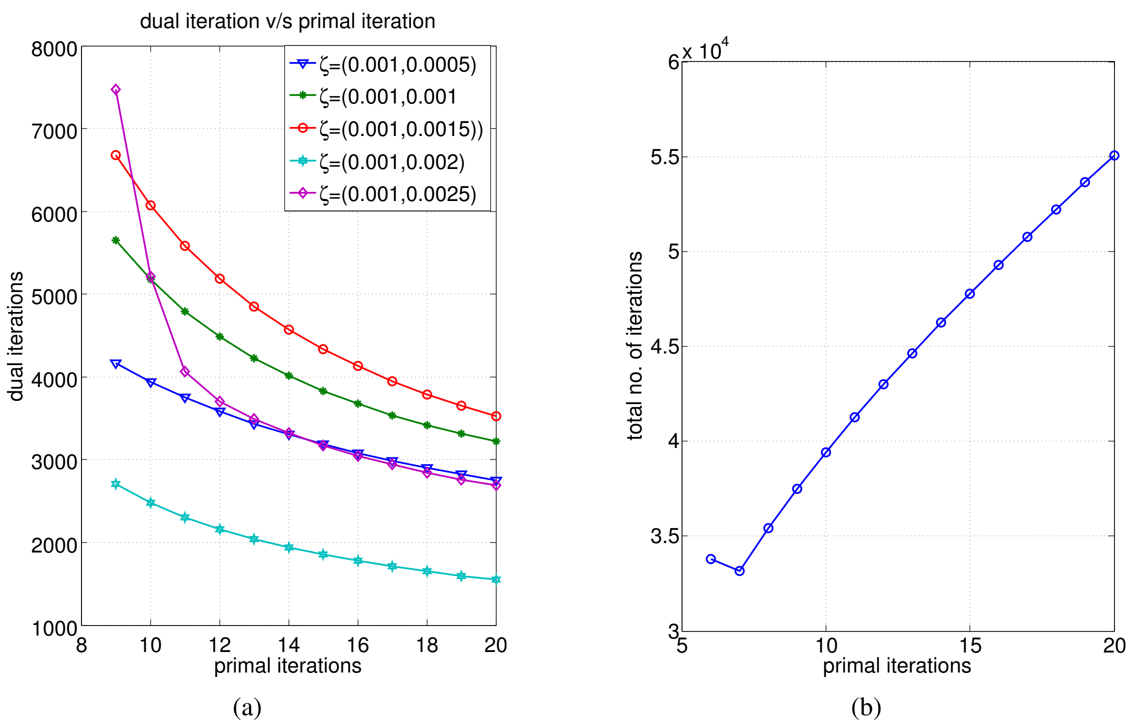


Figure 2.5: Inexact Dual Method: (a) Comparison of dual iterations for a fixed number K of primal iterations; (b) Dependency of the total number of primal and dual iterations as the number K of primal iterations varies.

Chapter 3

Regularized Iterative Stochastic Approximation Methods for Cartesian Stochastic Variational Inequalities

In Chapter 2 we presented two distributed algorithms for the generalized multiuser optimization problem. The interest of this chapter is to study distributed algorithms for stochastic variational inequalities with monotone maps. An important question clouding our interest is whether one can construct distributed algorithms abiding the suitable criteria we set in Chapter 1. In particular, the single level iteration one is especially desirable for reasons which will become apparent soon.

Within the framework of stochastic variational inequalities, stochastic approximation methods have been recently employed in [36]. The typical stochastic approximation procedure, first introduced by Robbins and Monro [68], works toward finding an extremum of a function $h(x)$ using the following iterative method:

$$x^{k+1} = x^k + a_k(\nabla h(x^k) + M^{k+1}),$$

where $a_k > 0$ is a stepsize and M^{k+1} is a martingale difference term. Under reasonable assumptions on the stochastic errors M^k, stochastic approximation methods ensure that $\{x^k\}$ converges almost surely to an optimal solution of the problem. Jiang and Xu [36] consider the use of stochastic approximation for strongly monotone and Lipschitz continuous maps in the realm of stochastic variational inequalities, rather than optimization problems. The use of stochastic approximation methods has a long tradition in stochastic optimization for both differentiable [68] and nondifferentiable problems [69], while a subset of more recent efforts include [70, 71, 72]. In contrast, our work builds on different deterministic algorithms, including Tikhonov regularization and proximal-point methods, and combines them with the stochastic approximation approach. At the same time, our convergence results require less stringent monotonicity assumptions on the map.

It should be remarked that Tikhonov-based regularization and Proximal point methods [39] methods have a long history in the solution of ill-posed optimization and variational problems [73, 35] (see Nesterov [64] and Nemirovski [63] for recent work o on proximal point and error bounds). Such methods, in general, require a solution of a regularized (well-posed) problem and an iterative process is often needed to obtain the solution. In both Tikhonov regularization and proximal point methods, two nested iterative procedures are involved, where the outer procedure updates a parameter after an increasingly accurate solution of an inner subproblem is available. In networked stochastic regimes, this is challenging for two reasons: (1) First, obtaining increasingly accurate solutions of stochastic variational problems via simulation techniques requires significant effort; and (2) Second, assessing solution quality formally requires validation analysis to be conducted over the network, a somewhat challenging task. We obviate this challenge by considering algorithms where the parameter is updated after every iteration and, thus, the update of the steplength and the regularization parameter is synchronized. Such an approach is popularly referred to as *iterative regularization.* While there have been efforts to use such techniques for optimization problems (cf. [73]), there has been noticeably less in the realm of variational inequalities, barring [74] and more recently [6, 20, 75]. However, much of this work has been restricted to deterministic regimes. In a stochastic regime, Borkar [76] examined two timescale stochastic approximation methods. The present work emphasizes iterative regularization for stochastic variational inequalities with monotone maps. We present and analyze two stochastic iterative regularization methods. Notably, distributed iterative regularization techniques, the deterministic counterparts of these methods, have been examined in the context of monotone Nash games [20, 75].

1. *Stochastic iterative Tikhonov regularization method*:

   We consider a stochastic iterative Tikhonov regularization method for monotone stochastic variational inequalities where the steplength and regularization parameter are updated at every iteration. Partially coordinated generalizations are presented where users independently select stepsize and regularization sequences. Under some restrictions on the deviations across the users' choices, we establish convergence properties of the method in almost sure sense.

2. *Stochastic iterative proximal-point method*:

   An alternative to the stochastic iterative Tikhonov method lies in a stochastic iterative proximal-point method where the steplength and prox-parameter are updated at *every iteration*. As in the case of iterative Tikhonov method, we present convergence results for a partially coordinated implementation. Our convergence results are established for strictly monotone mappings.

The remainder of the chapter is organized as follows. In Section 3.1 we describe the basic framework of stochastic approximation and the supporting convergence results. In Section 3.2, we propose and analyze a stochastic iterative Tikhonov regularization method. Analogous results for a stochastic iterative proximal point method are provided in Section 3.3. Section 3.4, the performance of our methods and their relative sensitivity to parameters is examined in the context of networked rate allocation game. We conclude the chapter with some remarks in Section 3.5.

3.1 Stochastic approximation approach

Consider a variational inequality problem, denoted by $\mathrm{VI}(K,F)$, where the mapping $F : K \to \mathbb{R}^n$ and the set K are given by

$$F(x) \triangleq \begin{pmatrix} \nabla_{x_1}\mathbb{E}[f_1(x,\xi_1)] \\ \vdots \\ \nabla_{x_N}\mathbb{E}[f_N(x,\xi_N)] \end{pmatrix}, \qquad K = \prod_{i=1}^{N} K_i. \tag{3.1}$$

Note that the set K is closed and convex set in $\mathbb{R}^n$, whenever the sets K_i are closed and convex. Recall that $\mathrm{VI}(K,F)$ requires determining a vector $x^* \in K$ such that

$$(x-x^*)^T F(x^*) \geq 0 \quad \text{for all } x \in K. \tag{3.2}$$

When the expectation is over a general measure space, analytical forms of the expectation are often hard to obtain. In such settings, stochastic approximation methods assume relevance.

Towards this end, consider the Robbins-Monro stochastic approximation method for solving the stochastic variational inequality $\mathrm{VI}(K,F)$ in (3.1)–(3.2), given by

$$x^{k+1} = \Pi_K[x^k - \alpha_k(F(x^k)+w^k)] \quad \text{for } k \geq 0, \tag{3.3}$$

where $x^0 \in K$ is a random initial vector that is independent of the random variables ξ_i for all i and such that $\mathbb{E}[\|x^0\|^2]$ is finite. The vector $F(x^k)$ is the true value of $F(x)$ at $x = x^k$, $\alpha_k > 0$ is the stepsize, while the vector w^k is the stochastic error given by

$$w^k = -F(x^k) + \tilde{F}(x^k,\xi^k),$$

with

$$\tilde{F}(x^k,\xi^k) \triangleq \begin{pmatrix} \nabla_{x_1} f_1(x^k,\xi_1^k) \\ \vdots \\ \nabla_{x_N} f_N(x^k,\xi_N^k) \end{pmatrix} \qquad \text{and} \qquad \xi^k \triangleq \begin{pmatrix} \xi_1^k \\ \vdots \\ \xi_N^k \end{pmatrix}.$$

The projection method (3.3) is shown to be convergent when the mapping F is strongly monotone and Lipschitz continuous in [36]. In this chapter, we examine how the use of regularization methods can alleviate the strong monotonicity requirement while maintaining single timescale and distributed structure of the algorithm.

In our analysis we use some well-known results on supermartingale convergence, which we provide for convenience. The following result is from [73], Lemma 10, page 49.

Lemma 8. *Let V_k be a sequence of non-negative random variables adapted to σ-algebra $\mathcal{F}_k$ and such that almost surely*

$$\mathbb{E}[V_{k+1} \mid \mathcal{F}_k] \leq (1-u_k)V_k + \beta_k \qquad \text{for all } k \geq 0,$$

where $0 \leq u_k \leq 1$, $\beta_k \geq 0$, and

$$\sum_{k=0}^{\infty} u_k = \infty, \quad \sum_{k=0}^{\infty} \beta_k < \infty, \quad \lim_{k\to\infty} \frac{\beta_k}{u_k} \to 0.$$

Then, $V_k \to 0$ a.s.

The result of the following lemma can be found in [73], Lemma 11, page 50.

Lemma 9. *Let V_k, u_k, β_k and γ_k be non-negative random variables adapted to σ-algebra $\mathcal{F}_k$. If almost surely $\sum_{k=0}^{\infty} u_k < \infty$, $\sum_{k=0}^{\infty} \beta_k < \infty$, and*

$$\mathbb{E}[V_{k+1} \mid \mathcal{F}_k] \leq (1+u_k)V_k - \gamma_k + \beta_k \quad \text{for all } k \geq 0,$$

then almost surely $\{V_k\}$ is convergent and $\sum_{k=0}^{\infty} \gamma_k < \infty$.

3.2 Stochastic iterative Tikhonov methods

In this section, we propose and analyze a stochastic iterative Tikhonov algorithm for solving the variational inequality VI(K,F) in (3.1)–(3.2). We consider the case when the mapping F is monotone over the set K, i.e., F is such that

$$(F(x)-F(y))^T(x-y) \geq 0 \qquad \text{for all } x, y \in K.$$

As we have already seen in previous chapter, a possible approach for addressing monotone variational problems is through a Tikhonov regularization method [35, Ch. 12] (cf. [38, 37]). In the context of variational inequalities, this avenue typically requires solving a sequence of perturbed variational inequality problems. In particular, the kth problem in the sequence is the variational inequality $\mathrm{VI}(K, F+\varepsilon_k \mathbf{I})$, where the mapping $F+\varepsilon_k \mathbf{I}$ is a perturbation of the original mapping F given by a positive scalar ε_k. In this way, each of the variational inequality problems $\mathrm{VI}(K, F+\varepsilon_k \mathbf{I})$ is strongly monotone and, hence, it has a unique solution denoted by $y^k \in K$ (see Theorem 2.3.3 in [35]). Under suitable conditions, it can be seen that the Tikhonov sequence $\{y^k\}$ satisfies $\lim_{k\to\infty} y^k = x^*$, where x^* is the least norm solution of $\mathrm{VI}(K,F)$ (see Theorem 12.2.3 in [35]). Thus, to reach a solution of $\mathrm{VI}(K,F)$, one has to solve a sequence $\mathrm{VI}(K, F+\varepsilon_k \mathbf{I})$ of variational inequality problems along some diminishing sequence $\{\varepsilon_k\}$. However, in the current setting, determining a solution y^k for a regularized problem $\mathrm{VI}(K, F+\varepsilon_k \mathbf{I})$ requires either the exact or approximate solution of a strongly monotone stochastic variational inequality (see Section 12.2 in [35]).

In deterministic regimes, a solution to the regularized Tikhonov subproblem may be obtained in a distributed fashion via a projection method. However, in stochastic regimes, this is a more challenging proposition. While an almost sure convergence theory for a stochastic approximation method for strongly monotone variational problems is provided in [36], termination criteria are generally much harder to provide. As a consequence, one often provides confidence intervals in practice by generating a fixed number of sample paths. Furthermore, the convergence theory of Tikhonov-based methods necessitates that the solutions to the subproblem be computed with increasing accuracy. Implementing such algorithms in a stochastic regime is significantly harder since simulation-based methods are employed to obtain confidence intervals for each regularized problem, which require that these intervals get increasingly tighter. In the numerical results, we revisit this challenge by considering the behavior of the standard regularization methods (operating in two nested iterative updates).

Accordingly, we consider an alternative iterative method that avoids solving a sequence of variational inequality problems; instead, each user takes a *single* projection step associated with his regularized problem. By imposing appropriate assumptions on the steplength and regularization sequences, we may recover convergence to the least-norm Nash equilibrium. To summarize, our intent lies in developing algorithms that are characterized by (a) a single iterative process; (b) a distributed architecture that can accommodate computation of equilibria; and (c) the ability to accommodate uncertainty via expected-value objectives. An important characteristic of our distributed methods is that users can autonomously choose their parameters within from a provided set. Thus, we consider a situation where users choose their individual stepsize and regularization sequence, leading to the following coupled user-specific Tikhonov updates:

$$x_i^{k+1} = \Pi_{K_i}[x_i^k - \alpha_{k,i}(F_i(x^k) + \varepsilon_{k,i} x_i^k) + w_i^k]. \tag{PITR}$$

Note that $x_i^0 \in K_i$ is a random initial point with a finite expectation $\mathbb{E}[\|x_i^0\|^2]$ and $F_i(x^k)$ denotes the ith component of the mapping $F(x)$ evaluated at x^k. The vector w_i^k is a stochastic error for user i in evaluating $F_i(x^k)$, while $\alpha_{k,i}$ is the stepsize and $\varepsilon_{k,i}$ is the regularization parameter chosen by user i at the kth iteration. The iterate updates can be compactly written as

$$x^{k+1} = \Pi_K[x^k - D(\alpha_k)(F(x^k) + D(\varepsilon_k)x^k + w^k)], \tag{3.4}$$

where $F = (F_1, \ldots, F_N)$, $K = \prod_{i=1}^N K_i$ and $w^k = (w_1^k, \ldots, w_N^k)$, while $D(\alpha_k)$ and $D(\varepsilon_k)$ denote the diagonal matrices with diagonal entries $\alpha_{k,i}$ and $\varepsilon_{k,i}$, respectively. The method specified by (PITR) (and its compact version (3.4)) is referred to as a **p**artially coordinated **i**terative **T**ikhonov **r**egularization (PITR) method. It is motivated by the need to allow users to choose their steplength and regularization parameter, namely $\alpha_{k,i}$ and $\varepsilon_{k,i}$, while abiding by a coordination requirement.

Typically, an iterative Tikhonov method is studied by at first analyzing the behavior of the Tikhonov sequence $\{y^k\}$, where each y^k is a (unique) solution to $\mathrm{VI}(K, F + \bar{\varepsilon}_k \mathbf{I})$ and the sequence $\{y^k\}$ is obtained as the parameter $\bar{\varepsilon}_k \geq 0$ is let to go to zero. Under certain conditions the Tikhonov sequence $\{y^k\}$ converges to the smallest norm solution of $\mathrm{VI}(K, F)$. Then, the sequence of iterates $\{x^k\}$ is related to the Tikhonov sequence to assert the convergence of the iterates x^k.

We adopt the same approach. However, we cannot directly use the existing results for Tikhonov sequence $\{y^k\}$ such as, for example, those given in Chapter 12.2 of [35]. In particular, arising from user-specific Tikhonov regularization parameters $\varepsilon_{k,i}$ in (3.4), our variational inequalities have the form[1] $\mathrm{VI}(K, F + D(\varepsilon_k))$ instead of $\mathrm{VI}(K, F + \bar{\varepsilon}_k \mathbf{I})$ (which would be obtained if all the users choose the same regularization parameter $\varepsilon_{k,i} = \bar{\varepsilon}_k$). In the next two subsections, we develop a necessary result for Tikhonov sequence and investigate the convergence of the method.

3.2.1 Tikhonov sequence

Here, we analyze the behavior of our Tikhonov sequence $\{y^k\}$ as each user lets its regularization parameter $\varepsilon_{k,i}$ go to zero with $y^k \in \mathrm{SOL}(K, F + D(\varepsilon_k))$. Recall that $D(\varepsilon_k)$ is the diagonal matrix with diagonal entries $\varepsilon_{k,i} > 0$ and note that each $\mathrm{VI}(K, F + D(\varepsilon_k))$ is strongly monotone. Thus, the sequence $\{y^k\}$ is uniquely determined by the choice of user sequences $\{\varepsilon_{k,i}\}$, $i = 1, \ldots, N$. For the

[1] Note the slight abuse of notation; $D(\varepsilon_k)$ also denotes the mapping $D(\varepsilon_k)\mathbf{I}$ whenever it is used as $\mathrm{VI}(K, F + D(\varepsilon_k))$.

sequence $\{y^k\}$, we have the following result.

Lemma 10. *Let the set $K \subseteq \mathbb{R}^n$ be closed and convex, and let the map $F : K \to \mathbb{R}^n$ be continuous and monotone over K. Assume that $SOL(K,F)$ is nonempty. Let the sequences $\{\varepsilon_{k,i}\}$, $i = 1,\dots,N$, be monotonically decreasing to zero and such that $\limsup_{k\to\infty} \frac{\varepsilon_{k,\max}}{\varepsilon_{k,\min}} < \infty$, where $\varepsilon_{k.max} = \max_i \varepsilon_{k,i}$ and $\varepsilon_{k,\min} = \min_i \varepsilon_{k,i}$. Then, for the Tikhonov sequence $\{y^k\}$ we have*

(a) $\{y^k\}$ is bounded and every accumulation point of $\{y^k\}$ is a solution of $VI(K,F)$;

(b) The following inequality holds

$$\|y^k - y^{k-1}\| \le M_y \frac{\varepsilon_{k-1,\max} - \varepsilon_{k,\min}}{\varepsilon_{k,\min}} \|y^{k-1}\| \qquad \text{for all } k \ge 1,$$

where M_y is a norm bound on the Tikhonov sequence, i.e., $\|y^k\| \le M_y$ for all $k \ge 0$;

(c) If $\limsup_{k\to\infty} \frac{\varepsilon_{k,\max}}{\varepsilon_{k,\min}} \le 1$, then $\{y^k\}$ converges to the smallest norm solution of $VI(K,F)$.

Proof. (a) Since $\text{SOL}(K,F) \neq \emptyset$, by letting x^* be any solution of $\text{VI}(K,F)$ we have

$$(x - x^*)^T F(x^*) \ge 0 \qquad \text{for all } x \in K. \tag{3.5}$$

Since $y^k \in K$ solves $\text{VI}(K, F + D(\varepsilon_k))$ for each $k \ge 0$ we have

$$(y - y^k)^T (F(y^k) + D(\varepsilon_k) y^k) \ge 0 \qquad \text{for all } y \in K \text{ and } k \ge 0. \tag{3.6}$$

By letting $x = y^k$ in Eq. (3.5) and $y = x^*$ in Eq. (3.6), we obtain for all $k \ge 0$,

$$(y^k - x^*)^T F(x^*) \ge 0 \quad \text{and} \quad (x^* - y^k)^T (F(y^k) + D(\varepsilon_k) y^k) \ge 0.$$

By the monotonicity of F we have $(y^k - x^*)^T (F(x^*) - F(y^k)) \le 0$, implying that

$$(x^* - y^k)^T D(\varepsilon_k) y^k \ge 0.$$

By rearranging the terms in above expression we have

$$(x^*)^T D(\varepsilon_k) y^k \ge (y^k)^T D(\varepsilon_k) y^k \ge \varepsilon_{k,\min} \|y^k\|^2,$$

where $\varepsilon_{k,\min} = \min_{1\le i\le N} \varepsilon_{k,i}$. By using the Cauchy-Schwartz inequality, we see that

$$\varepsilon_{k,\max} \|x^*\| \, \|y^k\| \ge (x^*)^T D(\varepsilon_k) y^k,$$

where $\varepsilon_{k,\max} = \max_{1\le i\le N} \varepsilon_{k,i}$. Combining the preceding two inequalities, we obtain

$$\|y^k\| \le \frac{\varepsilon_{k,\max}}{\varepsilon_{k,\min}}\|x^*\|. \tag{3.7}$$

Let $\limsup_{k\to\infty} \frac{\varepsilon_{k,\max}}{\varepsilon_{k,\min}} = c$. Since c is finite (by our assumption), it follows that the sequence $\{y^k\}$ is bounded. By choosing any accumulation point $\tilde{y}$ of $\{y^k\}$ and letting $k \to \infty$ in Eq. (3.6) over a corresponding convergent subsequence of $\{y^k\}$, in view of continuity of F and $\varepsilon_{k,i} \to 0$ as $k \to \infty$, we conclude that

$$(y-\tilde{y})^T F(\tilde{y}) \ge 0 \qquad \text{for all } y \in K.$$

Thus, every accumulation point $\tilde{y}$ of $\{y^k\}$ is a solution to $\mathrm{VI}(K,F)$.
(b) Now, we establish the inequality satisfied by the Tikhonov sequence $\{y^k\}$. Since y^k solves $\mathrm{VI}(K, F+D(\varepsilon_k))$ for each $k\ge 0$, we have for $k \ge 1$,

$$(y^{k-1}-y^k)^T(F(y^k)+D(\varepsilon_k)y^k) \ge 0 \ \text{ and } \ (y^k-y^{k-1})^T(F(y^{k-1})+D(\varepsilon_{k-1})y^{k-1}) \ge 0.$$

By adding the preceding relations, we obtain

$$(y^{k-1}-y^k)^T(F(y^k)-F(y^{k-1}))+(y^{k-1}-y^k)^T(D(\varepsilon_k)y^k-D(\varepsilon_{k-1})y^{k-1}) \ge 0.$$

By the monotonicity of the mapping F, it follows

$$(y^{k-1}-y^k)^T(D(\varepsilon_k)y^k-D(\varepsilon_{k-1})y^{k-1}) \ge 0,$$

and thus

$$(y^{k-1}-y^k)^T(D(\varepsilon_k)y^k-D(\varepsilon_k)y^{k-1}+D(\varepsilon_k)y^{k-1}-D(\varepsilon_{k-1})y^{k-1}) \ge 0.$$

By rearranging the terms in the above expression, we obtain

$$(y^{k-1}-y^k)^T(D(\varepsilon_k)-D(\varepsilon_{k-1}))y^{k-1} \ge (y^{k-1}-y^k)^T D(\varepsilon_k)(y^{k-1}-y^k) \ge \varepsilon_{k,\min}\|y^k-y^{k-1}\|^2.$$

In the view of the Cauchy-Schwartz inequality, the left hand side is bounded from above as

$$\begin{aligned}(y^{k-1}-y^k)^T(D(\varepsilon_k)-D(\varepsilon_{k-1}))y^{k-1} &\le \|y^{k-1}-y^k\|\,\left\|(D(\varepsilon_k)-D(\varepsilon_{k-1}))\,y^{k-1}\right\| \\ &\le (\varepsilon_{k-1,\max}-\varepsilon_{k,\min})\,\|y^{k-1}-y^k\|\|y^{k-1}\|,\end{aligned}$$

where we use the monotonically decreasing property of the regularization sequences $\{\varepsilon_{k,i}\}$ $i = 1,\ldots,N$, to bound the norm $\|D(\varepsilon_k) - D(\varepsilon_{k-1})\|$. Combining the preceding relations we obtain

$$\|y^k - y^{k-1}\| \leq \|y^{k-1}\| \frac{\varepsilon_{k-1,\max} - \varepsilon_{k,\min}}{\varepsilon_{k,\min}}. \tag{3.8}$$

From the part (a) we have that the Tikhonov sequence is bounded. Let $M_y > 0$ be such that $\|y^k\| \leq M_y$ for all k. Then, from relation (3.8) we obtain

$$\|y^k - y^{k-1}\| \leq M_y \frac{\varepsilon_{k-1,\max} - \varepsilon_{k,\min}}{\varepsilon_{k,\min}} \qquad \text{for all } k \geq 1.$$

(c) Suppose now $\limsup_{k\to\infty} \frac{\varepsilon_{k,\max}}{\varepsilon_{k,\min}} \leq 1$. Then, by part (a), the sequence $\{y^k\}$ is bounded. Furthermore, in view of relation (3.7) (where the solution x^* is arbitrary), it follows that every accumulation point $\tilde{y}$ of $\{y^k\}$ satisfies

$$\|\tilde{y}\| \leq \limsup_{k\to\infty} \frac{\varepsilon_{k,\max}}{\varepsilon_{k,\min}} \|x^*\| \leq \|x^*\| \qquad \text{for all } x^* \in \text{SOL}(K,F). \tag{3.9}$$

Therefore, every accumulation point $\tilde{y}$ of $\{y^k\}$ is bounded in norm from above by the norm of any solution to $\text{VI}(K,F)$. Since K is closed and convex and $F : K \to \mathbb{R}^n$ is continuous and monotone over K, the solution set $\text{SOL}(K,F)$ is closed and convex [35, Vol. 1, Theorem 2.3.5, pg. 158]. By the strong convexity of Euclidean norm $\|x\|^2$, the smallest norm solution $x^* \in \text{SOL}(K,F)$ must exist and it is unique. In view of relation (3.9), it follows the norm of each accumulation point $\tilde{y}$ of $\{y^k\}$ is bounded from above by the least-norm solution of $\text{VI}(K,F)$. Also, by part (a), we have $\tilde{y} \in \text{SOL}(K,F)$ for every accumulation point $\tilde{y}$, thus implying that the sequence $\{y^k\}$ must converge to the smallest norm solution of $\text{VI}(K,F)$. ■

Lemma 10 plays a key role in the convergence analysis of the stochastic iterative Tikhonov method (PITR). Aside from this, Lemma 10 may be of its own interest as it extends the existing results for Tikhonov regularization to the case when the regularization mapping is a time varying diagonal matrix as opposed to being the identity mapping as in the standard literature ([77, 35]).

3.2.2 Almost sure convergence of stochastic iterative Tikhonov method

We now focus on the method in (PITR). We introduce some notation and state assumptions on the stochastic errors w^k that are standard in stochastic approximation methods. Specifically, throughout this section and the remainder of the chapter, we use $\mathcal{F}_k$ to denote the σ-field generated by the initial point x^0 and errors w^ℓ for $\ell = 0, 1, \ldots, k$, i.e., $\mathcal{F}_0 = \{x^0\}$ and

$$\mathcal{F}_k = \{x^0,\ (w^\ell,\ \ell = 0, 1, \ldots, k-1)\} \qquad \text{for } k \geq 1.$$

Now, we specify our assumptions for $\mathrm{VI}(K, F)$ in (3.1)–(3.2) and the stochastic errors w^k.

Assumption 5. *Let the following hold:*

(a) The sets $K_i \subseteq \mathbb{R}^{n_i}$ are closed and convex;

(b) The mapping $F : K \to \mathbb{R}^n$ is monotone and Lipschitz continuous over K with a constant L;

(c) The stochastic error is such that $\mathbb{E}[w^k \mid \mathcal{F}_k] = 0$ for all $k \geq 0$ almost surely.

Expectedly, convergence of the method (PITR) does rely on some coordination across steplengths and the regularization parameters. Specifically, we impose the following conditions.

Assumption 6. *Let $\alpha_{k,\max} = \max_{1\leq i\leq N}\{\alpha_{k,i}\}$, $\alpha_{k,\min} = \min_{1\leq i\leq N}\{\alpha_{k,i}\}$, $\varepsilon_{k,\max} = \max_{1\leq i\leq N}\{\varepsilon_{k,i}\}$, $\varepsilon_{k,\min} = \min_{1\leq i\leq N}\{\varepsilon_{k,i}\}$. Let $\{\varepsilon_{k,i}\}$ be a monotonically decreasing sequence for each i. Furthermore, with L being the Lipschitz constant of mapping F, let the following hold:*

(a) $\lim_{k\to\infty} \frac{\alpha_{k,\max}}{\alpha_{k,\min}} \frac{\alpha_{k,\max}}{\varepsilon_{k,\min}} (\varepsilon_{k,\max} + L)^2 = 0$ *and* $\lim_{k\to\infty} \frac{\alpha_{k,\max} - \alpha_{k,\min}}{\alpha_{k,\min}\varepsilon_{k,\min}} = 0$*;*

(b) $\lim_{k\to\infty} \alpha_{k,\min}\varepsilon_{k,\min} = 0$ *and* $\lim_{k\to\infty} \varepsilon_{k,i} = 0$ *for all i;*

(c) $\sum_{k=0}^{\infty} \alpha_{k,\min}\varepsilon_{k,\min} = \infty$*;*

(d) $\sum_{k=1}^{\infty} \frac{(\varepsilon_{k-1,\max} - \varepsilon_{k,\min})^2}{\varepsilon_{k,\min}^2} \left(1 + \frac{1}{\alpha_{k,\min}\varepsilon_{k,\min}}\right) < \infty$*;*

(e) $\lim_{k\to\infty} \frac{(\varepsilon_{k-1,\max} - \varepsilon_{k,\min})^2}{\varepsilon_{k,\min}^3 \alpha_{k,\min}} \left(1 + \frac{1}{\alpha_{k,\min}\varepsilon_{k,\min}}\right) = 0$*;*

(f) $\lim_{k\to\infty} \frac{\alpha_{k,\max}}{\varepsilon_{k,\min}} \mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k] = 0$ *and* $\sum_{k=0}^{\infty} \alpha_{k,\max}^2 \mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k] < \infty$ *a.s.*

When all the stepsizes $\alpha_{k,i}$ and the regularization parameters $\varepsilon_{k,i}$ across the users are the same, the conditions in Assumption 6 are a combination of the conditions typically assumed for deterministic Tikhonov algorithms and the stepsize conditions imposed in stochastic approximation methods. Later in forthcoming Lemma 11, we demonstrate that Assumption 6 can be satisfied by a simple choice of steplength and regularization sequences of the form $(k + \eta_i)^{-a}$ and $(k + \zeta_i)^{-b}$.

In the following proposition, using Assumption 6, we prove that the random sequence $\{x^k\}$ of the method (PITR) and the Tikhonov sequence $\{y^k\}$ associated with the problems $\mathrm{VI}(K, F + D(\varepsilon_k))$,

$k \geq 0$, have the same accumulation points *a.s.* Assumption 6 basically provides the conditions on the sequences $\{\varepsilon_{k,i}\}$ and $\{\alpha_{k,i}\}$ ensuring that the sequence $\{\|x^k - y^{k-1}\|^2\}$ is a convergent supermartingale.

Proposition 7. *Let Assumptions 5 and 6 hold. Also, assume that $SOL(K,F)$ is nonempty. Let the sequence $\{x^k\}$ be generated by stochastic iterative Tikhonov algorithm* (PITR). *Then, we have*

$$\lim_{k\to\infty} \|x^k - y^{k-1}\| = 0 \qquad a.s.$$

Proof. By using the relation $y_i^k = \Pi_{K_i}[y_i^k - \alpha_{k,i}(F_i(y^k) + \varepsilon_{k,i} y_i^k)]$ and the non-expansive property of the projection operator, we have

$$\begin{aligned}\|x_i^{k+1} - y_i^k\|^2 &= \|\Pi_{K_i}[x_i^k - \alpha_{k,i}(F_i(x^k) + \varepsilon_{k,i} x_i^k + w_i^k)] - \Pi_{K_i}[y_i^k - \alpha_{k,i}(F_i(y^k) + \varepsilon_{k,i} y_i^k)]\|^2 \\ &\leq \|x_i^k - \alpha_{k,i}(F_i(x^k) + \varepsilon_{k,i} x_i^k + w_i^k) - y_i^k + \alpha_{k,i}(F_i(y^k) + \varepsilon_{k,i} y_i^k)\|^2.\end{aligned}$$

Further, on expanding the expression on left of the preceding relation it can be verified that

$$\begin{aligned}\|x_i^{k+1} - y_i^k\|^2 \leq{}& \|x_i^k - y_i^k\|^2 - 2\alpha_{k,i}(x_i^k - y_i^k)^T(F_i(x^k) - F_i(y^k)) - 2\alpha_{k,i}\varepsilon_{k,i}\|x_i^k - y_i^k\|^2 \\ &- 2\alpha_{k,i}(x_i^k - y_i^k)^T w_i^k + \alpha_{k,i}^2\|F_i(x^k) - F_i(y^k) + w_i^k + \varepsilon_{k,i}(x_i^k - y_i^k)\|^2. \end{aligned} \tag{3.10}$$

The last term in the inequality can be expanded as

$$\begin{aligned}&\|F_i(x^k) - F_i(y^k) + w_i^k + \varepsilon_{k,i}(x_i^k - y_i^k)\|^2 = \|F_i(x^k) - F_i(y^k)\|^2 + \|w_i^k\|^2 + \varepsilon_{k,i}^2\|x_i^k - y_i^k\|^2 \\ &+ 2(F_i(x^k) - F_i(y^k))^T w_i^k + 2\varepsilon_{k,i}((x_i^k - y_i^k)^T w_i^k + (F_i(x^k) - F_i(y^k))^T(x_i^k - y_i^k)).\end{aligned} \tag{3.11}$$

Now, we take the expectation of (3.10) and (3.11) conditional on the past $\mathcal{F}_k$, and use $\mathbb{E}[w_i^k \mid \mathcal{F}_k] = 0$ (cf. Assumption 5(c)). By combining the resulting two relations we get

$$\begin{aligned}\mathbb{E}[\|x_i^{k+1} - y_i^k\|^2 \mid \mathcal{F}_k] \leq{}& (1 - 2\alpha_{k,i}\varepsilon_{k,i} + \alpha_{k,i}^2\varepsilon_{k,i}^2)\|x_i^k - y_i^k\|^2 + \alpha_{k,i}^2(\|F_i(x^k) - F_i(y^k)\|^2 + \mathbb{E}[\|w_i^k\|^2 \mid \mathcal{F}_k]) \\ &+ 2\alpha_{k,i}^2\varepsilon_{k,i}(x_i^k - y_i^k)^T(F_i(x^k) - F_i(y^k)) - 2\alpha_{k,i}(x_i^k - y_i^k)^T(F_i(x^k) - F_i(y^k)).\end{aligned}$$

Summing over all i and using $\alpha_{k,\min} \leq \alpha_{k,i} \leq \alpha_{k,\max}$, $\varepsilon_{k,\min} \leq \varepsilon_{k,i} \leq \varepsilon_{k,\max}$ together with the Lipschitz continuity of F yields

$$\begin{aligned}\mathbb{E}[\|x^{k+1}-y^k\|^2 \mid \mathcal{F}_k] \le (1-2\alpha_{k,\min}\varepsilon_{k,\min}+\alpha_{k,\max}^2\varepsilon_{k,\max}^2+\alpha_{k,\max}^2L^2)\|x^k-y^k\|^2 \\ +\alpha_{k,\max}^2\mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k]+2\sum_{i=1}^N\alpha_{k,i}^2\varepsilon_{k,i}(x_i^k-y_i^k)^T(F_i(x^k)-F_i(y^k)) \\ -2\sum_{i=1}^N\alpha_{k,i}(x_i^k-y_i^k)^T(F_i(x^k)-F_i(y^k)).\end{aligned} \tag{3.12}$$

Next we estimate the last two sums in (3.12) with the inner product terms $(x_i^k-y_i^k)^T(F_i(x^k)-F_i(y^k))$. The first sum involving $\alpha_{k,i}^2\varepsilon_{k,i}$ can be estimated as follows:

$$\sum_{i=1}^N 2\alpha_{k,i}^2\varepsilon_{k,i}(x_i^k-y_i^k)^T(F_i(x^k)-F_i(y^k)) \le \alpha_{k,\max}^2\varepsilon_{k,\max}\sum_{i=1}^N\|x_i^k-y_i^k\|\|F_i(x^k)-F_i(y^k)\|.$$

By Hölder's inequality, we have $\sum_{i=1}^N\|x_i^k-y_i^k\|\|F_i(x^k)-F_i(y^k)\| \le \|x^k-y^k\|\|F(x^k)-F(y^k)\|$, which through the use of Lipschitz continuity of F yields

$$\sum_{i=1}^N 2\alpha_{k,i}^2\varepsilon_{k,i}(x_i^k-y_i^k)^T(F_i(x^k)-F_i(y^k)) \le \alpha_{k,\max}^2\varepsilon_{k,\max}L\|x^k-y^k\|^2. \tag{3.13}$$

Adding and subtracting $\alpha_{k,\min}(x_i^k-y_i^k)^T(F_i(x^k)-F_i(y^k))$ in the last term of (3.12) we have

$$\begin{aligned}-\sum_{i=1}^N 2\alpha_{k,i}(x_i^k-y_i^k)^T(F_i(x^k)-F_i(y^k)) \le -2\alpha_{k,\min}(x^k-y^k)^T(F(x^k)-F(y^k)) \\ +2(\alpha_{k,\max}-\alpha_{k,\min})\sum_{i=1}^N\|x_i^k-y_i^k\|\|F_i(x^k)-F_i(y^k)\|.\end{aligned}$$

Using monotonicity of F we have $(x^k-y^k)^T(F(x^k)-F(y^k)) \ge 0$. Further, by letting $\delta_k \triangleq \alpha_{k,\max}-\alpha_{k,\min}$, and using Hölder's inequality and Lipschitz continuity of F, we get

$$-\sum_{i=1}^N 2\alpha_{k,i}(x_i^k-y_i^k)^T(F_i(x^k)-F_i(y^k)) \le 2\delta_k L\|x^k-y^k\|^2, \tag{3.14}$$

Combining relations (3.12), (3.13) and (3.14), we obtain

$$\begin{aligned}\mathbb{E}[\|x^{k+1}-y^k\|^2 \mid \mathcal{F}_k] \le (1-2\alpha_{k,\min}\varepsilon_{k,\min}+\alpha_{k,\max}^2\varepsilon_{k,\max}^2+\alpha_{k,\max}^2L^2)\|x^k-y^k\|^2 \\ +\alpha_{k,\max}^2\mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k]+(2\alpha_{k,\max}^2\varepsilon_{k,\max}L+2\delta_k L)\|x^k-y^k\|^2.\end{aligned}$$

Letting $q_k \triangleq 1-2\alpha_{k,\min}\varepsilon_{k,\min}+\alpha_{k,\max}^2(\varepsilon_{k,\max}+L)^2+2\delta_k L$, we can write

$$\mathbb{E}[\|x^{k+1}-y^k\|^2 \mid \mathcal{F}_k] \leq q_k\|x^k-y^k\|^2+\alpha_{k,\max}^2\mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k]. \tag{3.15}$$

Now, we relate $\|x^k-y^k\|$ to $\|x^k-y^{k-1}\|$. By the triangle inequality, we have $\|x^k-y^k\| \leq \|x^k-y^{k-1}\|+\|y^{k-1}-y^k\|$, while from Lemma 10 we have

$$\|y^k-y^{k-1}\| \leq M_y\frac{\varepsilon_{k-1,\max}-\varepsilon_{k,\min}}{\varepsilon_{k,\min}} \qquad \text{for all } k \geq 1.$$

Therefore, it follows that

$$\begin{aligned}\|x^k-y^k\|^2 &\leq \|x^k-y^{k-1}\|^2+\|y^k-y^{k-1}\|^2+2\|x^k-y^{k-1}\|\|y^k-y^{k-1}\| \\ &\leq \|x^k-y^{k-1}\|^2+\left(M_y\frac{\varepsilon_{k-1,\max}-\varepsilon_{k,\min}}{\varepsilon_{k,\min}}\right)^2+2M_y\frac{\varepsilon_{k-1,\max}-\varepsilon_{k,\min}}{\varepsilon_{k,\min}}\|x^k-y^{k-1}\|.\end{aligned}$$

Further we use Cauchy-Schwartz inequality to estimate the last term as follows:

$$\begin{aligned}M_y\frac{\varepsilon_{k-1,\max}-\varepsilon_{k,\min}}{\varepsilon_{k,\min}}\|x^k-y^{k-1}\| &= 2\sqrt{\alpha_{k,\min}\varepsilon_{k,\min}}\|x^k-y^{k-1}\|\frac{\varepsilon_{k-1,\max}-\varepsilon_{k,\min}}{\sqrt{\alpha_{k,\min}\varepsilon_{k,\min}}\,\varepsilon_{k,\min}}M_y \\ &\leq \alpha_{k,\min}\varepsilon_{k,\min}\|x^k-y^{k-1}\|^2+\frac{(\varepsilon_{k-1,\max}-\varepsilon_{k,\min})^2}{\alpha_{k,\min}\varepsilon_{k,\min}^3}M_y^2\end{aligned}$$

Using this in the preceding relation we obtain

$$\|x^k-y^k\|^2 \leq (1+\alpha_{k,\min}\varepsilon_{k,\min})\|x^k-y^{k-1}\|^2+\left(M_y\frac{\varepsilon_{k-1,\max}-\varepsilon_{k,\min}}{\varepsilon_{k,\min}}\right)^2\left(1+\frac{1}{\alpha_{k,\min}\varepsilon_{k,\min}}\right). \tag{3.16}$$

Combining the relations of (3.15) and (3.16) we obtain the following estimate:

$$\begin{aligned}\mathbb{E}[\|x^{k+1}-y^k\|^2 \mid \mathcal{F}_k] \leq & q_k(1+\alpha_{k,\min}\varepsilon_{k,\min})\|x^k-y^{k-1}\|^2+\alpha_{k,\max}^2\mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k] \\ & +q_k\left(M_y\frac{\varepsilon_{k-1,\max}-\varepsilon_{k,\min}}{\varepsilon_{k,\min}}\right)^2\left(1+\frac{1}{\alpha_{k,\min}\varepsilon_{k,\min}}\right).\end{aligned} \tag{3.17}$$

Next, we estimate the coefficient of $\|x^k-y^{k-1}\|^2$ in (3.17). Recalling the definition $q_k = 1-2\alpha_{k,\min}\varepsilon_{k,\min}+\alpha_{k,\max}^2(\varepsilon_{k,\max}+L)^2+2\delta_k L$, we show that $q_k \in (0,1)$ for all k large enough. Note that we can write

$$q_k = 1 - \alpha_{k,\min}\varepsilon_{k,\min}\left(2 - \frac{\alpha_{k,\max}^2}{\alpha_{k,\min}\varepsilon_{k,\min}}(\varepsilon_{k,\max}+L)^2 - \frac{2\delta_k L}{\alpha_{k,\min}\varepsilon_{k,\min}}\right).$$

By Assumption 6(a) we have

$$\frac{\alpha_{k,\max}^2}{\alpha_{k,\min}\varepsilon_{k,\min}}(\varepsilon_{k,\max}+L)^2 + \frac{2\delta_k L}{\alpha_{k,\min}\varepsilon_{k,\min}} \to 0,$$

implying that there exists a large enough integer $\tilde{k} \geq 0$ such that

$$\frac{\alpha_{k,\max}^2}{\alpha_{k,\min}\varepsilon_{k,\min}}(\varepsilon_{k,\max}+L)^2 + \frac{2\delta_k L}{\alpha_{k,\min}\varepsilon_{k,\min}} \leq c \qquad \text{for all } k \geq \tilde{k} \text{ and some } c \in (0,1). \tag{3.18}$$

Thus,

$$1 \leq 2 - \frac{\alpha_{k,\max}^2}{\alpha_{k,\min}\varepsilon_{k,\min}}(\varepsilon_{k,\max}+L)^2 - \frac{2\delta_k L}{\alpha_{k,\min}\varepsilon_{k,\min}} \leq 2 \qquad \text{for all } k \geq \tilde{k},$$

implying that for q_k we have

$$1 - 2\alpha_{k,\min}\varepsilon_{k,\min} \leq q_k \leq 1 - \alpha_{k,\min}\varepsilon_{k,\min} \qquad \text{for all } k \geq \tilde{k}.$$

Furthermore, since $\alpha_{k,\min}\varepsilon_{k,\min} \to 0$ by Assumption 6(b), we can choose $\tilde{k}$ large enough so that $q_k \in (0,1)$ for $k \geq \tilde{k}$. Hence, for $k \geq \tilde{k}$ we obtain $0 \leq q_k(1+\alpha_{k,\min}\varepsilon_{k,\min}) \leq q_k + \alpha_{k,\min}\varepsilon_{k,\min}$ and using the definition of q_k we further have for $k \geq \tilde{k}$,

$$\begin{aligned} 0 \leq q_k(1+\alpha_{k,\min}\varepsilon_{k,\min}) &\leq 1 - \alpha_{k,\min}\varepsilon_{k,\min}\left(1 - \frac{\alpha_{k,\max}^2}{\alpha_{k,\min}\varepsilon_{k,\min}}(\varepsilon_{k,\max}+L)^2 - \frac{2\delta_k L}{\alpha_{k,\min}\varepsilon_{k,\min}}\right) \\ &\leq 1 - \alpha_{k,\min}\varepsilon_{k,\min}(1-c), \end{aligned} \tag{3.19}$$

where the last inequality follows from (3.18). Using relations (3.19) and (3.17), we obtain

$$\mathbb{E}[\|x^{k+1} - y^k\|^2 \mid \mathcal{F}_k] \leq (1-u_k)\|x^k - y^{k-1}\|^2 + v_k \qquad \text{for all } k \geq \tilde{k},$$

where $u_k \triangleq (1-c)\alpha_{k,\min}\varepsilon_{k,\min}$ and

$$v_k = q_k\left(M_y \frac{\varepsilon_{k-1,\max} - \varepsilon_{k,\min}}{\varepsilon_{k,\min}}\right)^2\left(1 + \frac{1}{\alpha_{k,\min}\varepsilon_{k,\min}}\right) + \alpha_{k,\max}^2 \mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k].$$

We now verify that the conditions of Lemma 8 are satisfied for $k \geq \tilde{k}$. Since $c < 1$, from (3.19) we have $0 \leq u_k \leq 1$ for all $k \geq \tilde{k}$, while from Assumption 6(c) we have $\sum_{k=\tilde{k}}^{\infty} u_k = \infty$. Under stepsize conditions Assumption 6(d)–(f), it can be verified that $\lim_{k\to\infty} \frac{v_k}{u_k} = 0$ and $\sum_{k=0}^{\infty} v_k < \infty$. Thus, the conditions of Lemma 8 are satisfied for $k \geq \tilde{k}$. Noting that Lemma 8 applies to a process delayed by a deterministic time-offset, we can conclude that $\|x^k - y^{k-1}\| \to 0$ *a.s.* ■

As an immediate consequence of Proposition 7 and the properties of Tikhonov sequence established in Lemma 10, we have the following result.

Proposition 8. *Let Assumptions 5 and 6 hold. Also, assume that SOL(K,F) is nonempty. Then, for the sequence $\{x^k\}$ generated by stochastic iterative Tikhonov algorithm* (PITR)*, we have*

(a) If $\limsup_{k\to\infty} \frac{\varepsilon_{k,\max}}{\varepsilon_{k,\min}} < \infty$, then $\{x^k\}$ is bounded and every accumulation point of $\{x^k\}$ is a solution of VI(K,F).

(b) If $\limsup_{k\to\infty} \frac{\varepsilon_{k,\max}}{\varepsilon_{k,\min}} \leq 1$, then $\{x^k\}$ converges to the smallest-norm solution of VI(K,F).

A further extension of Proposition 7 is obtained when the mapping F is strictly monotone over the set K. In this case, the uniqueness of solution of VI(K,F) is guaranteed provided a solution exists. Hence, from Lemma 10(a) we have $\{y^k\}$ converging to the unique solution of VI(K,F), which in view of Proposition 7 implies that $\{x^k\}$ is converging to the solution *a.s.* This result is precisely presented in the following corollary.

Corollary 3. *Let Assumption 5 hold with F being strictly monotone over the set K. Also let Assumption 6 hold, and assume that SOL(K,F) is nonempty. Then, the sequence $\{x^k\}$ generated by iterative Tikhonov method* (PITR) *converges to the unique solution of VI(K,F) a.s.*

We conclude this section by providing an example of steplength and regularization sequences that satisfy the conditions of Assumption 6(a)–(e).

Lemma 11. *Consider $\alpha_{k,i} = (k+\eta_i)^{-a}$ and $\varepsilon_k = (k+\zeta_i)^{-b}$ for $k \geq 0$, where each η_i and ζ_i are selected from a uniform distribution on the intervals $[\underline{\eta}, \bar{\eta}]$ and $[\underline{\zeta}, \bar{\zeta}]$, respectively, for some $0 < \underline{\eta} < \bar{\eta}$ and $0 < \underline{\zeta} < \bar{\zeta}$. Let $a, b \in (0,1)$, $a+b < 1$, and $a > b$. Then $\{\alpha_{k,i}\}$ and $\{\varepsilon_{k,i}\}$ satisfy Assumption 6(a)–(e).*

Proof. The first limit condition in Assumption 6(a) holds trivially as we see that for $a > b$,

$$\lim_{k\to\infty} \frac{\alpha_{k,\max}}{\alpha_{k,\min}} \frac{\alpha_{k,\max}}{\varepsilon_{k,\min}} (\varepsilon_{k,\max} + L)^2 = \lim_{k\to\infty} \frac{(k+\eta_{\min})^{-a}}{(k+\eta_{\max})^{-a}} \frac{(k+\eta_{\min})^{-a}}{(k+\zeta_{\max})^{-b}} ((k+\zeta_{\min})^{-b} + L)^2,$$

where $\eta_{\max} = \max_{1\le i\le N}\{\eta_i\}$, $\eta_{\min} = \min_{1\le i\le N}\{\eta_i\}$, $\zeta_{\max} = \max_{1\le i\le N}\{\zeta_i\}$, and $\zeta_{\min} = \min_{1\le i\le N}\{\zeta_i\}$. We further have

$$\lim_{k\to\infty} \frac{(k+\eta_{\min})^{-a}}{(k+\eta_{\max})^{-a}} = 1,$$

implying

$$\lim_{k\to\infty} \frac{(k+\eta_{\min})^{-a}}{(k+\eta_{\max})^{-a}} \frac{(k+\eta_{\min})^{-a}}{(k+\zeta_{\max})^{-b}}((k+\zeta_{\min})^{-b}+L)^2 = \lim_{k\to\infty} \frac{(k+\eta_{\min})^{-a}}{(k+\zeta_{\max})^{-b}}((k+\zeta_{\min})^{-b}+L)^2 = 0,$$

where the last equality follows by $a > b$. The second condition of Assumption 6(a) can be seen to follow by noticing that the argument of the limit can be written as

$$\begin{aligned}\frac{\alpha_{k,\max}-\alpha_{k,\min}}{\alpha_{k,\min}\varepsilon_{k,\max}} &= \frac{(k+\eta_{\min})^{-a}-(k+\eta_{\max})^{-a}}{(k+\eta_{\max})^{-a}(k+\zeta_{\max})^{-b}} = \frac{\frac{(k+\eta_{\min})^{-a}}{(k+\eta_{\max})^{-a}}-1}{(k+\zeta_{\max})^{-b}} \\ &= \frac{\left(1-\frac{\eta_{\max}-\eta_{\min}}{k+\eta_{\max}}\right)^{-a}-1}{k^{-b}(1+\frac{\zeta_{\max}}{k})^{-b}} \approx \frac{1+a\frac{\eta_{\max}-\eta_{\min}}{k+\eta_{\max}}+O(1/k^2)-1}{k^{-b}(1+\frac{\zeta_{\max}}{k})^{-b}} = O(1/k^{1-b}).\end{aligned}$$

As $k\to\infty$, the required result follows. Also, Assumption 6(b) and (c) hold since $\alpha_{k,\min}\varepsilon_{k,\min} = k^{-a-b}(1+\eta_{\max}/k)^{-a}(1+\zeta_{\max}/k)^{-b} > k^{-1}$. Under the given form of $\varepsilon_{k,i}$ and $\alpha_{k,i}$ the expression in the summation of Assumption 6(d) becomes

$$\begin{aligned}&\frac{((k-1+\zeta_{\min})^{-b}-(k+\zeta_{\max})^{-b})^2}{(k+\zeta_{\max})^{-2b}}\left(1+\frac{1}{(k+\eta_{\max})^{-a}(k+\zeta_{\max})^{-b}}\right) \\ &\le 2\frac{((1+(\zeta_{\min}-1)/k)^{-b}(1+\zeta_{\max}/k)^{b}-1)^2}{(k+\eta_{\max})^{-a}(k+\zeta_{\max})^{-b}},\end{aligned}$$

where the inequality follows from the fact that

$$\frac{1}{(k+\eta_{\max})^{-a}(k+\zeta_{\max})^{-b}} \ge 1 \qquad \text{for } k\ge 1.$$

Using the expansion of $(1-x)^{-b}$ for x small and ignoring higher order terms, we have

$$\begin{aligned}\left((1+(\zeta_{\min}-1)/k)^{-b}(1+\zeta_{\max}/k)^{b}-1\right)^2 &\approx \left(\left(1-b\frac{\zeta_{\min}-1}{k}\right)\left(1+b\frac{b\zeta_{\max}}{k}\right)-1\right)^2 \\ &\approx \frac{b^2(\zeta_{\max}-\zeta_{\min}+1)^2}{k^2}.\end{aligned}$$

Also for k large enough, we have $(k+\eta_{\max})^{-a}(k+\zeta_{\max})^{-b} \approx k^{-a-b}$. Thus we have

$$2\frac{((1+(\zeta_{\min}-1)/k)^{-b}(1+\zeta_{\max}/k)^{b}-1)^2}{(k+\eta_{\max})^{-a}(k+\zeta_{\max})^{-b}} \approx 2\frac{b^2(\zeta_{\max}-\zeta_{\min}+1)^2}{k^{2-a-b}} = O(k^{-(1+\delta)}),$$

where in the equality we use $a+b<1$ and $\delta = 1-(a+b)>0$. Following a similar argument, it can be verified that the term in Assumption 6(e) reduces to

$$\frac{((k-1+\zeta_{\min})^{-b}-(k+\zeta_{\max})^{-b})^2}{(k+\zeta_{\max})^{-2b}}\left(1+\frac{1}{(k+\eta_{\max})^{-a}(k+\zeta_{\max})^{-b}}\right)\frac{1}{(k+\eta_{\max})^{-a}(k+\zeta_{\max})^{-b}}$$
$$\approx 2\frac{b^2(\zeta_{\max}-\zeta_{\min}+1)^2}{k^{2-a-b}}\frac{1}{k^{-a-b}},$$

and the limit in Assumption 6(e) follows from $a+b<1$. ■

Note that Assumption 6(f) is immediately satisfied when $\mathbb{E}[\|w^k\|^2|\mathcal{F}_k]$ is uniformly bounded by some constant and the steplength and regularization sequences are chosen as per Lemma 11.

3.3 Stochastic iterative proximal-point methods

An alternative to using iterative Tikhonov regularization techniques is available through proximal-point methods, a class of techniques that appear to have been first studied by Martinet [39], and subsequently by Rockafellar [78]. A more recent description in the context of maximal-monotone operators can be found in [35]. In the standard proximal-point methods, the convergence to a single solution of $\mathrm{VI}(K,F)$ is obtained through the addition of a proximal term $\theta(x_k - x_{k-1})$, where θ is a fixed positive parameter. In effect, $x_k = \mathrm{SOL}(K, F+\theta(\mathbf{I}-x_{k-1})$ and convergence may be guaranteed under suitable assumptions.

A crucial shortcoming of standard proximal-point methods lies in the need to solve a sequence of variational problems. Analogous to our efforts in constructing an iterative Tikhonov regularization technique, we consider an iterative proximal-point method. In such a method, the centering term x_{k-1} is updated after *every* projection step rather than when it obtains an accurate solution of $\mathrm{VI}(K,F+\theta(\mathbf{I}-x_{k-1}))$.

Before providing a detailed analysis of the convergence properties of this method, we examine the relationship between the proposed iterative proximal point method and the standard gradient projection method, in the context of variational inequalities. An iterative proximal-point method for $\mathrm{VI}(K,F)$ necessitates an update given by

$$x^{k+1} = \Pi_K[x^k - \gamma_k(F(x^k) + \theta(x^k - x^{k-1}))] = \Pi_K[x^k(\theta) - \gamma_k F(x^k)],$$

where $x^k(\theta) \triangleq (1-\gamma_k\theta)x^k + \gamma_k\theta x^{k-1}$. Therefore, when $\theta \equiv 0$ and $\gamma_k \to 0$, the method reduces to the standard gradient projection method. More generally, one can view the proximal-point method as employing a convex combination of the old iterate x^{k-1} and x^k instead of x^k in the standard gradient method. In our algorithm, we allow θ to vary at every iteration, i.e., we employ a sequence θ_k which can grow to $+\infty$ but at a sufficiently slow rate.

Analogous to the partially coordinated iterative Tikhonov (PITR) method, we consider a limited coordination generalization of the iterative proximal-point method (PIPP) where users independently choose their individual stepsizes. More precisely we have the following algorithm:

$$x_i^{k+1} = \Pi_{K_i}[x_i^k - \alpha_{k,i}(F_i(x^k) + \theta_{k,i}(x_i^k - x_i^{k-1}) + w_i^k)] \quad \text{for } i = 1,\ldots,N, \tag{PITR}$$

where $\alpha_{k,i}$ is the stepsize and $\theta_{k,i}$ is the centering term parameter chosen by the ith user at the kth iteration. We make the following assumption on user steplengths and parameters $\theta_{k,i}$.

Assumption 7. *Let* $\alpha_{k,\max} = \max_{1\le i\le N}\{\alpha_{k,i}\}$, $\alpha_{k,\min} = \min_{1\le i\le N}\{\alpha_{k,i}\}$, $\theta_{k,\max} = \max_{1\le i\le N}\{\theta_{k,i}\}$, $\theta_{k,\min} = \min_{1\le i\le N}\{\theta_{k,i}\}$*, and let the following hold:*

(a) $\alpha_{k,\max}\theta_{k,\max} \le \left(1 + 2\alpha_{k,\max}^2 L^2\right)\alpha_{k-1,\min}\theta_{k-1,\min}$ *for all* $k \ge 1$*, and*

$$\lim_{k\to\infty} \frac{\alpha_{k,\max}^2\theta_{k,\max}^2}{\alpha_{k,\min}\theta_{k,\min}} = c \quad \textit{with } c \in [0, 1/2);$$

(b) $\sum_{k=0}^{\infty}\alpha_{k,i} = \infty$ *and* $\sum_{k=0}^{\infty}\alpha_{k,i}^2 < \infty$ *for all i;*

(c) $\sum_{k=0}^{\infty}\left(\alpha_{k,\max} - \alpha_{k,\min}\right) < \infty$.

(d) $\sum_{k=0}^{\infty}\alpha_{k,\max}^2\mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k] < \infty$ *almost surely.*

Later on, after our convergence results of this section, we will provide an example for the stepsizes and prox-parameters satisfying Assumption 7.

Our main result is given in the following proposition, where by using Assumption 7 we show almost sure convergence of the method. In addition, we assume that the mapping F is strictly monotone over the set K, i.e.,

$$(F(x) - F(y))^T(x - y) > 0 \qquad \text{for all } x, y \in K \text{ with } x \ne y.$$

Proposition 9. *Let Assumption 5 hold with F being strictly monotone. Assume that $SOL(K,F)$ is nonempty. Also, let the steplengths and the prox-parameters satisfy Assumption 7. Then, the sequence $\{x^k\}$ generated by method* (PITR) *converges almost surely to the solution of $VI(K,F)$.*

Proof. Since F is strictly monotone and $\mathrm{SOL}(K,F)$ is nonempty, $\mathrm{VI}(K,F)$ must have a unique solution, denoted by $x^* = (x_1^*, \ldots, x_N^*)$ (cf. Theorem 2.3.3 in [35]). By using $x_i^* = \Pi_{K_i}[x_i^* - \alpha_{k,i}F_i(x^*)]$ for all i and the nonexpansive property of the Euclidean projection operator, we bound the term $\|x_i^{k+1} - x_i^*\|$ as follows:

$$\begin{aligned}\|x_i^{k+1} - x_i^*\|^2 &= \|\Pi_{K_i}[x_i^k - \alpha_{k,i}(F_i(x^k) + \theta_{k,i}(x_i^k - x_i^{k-1}) + w_i^k)] - \Pi_{K_i}[x_i^* - \alpha_{k,i}F_i(x^*)]\|^2 \\ &\leq \left\|(x_i^k - x_i^*) - \alpha_{k,i}\left(F_i(x^k) - F_i(x^*) - \theta_{k,i}(x_i^k - x_i^{k-1}) - w_i^k\right)\right\|^2.\end{aligned}$$

Further, the right hand side of preceding relation can be expanded as

$$\begin{aligned}\mathrm{RHS} = {} & \|x_i^k - x_i^*\|^2 + \alpha_{k,i}^2\|F_i(x^k) - F_i(x^*)\|^2 + \alpha_{k,i}^2\|w_i^k\|^2 + (\alpha_{k,i}\theta_{k,i})^2\|x_i^k - x_i^{k-1}\|^2 \\ & - 2\alpha_{k,i}(x_i^k - x_i^*)^T(F_i(x^k) - F_i(x^*)) - 2\alpha_{k,i}\theta_{k,i}(x_i^k - x_i^*)^T(x_i^k - x_i^{k-1}) - 2\alpha_{k,i}(x_i^k - x_i^*)^T w_i^k \\ & + 2\alpha_{k,i}^2\theta_{k,i}(F_i(x^k) - F_i(x^*))^T(x_i^k - x_i^{k-1}) + 2\alpha_{k,i}^2(F_i(x^k) - F_i(x^*))^T w_i^k + 2\alpha_{k,i}^2\theta_{k,i}(x_i^k - x_i^{k-1})^T w_i^k.\end{aligned}$$

Taking expectation and using $\mathbb{E}[w_i^k \mid \mathcal{F}_k] = 0$ (Assumption 5(c)), we obtain

$$\begin{aligned}\mathbb{E}[\|x_i^{k+1} - x_i^*\|^2 \mid \mathcal{F}_k] \leq {} & \|x_i^k - x_i^*\|^2 + \alpha_{k,i}^2\|F_i(x^k) - F_i(x^*)\|^2 + \alpha_{k,i}^2\mathbb{E}[\|w_i^k\|^2 \mid \mathcal{F}_k] \\ & + (\alpha_{k,i}\theta_{k,i})^2\|x_i^k - x_i^{k-1}\|^2 - 2\alpha_{k,i}(x_i^k - x_i^*)^T(F_i(x^k) - F_i(x^*)) \\ & - 2\alpha_{k,i}\theta_{k,i}(x_i^k - x_i^*)^T(x_i^k - x_i^{k-1}) + 2\alpha_{k,i}^2\theta_{k,i}(F_i(x^k) - F_i(x^*))^T(x_i^k - x_i^{k-1}).\end{aligned}$$

Let $\alpha_{k,\max} = \max_{1\leq i\leq N}\{\alpha_{k,i}\}$, $\alpha_{k,\min} = \min_{1\leq i\leq N}\{\alpha_{k,i}\}$, $\theta_{k,\max} = \max_{1\leq i\leq N}\{\theta_{k,i}\}$ and $\theta_{k,\min} = \min_{1\leq i\leq N}\{\theta_{k,i}\}$. Summing over all i and using Lipschitz continuity of F (Assumption 5(b)) we arrive at

$$\mathbb{E}[\|x^{k+1}-x^*\|^2 \mid \mathcal{F}_k] \leq (1+\alpha_{k,\max}^2 L^2)\|x^k-x^*\|^2 + \alpha_{k,\max}^2 \mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k]$$
$$+(\alpha_{k,\max}\theta_{k,\max})^2\|x^k-x^{k-1}\|^2 \underbrace{-2\sum_{i=1}^N \alpha_{k,i}(x_i^k-x_i^*)^T(F_i(x^k)-F_i(x^*))}_{\textbf{Term 1}}$$
$$\underbrace{-2\sum_{i=1}^N \alpha_{k,i}\theta_{k,i}(x_i^k-x_i^*)^T(x_i^k-x_i^{k-1})}_{\textbf{Term 2}}$$
$$\underbrace{+2\sum_{i=1}^N \alpha_{k,i}^2\theta_{k,i}(F_i(x^k)-F_i(x^*))^T(x_i^k-x_i^{k-1})}_{\textbf{Term 3}}. \tag{3.20}$$

By adding and subtracting $2\alpha_{k,\min}(x_i^k-x_i^*)^T(F_i(x^k)-F_i(x^*))$ to each term of Term 1 we see that

$$\begin{aligned}\textbf{Term 1} &\leq -2\alpha_{k,\min}(x^k-x^*)^T(F(x^k)-F(x^*)) + 2(\alpha_{k,\max}-\alpha_{k,\min})\sum_{i=1}^N \|x_i^k-x_i^*\|\|F_i(x^k)-F_i(x^*)\| \\ &\leq -2\alpha_{k,\min}(x^k-x^*)^T(F(x^k)-F(x^*)) + 2(\alpha_{k,\max}-\alpha_{k,\min})L\|x^k-x^*\|^2, \end{aligned} \tag{3.21}$$

where the first inequality follows by using Cauchy-Schwartz inequality, while the second inequality follows from Hölder's inequality and Lipschitz continuity of F.

We now estimate Term 2. Since $2(x-y)^T(x-z) = \|x-y\|^2+\|x-z\|^2-\|y-z\|^2$, we have

$$\begin{aligned}\textbf{Term 2} &= -\sum_{i=1}^N \alpha_{k,i}\theta_{k,i}\left[\|x_i^k-x_i^*\|^2+\|x_i^k-x_i^{k-1}\|^2-\|x_i^{k-1}-x_i^*\|^2\right] \\ &\leq -\alpha_{k,\min}\theta_{k,\min}\sum_{i=1}^N\left[\|x_i^k-x_i^*\|^2+\|x_i^k-x_i^{k-1}\|^2\right]+\alpha_{k,\max}\theta_{k,\max}\sum_{i=1}^N\|x_i^{k-1}-x_i^*\|^2 \\ &= -\alpha_{k,\min}\theta_{k,\min}\left[\|x^k-x^*\|^2+\|x^k-x^{k-1}\|^2\right]+\alpha_{k,\max}\theta_{k,\max}\|x^{k-1}-x^*\|^2. \end{aligned} \tag{3.22}$$

We now consider Term 3. Using $2x^T y \le \|x\|^2 + \|y\|^2$ and Lipschitz continuity of F, we obtain

$$
\begin{aligned}
\textbf{Term 3} &\le \sum_{i=1}^{N} \alpha_{k,i}^2 \left(\|F_i(x^k) - F_i(x^*)\|^2 + \theta_{k,i}^2 \|x_i^k - x_i^{k-1}\|^2 \right) \\
&\le \alpha_{k,\max}^2 \left(\|F(x^k) - F(x^*)\|^2 + \theta_{k,\max}^2 \|x^k - x^{k-1}\|^2 \right) \\
&\le \alpha_{k,\max}^2 \left(L^2 \|x^k - x^*\|^2 + \theta_{k,\max}^2 \|x^k - x^{k-1}\|^2 \right).
\end{aligned} \tag{3.23}
$$

Combining (3.20) with (3.21), (3.22), and (3.23), we obtain

$$
\begin{aligned}
\mathbb{E}[\|x^{k+1} - x^*\|^2 \mid \mathcal{F}_k] \le{}& \left(1 + 2\alpha_{k,\max}^2 L^2 + 2(\alpha_{k,\max} - \alpha_{k,\min})L\right) \|x^k - x^*\|^2 \\
&+ \alpha_{k,\max}\theta_{k,\max} \|x^{k-1} - x^*\|^2 - \alpha_{k,\min}\theta_{k,\min} \|x^k - x^*\|^2 \\
&- \alpha_{k,\min}\theta_{k,\min} \left(1 - \frac{2\alpha_{k,\max}^2 \theta_{k,\max}^2}{\alpha_{k,\min}\theta_{k,\min}} \right) \|x^k - x^{k-1}\|^2 \\
&- 2\alpha_{k,\min}(x^k - x^*)^T (F(x^k) - F(x^*)) + \alpha_{k,\max}^2 \mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k],
\end{aligned} \tag{3.24}
$$

By Assumption 7(a) we have

$$
\begin{aligned}
\alpha_{k,\max}\theta_{k,\max} &\le \left(1 + 2\alpha_{k,\max}^2 L^2\right) \alpha_{k-1,\min}\theta_{k-1,\min} \\
&\le \left(1 + 2\alpha_{k,\max}^2 L^2 + 2(\alpha_{k,\max} - \alpha_{k,\min})L\right) \alpha_{k-1,\min}\theta_{k-1,\min}.
\end{aligned}
$$

Using this, moving the term $-\alpha_{k,\min}\theta_{k,\min}\|x^k - x^*\|^2$ on the other side of inequality (3.24), and noting that

$$
\frac{2\alpha_{k,\max}^2 \theta_{k,\max}^2}{\alpha_{k,\min}\theta_{k,\min}} \le d \qquad \text{for some } d \in (0,1) \text{ and for } k \ge \tilde{k},
$$

with sufficiently large $\tilde{k}$ (since $\frac{2\alpha_{k,\max}^2 \theta_{k,\max}^2}{\alpha_{k,\min}\theta_{k,\min}} \to 2c$ with $2c < 1$ by Assumption 7(a)). We further see that for $k \ge \tilde{k}$,

$$
\begin{aligned}
&\mathbb{E}[\|x^{k+1} - x^*\|^2 \mid \mathcal{F}_k] + \alpha_{k,\min}\theta_{k,\min}\|x^k - x^*\|^2 \\
&\le \left(1 + 2\alpha_{k,\max}^2 L^2 + 2(\alpha_{k,\max} - \alpha_{k,\min})L\right) \left(\|x^k - x^*\|^2 + \alpha_{k-1,\min}\theta_{k-1,\min}\|x^{k-1} - x^*\|^2 \right) \\
&- \alpha_{k,\min}\theta_{k,\min}(1 - d) \|x^k - x^{k-1}\|^2 \\
&- 2\alpha_{k,\min}(x^k - x^*)^T (F(x^k) - F(x^*)) + \alpha_{k,\max}^2 \mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k].
\end{aligned} \tag{3.25}
$$

It remains to show that the sequence $\{\|x^{k+1} - x^*\|\}$ converges to zero. This can be done by applying Lemma 9 to relation (3.25) with the following identification:

$$V_k = \|x^k - x^*\|^2 + \alpha_{k-1,\min}\theta_{k-1,\min}\|x^{k-1} - x^*\|^2, \qquad u_k = 2\alpha_{k,\max}^2 L^2 + 2(\alpha_{k,\max} - \alpha_{k,\min})L,$$

$$\gamma_k = \alpha_{k,\min}\theta_{k,\min}(1-d)\|x^k - x^{k-1}\|^2 + 2\alpha_{k,\min}(x^k - x^*)^T(F(x^k) - F(x^*)),$$

$$\beta_k = \alpha_{k,\max}^2 \mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k].$$

To use the lemma, we need to verify that $\gamma_k \geq 0$, $\sum_{k=0}^{\infty} u_k < \infty$ and $\sum_{k=0}^{\infty} \beta_k < \infty$. Note that $\gamma_k > 0$ for all $k \geq \tilde{k}$ since $d \in (0,1)$ and F is monotone. The condition $\sum_{k=0}^{\infty} u_k < \infty$ holds by our assumption that $\sum_{k=0}^{\infty} \alpha_{k,i}^2 < \infty$ for all i (Assumption 7(b)), while $\sum_{k=0}^{\infty} \beta_k < \infty$ holds by Assumption 7(d). Thus, according to Lemma 9 (that holds for all k large enough) we have for the solution x^*,

$$\{\|x^k - x^*\|^2 + \alpha_{k-1,\min}\theta_{k-1,\min}\|x^{k-1} - x^*\|^2\} \text{ converges } a.s., \tag{3.26}$$

$$\sum_{k=0}^{\infty} \alpha_{k,\min}(x^k - x^*)^T(F(x^k) - F(x^*)) < \infty \ \ a.s. \tag{3.27}$$

The expression (3.26) implies that the sequence $\{x^k\}$ is bounded *a.s.* and has accumulation points *a.s.* Since K is closed and $\{x^k\} \subset K$, it follows that all the accumulation points of $\{x^k\}$ belong to K. By (3.27) and the relation $\sum_{k=0}^{\infty} \alpha_{k,\min} = \infty$ (see Assumption 7(b)) it follows that $(x^k - x^*)^T(F(x^k) - F(x^*)) \to 0$ along a subsequence *a.s.* This and strict monotonicity of F imply that $\{x^k\}$ has one accumulation point, say $\tilde{x}$, that must coincide with the solution x^*. By relation (3.26) it follows that the entire sequence must converge to the solution x^* *a.s.* ■

Consider now the case when we have uniformity in user stepsize and prox-parameter. Precisely, let each user i implement the following update rule:

$$x_i^{k+1} = \Pi_{K_i}[x_i^k - \alpha_k(F_i(x^k) + \theta_k(x_i^k - x_i^{k-1}) + w_i^k)], \tag{IPP}$$

where $\theta_k > 0$ is the prox-parameter and α_k is a stepsize for all players at iteration k. We refer to this version of the method as Iterative Proximal Point (IPP) algorithm, to differentiate it from its partially-coordinated version (PITR) where the users have some freedom in selecting the parameters.

Almost sure convergence of the sequence $\{x^k\}$ generated using (IPP) can be obtained as a corollary of Proposition 9.

Corollary 4. *Let Assumption 5 hold, where F is strictly monotone. Assume that $SOL(K,F)$ is nonempty. Also, let the steplengths and the prox-parameters satisfy Assumption 7 with $\alpha_{k,i} = \alpha_k$*

and $\theta_{k,i} = \theta_k$ *for all i. Then, the sequence* $\{x^k\}$ *generated by method* (IPP) *converges almost surely to the solution of* $VI(K,F)$*.*

Proposition 9 holds if the stepsize and prox-parameter sequences satisfy Assumption 7 for all $k \geq \tilde{k}$ where $\tilde{k}$ is some positive integer. We next discuss some examples for choices of stepsize sequence $\{\alpha_{k,i}\}$ and prox-parameter sequence $\{\theta_{k,i}\}$ that satisfy Assumption 7(a)–(c) for sufficiently large indices k. Let

$$\alpha_{k,i} = (k+\eta_i)^{-a} \qquad \text{and} \qquad \theta_{k,i} = (k+\eta_i)^{-b},$$

for some scalars a and b such that $a \in (1/2, 1]$ and $a+b > 0$. The scalars η_i are random with uniform distribution over an interval $[\underline{\eta}, \bar{\eta}]$ for some $0 < \underline{\eta} < \bar{\eta}$. Then, Assumption 7(a) holds if

$$\frac{\alpha_{k,\max}\theta_{k,\max}}{\alpha_{k-1,\min}\theta_{k-1,\min}} \leq 1 + 2\alpha_{\max,k}^2 L^2. \tag{3.28}$$

If $\eta_{\max} \triangleq \max_{1\leq i\leq N}\{\eta_i\}$ and $\eta_{\min} \triangleq \min_{1\leq i\leq N}\{\eta_i\}$, we have that

$$\frac{\alpha_{k,\max}\theta_{k,\max}}{\alpha_{k-1,\min}\theta_{k-1,\min}} = \frac{(k+\eta_{\min})^{-(a+b)}}{(k-1+\eta_{\max})^{-(a+b)}} = \left(1+\frac{\eta_{\min}}{k}\right)^{-(a+b)}\left(1-\frac{1-\eta_{\max}}{k}\right)^{a+b}.$$

Using $(1+x)^n \approx 1+nx$ for $|x| << 1$, then for sufficiently large k, we have

$$\frac{\alpha_{k,\max}\theta_{k,\max}}{\alpha_{k-1,\min}\theta_{k-1,\min}} \approx \left(1-\frac{(a+b)\eta_{\min}}{k}\right)\left(1-\frac{(a+b)(1-\eta_{\max})}{k}\right) \approx 1-\frac{(a+b)(1+\eta_{\min}-\eta_{\max})}{k}.$$

Thus for $\eta_{\max} - \eta_{\min} < 1$ we have,

$$\lim_{k\to\infty} \frac{\alpha_{k,\max}\theta_{k,\max}}{\alpha_{k-1,\min}\theta_{k-1,\min}} \nearrow 1.$$

The preceding relation combined with the fact $1+2\alpha_{\max,k}^2 L^2 > 1$ implies that relation (3.28) holds for sufficiently large k.

We now consider the limit in the second part of Assumption 7(a). We have

$$\lim_{k\to\infty} \frac{\alpha_{k,\max}^2\theta_{k,\max}^2}{\alpha_{k,\min}\theta_{k,\min}} = \lim_{k\to\infty} \frac{(k+\eta_{\min})^{-2(a+b)}}{(k+\eta_{\max})^{-(a+b)}} = 0,$$

where the zero-limit follows by $a+b > 0$. The conditions of Assumption 7(b) hold trivially for

$a \in (1/2, 1]$. For the condition of Assumption 7(c), we have

$$\begin{aligned}\alpha_{k,\max} - \alpha_{k,\min} &= (k+\eta_{\min})^{-a} - (k+\eta_{\max})^{-a} = (k+\eta_{\max})^{-a}\left(\frac{(k+\eta_{\min})^{-a}}{(k+\eta_{\max})^{-a}} - 1\right) \\ &= (k+\eta_{\max})^{-a}\left(\left(1 - \frac{\eta_{\max}-\eta_{\min}}{k+\eta_{\max}}\right)^{-a} - 1\right) \\ &\approx (k+\eta_{\max})^{-a}\left(1 + a\frac{\eta_{\max}-\eta_{\min}}{k+\eta_{\max}} + O(1/k^2) - 1\right) = O(1/k^{1+a}),\end{aligned}$$

which is summable for $a > 0$.

3.4 Case study

In this section, we examine the sensitivity of the proposed iterative Tikhonov method and the proximal-point method to algorithm parameters. More specifically, in Section 3.4.1, we describe the player payoffs and strategy sets as well as the network constraints employed in the case study. The sensitivity of the methods to algorithm parameters is examined in Section 3.4.2. Finally, Sections 3.4.3 and 3.4.4 provide comparisons with their standard (two-loop) counterparts as well as sample-average approximation methods.

3.4.1 Network and user data

We re-consider the spatial network of Chapter 2 and present it again in Fig. 3.1 for completeness. Suppose now there are N selfish users that compete over the network. Each user is characterized by a user-specific utility and faces a congestion cost that is a function of the aggregate flow in a link. Such a problem captures traffic and communication networks where the congestion cost may manifest itself through link-specific delays [5, 6]. The ith user's cost function $f_i(x_i, \xi_i, \omega_i)$ is a function of flow decisions x_i and is parameterized by the uncertainty, denoted by (ξ_i, ω_i). It is defined as

$$f_i(x_i, \xi_i, \omega_i) \triangleq -\xi_i \log(1 + x_i + \omega_i). \tag{3.29}$$

Each user selects an origin-destination pair of nodes on this network and faces congestion based on the links traversed along the prescribed path connecting the selected origin-destination nodes. We assume that the network links are indexed by an index set $\mathscr{L}$ and we consider a congestion cost of the form:

$$c(x,\varsigma) = \varsigma \sum_{i=1}^{N} \sum_{l\in\mathscr{L}} x_{li} \left(\sum_{j=1}^{N} x_{lj} \right), \tag{3.30}$$

where x_{lj} denotes the flow of user j on link l and ς is a random scaling parameter. For all $i \in \{1,\ldots,N\}$ and $l \in \mathscr{L}$, x_{li} is given by

$$x_{li} = \begin{cases} x_i & \text{if user } i \text{ uses link } l, \\ 0 & \text{otherwise.} \end{cases}$$

Let A denote the adjacency matrix that specifies the set of links traversed by the traffic generated by a particular user. More precisely, for every link $l \in \mathscr{L}$ and user i, we have $A_{li} = 1$ if link l carries flow of user i, and $A_{li} = 0$ otherwise. Throughout this section, we consider the network with 9 links and 5 users, as given in Fig. 2.1. The simulation results are reported in tables, where we use $U(t,\tau)$ to denote the uniform distribution over an interval $[t,\tau]$ for $t < \tau$.

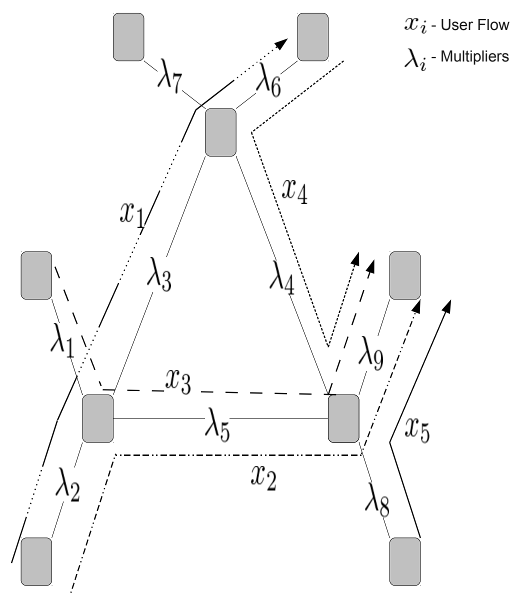


Figure 3.1: A network with 5 users and 9 links.

Table 3.1 summarizes the traffic in the network as generated by the users and provides the uniform distribution for the parameters $k_i(\xi_i)$ and noise ω_i of the user objectives. In addition we assume that congestion scaling parameter is uniformly distributed i.e., $\varsigma \sim U(1/2,1)$.

The strategy sets are coupled through an expected-value constraint of the form $\sum_{i=1}^{N} A_{li}x_i \leq \mathbb{E}[C_l(\zeta_l)]$ for all $l \in \mathscr{L}$ where $C_l(\zeta_l)$ is the random aggregate traffic through link l. The constraint can be compactly written as $Ax \leq \mathbb{E}[C(\zeta)]$, where $\zeta = (\zeta_1,\ldots,\zeta_9)^T$ and $C(\zeta)$ is the random link

Table 3.1: Network and user data

User(i)	Links traversed	ξ_i	ω_i
1	L2, L3, L6	$U(0,10)$	$U(0,1)$
2	L2, L5, L9	$U(0,10)$	$U(0,1)$
3	L1, L5, L9	$U(0,10)$	$U(0,1)$
4	L6, L4, L9	$U(0,10)$	$U(0,1)$
5	L8, L9	$U(0,10)$	$U(0,1)$

capacity vector with $C(\zeta) \sim U(\bar{C}-1, \bar{C}+1)$ and $\bar{C} = (10,15,20,10,15,20,20,15,25)$. In the resulting $N-$player stochastic Nash game, given x_{-i}, the ith player solves the following parametrized convex program:

$$\begin{aligned} \min_{x_i \in K_i} \quad & \mathbb{E}[f_i(x_i,\xi_i,\omega_i) + c(x,\varsigma)] \\ \text{s.t.} \quad & Ax \leq \mathbb{E}[C(\zeta)], \end{aligned} \tag{3.31}$$

Since, the strategy sets are coupled by a set of shared constraints, the associated game is a generalized Nash game with shared constraints. Suppose x denotes an equilibrium of this shared constraint game. Then, under convexity assumptions on the player problems and the polyhedrality of the shared constraints, there exist vectors $\lambda^1, \ldots, \lambda^N$, such that

$$\begin{aligned} & (y_i - x_i)^T \nabla_{x_i} \mathbb{E}[f_i(x_i,\xi_i,\omega_i) + c(x,\varsigma) \\ & \qquad + \sum_{l=1}^{|\mathscr{L}|} \lambda_l^i (A_{l,i}x_i - C_l(\zeta))] \geq 0, \ \forall y_i \in K_i, \ i = 1,\ldots,N, \\ & 0 \leq \lambda_l^i \perp \sum_{i=1}^{N} \mathbb{E}[A_{li}x_i - C_l(\zeta)] \leq 0, \\ & \qquad l = 1,\ldots,|\mathscr{L}|, \ i = 1,\ldots,N. \end{aligned} \tag{3.32}$$

While the equilibria of this shared-constraint game are wholly captured by the solution set of a quasi-variational inequality, a subset of equilibria (referred to as *variational equilibria* (VE)) is characterized by common Lagrange multipliers associated with the shared constraint; more specifically, a VE is given by an x that solves (3.32) with $\lambda = \lambda^1 = \lambda^2 = \ldots = \lambda^N$. A variational equilibrium is obtainable by solving a suitably defined variational inequality problem. Defining such a problem requires introducing a uniform pricing mechanism, controlled by the network administrator, that allows for relaxing the shared constraints. If such a price is denoted by a Lagrange multiplier λ, player i solves:

$$\min_{x_i \in K_i} \quad \mathbb{E}[f_i(x_i, \xi_i, \omega_i) + c(x, \varsigma) + \sum_{l=1}^{|\mathscr{L}|} \lambda_l (A_{l,i} x_i - C_l(\zeta))],$$

where the expectation is with respect to ξ_i, ω_i, ς and ζ and the price vector $\lambda \triangleq (\lambda_l)_{l=1}^{|\mathscr{L}|}$ such that $\lambda \in \mathbb{R}_+^{|\mathscr{L}|}$ and satisfies the complementarity relationship given by

$$0 \leq \lambda_i \perp A_{l,i} x_i - \mathbb{E}[C_l(\zeta)] \leq 0, \qquad l = 1, \ldots, |\mathscr{L}|.$$

It is important to note that in this modified definition of player i's payoff, the decision of other players x_{-i} and the price vector λ are to be viewed as parameters. Since $f_i(x_i, \xi, \omega_i) = -\xi_i \log(1 + x_i + w_i)$, the gradient vector for player i is:

$$\begin{aligned} F_i(x, \lambda) &\triangleq \nabla_{x_i} \mathbb{E}[f_i(x_i, \xi_i, \omega_i) + c(x, \varsigma) + \sum_{l=1}^{|\mathscr{L}|} \lambda_l (A_{l,i} x_i - C_l(\zeta))] \\ &= \mathbb{E}\left[-\frac{\xi_i}{1 + x_i + \omega_i} \right] + \mathbb{E}\left[2\varsigma \sum_{j=1}^{N} (A^T A)_{ij} x_j + \sum_{l=1}^{|\mathscr{L}|} \lambda_l A_{l,i} \right]. \end{aligned}$$

Furthermore, the expected violation of the shared constraints is denoted by the mapping $\Lambda(x, \lambda)$ which is defined as

$$\Lambda(x, \lambda) \triangleq \begin{pmatrix} \sum_{i=1}^{N} (\mathbb{E}[C_1(\zeta)] - A_{1i} x_i) \\ \vdots \\ \sum_{i=1}^{N} (\mathbb{E}[C_{|\mathscr{L}|}(\zeta)] - A_{|\mathscr{L}|i} x_i) \end{pmatrix}.$$

Based on [79, Th. 3.1], under the convexity assumptions on the player problems, x is an equilibrium of the shared-constraint Nash game at which (3.32) holds with $\lambda = \lambda^1 = \lambda^2 = \ldots = \lambda^N$ if and only if (x, λ) solves $\mathrm{VI}(K \times \mathbb{R}_+^{|\mathscr{L}|}, \Phi)$, where $\Phi(z) = \Phi(x, \lambda) \triangleq (F_1^T, \ldots, F_N^T, \Lambda^T)^T$.

It remains to verify that the mapping Φ is monotone over the set $K \times \mathbb{R}_+^{|\mathscr{L}|}$. Its Jacobian is given by

$$\nabla \Phi(z) = \begin{pmatrix} \tilde{H}(x) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} 2\mathbb{E}[\varsigma] A^T A & A^T \\ -A & \mathbf{0} \end{pmatrix},$$

where $\tilde{H}(x) = \mathbb{E}[H(x,\xi,\omega)]$ and

$$H(x,\xi,\omega) = \begin{pmatrix} \frac{\xi_1}{(1+x_1+\omega_1)^2} & & \\ & \ddots & \\ & & \frac{\xi_N}{(1+x_N+\omega_N)^2} \end{pmatrix}.$$

Since the matrix $\tilde{H}(x)$ arises as the Hessian of user-specific utilities which are independent across users, accordingly, the Hessian is a diagonal positive definite matrix. For $z \in K \times \mathbb{R}_+^{|\mathscr{L}|}$, consider $z^T \nabla\Phi(z) z$, we have $z^T \nabla\Phi(z) z = x^T \tilde{H}(x) x + 2\mathbb{E}[\varsigma] x^T A^T A x \geq 0$, where we use $\mathbb{E}[\varsigma] = 3/4$ as $\varsigma \sim$ U(1/2,1). Thus from the positive semidefiniteness of $\nabla\Phi(z)$ it follows that Φ is monotone over $K \times \mathbb{R}_+^{|\mathscr{L}|}$. Given the absence of strong monotonicity and the lack of compactness of $K \times \mathbb{R}_+^{|\mathscr{L}|}$, existence and uniqueness claims, while not immediate, can be derived (cf. [80]). For the present, we assume that an equilibrium does indeed exist. Given these definitions, the standard projection method is defined as

$$\begin{aligned} x_i^{k+1} &= \Pi_{K_i}[x_i^k - \alpha_{k,i} F_i(x,\lambda)], && \text{for all } i = 1,\ldots,N, \\ \lambda_l^{k+1} &= \Pi_{\mathbb{R}_+}[\lambda_l^k - \alpha_{k,N+l} \Lambda_l(x,\lambda)], && \text{for all } l = 1,\ldots,|\mathscr{L}|, \end{aligned}$$

where users and links have their steplength sequences and the regularized counterparts may be appropriately defined. Compactly, the preceding iterations may be written as

$$z^{k+1} = \Pi_{K \times \mathbb{R}_+^{|\mathscr{L}|}}[z^k - D(\alpha_k)\Phi(z^k)],$$

where $\alpha_k = (\alpha_{k,1},\ldots,\alpha_{k,N+|\mathscr{L}|})$ and $z \triangleq (x,\lambda)$. It is worth noting that our convergence theory for PIPP requires strict monotonicity of Φ while our mapping is merely monotone; yet our numerical results suggests that PIPP still performs well.

We now describe our experimental setup. Unless mentioned otherwise, we terminate each of our simulations after 10,000 iterations and obtain 95% confidence intervals by using 100 sample-paths. We report the confidence intervals at a 95% level for the normed error between the terminating iterate and equilibrium solution, i.e., the confidence interval for $\|z^k - z^*\|$ where $k = 10,000$. For the ITR method, the update rule for the regularization parameter is taken to be of the form $\varepsilon_k = (1000+k)^{-a}$ where $k \geq 1$ is the current iterate. When implementing the IPP method, the proximal parameter is updated using $\theta_k = (1000+k)^c$ where $k \geq 1$. The steplength α_k updated as $\alpha_k = (1000+k)^{-b}$, is chosen to be the same for both ITR and IPP methods. In both partially coordinated methods, for $i = 1,\ldots,N+|\mathscr{L}|$, we let $\alpha_{k,i} = (1000+k+\delta_i)^{-b}$ with $\varepsilon_{k,i} = (1000+k+\delta_i)^{-a}$ for

PITR and $\theta_{k,i} = (1000 + k + \delta_i)^{-b}$ for PIPP, where $\delta_i \sim U(-500, 500)$. Finally, z^* denotes an approximate solution of VI($K \times \mathbb{R}_+^{|\mathscr{L}|}, \Phi$) computed by solving a sample-average approximation (SAA) problem using the nonlinear programming solver `knitro` [67] on Matlab 7. Note that $x \in [0, 1 + C_{\max}]$, where $C_{\max} = \max_{l \in \mathscr{L}} \{\bar{C}_l\}$ and $Ax \leq \mathbb{E}[C(\zeta)]$. For a sample size of 200, with 2000 replications for each sample, we observe that $\|z^*\| = 0.808$. We now summarize our numerical algorithm for users iterate update.

1. For each sample path, at the beginning of each iteration, we draw a random sample of ξ, ω, ς and $C(\zeta)$ from their respective distributions.
2. Using this random sample and its current iterate, each user generates a sample of a gradient.
3. The next iterate is generated using the proposed algorithm with the sampled gradient combined with appropriate stepsize and regularization parameter.
4. Repeat steps 1–3 for $\tilde{k}$ iteration; record the error $\|z^{\tilde{k}} - z^*\|$ for a particular sample.
5. The sample mean error and a 95% confidence interval is obtained and reported.

3.4.2 Sensitivity to parameters

We consider cases when the regularization parameter sequence is driven to zero at different rates. Specifically, we choose ε_k as $(1000+k)^{-a}$ for ITR and for $i = 1, \ldots, N + \mathscr{L}$, $\varepsilon_{k,i}$ is $(1000 + k + \delta_i)^{-a}$ where $\delta_i \sim U(-500, 500)$ for PITR. The user and link stepsizes α_k are set to $(1000+k)^{-0.54}$ for ITR and $\alpha_{k,i} = (1000 + k + \delta_i)^{-0.54}$ for PITR. Table 3.2 compares the 95% confidence interval for normed error $\|z^{\tilde{k}} - z^*\|$ of ITR method to that of PITR method, as a function of the parameter a of the regularization stepsizes ε_k and $\varepsilon_{k,i}$. It can be seen that as a increases, the confidence intervals tend to be tighter upon termination. Further, we also observe that when users and links choose their steplengths independently, the resulting confidence intervals appear to be slightly better.

Next, we examine the performance of iterative proximal-point methods. Table 3.3 compares the performance of IPP and PIPP methods when the rate of decay or growth of the prox parameter is varied keep the users and links steplength update rule fixed. Specifically, we let $\theta_k = (1000+k)^c$ with $\alpha_k = (1000+k)^{-0.54}$ for IPP, and for $i = 1, \ldots, N + |\mathscr{L}|$, $\theta_{k,i} = (1000 + k + \delta_i)^c$ with $\alpha_k = (1000 + k + \delta_i)^{-0.54}$ and $\delta_i \sim U(-500, 500)$ for PIPP. Note that when $c > 0$ we have $\theta \nearrow \infty$. No clear relationship can be observed between c (rate control parameter) and the recorded accuracy upon termination though it seems that letting $c > 0$ results in a slightly better accuracy. We also note that limited coordination has minimal impact on the obtained confidence intervals.

Table 3.2: Varying a in regularization term of the form k^{-a} for a fixed choice for stepsizes

Width of confidence intervals		
a	ITR	PITR
0.25	1.29e-02	1.09e-02
0.30	1.19e-02	1.08e-02
0.35	1.15e-02	1.07e-02
0.40	1.14e-02	1.07e-02
0.45	1.10e-02	1.06e-02

Table 3.3: Varying c in prox-parameter of the form k^{-c} for a fixed choice of the stepsizes

Width of confidence intervals		
c	IPP	PIPP
-0.35	1.23e-02	1.13e-02
-0.15	1.18e-02	1.04e-02
0	1.16e-02	1.17e-02
0.15	9.46e-02	1.07e-02
0.35	1.04e-02	9.62e-02

Now, we examine the behavior of ITR and IPP methods by changing common parameters.

1. *Impact of steplength, regularization and proximal-parameter sequences:* We begin by examining the impact on the width of confidence interval of the rate at which stepsize α_k decays to zero in both ITR and IPP methods. Further, we report the computational time to achieve the desired level of accuracy. The table to the right in Table 3.4 compares the width of confidence intervals in both methods. It can be observed that as the decay rates increase, IPP performs slightly better than ITR in terms of accuracy upon termination. Since there is relatively limited impact on ITR methods from changing the decay rate of the regularization parameter (see Table 3.2), given that $a+b<1$ choosing b as close to 1 appears to be advantageous. Note the slight change in the update rule of ε_k so as to accommodate a larger range of variability for b. In Table 3.4 (right), we list the computation times required by ITR and IPP to achieve the corresponding level of accuracy of right table in Table 3.4. Combining the results of Table 3.4 with that of Table 3.2 it can be concluded that for iterative Tikhonov methods, it might be useful to choose a faster decay rate for α_k.

2. *Varying coordination requirements:* A worthwhile question in examining limited coordination generalizations is the extent to which disparity in steplength and parameter sequences impacts the overall confidence width. In Table 3.5, we tabulate the performance of varying the coordination amongst users and links by changing the deviation in their individual

Table 3.4: ITR vs. IPP: Varying b in $\alpha_k = k^{-b}$ with $\varepsilon_k = (1000+k)^{-0.25}$ and $\theta_k = (1000+k)^{0.35}$.

Width of confidence intervals		
b	ITR	IPP
0.54	1.16e-02	1.03e-02
0.59	8.89e-03	8.88e-03
0.64	7.20e-03	7.36e-03
0.69	7.08e-03	5.53e-03
0.74	4.84e-03	4.48e-03

Computational time in seconds		
b	ITR	IPP
0.54	164.71	128.76
0.59	165.98	129.47
0.64	165.95	129.62
0.69	165.85	129.67
0.74	165.94	129.54

Table 3.5: Performance of PITR and PIPP for various levels of coordination δ_i.

Delta	95 % Confidence Interval	
δ_i	PITR	PIPP
$U(-50,50)$	$1.19e-02$	$9.08e-03$
$U(-100,100)$	$1.08e-02$	$9.49e-03$
$U(-200,200)$	$9.74e-03$	$1.07e-02$
$U(-500,500)$	$1.14e-02$	$8.25e-03$

steplengths and parameters which is controlled by changing the size of the support of the uniform distribution governing parameter δ_i. Notably, our tests show that within the range of testing conducted, there is relatively minor impact associated with limited coordination.

3.4.3 Comparison with standard Tikhonov and proximal-point methods

Our methods are motivated by the observation that regularization-based algorithms that rely on obtaining increasingly accurate solutions to a sequence of problems and such techniques cannot be easily extended to regimes where the subproblems are stochastic, particularly when relying on simulation-based methods. Naturally, getting solutions of increasing accuracy requires increasing replication lengths at a much faster rate, making such approaches computationally impractical. Iterative regularization methods obviate this challenge by requiring a replication in which the regularization parameter is updated during the replication.

In this subsection, we detail the insights drawn from a rudimentary bounded complexity implementation of the standard Tikhonov and proximal-point methods (with two-nested loops). In effect, we obtain solutions of fixed accuracy and not increasing accuracy. More specifically, we examined the behavior of a Tikhonov regularization method where a sequence of subproblems was solved and the method was initiated with $\varepsilon = 1$ and was terminated when the regularization parameter ε dropped below $(11000)^{-0.35}$ (terminating value of the base case of corresponding iterative

Tikhonov method). Note that the regularized subproblem was solved via a simulation method in which the averaged solution over a fixed number of sample-paths was employed (100 in this case), each of which required $10,000$ steps. In the context of the proximal-point method, we terminate the method when the normed error for numerically obtained solution and actual solution drops below 1e-2. Note that both of these are heuristics and other parameter settings were examined to get a flavor for the behavior of such methods, as well as to make the comparison as fair as possible. Table 3.6 tabulates the performance for both methods in terms of width of confidence interval (table on the left) and the computational time required to reach the desired level of accuracy (table on the right). On comparing with corresponding data of iterative methods in Table 3.4 we notice that the both Tikhonov and proximal-point methods display almost identical performance in terms of level of accuracy upon termination but require significant effort to do so, especially Tikhonov methods.

Table 3.6: Tikhonov v/s proximal point method: Varying b in $\alpha_k = k^{-b}$.

Width of confidence intervals		
b	Tikhonov	Proximal Point
0.54	1.58e-02	1.62e-02
0.59	1.30e-02	1.38e-02
0.64	1.15e-02	1.03e-02
0.69	8.57e-03	8.10e-03
0.74	6.85e-03	7.27e-03

Computational time in seconds	
Tikhonov	Proximal Point
297.62	183.08
360.30	138.10
360.32	137.62
360.36	137.56
360.26	137.46

3.4.4 Comparison with sample average approximation (SAA) techniques

In this subsection, we investigate the performance of SAA methods with one of the candidate methods namely, PIPP. We report a 95% confidence interval SAA obtained for a sample of size 100 for various replication levels to approximate the problem. For each sample, the replicated averaged problem is solved using `knitro` and for $i = 1, \ldots, N + |\mathscr{L}|$, PIPP is implemented with $\alpha_{k,i} = (1000 + k + \delta_i)^{-0.54}$ and $\theta_{k,i} = (1000 + k + \delta_i)^{0.35}$ where $\delta_i \sim U(-500, 500)$. Table 3.7 demonstrates the performance of SAA with accuracy level reached for various number of replications while Table 3.8 compares the computational effort required measured in time (seconds) by SAA to reach the accuracy level to that of PIPP.

Table 3.7: Performance for SAA

Width of confidence intervals for 100 samples	
Replication per sample	SAA
1000	3.86e-03
2000	2.32e-03
5000	1.65e-03

Table 3.8: Comparison of performance for SAA and PIPP

PIPP Iteration / SAA Replication	Computational time in seconds	
	SAA	PIPP
1000	1438.93	20.65
2000	3598.07	41.32
5000	14,645.32	103.27

3.5 Concluding remarks

In this chapter we proposed and investigated two algorithms for computing solutions to stochastic variational inequalities when the mappings are not necessarily strongly monotone. The work in this chapter is related to the past work by Jiang and Xu [36] who considered how stochastic approximation procedures could address stochastic variational inequalities with strongly monotone mappings. Yet, these methods cannot easily contend with weaker requirements (such as strict monotonicity or monotonicity) while retaining the single iteration structure. A simple regularization-based extension leads to a two-level method, that is usually harder to implement in networked settings.

Accordingly, this chapter makes the following contributions. First, we present *single-loop iterative* counterparts of standard Tikhonov and proximal-point methods that obviate the need to solve a sequence of subproblems. Instead, we present a stochastic iterative Tikhonov regularization method and a stochastic iterative proximal-point method in which the regularization parameter in the former and the centering parameter in the latter are updated at *every* iteration. Suitable conditions on the parameter sequences are established for guaranteeing the almost-sure convergence of the resulting methods. Notably, the iterative proximal-point method also allows for raising the proximal-parameter at every step.

The chapter concludes with a detailed study of the computational performance of these methods on a networked monotone stochastic rate allocation game. Through this case study, we observe that the methods perform better when the steplength sequences are driven to zero at a faster rate but are less sensitive to changing the decay rates of the regularization and proximal parameter sequences. Notably, partial coordination of steplength choices has minimal impact on the accuracy of the solution. Finally, naive implementations of standard Tikhonov and proximal-point methods prove illuminating; both Tikhonov and proximal methods provide accurate solutions but at a significant computational expense.

Chapter 4

Network Aggregative Nash Games

In this chapter we focus on a subclass of Nash games, referred to as aggregative games. An aggregative game is a non-cooperative Nash game in which each player's payoff is parametrized by its action and the aggregate of the actions taken by all players [81]. Nash-Cournot games represent one instance of such games; here, firms make quantity bids that fetch a price based aggregate quantity sold, implying that the payoff of any player is a function of the aggregate quantity sold [17, 48]. This chapter considers aggregative games wherein the players[1] compete over a network. However, distributed computation of equilibria in such games is complicated by two challenges. First, the connectivity graphs of the underlying network may evolve over time. Second, in many settings, agents do not have ready access to aggregate decisions, implying that agents cannot compute their exact payoffs (or their gradients). Consequently, standard gradient-based or best-response schemes cannot be directly implemented since agents do not have ready access to the aggregate. Accordingly, in this chapter, we propose two distributed agreement-based algorithms which allow agents to build estimates of the aggregate and consequently compute an equilibrium of aggregative games. Of these, the first is a synchronous algorithm where all agents update simultaneously, while the second is a gossip-based algorithm that allows for asynchronous computation:

(a) *Synchronous distributed algorithm:* At each epoch, every agent performs a "learning step" to update its estimate of the aggregate using the information obtained through the time-varying states of its neighbors. All agents exchange information and perform decision updates simultaneously. This algorithm builds on the ideas of the method developed in [82] for distributed optimization problems.

(b) *Asynchronous distributed algorithm:* In contrast, the asynchronous algorithm uses a gossip-based protocol for information exchange. In the gossip-based algorithm, only a pair of randomly selected agent exchange their information and update their estimates of the aggregate

[1] Recall that we also refer player to as agent.

and their decisions. The algorithm combines our synchronous method in (a) with the gossip technique proposed in [83] for the agreement (consensus) problem [2].

We investigate the convergence behavior of these algorithms under a diminishing stepsize rule, and provide error bounds under a constant steplength regime. Additionally, the results are supported numerics derived from application of the proposed schemes on a class of networked Nash-Cournot games. The novelty of this work is in our examination of distributed (neighbor-based) algorithms for computation of a Nash equilibrium point for aggregative Nash games, while the majority of preceding efforts on such algorithms have been spent towards solving feasibility and optimization problems.

The distributed algorithms presented in this chapter draw inspiration from the seminal work in [49], where a distributed method for optimization has been developed by allowing agents to communicate locally with their neighbors over a time-varying communication network. This idea has attracted a lot of attention recently in an effort to extend the algorithm of [49] to more general and broader range of problems [85, 86, 87, 88, 89, 90, 91, 92, 93, 94]. Much of the aforementioned work focuses on optimizing the sum of local objective function [85, 86, 87, 88, 89, 90, 91] in a multi-agent networks, while a subset of recent work considered the min-max optimization problem [95, 96], where the objective is to minimize the maximum cost incurred by any agent in the network. Notably, extensions of consensus based algorithms have also been studied in the domain of distributed regression [93], estimation and inference tasks [92, 94]. The work in this paper extends the realm of such algorithm to capture competitive aspect of multi-agent networks.

In section 4.1, we describe the problem of interest, state our assumptions and the equilibrium conditions of the game. A synchronous distributed algorithm is proposed in section 4.2 and convergence theory is provided. An asynchronous gossip-based variant of this algorithm is described in section 4.3 and is supported by convergence theory and error analysis. In section 4.4, we present an extension of aggregative games and suitably adapt the distributed synchronous and asynchronous algorithm to address this generalization. We present some numerical results in section 4.5 and, finally, conclude in section 4.6.

4.1 Problem formulation and background

In this section we introduce an aggregative game of our interest and provide its sufficient equilibrium conditions. The players in this game are assumed to have local interactions with each

[2] A subset of this work appears in [84].

other over time, where these interactions are modeled by time-varying connectivity graphs. We also discuss some auxiliary results for the players' connectivity graphs and present our distributed algorithm for equilibrium computation.

Consider a set of N players (or agents) indexed by $1,\ldots,N$, and let $\mathcal{N} = \{1,\ldots,N\}$. The ith player is characterized by a strategy set $K_i \subseteq \mathbb{R}^n$ and a payoff function $f_i(x_i,\bar{x})$, which depends on player i decision x_i and the aggregate $\bar{x} = \sum_{i=1}^N x_i$ of all player decisions. To formalize the game, let $\bar{K}$ denote the Minkowski sum of the sets K_i:

$$\bar{K} = \sum_{i=1}^N K_i. \tag{4.1}$$

In a generic aggregative game, player i faces the following parametrized optimization problem:

$$\begin{aligned} &\text{minimize} && f_i(x_i,\bar{x}) \\ &\text{subject to} && x_i \in K_i, \end{aligned} \tag{4.2}$$

where $K_i \subseteq \mathbb{R}^n$ and $\bar{x}$ is the aggregate of the agent's decisions x_i, i.e.,

$$\bar{x} = \sum_{j=1}^N x_j, \qquad \bar{x} \in \bar{K}, \tag{4.3}$$

with $\bar{K} \subseteq \mathbb{R}^n$ as given in (4.1), and $f_i : K_i \times \bar{K} \to \mathbb{R}$. The set K_i and the function f_i are assumed to be known by agent i only.

4.1.1 Equilibrium conditions and assumptions

To articulate sufficiency conditions, we make the following assumptions on the constraint sets K_i and the functions f_i.

Assumption 8. *For each $i = 1,\ldots,N$, the set $K_i \subset \mathbb{R}^n$ is compact and convex. Each function $f_i(x_i,y)$ is continuously differentiable in (x_i,y) over some open set containing the set $K_i \times \bar{K}$, while each function $x_i \mapsto f_i(x_i,\bar{x})$ is convex over the set K_i.*

Under Assumption 8, the (sufficient) equilibrium conditions of the Nash game in (4.2) can be specified as a variational inequality problem VI(K,ϕ) (cf. [35]). Recall that VI(K,ϕ) requires determining a point $x^* \in K$ such that

$$(x - x^*)^T \phi(x^*) \geq 0 \qquad \text{for all } x \in K,$$

where

$$\phi(x) \triangleq \begin{pmatrix} \nabla_{x_1} f_1(x_1, x_1 + \sum_{j=2}^{N} x_j) \\ \vdots \\ \nabla_{x_N} f_N(x_N, \sum_{j=1}^{N-1} x_j + x_N) \end{pmatrix}, \qquad K = \prod_{i=1}^{N} K_i, \tag{4.4}$$

with $x \triangleq (x_1^T, \ldots, x_N^T)^T$, $x_i \in K_i$ for all i. Note that, by Assumption 8, the set K is a compact and convex set in $\mathbb{R}^{nN}$, and the mapping $\phi : K \to \mathbb{R}^{nN}$ is continuous. To emphasize the particular form of the mapping ϕ, we define $F_i(x_i, \bar{x})$ as follows:

$$F_i(x_i, \bar{x}) = \nabla_{x_i} f_i(x_i, \bar{x}) \qquad \text{for all } i = 1, \ldots, N. \tag{4.5}$$

The mapping $F(x, u)$ is given by

$$F(x, u) \triangleq \begin{pmatrix} F_1(x_1, u) \\ \vdots \\ F_N(x_N, u) \end{pmatrix}, \tag{4.6}$$

where the component maps $F_i : K_i \times \bar{K} \to \mathbb{R}^n$ are given by (4.5). With this notation, we have

$$\phi(x) = F(x, \bar{x}) \qquad \text{for all } x \in K. \tag{4.7}$$

Next, we make an assumption on the mapping $\phi(x)$.

Assumption 9. *The mapping $\phi(x)$ is strictly monotone over K, i.e.,*

$$(\phi(x) - \phi(x'))^T (x - x') > 0, \quad \textit{for all } x, x' \in K.$$

Together with the compactness of K, this assumption allows one to claim existence and uniqueness of a Nash equilibrium.

Proposition 10. *Consider the aggregative Nash game defined in* (4.2)*. Suppose Assumptions 8 and 9 hold. Then, the game admits a unique Nash equilibrium.*

Proof. By Assumption 8, the set K is compact and ϕ is continuous. It follows from Corollary

2.2.5 [35] that $\text{VI}(K,\phi)$ has a solution. By the strict monotonicity of $\phi(x)$, $\text{VI}(K,\phi)$ has at most one solution based on Theorem 2.3.3 [35] and uniqueness follows. ■

Strict monotonicity assumptions on the mapping are seen to hold in a range of practical problem settings, including Nash-Cournot games [20], rate allocation problems [5, 7, 6], amongst others.

We now state our assumptions on the mappings F_i, which are related to the coordinate mappings of ϕ in (4.4).

Assumption 10. *Each mapping $F_i(x_i,u)$ is uniformly Lipschitz continuous in u over $\bar{K}$, for every fixed $x_i \in K_i$ i.e., for some $L_{-i} > 0$ and for all $u,z \in \bar{K}$,*

$$\|F_i(x_i,u) - F_i(x_i,z)\| \leq L_{-i}\|u - z\|.$$

One would naturally question whether such assumptions are seen to hold in practical instances of aggregative games. We will show in section 4.5 that the assumptions are satisfied for the Nash-Cournot game of Example 2.

Before proceeding, it is worthwhile to recall the motivation for the present work. In the context of continuous-strategy Nash games, when the mapping ϕ satisfies a suitable monotonicity property over K, then a range of distributed projection-based schemes [35, 7, 5, 13, 33] and their regularized variants schemes [6, 75, 20] can be constructed. In all of these instances, every agent should be able to observe the aggregate $\bar{x}$ of the agent decisions. In this paper, we assume that this aggregate *cannot be observed* and *no central entity exists* that can provide this quantity at any time. Yet, when agents are connected in some manner, then a given agent can communicate locally with their neighbors and generate estimates of the aggregate decisions. Under this restriction, *we are interested in designing algorithms for computing an equilibrium of an aggregative Nash game* (4.2).

4.2 Distributed synchronous algorithm

In this section we develop a distributed synchronous algorithm for equilibrium computation of the game in (4.2) that relies on agents constructing an estimate by *mixing* information drawn from local neighbors and making a subsequent projection step. In Section 4.2.1, we describe the scheme and provide some preliminary results in Section 4.2.2. The section concludes in Section 4.2.3 with an analysis of the convergence of the proposed scheme.

4.2.1 Outline of algorithm

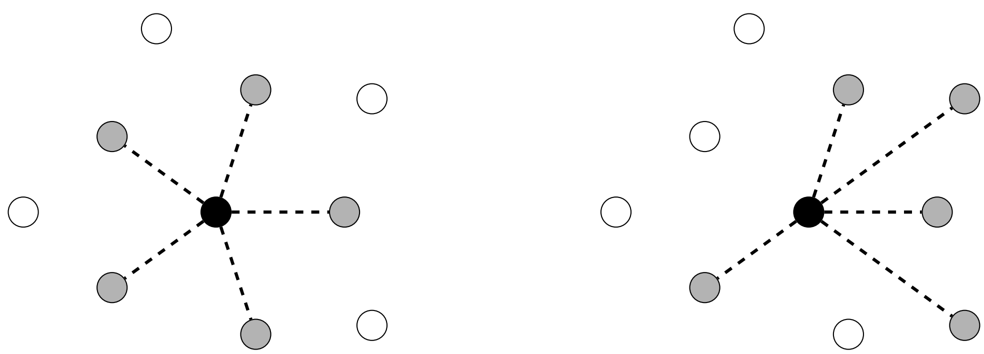
Our algorithm equips each agent in the network with a protocol that mandates that every agent exchange information with its neighbors, and subsequently update its decision and the estimate of the aggregate decisions, simultaneously. We employ a synchronous time model which can contend with a time varying connectivity graph. Consequently, in this section we consider a time varying network to model agent's communications in time. More specifically, let $\mathcal{E}_k$ be the set of underlying directed edges between agents and let $\mathcal{G}_k = (\mathcal{N}, \mathcal{E}_k)$ denote the connectivity graph at time k. Let $\mathcal{N}_i(k)$ denote the set of agents who are immediate neighbors of agent i at time k that can send information to i, assuming that $i \in \mathcal{N}_i(k)$ for all $i \in \mathcal{N}$ and all $k \geq 0$. Mathematically, $\mathcal{N}_i(k)$ can be expressed as:

$$\mathcal{N}_i(k) = \{j : (j,i) \in \mathcal{E}_k\}.$$

We make the following assumption on the graph $\mathcal{G}_k = (\mathcal{N}, \mathcal{E}_k)$.

Assumption 11. *There exists an integer $Q \geq 1$ such that the graph $(\mathcal{N}, \bigcup_{\ell=1}^{Q} \mathcal{E}_{\ell+k})$ is strongly connected for all $k \geq 0$.*

This assumption ensures that the intercommunication intervals are bounded for agents that communicate directly; i.e., every agent sends information to each of its neighboring agents at least once every Q time intervals. This assumption has been commonly used in distributed algorithms on networks, starting with [49].



(a) Connectivity graph at one instant

(b) Connectivity graph at another instant

Figure 4.1: A depiction of an (undirected) communication network.

Due to incomplete information at any point, an agent only has an estimate of $\bar{x}$ in contrast to the actual $\bar{x}$. We describe how an agent may build this estimate. Let x_i^k be the iterate and v_i^k be the estimate of the average of the decisions $x_1^k, \ldots, x_N^k$ for agent i at the end of the kth iteration.

At the beginning of the $(k+1)$st iteration, agent i receives the estimates v_j^k from its neighbors $j \in \mathcal{N}_i(k+1)$. Using this information, agent i aligns its intermediate estimate according to the following rule:

$$\hat{v}_i^k = \sum_{j \in \mathcal{N}_i(k)} w_{ij}(k) v_j^k, \tag{4.8}$$

where $w_{ij}(k)$ is the nonnegative weight that agent i assigns to agent j's estimate. By specifying $w_{ij} = 0$ for $j \notin \mathcal{N}_i(k)$ we can write:

$$\hat{v}_i^k = \sum_{j=1}^{N} w_{ij}(k) v_j^k \qquad \text{with} \qquad v_j^0 = x_j^0 \text{ for all } j = 1, \ldots, N.$$

Using this aligned average estimate $\hat{v}_i^k$ and its own iterate x_i^k, agent i updates its iterate and average estimate as follows:

$$x_i^{k+1} = \Pi_{K_i}[x_i^k - \alpha_k F_i(x_i^k, N\hat{v}_i^k)], \tag{4.9}$$

$$v_i^{k+1} = \hat{v}_i^k + x_i^{k+1} - x_i^k, \tag{4.10}$$

where, α_k is the stepsize, Π_{K_i} denotes the Euclidean projection onto the set K_i and F_i is as defined in (4.5). The quantity $N\hat{v}_i^k$ in (4.9) is the aggregate estimate that agent i uses instead of the true estimate $\sum_{i=1}^N x_i^k$ of the agent decisions at time k. Under suitable conditions on the agents weights $w_{ij}(k)$ and the stepsize α_k, the iterate vector $(x_1^k, \ldots, x_N^k)$ can converge to a Nash equilibrium point $(x_1^*, \ldots, x_N^*)$ and the estimates $N\hat{v}_i^k$ in (4.9) will converge to the true aggregate value $\sum_{i=1}^N x_i^*$ at the equilibrium. These assumptions are given below.

Assumption 12. *For all $i \in \mathcal{N}$ and all $k \geq 0$, the following hold:*

(i) $w_{ij}(k) \geq \delta$ *for all* $j \in \mathcal{N}_i(k)$ *and* $w_{ij}(k) = 0$ *for* $j \notin \mathcal{N}_i(k)$;

(ii) $\sum_{j=1}^N w_{ij}(k) = 1$ *for all* i;

(iii) $\sum_{i=1}^N w_{ij}(k) = 1$ *for all* j.

Assumption 13. *The stepsize α_k is chosen such that the following hold:*

(i) The sequence $\{\alpha_k\}$ *is monotonically non-increasing i.e.,* $\alpha_{k+1} \leq \alpha_k$ *for all* k;

(ii) $\sum_{k=0}^{\infty} \alpha_k = \infty$;

(iii) $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.

Such an assumption is trivially satisfied for a stepsize update of the form $\alpha_k = (k+1)^{-b}$ where $0.5 < b \leq 1$.

4.2.2 Preliminary results

We next provide some auxiliary results for the weight matrices and the estimates generated by the method. Let $W(k)$ be the weight matrix with entries $w_{ij}(k)$. We introduce the transition matrices $\Phi(k,s)$ from time s to $k > s$, as follows:

$$\Phi(k,s) = W(k)W(k-1)\cdots W(s+1)W(s) \qquad \text{for } 0 \leq s < k,$$

where $\Phi(k,k) = W(k)$ for all k. Let $[\Phi(k,s)]_{ij}$ denote the (i,j)th entry of the matrix $\Phi(k,s)$, and let $\mathbf{1} \in \mathbb{R}^N$ be the column vector with all entries equal to 1. We next state a result on the convergence properties of the matrix $\Phi(k,s)$. The result can be found in [97] (Corollary 1).

Lemma 12 ([97] Corollary 1)**.** *Let Assumptions 11 and 12 hold. Then*

(i) $\lim_{k\to\infty} \Phi(k,s) = \frac{1}{N}\mathbf{1}\mathbf{1}^T$ *for all* $s \geq 0$.

(ii) The convergence rate of $\Phi(k,s)$ *is geometric; specifically, we have* $\left|[\Phi(k,s)]_{ij} - \frac{1}{N}\right| \leq \theta\beta^{k-s}$ *for all* $k \geq s \geq 0$ *and for all* i *and* j, *where* $\theta = (1 - \frac{\delta}{4N^2})^{-2}$ *and* $\beta = (1 - \frac{\delta}{4N^2})^{\frac{1}{Q}}$.

Now, we state some results which will allow us to claim the convergence of the algorithm. These results involve the average y^k of the estimates $v_i^k, i \in \mathcal{N}$, given by

$$y^k = \frac{1}{N}\sum_{i=1}^{N} v_i^k \quad \text{for all } k \geq 0. \tag{4.11}$$

As we will see, y^k will play a key role in establishing the convergence of the iterates produced by the algorithm in (4.9)–(4.10). One important property of y^k is that we have $y^k = \frac{1}{N}\sum_{j=1}^{N} x_j^k$ for all $k \geq 0$. Thus, y^k not only captures the average belief of the agents in the network but it also represents the true average information. This property of the true average y^k has been shown in [93] within the proof of Lemma 5.2 for a different setting, and it is given in the following lemma for sake of clarity.

Lemma 13. *Let* $W(k)$ *be such that* $\sum_{j=1}^{N}[W(k)]_{ji} = 1$ *for every* i *and* k. *Then,* $y^k = \frac{1}{N}\sum_{i=1}^{N} x_i^k$ *for all* $k \geq 0$, *where* y^k *is defined by* (4.11).

Proof. It suffices to show that for all $k \geq 0$,

$$\sum_{j=1}^{N} v_j^k = \sum_{j=1}^{N} x_j^k. \tag{4.12}$$

We show this by induction on k. For $k = 0$ relation (4.12) holds trivially, as we have initialized the beliefs with $v_j^0 = x_j^0$ for all j. Assuming relation (4.12) holds for $k-1$, as the induction step, we have

$$\begin{aligned}
\sum_{j=1}^{N} v_j^k &= \sum_{j=1}^{N} (\hat{v}_j^{k-1} + x_j^k - x_j^{k-1}) \\
&= \sum_{j=1}^{N} \sum_{i=1}^{N} [W(k-1)]_{ji} v_i^{k-1} + \sum_{j=1}^{N} (x_j^k - x_j^{k-1}) \\
&= \sum_{i=1}^{N} v_i^{k-1} + \sum_{j=1}^{N} (x_j^k - x_j^{k-1}),
\end{aligned}$$

where the first equality follows from (4.10), the second inequality is a consequence of the mixing relationship articulated by (4.8), and the last equality follows from $\sum_{j=1}^{N} [W(k)]_{ji} = 1$ for every i and k. Furthermore, using the induction hypothesis, we have $\sum_{j=1}^{N} (x_j^k - x_j^{k-1}) = \sum_{j=1}^{N} x_j^k - \sum_{j=1}^{N} v_j^{k-1}$, thus implying that $\sum_{j=1}^{N} v_j^k = \sum_{j=1}^{N} x_j^k$. ■

As a consequence of Lemma 13, Assumption 8 and Assumption 10, we have the following result which will be often used in the sequel.

Corollary 5. *Let $W(k)$ satisfy the assumption of Lemma 13. Also, let Assumptions 8 and 10(ii) hold. Then, there exists a constant C such that*

$$\|F_i(x_i^k, Ny^k)\| \leq C, \qquad \|F_i(x_i^k, N\hat{v}_i^k)\| \leq C \qquad \textit{for all } i \textit{ and } k \geq 0.$$

Proof. By Lemma 13, we have $Ny^k = \sum_{i=1}^{N} x_i^k = \bar{x}_k \in \bar{K}$, where $\bar{K}$ is compact since each K_i s compact (Assumption 8). Since each F_i is continuous over $K_i \times \bar{K}$, the first inequality follows. To show that $\{F_i(x_i^k, N\hat{v}_i^k)\}$ is bounded, we write

$$\|F_i(x_i^k, N\hat{v}_i^k)\| \leq \|F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^k, Ny^k)\| + \|F_i(x_i^k, Ny^k)\|.$$

Using the Lipschitz property of F_i of Assumption 10, we obtain

$$\|F_i(x_i^k, N\hat{v}_i^k)\| \leq L_{-i} N \|\hat{v}_i^k - y^k\| + \|F_i(x_i^k, Ny^k)\|.$$

Let $\hat{K}$ be the convex hull of the union set $\cup_i K_i$. Note that $\hat{v}_i^k, y^k \in \hat{K}$ for all k and that $\hat{K}$ is compact (since each K_i is compact). Thus, $\{\|\hat{v}_i^k - y^k\|\}$ is bounded. As already established, $\{F_i(x_i^k, Ny^k)\}$ is also bounded, implying that $\{F_i(x_i^k, N\hat{v}_i^k)\}$ is bounded as well. ■

In the following lemma, we establish some error bounds for the norms $\|y^k - \hat{v}_i^k\|$ which play important role in our analysis.

Lemma 14. *Let Assumptions 8–12 hold, and let y^k be defined by* (4.11). *Then, we have*

$$\|y^k - \hat{v}_i^k\| \leq \theta\beta^k M + \theta NC \sum_{s=1}^{k} \beta^{k-s}\alpha_{s-1} \qquad \text{for all } i \in \mathcal{N} \text{ and all } k \geq 1,$$

where $\hat{v}_i^k$ is defined in (4.8), $\theta = (1 - \frac{\delta}{4N^2})^{-2}$, $\beta = (1 - \frac{\delta}{4N^2})^{\frac{1}{Q}}$, $M = \sum_{j=1}^{N} \max_{x_j \in K_j} \|x_j\|$ *and C is as in Corollary 5.*

Proof. Using the definitions of v_i^{k+1} and $\hat{v}_i^k$ given in Eqs. (4.10) and (4.8), respectively, we have

$$v_i^{k+1} = \sum_{j=1}^{N} w_{ij}(k) v_j^k + x_i^{k+1} - x_i^k,$$

which through an iterative recursion leads to

$$\begin{aligned}
v_i^{k+1} &= \sum_{j=1}^{N} w_{ij}(k) \left(\sum_{\ell=1}^{N} w_{j\ell}(k-1) v_\ell^{k-1} + x_j^k - x_j^{k-1} \right) + x_i^{k+1} - x_i^k \\
&= \sum_{\ell=1}^{N} [\Phi(k,k-1)]_{i\ell} v_\ell^{k-1} + \sum_{j=1}^{N} [\Phi(k,k)]_{ij} \left(x_j^k - x_j^{k-1} \right) + x_i^{k+1} - x_i^k \\
&= \cdots \\
&= \sum_{\ell=1}^{N} [\Phi(k,0)]_{i\ell} v_\ell^0 + \sum_{s=1}^{k} \left(\sum_{j=1}^{N} [\Phi(k,s)]_{ij} (x_j^s - x_j^{s-1}) \right) + x_i^{k+1} - x_i^k.
\end{aligned}$$

The preceding relation can be rewritten as:

$$v_i^{k+1} - x_i^{k+1} + x_i^k = \sum_{\ell=1}^{N} [\Phi(k,0)]_{i\ell} v_\ell^0 + \sum_{s=1}^{k} \left(\sum_{j=1}^{N} [\Phi(k,s)]_{ij} (x_j^s - x_j^{s-1}) \right).$$

By the definition of v_i^{k+1} in Eq. (4.10), we have $\hat{v}_i^k = v_i^{k+1} - x_i^{k+1} + x_i^k$, through which we get

$$\hat{v}_i^k = \sum_{\ell=1}^{N} [\Phi(k,0)]_{i\ell} v_\ell^0 + \sum_{s=1}^{k} \left(\sum_{j=1}^{N} [\Phi(k,s)]_{ij} (x_j^s - x_j^{s-1}) \right). \tag{4.13}$$

Now, consider y^k which may be written as follows:

$$y^k = y^{k-1} + (y^k - y^{k-1}) = \cdots = y^0 + \sum_{s=1}^{k} (y^s - y^{s-1}).$$

By Lemma 13 we have $y^s = \frac{1}{N} \sum_{j=1}^{N} x_j^s$ for all $s \geq 0$, which implies

$$y^k = y^0 + \sum_{s=1}^{k} \sum_{j=1}^{N} \frac{1}{N} \left(x_j^s - x_j^{s-1} \right) = \sum_{\ell=1}^{N} \frac{1}{N} v_\ell^0 + \sum_{s=1}^{k} \sum_{j=1}^{N} \frac{1}{N} \left(x_j^s - x_j^{s-1} \right), \tag{4.14}$$

where the last equality follows by the definition of y^0 (see (4.11)).

From relations (4.13) and (4.14) we have

$$\begin{aligned}
\|y^k - \hat{v}_i^k\| &= \left\| \sum_{\ell=1}^{N} \left(\frac{1}{N} - [\Phi(k,0)]_{i\ell} \right) v_\ell^0 + \sum_{s=1}^{k} \sum_{j=1}^{N} \left(\frac{1}{N} - [\Phi(k,s)]_{ij} \right) \left(x_j^s - x_j^{s-1} \right) \right\| \\
&\leq \sum_{\ell=1}^{N} \left| \frac{1}{N} - [\Phi(k,0)]_{i\ell} \right| \left\| v_\ell^0 \right\| + \sum_{s=1}^{k} \sum_{j=1}^{N} \left| \frac{1}{N} - [\Phi(k,s)]_{ij} \right| \left\| x_j^s - x_j^{s-1} \right\| \\
&\leq \sum_{\ell=1}^{N} \beta^k \left\| v_\ell^0 \right\| + \sum_{s=1}^{k} \sum_{j=1}^{N} \beta^{k-s} \left\| x_j^s - x_j^{s-1} \right\|
\end{aligned} \tag{4.15}$$

where the last inequality follows from $\left| \frac{1}{N} - [\Phi(k,s)]_{ij} \right| \leq \theta \beta^{k-s}$ for all $0 \leq s \leq k$ (cf. Lemma 12).

Now, we estimate $\|x_i^s - x_i^{s-1}\|$. From relation (4.9) we see that for any $s \geq 1$,

$$\begin{aligned}
\|x_i^s - x_i^{s-1}\| &= \|\Pi_{K_i}[x_i^{s-1} - \alpha_{s-1} F_i(x_i^{s-1}, N\hat{v}_i^{s-1})] - x_i^{s-1}\| \\
&\leq \|x_i^{s-1} - \alpha_{s-1} F_i(x_i^{s-1}, N\hat{v}_i^{s-1}) - x_i^{s-1}\| \\
&= \alpha_{s-1} \|F_i(x_i^{s-1}, N\hat{v}_i^{s-1})\| \\
&\leq C\alpha_{s-1},
\end{aligned} \tag{4.16}$$

where the first inequality follows by the non-expansive property of projection map, and the last inequality follows by Corollary 5. Combining (4.16) and (4.15), we have

$$\|y^k - \hat{v}_i^k\| \leq \theta\beta^k \sum_{\ell=1}^{N} \|v_\ell^0\| + \theta N \sum_{s=1}^{k} \beta^{k-s}\alpha_{s-1}C \leq \theta\beta^k M + \theta NC \sum_{s=1}^{k} \beta^{k-s}\alpha_{s-1},$$

where in the last inequality, we use $v_\ell^0 = x_\ell^0 \in K_\ell$ and $M = \sum_{\ell=1}^{N} \max_{x_\ell \in K_\ell} \|x_\ell^0\|$, which is finite since each K_ℓ is a compact set (cf. Assumption 8). ■

4.2.3 Convergence of Algorithm

In this subsection, under our assumptions, we prove that the sequence produced by the proposed algorithm does indeed converge to the unique Nash equilibrium, which exists by Proposition 10. Our next proposition provides the main convergence result for the algorithm. Prior to providing this result, we state a lemma that will be employed in proving the required result [82, Lemma 3.1(b)].

Lemma 15. *[82, Lemma 3.1(b)] Let $\{\zeta_k\}$ be a non-negative scalar sequence. If $\sum_{k=0}^{\infty} \zeta_k < \infty$ and $0 < \beta < 1$, then $\sum_{k=0}^{\infty} \left(\sum_{s=0}^{k} \beta^{k-s}\zeta_s\right) < \infty$.*

In what follows, we use x^k to denote the vector with components x_i^k, $i = 1, \ldots, N$, i.e., $x^k = (x_1^k, \ldots, x_N^k)$ and, similarly, we write x^* for the vector $(x_1^*, \ldots, x_N^*)$.

Proposition 11. *Let Assumptions 8–13 hold. Then, the sequence $\{x^k\}$ generated by the method* (4.9)–(4.10) *converges to the (unique) solution x^* of VI(K, ϕ).*

Proof. By Proposition 10, VI(K, ϕ) has a unique solution $x^* \in K$. When x^* solves the variational inequality problem VI(K, ϕ), the following relation holds $x^* = \Pi_{K_i}[x_i^* - \alpha_k F_i(x_i^*, \bar{x}^*)]$ (see [35, Proposition 1.5.8, p. 83]). From this relation and the non-expansive property of projection operator, we see that

$$\begin{aligned}
\|x_i^{k+1} - x_i^*\|^2 &= \|\Pi_{K_i}[x_i^k - \alpha_k F_i(x_i^k, N\hat{v}_i^k)] - x_i^*\|^2 \\
&= \|\Pi_{K_i}[x_i^k - \alpha_k F_i(x_i^k, N\hat{v}_i^k)] - \Pi_{K_i}[x_i^* - \alpha_k F_i(x_i^*, \bar{x}^*)]\|^2 \\
&\leq \|x_i^k - x_i^* - \alpha_k(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*))\|^2.
\end{aligned}$$

By expanding the last term, we obtain the following expression:

$$\|x_i^{k+1} - x_i^*\|^2 \leq \|x_i^k - x_i^*\|^2 + \alpha_k^2 \underbrace{\|F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*)\|^2}_{\textbf{Term1}} \\ - 2\alpha_k \underbrace{(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*))^T (x_i^k - x_i^*)}_{\textbf{Term2}}. \tag{4.17}$$

To estimate **Term1**, we use the triangle inequality and the identity $(a+b)^2 \leq 2(a^2+b^2)$, which yields

$$\textbf{Term1} \leq 2\|F_i(x_i^k, N\hat{v}_i^k)\|^2 + 2\|F_i(x_i^*, \bar{x}^*)\|^2 \leq \tilde{C} \qquad \text{with} \quad \tilde{C} = 2C^2 + 2 \max_{(x_i,\bar{x}) \in K_i \times \bar{K}} \|F_i(x_i, \bar{x})\|^2,$$

where C is such that $\|F_i(x_i^k, N\hat{v}_i^k)\| \leq C$ for all i and k (cf. Corollary 5) and $\max_{(x_i,\bar{x}) \in K_i \times \bar{K}} \|F_i(x_i, \bar{x})\|$ is finite by Assumption 8. Next, we consider **Term2**. By adding and subtracting $F_i(x_i^k, Ny^k)$ in **Term2**, where y^k is defined by (4.11), we have

$$\textbf{Term2} = \Big(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^k, Ny^k)\Big)^T (x_i^k - x_i^*) + \Big(F_i(x_i^k, Ny^k) - F_i(x_i^*, \bar{x}^*)\Big)^T (x_i^k - x_i^*).$$

By applying the Cauchy-Schwartz inequality, i.e. $a^T b \geq -\|a\|\|b\|$, to the first term on the right hand side of the preceding relation and the Lipschitz continuity of $F_i(x_i, u)$ in u (cf. Assumption 10), we see that

$$\begin{aligned}(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^k, Ny^k))^T (x_i^k - x_i^*) &\geq -\|F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^k, Ny^k)\| \cdot \|x_i^k - x_i^*\| \\ &\geq -L_{-i} N \|\hat{v}_i^k - y^k\| \cdot \|x_i^k - x_i^*\| \\ &\geq -2L_{-i} MN \|\hat{v}_i^k - y^k\|,\end{aligned}$$

where in the last inequality we use $x_i^k, x_i^* \in K_i$ and the compactness of K_i (cf. Assumption 8) and $M \geq \max_{x_i \in K_i} \|x_i\|$ for all i. Therefore, we have

$$\textbf{Term2} \geq -2L_{-i} MN \|\hat{v}_i^k - y^k\| + \Big(F_i(x_i^k, Ny^k) - F_i(x_i^*, \bar{x}^*)\Big)^T (x_i^k - x_i^*).$$

By substituting the preceding estimates of **Term1** and **Term2** in (4.17), we obtain

$$\|x_i^{k+1} - x_i^*\|^2 \leq \|x_i^k - x_i^*\|^2 + \tilde{C}\alpha_k^2 + 4\alpha_k L_{-i} MN \|\hat{v}_i^k - y^k\| \\ - 2\alpha_k \Big(F_i(x_i^k, Ny^k) - F_i(x_i^*, \bar{x}^*)\Big)^T (x_i^k - x_i^*).$$

Summing over all agents from $i = 1$ to $i = N$, yields

$$\sum_{i=1}^{N} \|x_i^{k+1} - x_i^*\|^2 \leq \sum_{i=1}^{N} \|x_i^k - x_i^*\|^2 + N\tilde{C}\alpha_k^2 + 4\alpha_k MN \sum_{i=1}^{N} L_{-i}\|\hat{v}_i^k - y^k\|$$
$$- 2\alpha_k \sum_{i=1}^{N} \left(F_i(x_i^k, Ny^k) - F_i(x_i^*, \bar{x}^*)\right)^T (x_i^k - x_i^*).$$

Using $Ny^k = \sum_{i=1}^N x_i^k$ (see Lemma 13) and letting $\bar{x}^k = \sum_{i=1}^N x_i^k$, we have for all $k \geq 0$,

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 + N\tilde{C}\alpha_k^2 + 4\alpha_k MN \sum_{i=1}^{N} L_{-i}\|\hat{v}_i^k - y^k\|$$
$$- 2\alpha_k \left(\phi(x^k) - \phi(x^*)\right)^T (x^k - x^*), \tag{4.18}$$

where we also use the fact that $F_i(x_i, \bar{x})$ is a coordinate map for the mapping $\phi(x) = F(x, \bar{x})$ (see (4.6) and (4.7)).

To claim the convergence of x^k to x^*, we apply Lemma 9 (for the deterministic sequences) to relation (4.18). To apply this lemma, since $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ by Assumption 13, we only need to prove

$$\sum_{k=0}^{\infty} \alpha_k \|\hat{v}_i^k - y^k\| < \infty \qquad \text{for all } i \in \mathcal{N}. \tag{4.19}$$

In view of Lemma 14, we have

$$\|y^k - \hat{v}_i^k\| \leq \theta\beta^k M + \theta NC \sum_{s=1}^{k} \beta^{k-s}\alpha_{s-1} \qquad \text{for all } i \in \mathcal{N} \text{ and all } k \geq 1,$$

so it suffices to prove that

$$\sum_{k=1}^{\infty} \alpha_k \left(\sum_{s=1}^{k} \beta^{k-s}\alpha_{s-1}\right) < \infty \qquad \text{and} \qquad \sum_{k=1}^{\infty} \alpha_k \beta^k < \infty.$$

Using $\alpha_k \leq \alpha_s$ for all $k \geq s$ (Assumption 13), for the series $\sum_{k=1}^{\infty} \alpha_k \left(\sum_{s=1}^{k} \beta^{k-s}\alpha_{s-1}\right)$ we have

$$\sum_{k=1}^{\infty} \alpha_k \left(\sum_{s=1}^{k} \beta^{k-s}\alpha_{s-1}\right) = \sum_{k=1}^{\infty} \left(\sum_{s=1}^{k} \beta^{k-s}\alpha_k\alpha_{s-1}\right) \leq \sum_{k=1}^{\infty} \left(\sum_{s=1}^{k} \beta^{k-s}\alpha_{s-1}^2\right).$$

We now use Lemma 15, from which by letting $\zeta_s = \alpha_s^2$ we can see that $\sum_{k=1}^{\infty} \alpha_k \left(\sum_{s=1}^{k} \beta^{k-s}\alpha_{s-1}\right) <$

∞. To establish the convergence of $\sum_{k=0}^{\infty} \alpha_k \beta^k$, we note that $\alpha_k \leq \alpha_0$ (Assumption 13), implying that $\sum_{k=0}^{\infty} \alpha_k \beta^k \leq \alpha_0 \sum_{k=0}^{\infty} \beta^k < \infty$ since $0 < \beta < 1$. Thus, relation (4.19) is valid.

As relation (4.18) satisfies the conditions of (the deterministic case of) Lemma 9, it follows that

$$\{\|x^k - x^*\|\} \quad \text{converges}, \tag{4.20}$$

$$\sum_{k=0}^{\infty} \alpha_k (\phi(x^k) - \phi(x^*))^T (x^k - x^*) < \infty. \tag{4.21}$$

Since $\{x^k\} \subset K$ and K is compact (Assumption 8), $\{x^k\}$ has accumulation points in K. By (4.21) and $\sum_{k=0}^{\infty} \alpha_k = \infty$ it follows that $(\phi(x^k) - \phi(x^*))^T (x^k - x^*) \to 0$ along a subsequence, say $\{x^{k_\ell}\}$. This observation, together with the strict monotonicity of ϕ, implies that $\{x^{k_\ell}\} \to x^*$ as $\ell \to \infty$. By relation (4.20), the entire sequence $\{x^k\}$ must converge to x^*. ■

4.3 Distributed asynchronous algorithm

In this section, we propose a distributed gossip-based algorithm for computing an equilibrium of aggregative Nash game (4.2). A description of the algorithm and some preliminary results are provided in Section 4.3.1. The global convergence of the algorithm is examined in Section 4.3.2, while constant steplength error bounds are provided in Section 4.3.3.

4.3.1 Outline of algorithm

In the proposed algorithm, agents perform their estimate and iterate updates the same as in the synchronous algorithm (4.9)–(4.10), but the updates occur asynchronously. As a mechanism for generating asynchronous updates we employ the gossip model for agent communications [83]. Together with the asynchronous updates, we allow the agents to use uncoordinated stepsize values by letting each agent choose a stepsize based on its own information-update frequency. To accommodate these asynchronous updates and uncoordinated stepsize selections, we model the agent connectivity structure by an undirected static graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$, with node $i \in \mathcal{N}$ being agent i and $\mathcal{E}$ being the set of undirected edges among the agents. When $\{i, j\} \in \mathcal{E}$, the agents i and j can talk to each other. We let $\mathcal{N}_i$ denote the set of neighbors of agent i, i.e., $\mathcal{N}_i = \{j \mid \{i, j\} \in \mathcal{E}\}$. We use the following assumption for the graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$.

Assumption 14. *The undirected graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$ is connected.*

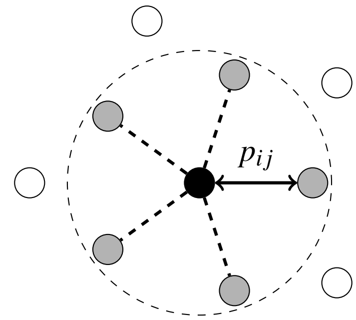


Figure 4.2: A depiction of a gossip communication.

We use a gossip protocol to model agent communication and exchange of the estimates of the aggregate $\bar{x}$. In this model, each agent is assumed to have a local clock which ticks according to a Poisson process with rate 1. At a tick of its clock, an agent i wakes up and contacts its neighbor $j \in \mathcal{N}_i$ with probability p_{ij}. The agents' clocks processes can be equivalently modeled as a single (virtual) clock which ticks according to a Poisson process with rate N. We assume that only one agent wakes up at each tick of the global clock, and we let Z^k denote kth tick time of the global Poisson process. We discretize time so that instant k corresponds to the time-slot $[Z^{k-1}, Z^k)$. At each time k, every agent i has its iterate x_i^k and estimate v_i^k of the average of the current aggregate. We let I^k denote the agent whose clock ticked at time k and we let J^k be the agent contacted by the agent I^k, where J^k is a neighbor of agent I^k, i.e., $J^k \in \mathcal{N}_{I^k}$. At time k, agents I^k and J^k exchange their estimates $v_{I^k}^k$ and $v_{J^k}^k$ and compute intermediate estimates:

$$\hat{v}_i^k = \frac{v_{I^k}^k + v_{J^k}^k}{2} \quad \text{for } i \in \{I^k, J^k\}, \tag{4.22}$$

and update their iterates and estimates of the aggregate average, as follows:

$$\left.\begin{array}{l} x_i^{k+1} = \Pi_{K_i}[x_i^k - \alpha_{k,i} F_i(x_i^k, N\hat{v}_i^k)] \\ v_i^{k+1} = \hat{v}_i^k + x_i^{k+1} - x_i^k \end{array}\right\} \qquad \text{for } i \in \{I^k, J^k\}, \tag{4.23}$$

where $\alpha_{k,i}$ is the stepsize for agent i and $F_i(x_i, y) = \nabla_{x_i} f_i(x_i, y)$. The other agents do nothing, i.e.,

$$\hat{v}_i^k = v_i^k, \quad x_i^{k+1} = x_i^k, \quad \text{and} \quad v_i^{k+1} = v_i^k \qquad \text{for } i \notin \{I^k, J^k\}. \tag{4.24}$$

As seen from the preceding update relations, the agents perform the same updates as in the synchronous algorithm (4.9)–(4.10), but instead of all agents updating, only two randomly selected agents update their estimates and iterates, while the other agents do not update.

We now rewrite the update steps more compactly. To capture the step in (4.22), we define the weight matrix $W(k)$:

$$W(k) = \mathbb{I} - \frac{1}{2}(e_{I^k} - e_{J^k})(e_{I^k} - e_{J^k})^T, \tag{4.25}$$

where $\mathbb{I}$ stands for the identity matrix, e_i is N-dimensional vector with ith entry equal to 1, and the other entries equal to 0. By using $W(k)$ we can rewrite the intermediate estimate update (4.22), as follows: for all $i = 1,\ldots,N$,

$$\hat{v}_i^k = \sum_{j=1}^{N} [W(k)]_{ij} v_j^k \quad \text{for all } k \geq 1, \quad \text{with} \quad v_i^0 = x_i^0, \tag{4.26}$$

where $x_i^0 \in K_i$, $i = 1,\ldots,N$, are initial (random) agent decisions. To rewrite the iterate x_i^{k+1} update (or no update) compactly for all agents, we let $\mathbb{1}_{\{i\in S\}}$ denote the indicator of the event $\{i \in S\}$. Then, the update relations in (4.23) and (4.24) can be written as:

$$x_i^{k+1} = \left(\Pi_{K_i}[x_i^k - \alpha_{k,i} F_i(x_i^k, N\hat{v}_i^k)] - x_i^k\right) \mathbb{1}_{\{i\in\{I^k,J^k\}\}} + x_i^k, \tag{4.27}$$

$$v_i^{k+1} = \hat{v}_i^k + x_i^{k+1} - x_i^k, \tag{4.28}$$

Note that only agents $i \in \{I^k, J^k\}$ update since $\mathbb{1}_{\{i\in\{I^k,J^k\}\}} = 0$ when $i \notin \{I^k, J^k\}$ and, hence, $x_i^{k+1} = x_i^k$ and $v_i^{k+1} = \hat{v}_i^k$ with $\hat{v}_i^k = v_i^k$ (by (4.26)).

We allow agents to use uncoordinated stepsizes that are based on the frequency of the agent updates. Specifically, agent i uses the stepsize $\alpha_{k,i} = \frac{1}{\Gamma_k(i)}$, where $\Gamma_k(i)$ denotes the number of updates that agent i has executed up to time k inclusively. These stepsizes are of the order of $\frac{1}{k}$ in a long run [87, 90]. To formalize this result, we need to introduce the probabilities of agents updates. We let p_i denote the probability of the event that agent i updates, i.e. $\{i \in \{I^k, J^k\}\}$, for which we have

$$p_i = \frac{1}{N}\left(1 + \sum_{j\in\mathcal{N}_i} p_{ji}\right) \quad \text{for all } i \in \mathcal{N},$$

where $p_{ji} > 0$ is the probability that agent i is contacted by its neighbor j. The long term estimates for $\alpha_{k,i}$ that we use in our analysis are given in the following lemma (cf. [90], Lemma 3).

Lemma 16. *Let Assumption 14 hold, and let $p_{\min} = \min_{i\in\mathcal{N}} p_i$ and $\alpha_{k,i} = 1/\Gamma_k(i)$ for all k and i. Then, for any $q \in (0, 1/2)$, there is a large enough $\tilde{k} = \tilde{k}(q,N)$ such that almost surely we have for all $k \geq \tilde{k}$ and $i \in \mathcal{N}$,*

$$\alpha_{k,i} \leq \frac{2}{kp_i}, \qquad \left|\alpha_{k,i} - \frac{1}{kp_i}\right| \leq \frac{2}{k^{3/2-q} p_{\min}^2}.$$

Another useful result is provided by [87, Theorem 1], which is stated below in a form suitable for our setting.

Lemma 17. *Let $\mathcal{G}(\mathcal{N},\mathcal{E})$ be a graph that satisfies Assumption 14. Let W be an $N\times N$ random stochastic matrix such that $\mathbb{E}[W]$ is doubly stochastic and $\mathbb{E}[W]_{ij}>0$ whenever $\{i,j\}\in\mathcal{E}$. Furthermore, let the diagonal elements of W be positive almost surely. Then, there exists a a scalar $\lambda<1$ such that*

$$\mathbb{E}\left[\left\|\left(W-\frac{1}{N}\mathbf{1}\mathbf{1}^TW\right)z\right\|^2\right]\leq\lambda\|z\|^2\qquad \text{for all } z\in\mathbb{R}^N.$$

The random matrices $W(k)$ in (4.25) are in fact doubly stochastic and thus, $\bar{W}=\mathbb{E}[W(k)]$ is doubly stochastic. Moreover, it can be easily seen that $\bar{W}_{ij}>0$ whenever $\{i,j\}\in\mathcal{E}$. In addition, $W(k)$ has positive diagonal entries. Hence, Lemma 17 applies to random matrices $W(k)$. However, since each $W(k)$ is in fact doubly stochastic, we have $\mathbf{1}^TW(k)=\mathbf{1}^T$, implying that $W(k)-\frac{1}{N}\mathbf{1}\mathbf{1}^TW(k)=W(k)-\frac{1}{N}\mathbf{1}\mathbf{1}^T$. Hence, using this observation and Lemma 17, we find that there exists $\lambda\in(0,1)$ such that for the matrix $D(k)=W(k)-\frac{1}{N}\mathbf{1}\mathbf{1}^T$ we have

$$\mathbb{E}[\|D(k)z\|^2]\leq\lambda\|z\|^2\qquad\text{for all } z\in\mathbb{R}^N. \tag{4.29}$$

By Jensen's inequality we have $|\mathbb{E}[X]|\leq\sqrt{\mathbb{E}[X^2]}$ for any random variable X (with a finite expectation), which when applied to relation (4.29) yields

$$\mathbb{E}[\|D(k)z\|]\leq\sqrt{\lambda}\,\|z\|\qquad\text{for all } z\in\mathbb{R}^N. \tag{4.30}$$

4.3.2 Convergence theory

In this section we establish the convergence of the asynchronous algorithm (4.26)–(4.28) with the agent specific diminishing stepsize of the form $\alpha_{k,i}=\frac{1}{\Gamma_k(i)}$. To take account of the history, we introduce $\mathcal{F}_k$ to denote the $\sigma-$algebra generated by the entire history up to k. More precisely

$$\mathcal{F}_k=\mathcal{F}_0\cup\{I^l,J^l;1\leq l\leq k-1\}\quad\text{for all } k\geq 2,$$

with $\mathcal{F}_1=\mathcal{F}_0\{x_i^0,i\in\mathcal{N}\}$. Thus, given $\mathcal{F}_k$, the vectors v_i^k and x_i^k are fully determined. First we state several result which we will use to claim the convergence of the algorithm, as well as to analyze the error bounds.

In what follows, we will use a vector-component based analysis. To this end, we introduce $[z]_\ell$ to denote the ℓ-th component of a vector $z\in\mathbb{R}^n$, with $\ell=1,\ldots,n$. A component-wise update of each v_i^{k+1} in (4.28) is given by: for all $i = 1, \ldots, N$,

$$[v_i^{k+1}]_\ell = \sum_{j=1}^{N} [W(k)]_{ij} [v_j^k]_\ell + [x_i^{k+1} - x_i^k]_\ell \quad \text{for } \ell = 1, \ldots, n.$$

We collect all ℓth coordinates of the vectors $v_1^k, \ldots, v_N^k$ and let $v^k(\ell) = ([v_1^k]_\ell, \ldots, [v_N^k]_\ell)^T$. We similarly do for the vectors $x_1^k, \ldots, x_N^k$ and let $x^k(\ell) = ([x_1^k]_\ell, \ldots, [x_N^k]_\ell)^T$. Using the vectors $v^k(\ell)$ and $x^k(\ell)$, we can rewrite the preceding relation as follows:

$$v^{k+1}(\ell) = W(k)v^k(\ell) + \zeta^{k+1}(\ell) \quad \text{with} \quad \zeta^{k+1}(\ell) = x^{k+1}(\ell) - x^k(\ell) \qquad \text{for all } \ell = 1, \ldots, n. \tag{4.31}$$

We have the following result for $v^{k+1}(\ell)$ for any ℓ.

Lemma 18. *Let Assumptions 8–10 and Assumption 14 hold. Then, for all $\ell = 1, \ldots, n$ and $k \geq 0$,*

$$\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\| \leq \|D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1})\| + \sqrt{2} C \max_i \alpha_{k,i},$$

where $D(k) = W(k) - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ and C is a constant as in Corollary 5.

Proof. We fix an arbitrary coordinate ℓ. By the decision update rule of (4.27), the ith coordinate of the vector $\zeta^{k+1}(\ell)$ is $[\zeta^{k+1}(\ell)]_i = [\left(\Pi_{K_i}[x_i^k - \alpha_{k,i} F_i(x_i^k, N\hat{v}_i^k)] - x_i^k\right) \mathbb{1}_{\{i \in \{I^k, J^k\}\}}]_\ell$. Since y^{k+1} is the average of the vectors v_i^{k+1}, from (4.31) for the ℓth coordinate of this vector we obtain

$$[y^{k+1}]_\ell = \frac{1}{N}\mathbf{1} v^{k+1}(\ell) = \frac{1}{N}\left(\mathbf{1}^T W(k) v^k(\ell) + \mathbf{1}^T \zeta^{k+1}(\ell)\right),$$

which together with (4.31) leads us to

$$v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1} = \left(W(k) - \frac{1}{N}\mathbf{1}\mathbf{1}^T W(k)\right) v^k(\ell) + \left(\mathbb{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\right) \zeta^{k+1}(\ell),$$

where $\mathbb{I}$ is the identity matrix. Note that each $W(k)$ is a doubly stochastic matrix i.e., $W(k)\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T W(k) = \mathbf{1}^T$. Thus, $\frac{1}{N}\mathbf{1}\mathbf{1}^T W(k) = \frac{1}{N}\mathbf{1}\mathbf{1}^T$. Furthermore, we have $\left(W(k) - \frac{1}{N}\mathbf{1}\mathbf{1}^T\right)\mathbf{1} = 0$, implying that

$$\left(W(k) - \frac{1}{N}\mathbf{1}\mathbf{1}^T\right)[y^k]_\ell \mathbf{1} = 0.$$

By combining the preceding two relations, using $\frac{1}{N}\mathbf{1}\mathbf{1}^T W(k) = \frac{1}{N}\mathbf{1}\mathbf{1}^T$, and letting $D(k) = W(k) -$

$\frac{1}{N}\mathbf{1}\mathbf{1}^T$, we obtain

$$v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1} = D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1}) + \left(\mathbb{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\right) \zeta^{k+1}(\ell).$$

Taking the norm, we obtain

$$\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\| \leq \|D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1})\| + \left\| \left(\mathbb{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\right) \zeta^{k+1}(\ell) \right\|. \tag{4.32}$$

We next estimate the last term in (4.32). The matrix $\mathbb{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ is a projection matrix (corresponds to the projection on the subspace orthogonal to the vector $\mathbf{1}$), so $\|\mathbb{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\| = 1$, implying that

$$\left\| \left(\mathbb{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\right) \zeta^{k+1}(\ell) \right\| \leq \left\| \mathbb{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T \right\| \|\zeta^{k+1}(\ell)\| = \|\zeta^{k+1}(\ell)\|. \tag{4.33}$$

From the definition of $\zeta^{k+1}(\ell)$ in (4.31), we see that

$$\|\zeta^{k+1}(\ell)\|^2 = \sum_{i \in \{I^k, J^k\}} \left\| \left[\Pi_{K_i}[x_i^k - \alpha_{k,i} F_i(x_i^k, N\hat{v}_i^k)] - x_i^k \right]_\ell \right\|^2.$$

Using the non-expansive property of the projection operator and $\alpha_{k,i}^2 \leq \max_i \alpha_{k,i}^2$, we have

$$\|\zeta^{k+1}(\ell)\|^2 \leq \sum_{i \in \{I^k, J^k\}} \|\alpha_{k,i}[F_i(x_i^k, N\hat{v}_i^k)]_\ell\|^2 \leq \max_i \alpha_{k,i}^2 \sum_{i \in \{I^k, J^k\}} \|F_i(x_i^k, N\hat{v}_i^k)\|^2.$$

By Corollary 5, $\|F_i(x_i^k, N\hat{v}_i^k)\| \leq C$ for all i,k and some $C > 0$. This and $|\{I^k, J^k\}| = 2$ imply $\|\zeta^{k+1}(\ell)\|^2 \leq 2C^2 \max_i \alpha_{k,i}^2$. By taking square roots we obtain $\|\zeta^{k+1}(\ell)\| \leq \sqrt{2} C \max_i \alpha_{k,i}$ which when combined with (4.32) and (4.33) yields

$$\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\| \leq \|D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1})\| + \sqrt{2} C \max_i \alpha_{k,i}.$$

■

Our result involves the average y^k of the estimates $v_i^k, i \in \mathcal{N}$, which will be important in establishing the convergence of the algorithm.

Lemma 19. *Let Assumptions 8–10 and Assumption 14 hold. Let v_i^k be given by* (4.26) *and* (4.28), *respectively, and let* $y^k = \frac{1}{N}\sum_{i=1}^N v_i^k$. *Then, we have*

$$\sum_{k=1}^{\infty} \frac{1}{k}\,\|v_i^k - y^k\|^2 < \infty \quad \textit{a.s. for all } i \in \mathcal{N}.$$

Proof. Using Lemma 16 we find that $\alpha_{k,i} \leq \frac{2}{kp_{\min}}$ almost surely for all k large enough, where $p_{\min} = \min_i p_i$. Thus, $\max_i \alpha_{k,i} \leq \frac{2}{kp_{\min}}$, and from Lemma 18, we obtain almost surely for all k large enough,

$$\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\| \leq \|D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1})\| + \frac{2\sqrt{2}C}{kp_{\min}}.$$

By taking the conditional expectation with respect to $\mathcal{F}_k$, we obtain almost surely for all k large enough,

$$\mathbb{E}[\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\| \mid \mathcal{F}_k] \leq \mathbb{E}[\|D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1})\| \mid \mathcal{F}_k] + \frac{2\sqrt{2}C}{kp_{\min}}. \tag{4.34}$$

Note that the expectation in the term on the right hand side is taken with respect to the randomness in the matrix $W(k)$ only. By relation (4.30), we have

$$\mathbb{E}[\|D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1})\| \mid \mathcal{F}_k] \leq \sqrt{\lambda}\|v^k(\ell) - [y^k]_\ell \mathbf{1}\|,$$

which combined with (4.34) yields almost surely for all k large enough,

$$\mathbb{E}[\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\| \mid \mathcal{F}_k] \leq \sqrt{\lambda}\|v^k(\ell) - [y^k]_\ell \mathbf{1}\| + \frac{2\sqrt{2}C}{kp_{\min}}. \tag{4.35}$$

By dividing both sides of (4.35) with $\frac{1}{k}$ and by using $\frac{1}{k+1} < \frac{1}{k}$ we find that almost surely for all k large enough,

$$\begin{aligned}\frac{1}{k+1}\mathbb{E}[\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\| \mid \mathcal{F}_k] &\leq \frac{\sqrt{\lambda}}{k}\|v^k(\ell) - [y^k]_\ell \mathbf{1}\| + \frac{2\sqrt{2}C}{k^2 p_{\min}} \\ &= \frac{1}{k}\|v^k(\ell) - [y^k]_\ell \mathbf{1}\| - \frac{1-\sqrt{\lambda}}{k}\|v^k(\ell) - [y^k]_\ell \mathbf{1}\| + \frac{2\sqrt{2}C}{k^2 p_{\min}}.\end{aligned}$$

Since $\lambda \in (0,1)$ we have $1 - \sqrt{\lambda} > 0$. Thus, by the supermartingale convergence of Lemma 9 (applied with an index-shift), we can conclude that

$$\sum_{k=1}^{\infty} \frac{1}{k} \|v^k(\ell) - [y^k]_\ell \mathbf{1}\| < \infty \quad \text{a.s.}$$

Recalling that $v^k(\ell) = ([v_1^k]_\ell, \ldots, [v_N^k]_\ell)^T$, the preceding relation implies that

$$\sum_{k=1}^{\infty} \frac{1}{k} \left| [v_i^k]_\ell - [y^k]_\ell \right| < \infty \quad \text{for all } i \in \mathcal{N} \text{ a.s.}$$

The coordinate index ℓ was arbitrary, so the relation is also true for every coordinate index $\ell = 1, \ldots, n$. In particular, since $\|v_i^k - y^k\| \leq \sum_{\ell=1}^n |[v_i^k]_\ell - [y^k]_\ell|$ we have

$$\sum_{k=1}^{\infty} \frac{1}{k} \left\| v_i^k - y^k \right\| \leq \sum_{k=1}^{\infty} \frac{1}{k} \sum_{\ell=1}^{n} |[v_i^k]_\ell - [y^k]_\ell| < \infty \quad \text{for all } i \in \mathcal{N} \text{ a.s.}$$

■

For the rest of the paper, we use x^k to denote the vector with components x_i^k, $i = 1, \ldots, N$, i.e., $x^k = (x_1^k, \ldots, x_N^k)$ and we write x^* for the vector $(x_1^*, \ldots, x_N^*)$. We now show the convergence of the algorithm. We have the following result, where x^* denotes the unique Nash equilibrium of the aggregative game in (4.2).

Proposition 12. *Let Assumptions 8–10 and Assumption 14 hold. Then, the sequence $\{x^k\}$ generated by the method* (4.26)–(4.28) *with the stepsize $\alpha_{k,i} = \frac{1}{\Gamma_k(i)}$ converges to the (unique) x^* of the game almost surely.*

Proof. Under strict monotonicity of the mapping and the compactness of K, uniqueness of the equilibrium follows from Proposition 10. Then, by the definition of x_i^{k+1} we have

$$\|x_i^{k+1} - x_i^*\|^2 = \left\| \left(\Pi_{K_i}[x_i^k - \alpha_{k,i} F_i(x_i^k, N\hat{v}_i^k)] - x_i^k \right) \mathbb{1}_{\{i \in \{I^k, J^k\}\}} + x_i^k - x_i^* \right\|^2.$$

Using $x_i^* = \Pi_{K_i}[x_i^* - \alpha_{k,i} F_i(x_i^*, \bar{x}^*)]$ and the non-expansive property of the projection operator, we have for $i \in \{I^k, J^k\}$,

$$\begin{aligned} \|x_i^{k+1} - x_i^*\|^2 \leq & \|x_i^k - \alpha_{k,i} F_i(x_i^k, N\hat{v}_i^k) - x_i^k - x_i^* + \alpha_{k,i} F_i(x_i^*, \bar{x}^*)\|^2 \\ = & \|x_i^k - x_i^*\|^2 + \alpha_{k,i}^2 \|F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*)\|^2 \\ & - 2\alpha_{k,i} (F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*))^T (x_i^k - x_i^*). \end{aligned}$$

By Lemma 16, for k large enough we have almost surely

$$\alpha_{k,i} \leq \frac{2}{kp_i}, \quad \left|\alpha_{k,i} - \frac{1}{kp_i}\right| \leq \frac{2}{k^{3/2-q}p_{\min}^2}.$$

Thus, we also have $\alpha_{k,i}^2 \leq \frac{4}{k^2p_i^2}$. Writing $\alpha_{k,i} = \left(\alpha_{k,i} - \frac{1}{kp_i}\right) + \frac{1}{kp_i}$ and using the preceding relations, we obtain almost surely for $i \in \{I^k, J^k\}$ and for all k large enough,

$$\begin{aligned}
\|x_i^{k+1} - x_i^*\|^2 = {} & \|x_i^k - x_i^*\|^2 + \alpha_{k,i}^2\|F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*)\|^2 \\
& - 2\left(\alpha_{k,i} - \frac{1}{kp_i}\right)(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*) \\
& - \frac{2}{kp_i}(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*) \\
\leq {} & \|x_i^k - x_i^*\|^2 + \frac{4}{k^2p_i^2}\|F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*)\|^2 \\
& + \frac{4}{k^{3/2-q}p_{\min}^2}\left|(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*)\right| \\
& - \frac{2}{kp_i}(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*).
\end{aligned} \tag{4.36}$$

By Corollary 5 and Assumption 8 we can see that $\|F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*)\|^2 \leq C_1$ for some scalar C_1, and for all i and k. Similarly, for the term in (4.36) involving the absolute value, we can see that $|(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*)| \leq C_2$ for some scalar C_2, and for all i and k. Substituting these estimates in (4.36), we obtain almost surely for all k large enough,

$$\|x_i^{k+1} - x_i^*\|^2 \leq \|x_i^k - x_i^*\|^2 + \frac{4C_1}{k^2p_i^2} + \frac{4C_2}{k^{3/2-q}p_{\min}^2} - \frac{2}{kp_i}(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*). \tag{4.37}$$

For the last term in the preceding relation, by adding and subtracting $F_i(x_i^k, Ny^k)$ and using $Ny^k =$

$\sum_{i=1}^{N} x_i^k = \bar{x}^k$ (cf. Lemma 13), we write

$$
\begin{aligned}
(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*))^T (x_i^k - x_i^*) =& (F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^k, Ny^k))^T (x_i^k - x_i^*) \\
&+ (F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T (x_i^k - x_i^*) \\
\geq& - \|F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^k, Ny^k)\| \, \|x_i^k - x_i^*\| \\
&+ (F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T (x_i^k - x_i^*) \\
\geq& - L_{-i} N \|\hat{v}_i^k - y^k\| M + (F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T (x_i^k - x_i^*)
\end{aligned}
$$

where we use the Lipschitz property of the mapping F_i (Assumption 10), while M is a constant such that $\max_{x_i, z_i \in K_i} \|x_i - z_i\| \leq M$ for all i. The vector $\hat{v}_i^k$ is a convex combination of v_j^k (cf. (4.26)), so by the convexity of the norm, we have $\|\hat{v}_i^k - y^k\| \leq \sum_{j=1}^{N} [W(k)]_{ij} \|v_j^k - y^k\|$, which yields

$$
\begin{aligned}
(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*, \bar{x}^*))^T (x_i^k - x_i^*) \geq& - L_{-i} NM \sum_{j=1}^{N} [W(k)]_{ij} \|v_j^k - y^k\| \\
&+ (F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T (x_i^k - x_i^*). \qquad (4.38)
\end{aligned}
$$

Finally, by combining relations (4.37) and (4.38) we obtain almost surely for $i \in \{I^k, J^k\}$ and for all k large enough,

$$
\begin{aligned}
\|x_i^{k+1} - x_i^*\|^2 \leq& \|x_i^k - x_i^*\|^2 + \frac{4C_1}{k^2 p_i^2} + \frac{4C_2}{k^{3/2-q} p_{\min}^2} + \frac{2}{k p_i} L_{-i} NM \sum_{j=1}^{N} [W(k)]_{ij} \|v_j^k - y^k\| \\
&- \frac{2}{k p_i} (F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T (x_i^k - x_i^*).
\end{aligned}
$$

Since $x_i^{k+1} = x_i^k$ when $i \notin \{I^k, J^k\}$, it follows that $\|x_i^{k+1} - x_i^*\|^2 = \|x_i^k - x_i^*\|^2$ for $i \notin \{I^k, J^k\}$. We combine these two cases with the fact that agent i updates with probability p_i and, thus obtain almost surely for all $i \in \mathcal{N}$ and for all k large enough,

$$
\begin{aligned}
\mathbb{E}[\|x_i^{k+1} - x_i^*\|^2 \mid \mathcal{F}_k] \leq& \|x_i^k - x_i^*\|^2 + \frac{4C_1}{k^2 p_i} + \frac{4C_2 p_i}{k^{3/2-q} p_{\min}^2} + \frac{2}{k} L_{-i} NM \sum_{j=1}^{N} [W(k)]_{ij} \|v_j^k - y^k\| \\
&- \frac{2}{k} (F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T (x_i^k - x_i^*). \qquad (4.39)
\end{aligned}
$$

Summing relations (4.39) over all $i = 1, \ldots, N$, using the fact that $W(k)$ is doubly stochastic and recalling that F_i are coordinate maps for F and $F(x, \bar{x})$ defines ϕ (cf. (4.6) and (4.7)), we further obtain for all k large enough almost surely

$$\mathbb{E}[\|x^{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq \|x^k - x^*\|^2 + \frac{4NC_1}{k^2 p_{\min}} + \frac{4NC_2 p_{\max}}{k^{3/2-q} p_{\min}^2} + \frac{2}{k} L_{-i} NM \sum_{j=1}^{N} \|v_j^k - y^k\| - \frac{2}{k}(\phi(x^k) - \phi(x^*))^T (x^k - x^*), \tag{4.40}$$

where $p_{\min} = \min_i p_i$ and $p_{\max} = \max_i p_i$. We now verify that we can apply the supermartingale convergence result (cf. Lemma 9) to relation (4.40). For $q \in (0, 1/2)$ we have

$$\sum_{k=1}^{\infty} \left(\frac{4NC_1}{k^2 p_{\min}} + \frac{4NC_2 p_{\max}}{k^{3/2-q} p_{\min}^2} \right) < \infty.$$

Further from Lemma 19 it follows that $\sum_{k=1}^{\infty} \sum_{i=1}^{N} \frac{1}{k} \|v_i^k - y^k\| < \infty$ almost surely. Thus, all conditions of Lemma 9 are satisfied (with a time-shift) and we conclude that

$$\{\|x^k - x^*\|\} \quad \text{converges } a.s., \tag{4.41}$$

$$\sum_{k=1}^{\infty} \frac{2}{k}(\phi(x) - \phi(x^*))^T (x^k - x^*) < \infty \quad a.s. \tag{4.42}$$

Since $\{x^k\} \subset K$ and K is compact (Assumption 8), it follows that $\{x^k\}$ has an accumulation point in K. By $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$, relation (4.42) implies that $(x^k - x^*)^T (\phi(x^k) - \phi(x^*)) \to 0$ along a subsequence almost surely, say $\{x^{k_\ell}\}$. Then, by the strict monotonicity of ϕ it follows that $\{x^{k_\ell}\} \to x^*$ as $\ell \to \infty$ almost surely. By (4.41), the entire sequence converges to x^* almost surely. ■

4.3.3 Error bounds for constant stepsize

In this section, we investigate the properties of the algorithm where agents employ a deterministic constant, albeit uncoordinated, stepsize. More specifically, our interest lies in establishing error bounds contingent on the deviation of stepsize across agents. Under this setting, the stepsize is $\alpha_{k,i} = \alpha_i$ in the update rule for agents' decisions in (4.27), which reduces to

$$x_i^{k+1} = \left(\Pi_{K_i}[x_i^k - \alpha_i F_i(x_i^k, N\hat{v}_i^k)] - x_i^k \right) \mathbb{1}_{\{i \in \{I^k, J^k\}\}} + x_i^k,$$

where α_i is a positive constant stepsize for agent i. It is worth mentioning that the estimate mixing rule of (4.22) and estimate update rule of (4.28) are invariant under this modification. Also, we allow agents to independently choose α_i, thereby maintaining the complete decentralization feature of the gossip algorithm. We begin by providing an updated estimate for the disagreement among the agents. Our result parallels the result of Lemma 19.

Lemma 20. *Let Assumptions 8–10 and 14 hold. Consider* $\{v_i^k\}$, $i = 1,\ldots,N$, *that are generated by algorithm in* (4.26)–(4.28) *with* $\alpha_{k,i} = \alpha_i$. *Then, for* $y^k = \frac{1}{N}\sum_{i=1}^N v_i^k$ *we have*

$$\limsup_{k\to\infty} \sum_{i=1}^N \mathbb{E}[\|v_i^k - y^k\|^2] \le \frac{2n\alpha_{\max}^2 C^2}{(1-\sqrt{\lambda})^2}, \qquad \limsup_{k\to\infty} \sum_{i=1}^N \mathbb{E}[\|v_i^k - y^k\|] \le \frac{\sqrt{2nN}\alpha_{\max} C}{1-\sqrt{\lambda}}, \qquad a.s.,$$

where $\alpha_{\max} = \max_i\{\alpha_i\}$, *C is the constant as in Corollary 5, and* λ *is as given in* (4.29).

Proof. We fix an arbitrary index ℓ. By Lemma 18 with $\alpha_{k,i} = \alpha_i$, we have for all $k \ge 0$,

$$\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\| \le \|D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1})\| + \sqrt{2}C\alpha_{\max}, \tag{4.43}$$

where $D(k) = W(k) - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ and $\alpha_{\max} = \max_i \alpha_i$. Note that by relation (4.30) we have

$$\mathbb{E}[\|D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1})\|] = \mathbb{E}\left[\mathbb{E}[\|D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1})\| \mid \mathcal{F}_k]\right] \le \sqrt{\lambda}\,\mathbb{E}[\|v^k(\ell) - [y^k]_\ell \mathbf{1}\|]. \tag{4.44}$$

Thus, by taking the expectation of booth sides in (4.43) we obtain

$$\mathbb{E}\left[\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\|\right] \le \sqrt{\lambda}\mathbb{E}\left[\|v^k(\ell) - [y^k]_\ell \mathbf{1}\|\right] + \sqrt{2}C\alpha_{\max} \qquad \text{for all } k \ge 0,$$

which by iterative recursion leads to

$$\mathbb{E}\left[\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\|\right] \le \left(\sqrt{\lambda}\right)^{k+1} \mathbb{E}\left[\|v^0(\ell) - [y^0]_\ell \mathbf{1}\|\right] + \sqrt{2}C\alpha_{\max} \sum_{s=0}^{k} (\sqrt{\lambda})^s \qquad \text{for all } k \ge 0.$$

Thus, by letting $k \to \infty$, we obtain the following limiting result

$$\limsup_{k\to\infty} \mathbb{E}[\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\|] \le \frac{\sqrt{2}C\alpha_{\max}}{1-\sqrt{\lambda}}. \tag{4.45}$$

By taking the squares of both sides in relation (4.43), we find

$$\|v^{k+1}(\ell) - [y^{k+1}]_\ell \mathbf{1}\|^2 \le \|D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1})\|^2 + 2\sqrt{2}C\alpha_{\max}\|D(k)(v^k(\ell) - [y^k]_\ell \mathbf{1})\| + 2C^2\alpha_{\max}^2.$$

Taking the expectation of both sides in the preceding relation and using estimate (4.44), we obtain

$$\begin{aligned}\mathbb{E}[\|v^{k+1}(\ell)-[y^{k+1}]_\ell\mathbf{1}\|^2] &\le \mathbb{E}[\|D(k)(v^k(\ell)-[y^k]_\ell\mathbf{1})\|^2]+2\sqrt{2}C\alpha_{\max}\sqrt{\lambda}\mathbb{E}[\|v^k(\ell)-[y^k]_\ell\mathbf{1}\|] \\ &\quad +2C^2\alpha_{\max}^2 \\ &\le \lambda\,\mathbb{E}[\|v^k(\ell)-[y^k]_\ell\mathbf{1}\|^2]+2\sqrt{2}C\alpha_{\max}\sqrt{\lambda}\mathbb{E}[\|v^k(\ell)-[y^k]_\ell\mathbf{1}\|] \\ &\quad +2C^2\alpha_{\max}^2\end{aligned} \tag{4.46}$$

where the last inequality follows by

$$\mathbb{E}[\|D(k)(v^k(\ell)-[y^k]_\ell\mathbf{1})\|^2]=\mathbb{E}\left[\mathbb{E}[\|D(k)(v^k(\ell)-[y^k]_\ell\mathbf{1})\|^2\mid\mathcal{F}_k]\right]\le\lambda\,\mathbb{E}[\|v^k(\ell)-[y^k]_\ell\mathbf{1}\|^2],$$

which is a consequence of relation (4.29). Since v_i^k and y^k are convex combinations of points in $K_1,\ldots,K_N$ and each K_i is compact, the sequence $\{\|v^{k+1}(\ell)-[y^{k+1}]_\ell\mathbf{1}\|\}$ is bounded, implying that so is the sequence $\{\mathbb{E}[\|v^{k+1}(\ell)-[y^{k+1}]_\ell\mathbf{1}\|]\}$. Thus, $\limsup_{k\to\infty}\mathbb{E}[\|v^{k+1}(\ell)-[y^{k+1}]_\ell\mathbf{1}\|^2]$ exists, and let us denote this limit by S. Letting $k\to\infty$ in relation (4.46) and using (4.45), we obtain

$$S\le\lambda S+\left(\frac{2\sqrt{\lambda}}{1-\sqrt{\lambda}}+1\right)2C^2\alpha_{\max}^2=\lambda S+\frac{1+\sqrt{\lambda}}{1-\sqrt{\lambda}}2C^2\alpha_{\max}^2,$$

which upon solving for S and recalling the notation yields

$$\limsup_{k\to\infty}\mathbb{E}[\|v^{k+1}(\ell)-[y^{k+1}]_\ell\mathbf{1}\|^2]\le\frac{1+\sqrt{\lambda}}{(1-\lambda)(1-\sqrt{\lambda})}2C^2\alpha_{\max}^2=\frac{1}{(1-\sqrt{\lambda})^2}2C^2\alpha_{\max}^2.$$

The preceding relation is true for any ℓ. Thus, since limsup is invariant under the index-shift, we have

$$\limsup_{k\to\infty}\sum_{\ell=1}^n\mathbb{E}[\|v^k(\ell)-[y^k]_\ell\mathbf{1}\|^2]\le\sum_{\ell=1}^n\limsup_{k\to\infty}\mathbb{E}[\|v^k(\ell)-[y^k]_\ell\mathbf{1}\|^2]\le\frac{2nC^2\alpha_{\max}^2}{(1-\sqrt{\lambda})^2},$$

and by the linearity of the expectation, it follows

$$\limsup_{k\to\infty}\mathbb{E}\left[\sum_{\ell=1}^n\|v^k(\ell)-[y^k]_\ell\mathbf{1}\|^2\right]\le\frac{2nC^2\alpha_{\max}^2}{(1-\sqrt{\lambda})^2}.$$

Recalling that vector $v^k(\ell)$ consists of the ℓth coordinates of the vectors $v_1^k,\ldots,v_N^k$ (i.e., $v^k(\ell)=([v_1^k]_\ell,\ldots,[v_N^k]_\ell)^T$), we see that $\|\sum_{\ell=1}^n\|v^k(\ell)-[y^k]_\ell\mathbf{1}\|^2=\sum_{i=1}^N\|v^k-y^k\|^2$. Hence, the preceding relation is equivalent to

$$\limsup_{k\to\infty}\sum_{i=1}^{N}\mathbb{E}[\|v_i^k - y^k\|^2] \le \frac{2nC^2\alpha_{\max}^2}{(1-\sqrt{\lambda})^2},$$

which is the first relation stated in the lemma. In particular, the preceding relation implies that

$$\limsup_{k\to\infty}\sqrt{\sum_{i=1}^{N}\mathbb{E}[\|v_i^k - y^k\|^2]} \le \frac{\sqrt{2n}C\alpha_{\max}}{1-\sqrt{\lambda}}. \tag{4.47}$$

On the other hand, by Holders' inequality we have

$$\sum_{i=1}^{N}\mathbb{E}[\|v_i^k - y^k\|] \le \sqrt{N}\sqrt{\sum_{i=1}^{N}\mathbb{E}[\|v_i^k - y^k\|^2]}.$$

from which by taking the limit as $k\to\infty$ and using (4.47), we obtain

$$\limsup_{k\to\infty}\sum_{i=1}^{N}\mathbb{E}[\|v_i^k - y^k\|] \le \frac{\sqrt{2nN}C\alpha_{\max}}{1-\sqrt{\lambda}}.$$

■

We now estimate the limiting error of the algorithm under the additional assumption of strong monotonicity of the mapping ϕ. For this result, we assume an additional Lipschitz property for the maps F_i, as given below.

Assumption 15. *Each mapping $F_i(x_i,u)$ is uniformly Lipschitz continuous in x_i over K_i, for every fixed $u\in\bar{K}$ i.e., for some $L_i>0$ and for all $x_i,y_i\in K_i$,*

$$\|F_i(x_i,u) - F_i(y_i,u)\| \le L_i\|x_i - y_i\|.$$

We have the following result.

Proposition 13. *Let Assumptions 8–10, 14, and 15 hold, and let the mapping ϕ be strongly monotone over the set K with a constant $\mu>0$, in the following sense:*

$$(\phi(x)-\phi(y))^T(x-y) \ge \mu\|x-y\|^2 \qquad \text{for all } x,y\in K.$$

Consider the sequence $\{x^k\}$ generated by the method (4.26)–(4.28) with $\alpha_{k,i}=\alpha_i$. Suppose that *the stepsizes α_i are such that*

$$0 < 1 - 2\mu p_{\min}\alpha_{\min} + 2p_{\max}(\max_i L_i)(\alpha_{\max} - \alpha_{\min}) < 1, \tag{4.48}$$

where $L_i, i = 1,\ldots,N$, are Lipshitz constants from Assumption 15, $\alpha_{\max} = \max_i \alpha_i$, $\alpha_{\min} = \min_i \alpha_i$, $p_{\max} = \max_i p_i$, and $p_{\min} = \min_i p_i$. Then, the following result holds

$$\limsup_{k\to\infty} \mathbb{E}[\|x^k - x^*\|^2] \leq \frac{p_{\max}\alpha_{\max}^2\left(2C^2N + BC\frac{\sqrt{2nN}}{1-\sqrt{\lambda}}\right)}{\mu p_{\min}\alpha_{\min} - p_{\max}(\max_i L_i)(\alpha_{\max} - \alpha_{\min})},$$

where x^* is the unique solution of VI(K,ϕ), C is as in Corollary 5, λ is as in (4.29)*, and $B = (\max_i L_{-i})NM$ with L_{-i}, $i = 1,\ldots,N$, being the Lipschitz constants from Assumption 10, and $M \geq \max_{x_i,z_i\in K_i}\|x_i - z_i\|$ for all i.*

Proof. Since the map is strongly monotone, there is a unique solution $x^* \in K$ to VI(K,ϕ) (see Theorem 2.3.3. in [35]). Then, by the definition of x_i^{k+1} we have

$$\|x_i^{k+1} - x_i^*\|^2 = \|\left(\Pi_{K_i}[x_i^k - \alpha_i F_i(x_i^k, N\hat{v}_i^k)] - x_i^k\right)\mathbb{1}_{\{i\in\{I^k,J^k\}\}} + x_i^k - x_i^*\|^2.$$

Using $x_i^* = \Pi_{K_i}[x_i^* - \alpha_i F_i(x_i^*,\bar{x}^*)]$ and the non-expansive property of the projection operator, we have for $i \in \{I^k, J^k\}$,

$$\begin{aligned}\|x_i^{k+1} - x_i^*\|^2 \leq& \|x_i^k - \alpha_i F_i(x_i^k, N\hat{v}_i^k) - x_i^k - x_i^* + \alpha_i F_i(x_i^*,\bar{x}^*)\|^2\\ =& \|x_i^k - x_i^*\|^2 + \alpha_i^2\|F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*,\bar{x}^*)\|^2\\ &- 2\alpha_i(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*,\bar{x}^*))^T(x_i^k - x_i^*).\end{aligned} \tag{4.49}$$

By using $(a+b)^2 \leq 2a^2 + 2b^2$ and Corollary 5 we can see that

$$\|F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*,\bar{x}^*)\|^2 \leq 4C^2.$$

We now approximate the inner product term by adding and subtracting $F_i(x_i^k, Ny^k)$ and using $y^k = \sum_{i=1}^N x_i^k = \bar{x}^k$ (see Lemma 13), to obtain

$$\begin{aligned}(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^*,\bar{x}^*))^T(x_i^k - x_i^*) \geq& -|(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^k, Ny^k))^T(x_i^k - x_i^*)|\\ &+ (F_i(x_i^k,\bar{x}^k) - F_i(x_i^*,\bar{x}^*))^T(x_i^k - x_i^*).\end{aligned}$$

By the Lipshitz property of the mapping F_i in Assumption 10, we have

$$|(F_i(x_i^k, N\hat{v}_i^k) - F_i(x_i^k, Ny^k))^T(x_i^k - x_i^*)| \leq L_{-i}N\|\hat{v}_i^k - y^k\|\,\|x_i^k - x_i^*\| \leq L_{-i}NM\|\hat{v}_i^k - y^k\|,$$

where $M \geq \max_{x_i,z_i}\|x_i - z_i\|$ for all i, which exists by compactness of each K_i. Upon combining the preceding estimates with (4.49), we obtain

$$\begin{aligned}\|x_i^{k+1} - x_i^*\|^2 \leq & \|x_i^k - x_i^*\|^2 + 4\alpha_i^2 C^2 + 2\alpha_i L_{-i}NM\|\hat{v}_i^k - y^k\| \\ & - 2\alpha_i(F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*).\end{aligned} \tag{4.50}$$

Now, we work with the last term in (4.50), by letting $\alpha_{\min} = \min_i \alpha_i$, and by adding and subtracting $2\alpha_{\min}(F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*)$, we can see that

$$\begin{aligned}\|x_i^{k+1} - x_i^*\|^2 \leq & \|x_i^k - x_i^*\|^2 + 4\alpha_i^2 C^2 + 2\alpha_i L_{-i}NM\|\hat{v}_i^k - y^k\| \\ & + 2(\alpha_i - \alpha_{\min})|(F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*)| \\ & - 2\alpha_{\min}(F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*).\end{aligned} \tag{4.51}$$

By using the Cauchy-Schwartz inequality and the Lipschitz property of F_i given in Assumption 15, we obtain

$$|(F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*)| \leq L_i\|x_i^k - x_i^*\|^2. \tag{4.52}$$

Further, by letting $\alpha_{\max} = \max_i \alpha_i$, from (4.51) and (4.52) by collecting the common terms we have for $i \in \{I^k, J^k\}$,

$$\begin{aligned}\|x_i^{k+1} - x_i^*\|^2 \leq & (1 + 2L_i(\alpha_{\max} - \alpha_{\min}))\|x_i^k - x_i^*\|^2 + 4\alpha_i^2 C^2 + 2\alpha_i L_{-i}NM\|\hat{v}_i^k - y^k\| \\ & - 2\alpha_{\min}(F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*).\end{aligned}$$

The fact that $x_i^{k+1} = x_i^k$ when $i \notin \{I^k, J^k\}$ implies that $\|x_i^{k+1} - x_i^*\|^2 = \|x_i^k - x_i^*\|^2$ for $i \notin \{I^k, J^k\}$. Next, we take the expectation in (4.53), whereby we combine the preceding two cases and take into account that agent i updates with probability p_i, and obtain for all $i \in \mathcal{N}$,

$$\begin{aligned}\mathbb{E}[\|x_i^{k+1} - x_i^*\|^2 \mid \mathcal{F}_k] \leq & (1 + 2p_i L_i(\alpha_{\max} - \alpha_{\min}))\|x_i^k - x_i^*\|^2 + 4p_i\alpha_i^2 C^2 \\ & + 2p_i\alpha_i L_{-i}NM\,\mathbb{E}[\|\hat{v}_i^k - y^k\| \mid \mathcal{F}_k] - 2p_i\alpha_{\min}(F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*).\end{aligned}$$

Since $\hat{v}_i^k$ is a convex combination of v_j^k, $j \in \mathcal{N}$ and the norm is a convex function, we have $\|\hat{v}_i^k -$

$y^k\| \le \sum_{j=1}^N [W(k)]_{ij} \|v_j^k - y^k\|$. Thus, $\mathbb{E}[\|\hat{v}_i^k - y^k\| \mid \mathcal{F}_k] \le \sum_{j=1}^N \mathbb{E}[W(k)]_{ij} \|v_j^k - y^k\|$. By using the preceding relation, $\min_i p_i = p_{\min}$, $p_{\max} = \max_i p_i$, and $\alpha_i \le \alpha_{\max}$, we arrive at the following relation for all $i \in \mathcal{N}$,

$$\mathbb{E}[\|x_i^{k+1} - x_i^*\|^2 \mid \mathcal{F}_k] \le (1 + 2p_{\max} \max_i L_i(\alpha_{\max} - \alpha_{\min}))\|x_i^k - x_i^*\|^2 + 4p_{\max}\alpha_{\max}^2 C^2$$
$$+ 2p_{\max}\alpha_{\max} B \sum_{j=1}^N \mathbb{E}[W(k)]_{ij}\|v_j^k - y^k\| - 2p_{\min}\alpha_{\min}(F_i(x_i^k, \bar{x}^k) - F_i(x_i^*, \bar{x}^*))^T(x_i^k - x_i^*), \quad (4.53)$$

with $B = (\max_i L_{-i})NM$.

Summing the relations in (4.53) over all $i = 1, \ldots, N$, recalling that F_i, $i = 1, \ldots, N$ are coordinate maps for the map F (see (4.6)), which in turn defines the mapping ϕ through (4.7), we further obtain

$$\mathbb{E}[\|x^{k+1} - x^*\|^2 \mid \mathcal{F}_k] \le (1 + 2p_{\max} \max_i L_i(\alpha_{\max} - \alpha_{\min}))\|x^k - x^*\|^2 + 4p_{\max}\alpha_{\max}^2 C^2 N$$
$$+ 2p_{\max}\alpha_{\max} B \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[W(k)]_{ij}\|v_j^k - y^k\| - 2p_{\min}\alpha_{\min}(\phi(x^k) - \phi(x^*))^T(x^k - x^*),$$

The matrix $\mathbb{E}[W(k)]$ is doubly stochastic, so we have

$$\sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[W(k)]_{ij}\|v_j^k - y^k\| = \sum_{j=1}^N \sum_{i=1}^N \mathbb{E}[W(k)]_{ij}\|v_j^k - y^k\| = \sum_{j=1}^N \|v_j^k - y^k\|.$$

Using this relation and the strong monotonicity of the mapping ϕ with a constant μ, gathering the common terms, and taking the total expectation, we obtain for all $k \ge 0$,

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] \le q\mathbb{E}[\|x^k - x^*\|^2] + 4p_{\max}\alpha_{\max}^2 C^2 N + 2p_{\max}\alpha_{\max} B \sum_{j=1}^N \mathbb{E}[\|v_j^k - y^k\|], \quad (4.54)$$

where $q = 1 - 2\mu p_{\min}\alpha_{\min} + 2p_{\max} \max_i L_i(\alpha_{\max} - \alpha_{\min})$. Note that by the condition

$$0 < 1 - 2\mu p_{\min}\alpha_{\min} + 2p_{\max} \max_i L_i(\alpha_{\max} - \alpha_{\min}) < 1$$

we have $0 < q < 1$. We further have $\{x^k\} \subseteq K$ for a compact set K, so the limit superior of $\mathbb{E}[\|x^k - x^*\|^2]$ exists. Thus, by taking the limit as $k \to \infty$ in relation (4.54) and using Lemma 20, we obtain

$$\limsup_{k\to\infty}\mathbb{E}[\|x^{k+1}-x^*\|^2] \le q\limsup_{k\to\infty}\mathbb{E}[\|x^k-x^*\|^2] + 4p_{\max}\alpha_{\max}^2 C^2 N + 2p_{\max}\alpha_{\max}B\frac{\sqrt{2nN}\,C\alpha_{\max}}{1-\sqrt{\lambda}},$$

which implies the stated result. ■

We have few comments on the result of Proposition 13, as follows. The error bound depends on the dimension n of the decision variables, the number N of players, the frequency with which players update their decisions (captioned by $p_{\min}$ and $p_{\max}$), and the network properties including the connectivity time bound B and the ability to propagate the information (captured by the value $1-\sqrt{\lambda}$). When the network parameters B and λ, and the players' update probabilities ($p_{\min}$ and $p_{\max}$) do not depend on N, the error bound grows linearly with the number N of players.

As a special case, consider the case when the agents employ an equal stepsize, i.e., $\alpha_{\min} = \alpha_{\max} = \alpha$ and α satisfies the following condition $0 < \alpha < \frac{1}{2\mu p_{\min}}$. Then, the result of Proposition 13 reduces to

$$\limsup_{k\to\infty}\mathbb{E}[\|x^k-x^*\|^2] \le \frac{p_{\max}\alpha\left(2C^2N + BC\frac{\sqrt{2nN}}{1-\sqrt{\lambda}}\right)}{\mu p_{\min}}.$$

As another special case, consider the case when all players have equal probabilities of updating, i.e., $p_{\min} = p_{\max} = p$. Then, we have the following result:

$$\limsup_{k\to\infty}\mathbb{E}[\|x^k-x^*\|^2] \le \frac{\alpha_{\max}^2\left(2C^2N + BC\frac{\sqrt{2nN}}{1-\sqrt{\lambda}}\right)}{\mu\alpha_{\min} - (\max_i L_i)(\alpha_{\max}-\alpha_{\min})},$$

When all players have equal probabilities of updating and all use equal stepsizes, i.e., $p_{\min} = p_{\max} = p$ and $\alpha_{\min} = \alpha_{\max} = \alpha$, then the condition of Proposition 13 reduces to $0 < \alpha < \frac{1}{2\mu p}$, and the bound further simplifies to:

$$\limsup_{k\to\infty}\mathbb{E}[\|x^k-x^*\|^2] \le \frac{\alpha}{\mu}\left(2C^2N + BC\frac{\sqrt{2nN}}{1-\sqrt{\lambda}}\right).$$

4.4 Extensions

It may have been observed that the proposed developments in the earlier two sections required that the agent decisions be of the same dimension. In this section, we extend the realm of (4.2) and generalize the algorithms presented in section 4.2 and section 4.3. To this end, consider the following aggregative game

$$\begin{aligned}
&\text{minimize} \qquad f_i(x_i, \sum_{i=1}^{N} h_i(x_i)) \\
&\text{subject to} \qquad x_i \in K_i,
\end{aligned} \tag{4.55}$$

where $K_i \subseteq \mathbb{R}^{n_i}$, $h_i : K_i \to \mathbb{R}^n$. The mappings g_i and h_i are considered to be information private to player i. Such an extension allows players decisions to have different dimensionality. To recover the problem articulated by (4.2), we set $h_i(x_i) = x_i$ with $K_i \subseteq \mathbb{R}^n$ for all i. We next discuss the generalization of the proposed distributed algorithms to solve problem (4.55).

4.4.1 Synchronous Algorithm

To make the synchronous algorithm suitable for the generalized problem in (4.55), the mixing step in (4.8) remains the same, but with a different initial condition. Namely, the mixing in (4.8) is initiated with

$$v_i^0 = h_i(x_i^0) \qquad \text{for all } i = 1, \ldots, N, \tag{4.56}$$

where $x_i^0 \in K_i$ are initial players' decisions. The iterate update of (4.9) and average estimate update of (4.10) modify to become:

$$x_i^{k+1} = \Pi_{K_i}[x_i^k - \alpha_k F_i(x_i^k, N\hat{v}_i^k)], \tag{4.57}$$

$$v_i^{k+1} = \hat{v}_i^k + h_i(x_i^{k+1}) - h_i(x_i^k), \tag{4.58}$$

where, α_k is the stepsize and the mapping F_i is given as

$$F_i\left(x_i, \sum_{i=1}^{N} h_i(x_i)\right) = \nabla_{x_i} f_i\left(x_i, \sum_{i=1}^{N} h_i(x_i)\right), \tag{4.59}$$

and $N\hat{v}_i^k$ in (4.57) is an estimate for the true value $\sum_{i=1}^{N} h_i(x_i)$. For the extended synchronous algorithm in the preceding discussion we have the following result.

Proposition 14. *Let Assumptions 8–13 hold for the mapping $\phi(x) = (\phi_1(x), \ldots, \pi_N(x))^T$ with coordinates $\phi_i(x) = \nabla_{x_i} f_i\left(x_i, \sum_{i=1}^{N} h_i(x_i)\right)$ and $x = (x_1^T, \ldots, x_N^T)^T$. Then, the sequence $\{x^k\}$ generated by the method* (4.57)–(4.58) *converges to the (unique) solution x^* of the game in* (4.55).

Proof. The proof mimics the proof of Proposition 11. ■

4.4.2 Asynchronous Algorithm

The gossip algorithm in section 4.3 can be modified as follows. The estimate mixing in (4.22) remains unchanged, but the initial condition is replaced with the one given in (4.56). The iterate update of (4.27) and average estimate update of (4.28) are modified, as follows:

$$x_i^{k+1} = (\Pi_{K_i}[x_i^k - \alpha_{k,i} F_i(x_i^k, N\hat{v}_i^k)] - x_i^k)\mathbb{1}_{\{i \in \{I^k, J^k\}\}} + x_i^k, \tag{4.60}$$

$$v_i^{k+1} = \hat{v}_i^k + h_i(x_i^{k+1}) - h_i(x_i^k), \tag{4.61}$$

where $\alpha_{k,i}$ is the stepsize for user i and the mapping F_i is as defined in (4.59). The following result establishes the convergence of the extended asynchronous algorithm.

Proposition 15. *Let Assumptions 8–10 and Assumption 14 hold. Then, the sequence $\{x^k\}$ generated by the method* (4.60)–(4.61) *with stepsize $\alpha_{k,i} = \frac{1}{\Gamma_k(i)}$ converges to the (unique) x^* of the game almost surely.*

Proof. With the initial condition specified by (4.56), the proof parallels the line of argument used in the proof of Proposition 12. ■

4.5 Numerics

In this section, we examine the performance of the proposed algorithms on a class of Nash-Cournot games. Such games represent an instance of aggregative Nash games and in section 4.5.1, we describe the player payoffs and strategy sets as well as verify that they satisfy the necessary assumptions. In section 4.5.2, we discuss the synchronous setting and present the results arising from applying our algorithms. In section 4.5.3, we turn our attention to asynchronous regime where we present our numerical experience of applying the gossip algorithm.

4.5.1 Nash-Cournot Game

We consider a networked Nash-Cournot games which is possibly amongst the best known examples of an aggregative game. Specifically, the aggregate in such games is the total sales which is the sum of production over all the players. The market price is set in accord with an inverse demand function which depends on the aggregate of the network. A formal description of such a game over a network is provided in Example 2. Before proceeding to describe our experimental setup, we show that Nash-Cournot games do indeed satisfy Assumptions 10 and 15, respectively, under some mild conditions on the cost and price functions. It is worth pointing that we have used Assumption 15 only for the error bound results for the asynchronous algorithm with a constant stepsize.

In the sequel, within the context of Example 2, we let $x_{il} = (g_{il}, s_{il})$ for all $l = 1, \ldots, \mathcal{L}$, $x_i = (x_{i1}, \ldots, x_{i\mathcal{L}})$ and $x = (x_1, \ldots, x_N)^T$. Further, we define coordinate maps $F_i(x_i, u)$, as follows:

$$F_i(x_i,u) = \begin{pmatrix} F_{i1}(x_{i1},u_1) \\ \vdots \\ F_{i\mathcal{L}}(x_{i\mathcal{L}},u_{\mathcal{L}}) \end{pmatrix}, \qquad F_{il}(x_{il},u_l) = \begin{pmatrix} c'_{il}(g_{il}) \\ -p_l(u_l) - p'_l(u_l)s_{il} \end{pmatrix}, \tag{4.62}$$

where the prime demotes the first derivative. We let $F(x,u) = (F_1(x_1,u)^T, \ldots, F_N(x_N,u)^T)^T$, and K_i denote the constraint set on player i decision, x_i, as given in Example 2.

We note that the Nash-Cournot game under the consideration satisfies Assumption 8 as long as the cost functions c_{il} are convex and the price functions $p_l(u_l)$ are concave for all i and l. Furthermore, the strict convexity condition of Assumption 9 is satisfied when, for example, all price functions p_l are strictly concave. This can be seen by observing that

$$(F(x,u) - F(\tilde{x},u))^T(x - \tilde{x}) = \sum_{l=1}^{\mathcal{L}}\sum_{i=1}^{N}(c'_{il}(g_{il}) - c'_{il}(\tilde{g}_{il}))(g_{il} - \tilde{g}_{il}) - \sum_{l=1}^{\mathcal{L}}\sum_{i=1}^{N} p'_l(u_l)(s_{il} - \tilde{s}_{il})^2.$$

Next, we show that the Lipschitzian requirements on the maps F_i of Assumption 10 holds under some mild assumptions on the cost and price functions in Nash-Cournot games.

Lemma 21. *Consider the Nash-Cournot game described in Example 2. Suppose that each $p_l(u_l)$ is concave and has Lipschitz continuous derivatives with a constant M_l (over a coordinate projection of $\bar{K}$ on the lth coordinate axis). Then, the following relation holds:*

$$\|F_i(x_i,u) - F_i(x_i,z)\| \le \sqrt{2}\sqrt{\sum_{l=1}^{\mathcal{L}}(C_l^2 + M_l^2 \mathrm{cap}_{il}^2)}\,\|u - z\| \qquad \textit{for all } u,z \in \bar{K}.$$

Proof. This result follows directly from the definition of the coordinate maps $F_{il}(x_{il}, u)$ and recalling that $x_{il} = (g_{il}, s_{il})$. In particular, for each i, ℓ we have

$$\begin{aligned}
\|F_{il}(x_{il}, u_l) - F_{il}(x_{il}, z_l)\| &= \sqrt{|c'_{il}(g_{il}) - c'_{il}(g_{il})|^2 + |p_l(u_l) + p'_l(u_l)s_{il} - p_l(z_l) - p'_l(z_l)s_{il}|^2} \\
&= \sqrt{|(p_l(u_l) - p_l(z_l)) + (p'_l(u_l) - p'_l(z_l))s_{il}|^2} \\
&\le \sqrt{2}\sqrt{|p_l(u_l) - p_l(z_l)|^2 + |p'_l(u_l) - p'_l(z_l)|^2 s_{il}^2},
\end{aligned}$$

where the inequality follows from $(a+b)^2 \le 2(a^2+b^2)$. Since $\bar{K}$ is compact and each p_l has continuous derivatives, it follows that there exists a constant C_l for every l such that

$$|p'_l(u_l)| \le C_l \qquad \text{for all } u_l \text{ with } u = (u_1, \ldots, u_{\mathcal{L}})^T \in \bar{K}.$$

Then, by using concavity of p_l, we can see that $|p_l(u_l) - p_l(z_l)| \le C_l |u_l - z_l|$ implying that

$$\begin{aligned}
\|F_{il}(x_{il}, u_l) - F_{il}(x_{il}, z_l)\| &\le \sqrt{2}\sqrt{C_l^2 |u_l - z_l|^2 + |p'_l(u_l) - p'_l(z_l)|^2 s_{il}^2} \\
&\le \sqrt{2}\sqrt{(C_l^2 + M_l^2 s_{il}^2)}|u_l - z_l|,
\end{aligned}$$

where the last inequality is obtained by using the Lipschitz property of the derivative $p'_l(u_l)$. From the structure of constraints we have $s_{il} \le \text{cap}_{il}$ yielding

$$\|F_{il}(x_{il}, u_l) - F_{il}(x_{il}, z_l)\| \le \sqrt{2}\sqrt{C_l^2 + M_l^2 \text{cap}_{il}^2}\,|u_l - z_l| \qquad \text{for all } l.$$

Further, by using Hölder's inequality, and recalling that $x_i = (x_{i1}, \ldots, x_{i\mathcal{L}})$ and $u = (u_1, \ldots, u_{\mathcal{L}})$, from the preceding relation we obtain

$$\|F_i(x_i, u) - F_i(x_i, z)\| = \sqrt{\sum_{l=1}^{\mathcal{L}} \|F_{il}(x_{il}, u_l) - F_{il}(x_{il}, z_l)\|^2} \le \sqrt{2}\sqrt{\sum_{l=1}^{\mathcal{L}} (C_l^2 + M_l^2 \text{cap}_{il}^2)}\,\|u - z\|.$$

■

Lemma 22. *Consider the Nash-Cournot game described in Example 2. Suppose that each c'_{il} is Lipschitz continuous with a constant L_{il} and $|p'_l(u)| \le \bar{p}_l$ for some scalar $\bar{p}_l$ and for all $u \in \bar{K}$. Then, the following relation holds for all i,*

$$\|F_i(x_i, u) - F_i(\tilde{x}_i, u)\| \le \sqrt{\sum_{l=1}^{\mathcal{L}} \left(L_{il}^2 + \bar{p}_l^2\right)}\,\|x_i - \tilde{x}_i\| \qquad \text{for all } x_i, \tilde{x}_i \in K_i.$$

Proof. First, we note that for each i, l,

$$\|F_{il}(x_{il}, u_l) - F_{il}(\tilde{x}_{il}, u_l)\| = \sqrt{|c'_{il}(g_{il}) - c'_{il}(\tilde{g}_{il})|^2 + |p'_l(u_l)(s_{il} - \tilde{s}_{il})|^2}$$
$$\leq \sqrt{L_{il}^2 |g_{il} - \tilde{g}_{il}|^2 + \bar{p}_l^2 |s_{il} - \tilde{s}_{il}|^2}.$$

Recalling our notation $x_{il} = (g_{il}, s_{il})$ and $\tilde{x}_{il} = (\tilde{g}_{il}, \tilde{s}_{il})$, and using Hölder's inequality, we find that

$$\|F_{il}(x_{il}, u_l) - F_{il}(\tilde{x}_{il}, u_l)\| \leq \sqrt{L_{il}^2 + \bar{p}_l^2}\, \|x_{il} - \tilde{x}_{il}\|.$$

Further, recalling that $x_i = (x_{i1}, \ldots, x_{i\mathcal{L}})$, $\tilde{x}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{i\mathcal{L}})$, and $u = (u_1, \ldots, u_{\mathcal{L}})$, the desired result follows from the preceding relation by using Hölders inequality. ■

In our numerical study, we consider a Nash-Cournot game being played over ten locations, i.e. $\mathcal{L} = 10$, in which all players have the same cost type and the ith player's optimization problem can be expressed as

$$\begin{aligned} &\text{minimize} && \sum_{l=1}^{10} (c_{il}(g_{il}) - p_l(\bar{s}_l) s_{il}) \\ &\text{subject to} && \sum_{l=1}^{10} g_{il} = \sum_{l=1}^{10} s_{il}, \\ & && g_{il}, s_{il} \geq 0, \quad g_{il} \leq \text{cap}_{il}, \qquad l = 1, \ldots, 10, \end{aligned} \tag{4.63}$$

where g_{il} and s_{il} denote player i's production and sales at location l, respectively, and $\bar{s}_l$ denotes the aggregate of all the players' decisions ($\bar{s}_l = \sum_{i=1}^N s_{il}$) at location l. The function $c_{il}(g_{il})$ is the cost for ith player at location l and it has the following form:

$$c_{il}(g_{il}) = a_{il} g_{il} + b_{il} g_{il}^2,$$

where a_{il} and b_{il} are scaling parameters for agent i. In our experiments, we draw a_{il} and b_{il} from a uniform distribution and fix them over the course of the entire simulation. More precisely, for $i = 1, \ldots, N$, and $l = 1, \ldots, 10$, we have $a_{il} \sim U(2, 12)$ and $b_{il} \sim U(2, 3)$, where $U(t, \tau)$ denotes the uniform distribution over an interval $[t, \tau]$ with $t < \tau$. The term $p_l(\bar{s}_l)$ captures the inverse demand function and takes the following form:

$$p_l(\bar{s}_l) = d_l - \bar{s}_l,$$

where d_l is a parameter for location l. The parameters d_l are also drawn randomly with a uniform distribution, $d_l \sim U(90,100)$ for all $l = 1,\dots,10$. Furthermore, we use $cap_{il} = 500$ for all $i = 1,\dots,N$ and for all $l = 1,\dots,10$. The affine price function gives rise to a strongly monotone map $\phi = F(x,\bar{x})$, which together with the compactness of the sets K_i, implies that this game has a unique Nash equilibrium.

4.5.2 Synchronous Algorithm

In this section, we investigate the performance of synchronous algorithm of section 4.2 for the computation of the equilibrium of aggregative game (4.63). We begin by describing our setting for the connectivity graph of the network of players, where each player is seen as a node in a graph. At each iteration k, we generate a symmetric $N \times N$ adjacency matrix A such that the underlying graph is connected. The entries of A are generated by performing the following steps:

(0) Let I denote the set of nodes that have already been generated;

(1) For each newly generated node j, select a node randomly $i \in I$ to establish an edge $\{i,j\}$ and set $[A]_{ij} = [A]_{ji} = 1$;

(2) Repeat step 1until $I = \{1,\dots,N\}$.

Given such an adjacency matrix A, we define a doubly stochastic symmetric weight matrix W such that

$$[W]_{ij} = \begin{cases} 0 & \text{if } A_{ij} = 0 \\ \delta & \text{if } A_{ij} = 1 \text{ and } i \neq j \\ 1 - \delta d(i) & \text{if } i = j, \end{cases}$$

where $d(i)$ represents the number of players communicating with player i, and

$$\delta = \frac{0.5}{\max_i\{d(i)\}}.$$

Using the adjacency matrix A and the weight matrix W players update their decision and their estimate of the average using (4.8)–(4.10). The stepsize rule for agent update is as follows:

$$\alpha_{k,i} = \frac{1}{k} \quad \text{for all } i = 1,\dots,N.$$

The algorithm is initiated at a random starting point, and it is terminated after a fixed number of iterations, denoted by $\tilde{k}$, for each sample path. We use a set of 50 sample paths for each simulation setting, and we report the mean of the sample errors, defined as:

$$\text{error}_{\tilde{k}} \frac{\max\limits_{i\in\{1,\ldots,N\}} \max\limits_{l\in\{1,\ldots,10\}} \left\{|g_{il}^{\tilde{k}} - g_{il}^*|, |s_{il}^{\tilde{k}} - s_{il}^*|\right\}}{\max\limits_{i\in\{1,\ldots,N\}} \max\limits_{l\in\{1,\ldots,10\}} \{|g_{il}^*|, |s_{il}^*|\}}, \tag{4.64}$$

where g_{il}^* and s_{il}^* are the decisions of agent i at the Nash equilibrium. The Nash equilibrium decisions g_{il}^* and s_{il}^* are computed using a constant steplength gradient projection algorithm assuming each agent has true information of the aggregate. Note that such an algorithm is guaranteed to converge under the strict convexity of the players' costs.

We investigate cases with 20 and 50 players in the network. In Table 4.1 and Table 4.2, we report the mean terminating error and confidence interval attained for different number $\tilde{k}$ of iterations, respectively. Some insights that can be drawn from the simulations are provided next:

- Expectedly, as seen in Table 4.1 and Table 4.2, the mean terminating error and width of the confidence interval decreases with increasing $\tilde{k}$ and increases with network size.

- The impact of the time-varying nature of the connectivity graph is explored by considering a static complete graph as a basis for comparison. In Table 4.3 and Table 4.4 we report the mean error and the confidence interval when the network is static. Under this setting, the agents have access to the true aggregate information throughout the run of the algorithm. Naturally, the performance of the algorithm on a static complete network is orders of magnitude better than that on a dynamic network. This deterioration in performance may be interpreted as the price of information from the standpoint of convergence.

4.5.3 Asynchronous Algorithm

We now demonstrate the performance of the asynchronous algorithm of section 4.3. We consider four instances of connectivity graphs which we describe next and, also, depict these graphs in Figure 4.3.

- *Cycle*: Every player has two neighbors;

- *Wheel*: There is one central player that is connected to every other player;

- *Grid*: Players on the vertex have two neighbors, players on the edge have three and everyone else has four neighbors. Each row in the grid consists of five players and there are $N/5$ rows where N is the size of the network;

- *Complete graph:* Every player has an edge connecting it to every other player.

For every type of connectivity graph, we initiate the algorithm from a random starting point and terminate it after $\tilde{k}$ iterations. A 95% confidence interval of the mean sample error at the termination is computed for a sample of size 50, where the sample-path error is defined as in (4.64). The players' stepsize rules that we use are:

$$\alpha_{k,i} = \begin{cases} \frac{9}{\Gamma_k(i)} & \text{for a diminishing stepsize} \\ \alpha_i & \text{for a constant stepsize,} \end{cases}$$

where α_i is randomly drawn from a uniform distribution, $\alpha_i \sim U(5e\text{-}3, 1e\text{-}2)$. We again investigate cases when there are 20 and 50 players in the network and derive the following insights:

- In Tables 4.5–4.6, we report the mean error and in Tables 4.7–4.8, we report the width of the confidence interval for various levels of $\tilde{k}$. The results are consistent with our theoretical findings, and they indicate a decrease in the mean error and the width of the confidence interval with increasing $\tilde{k}$. As expected, the mean error at the termination increases with the size of the network. It is worth mentioning the discrepancy in the value of $\tilde{k}$ across synchronous and asynchronous algorithm. Note that in the asynchronous algorithm, only two agents are performing updates and thus, for a network of size N, $\tilde{k}$ global iterations translates to $2\tilde{k}/N$ iterations per agent, approximately.

- On comparing the performance of the synchronous algorithm (cf. Tables 4.1–4.4) to that of the asynchronous algorithm (cf. Tables 4.5–4.8), we observe that the synchronous algorithm performs better than its asynchronous counterpart in terms of mean error and the confidence width at termination. This is expected as in the synchronous setting, the players' communicate more frequently and the network diffuses information faster than in the asynchronous setting.

- The nature of the connectivity graph plays an important role in the performance of the synchronous algorithm. However, such an influence in the asynchronous setting is less pronounced.

- In an effort to better understand the impact of connectivity, in Table 4.9, we compare the number of iterations[3] required for the player's to concur on the aggregate $\bar{g}^*$ within a threshold of 1e-3 when the network consists of $N = 20$ players. We also present a metric of

[3]The iteration number is the mean for 50 sample rounded to the smallest integer over-estimate.

connectivity density given by $p_{\min}/p_{\max}$ as well as the square root of the second largest eigenvalue of the expected weight matrix, i.e., $\sqrt{\lambda}$, which in effect determines the rate of information dissemination in the network. We note that the number of iterations needed to achieve the threshold error correlates with the value of $\sqrt{\lambda}$ and this prompts us to arrive at the following conclusion: Having a well-informed up-to-date neighbor is more important than having a denser connectivity. For instance, a wheel network has a poor connectivity of all the network type based on the $p_{\min}/p_{\max}$ criterion yet it has superior aggregate convergence to all but the complete network. In part, this is because the central agent in such a network updates throughout the course of the algorithm, allowing for good mixing of network wide information. In contrast, the cycle network though better connected yet cannot ensure good mixing of information, given that no agent has access to "good information." Similarly, a complete network provides each agent with an opportunity to communicate with every other agent and thus ensures good mixing of information. The grid network falls between the wheel and the cycle network in terms of availability of well-informed neighbors and thus the performance.

4.6 Summary and conclusions

This chapter focuses on a class of Nash games in which user interactions are seen through the aggregate sum of all players' actions. These agents are a part of network with limited connectivity which only allows for restricted local communication. We propose two classes of distributed algorithm, one synchronous and other asynchronous which abides the information exchange restriction for computation of equilibrium point. Our distributed synchronous algorithm can also contend with dynamic graph with varying connectivity. In contrast, our asynchronous algorithm allows for implementation with distributed architecture. Moreover, we establish error bounds on the deviation of user's decision from the equilibrium decision when a constant yet user specific stepsize is employed in the asynchronous algorithm. Our extension allows the users' decision to be an independent space. The contribution of our work can broadly be summarized as: (1) the development of synchronous and asynchronous distributed algorithm for aggregative games over graphs; (2) the establishment of the convergence of the algorithm (with agent specific stepsizes) to an equilibrium point; and (3) extension to more general classes of aggregative games. We also provide illustrative numerical results that support our theoretical findings.

Table 4.1: Dynamic network: Mean terminating error vs network size for various thresholds

N	$\tilde{k}=5e3$	$\tilde{k}=1e4$
20	9.22e-5	3.66e-5
50	8.38e-2	2.65e-3

Table 4.2: Dynamic network: Width of confidence interval of mean error

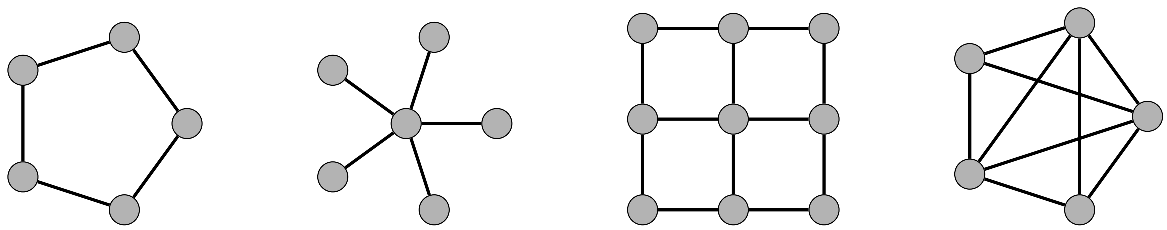
N	$\tilde{k}=5e3$	$\tilde{k}=1e4$
20	2.147e-4	9.33e-5
50	1.24e-1	2.78e-2

Table 4.3: Static network: Mean terminating error

N	$\tilde{k}=5e3$	$\tilde{k}=1e4$
20	3.66e-5	3.66e-5
50	6.23e-5	6.23e-5

Table 4.4: Static network: Width of confidence interval of mean error

N	$\tilde{k}=5e3$	$\tilde{k}=1e4$
20	6.99e-9	6.99e-9
50	1.61e-8	1.61e-8



(a) Cycle (b) Wheel (c) Grid (d) Complete graph

Figure 4.3: A depiction of communication networks used in simulations.

Table 4.5: Mean error after $\tilde{k}=5e4$ iterations for gossip algorithm

	Constant Step				Diminishing Step			
N	Cycle	Wheel	Grid	Complete	Cycle	Wheel	Grid	Complete
20	2.29e-3	3.66e-5	3.66e-5	3.66e-5	2.51e-2	1.01e-4	1.93e-3	4.64e-5
50	2.80e-1	6.76e-2	1.76e-1	1.26e-3	1.22	2.33e-2	8.41e-1	3.68e-3

Table 4.6: Mean error after $\tilde{k}=1e5$ iterations for gossip algorithm

	Constant Step				Diminishing Step			
N	Cycle	Wheel	Grid	Complete	Cycle	Wheel	Grid	Complete
20	3.78e-5	3.66e-5	3.66e-5	3.66e-5	3.93e-3	3.65e-5	1.69e-4	3.67e-5
50	1.65e-1	1.09e-3	9.19e-2	6.23e-5	7.63e-1	1.99e-3	4.57e-1	2.83e-4

Table 4.7: Width of confidence interval after $\tilde{k}=5e4$ iterations for gossip algorithm

	Constant Step				Diminishing Step			
N	Cycle	Wheel	Grid	Complete	Cycle	Wheel	Grid	Complete
20	1.87e-4	8.22e-7	5.34e-7	0e0	7.51e-2	4.76e-3	2.08e-2	3.23e-3
50	1.68e-2	6.33e-3	1.89e-2	1.77e-4	5.23e-1	7.23e-2	4.35e-1	2.87e-2

Table 4.8: Width of confidence interval after $\tilde{k} = 1e5$ iterations for gossip algorithm

	Constant Step				Diminishing Step			
N	Cycle	Wheel	Grid	Complete	Cycle	Wheel	Grid	Complete
20	2.4e-6	4.74e-10	4.74e-10	4.74e-10	4.40e-4	7.23e-7	2.07e-5	1.56e-7
50	1.35e-2	1.33e-4	1.40e-2	1.78e-8	6.91e-2	1.81e-4	4.15e-2	2.22e-5

Table 4.9: Number of iteration for concurrence of player's aggregate within an error of 1e-3

Network	$p_{\min}/p_{\max}$	λ	Iterations
Cycle	1	0.9994	48818
Wheel	1/19	0.1622	8324
Grid	5/7	0.3151	17950
Complete	1	1.0888e-08	5842

Appendix A

Bound on optimal value

Lemma 23. *Let Assumption 2 hold. Then, for each* $\nu > 0$*, we have*

$$0 \leq f(x_\nu^*) - f^* \leq \frac{\nu}{2}(D^2 - \|x_\nu^*\|^2) \qquad \textit{where } D = \max_{x \in X} \|x\|.$$

Proof. Under Assumption 2, both the original problem and the regularized problem have solutions. Since the regularized problem is strongly convex, the solution $x_\nu^* \in X$ is unique for every $\nu > 0$. Furthermore, we have

$$f_\nu(x_\nu^*) - f_\nu(x) \leq 0 \qquad \text{for all } x \in X.$$

Letting $x = x^*$ in the preceding relation, and using $f_\nu(x) = f(x) + \frac{\nu}{2}\|x\|^2$ and $f^* = f(x^*)$ we get

$$f(x_\nu^*) - f^* \leq \frac{\nu}{2}\left(\|x^*\|^2 - \|x_\nu^*\|^2\right).$$

Since $x^* \in X$ solves the original problem and $x_\nu^* \in X$, we have $0 \leq f(x_\nu^*) - f(x^*)$. Thus, from $f^* = f(x^*)$, using $D = \max_{x \in X} \|x\|$, it follows that $0 \leq f(x_\nu^*) - f^* \leq \frac{\nu}{2}\left(\|x^*\|^2 - \|x_\nu^*\|^2\right) \leq \frac{\nu}{2}(D^2 - \|x_\nu^*\|^2)$. ■

References

[1] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness, and stability," *Journal of the Operational Research Society*, vol. 49, pp. 237–252, 1998. *(Cited on pages 3, 7, and 8)*

[2] R. Srikant, *Mathematics of Internet Congestion Control*. Birkhauser, 2004. *(Cited on pages 3, 7, 8, 10, and 41)*

[3] S. Low and D. Lapsley, "Optimization flow control. i. basic algorithm and convergence," *Networking, IEEE/ACM Transactions on*, vol. 7, no. 6, pp. 861–874, dec. 1999. *(Cited on pages 3, 7, 38, and 41)*

[4] E. Altman, T. Başar, T. Jiménez, and N. Shimkin, "Competitive routing in networks with polynomial cost," in *INFOCOM (3)*, 2000. [Online]. Available: citeseer.comp.nus.edu.sg/article/altman00competitive.html pp. 1586–1593. *(Cited on page 3)*

[5] T. Alpcan and T. Başar, "A game-theoretic framework for congestion control in general topology networks," in *41th IEEE Conf. Decision and Control*, Las Vegas, NV, December 2002. *(Cited on pages 3, 4, 10, 13, 79, and 94)*

[6] H. Yin, U. V. Shanbhag, and P. G. Mehta, "Nash equilibrium problems with scaled congestion costs and shared constraints," *Automatic Control, IEEE Transactions on*, vol. 56, no. 7, pp. 1702 –1708, july 2011. *(Cited on pages 3, 10, 14, 57, 79, and 94)*

[7] T. Alpcan and T. Başar, "Distributed algorithms for Nash equilibria of flow control games," in *Advances in Dynamic Games*, ser. Annals of the International Society of Dynamic Games, vol. 7. Birkhäuser Boston, 2003. [Online]. Available: http://citeseer.ifi.unizh.ch/alpcan03distributed.html pp. 473–498. *(Cited on pages 4, 10, 13, and 94)*

[8] D. Dutta, A. Goel, and J. Heidemann, "Oblivious aqm and nash equilibria," *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE*, vol. 1, pp. 106–113 vol.1, March-3 April 2003. [Online]. Available: http://citeseer.ist.psu.edu/dutta03oblivious.html *(Cited on page 4)*

[9] M. Bloem, T. Alpcan, and T. Başar, "A stackelberg game for power control and channel allocation in cognitive radio networks," in *ValueTools '07: Proceedings of the 2nd international conference on Performance evaluation methodologies and tools*. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2007, pp. 1–9. *(Cited on page 4)*

[10] G. Scutari, F. Facchinei, J.-S. Pang, and D. Palomar, "Monotone games in communications: Theory, algorithms, and models," *submitted to IEEE Transactions on Information Theory*, April 2010. *(Cited on pages 4 and 10)*

[11] B. F. Hobbs and J. S. Pang, "Nash-Cournot equilibria in electric power markets with piecewise linear demand functions and joint constraints," *Oper. Res.*, vol. 55, no. 1, pp. 113–127, 2007. *(Cited on page 4)*

[12] B. F. Hobbs, "Linear complementarity models of Nash-Cournot competition in bilateral and poolco power markets," *IEEE Transactions on Power Systems*, vol. 16, no. 2, pp. 194–202, 2001. *(Cited on pages 4 and 10)*

[13] L. Pavel, "A noncooperative game approach to OSNR optimization in optical networks," *IEEE Transactions on Automatic Control*, vol. 51, no. 5, pp. 848–852, 2006. *(Cited on pages 4 and 94)*

[14] Y. Pan and L. Pavel, "Games with coupled propagated constraints in optical network with multi-link topologies," *Automatica*, vol. 45, pp. 871–880, 2009. *(Cited on pages 4, 10, and 13)*

[15] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff-based dynamics for multiplayer weakly acyclic games," *SIAM J. Control Optim.*, vol. 48, no. 1, pp. 373–396, 2009. [Online]. Available: http://dx.doi.org/10.1137/070680199 *(Cited on page 4)*

[16] J. S. Shamma and G. Arslan, "Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria," *IEEE Trans. Automat. Control*, vol. 50, no. 3, pp. 312–327, 2005. [Online]. Available: http://dx.doi.org/10.1109/TAC.2005.843878 *(Cited on page 4)*

[17] A. Cournot, "Recherches sur les prinicipes mathématiques de la théorie des richesses," 1838. *(Cited on pages 5, 11, 12, and 90)*

[18] K. Okuguchi and F. Szidarovszky, *The theory of oligopoly with multi-product firms*, ser. Lecture Notes in Economics and Mathematical Systems. Berlin: Springer-Verlag, 1990, vol. 342. *(Cited on pages 5 and 12)*

[19] C. Metzler, B. Hobbs, and J.-S. Pang, "Nash-cournot equilibria in power markets on a linearized dc network with arbitrage: Formulations and properties," *Networks and Spatial Theory*, vol. 3, no. 2, pp. 123–150, 2003. *(Cited on pages 5 and 12)*

[20] A. Kannan and U. V. Shanbhag, "Distributed computation of equilibria in monotone Nash games via iterative regularization techniques," *To appear in SIAM Journal of Optimization*, 2012. *(Cited on pages 5, 12, 13, 57, and 94)*

[21] R. Cornes and R. Hartley, "Aggregative public good games," *Journal of Public Economic Theory*, vol. 9, no. 2, pp. 201–219, Apr 2007. *(Cited on page 5)*

[22] B. D. Bernheim and M. D. Whinston, "Common agency," *Econometrica*, vol. 54, no. 4, pp. pp. 923–942, 1986. *(Cited on page 5)*

[23] R. Hartley and R. Cornes, "Joint production games and share functions," Department of Economics, Keele University, Keele Department of Economics Discussion Papers (1995-2001), 2000. *(Cited on page 5)*

[24] S. Shakkottai and R. Srikant, "Network optimization and control," *Foundations and Trends in Networking*, vol. 2, no. 3, pp. 271–379, 2007. *(Cited on page 7)*

[25] M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255 –312, jan. 2007. *(Cited on page 7)*

[26] X. Qin and R. Berry, "Distributed power allocation and scheduling for parallel channel wireless networks," in *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, 2005. WIOPT 2005. Third International Symposium on*, april 2005, pp. 77 – 85. *(Cited on page 7)*

[27] A. Nedić and A. Ozdaglar, "Subgradient methods in network resource allocation: Rate analysis," in *42nd Annual Conference on Information Sciences and Systems, CISS, Princeton*, 2008, pp. 1189–1194. *(Cited on pages 7 and 38)*

[28] A. Nedić and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM J. of Optimization*, vol. 19, no. 4, pp. 1757–1780, 2009. *(Cited on page 7)*

[29] A. Nedić and A. Ozdaglar, "Subgradient methods for Saddle-Point problems," *Journal of Optimization Theory and Applications*, vol. 142, no. 1, pp. 205–228, 2009. *(Cited on page 7)*

[30] J. Koshal, A. Nedić, and U. Shanbhag, "Distributed multiuser optimization: Algorithms and error analysis," in *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, dec. 2009, pp. 4372 –4377. *(Cited on page 9)*

[31] J. Koshal, A. Nedić, and U. Shanbhag, "Multiuser optimization: Distributed algorithms and error analysis," *SIAM J. Optim.*, vol. 21, no. 3, pp. 1046–1081, 2011. [Online]. Available: http://dx.doi.org/10.1137/090770102 *(Cited on page 9)*

[32] A. Rantzer, "Using game theory for distributed control engineering," Department of Automatic Control, Lund University, Sweden, Tech. Rep. ISRN LUTFD2/TFRT--7620--SE, July 2008, presented at Games 2008, 3rd World Congress of the Game Theory Society. *(Cited on pages 10 and 13)*

[33] L. Pavel, "An extension of duality to a game-theoretic framework," *Automatica*, vol. 43, pp. 226–237, 2007. *(Cited on pages 10, 13, and 94)*

[34] A. Kannan, U. Shanbhag, and H. Kim, "Strategic behavior in power markets under uncertainty," *To appear in Energy Systems*, no. To appear, 2011. *(Cited on page 10)*

[35] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer-Verlag Inc., New York, 2003, vol. 1 and 2. *(Cited on pages 10, 14, 16, 18, 56, 59, 60, 61, 64, 72, 74, 92, 94, 101, and 118)*

[36] H. Jiang and H. Xu, "Stochastic approximation approaches to the stochastic variational inequality problem," *IEEE Transactions on Automatic Control,*, vol. 53, no. 6, pp. 1462–1475, 2008. *(Cited on pages 10, 56, 59, 60, and 88)*

[37] A. Tikhonov and V. Ars'enine, *M'ethodes de resolution de probl'emes mal pos'es*. Moscow: 'Editions Mir, 1976, traduit du russe par Vladimir Kotliar. *(Cited on pages 10 and 59)*

[38] A. Tikhonov, "On the solution of incorrectly put problems and the regularisation method," in *Outlines Joint Sympos. Partial Differential Equations (Novosibirsk, 1963)*. Acad. Sci. USSR Siberian Branch, Moscow, 1963, pp. 261–265. *(Cited on pages 10 and 59)*

[39] B. Martinet, "Régularisation d'inéquations variationnelles par approximations successives," *Rev. Française Informat. Recherche Opérationnelle*, vol. 4, no. Ser. R-3, pp. 154–158, 1970. *(Cited on pages 10, 56, and 72)*

[40] J. Koshal, A. Nedić, and U. V. Shanbhag, "Single timescale regularized stochastic approximation schemes for monotone nash games under uncertainty," *Proceedings of the IEEE Conference on Decision and Control (CDC)*, 2010. *(Cited on pages 11 and 13)*

[41] J. Koshal, A. Nedić, and U. V. Shanbhag, "Single timescale stochastic approximation for stochastic nash games in cognitive radio systems," in *Digital Signal Processing (DSP), 2011 17th International Conference on*, july 2011, pp. 1–8. *(Cited on page 11)*

[42] W. Novshek, "On the existence of cournot equilibrium," *Review of Economic Studies*, vol. 52, no. 1, pp. 85–98, 1985. *(Cited on page 11)*

[43] M. K. Jensen, "Aggregative games," Tech. Rep., 2006. *(Cited on page 11)*

[44] D. Martimort and L. Stole, "Aggregate representations of aggregate games," Tech. Rep., 2011. *(Cited on page 11)*

[45] B. Schipper, "Pseudo-potential games," 2004. *(Cited on page 11)*

[46] P. Dubey, O. Haimanko, and A. Zapechelnyuk, "Strategic complements and substitutes, and potential games," *Games and Economic Behavior*, vol. 54, no. 1, pp. 77–94, January 2006. *(Cited on page 11)*

[47] D. Monderer and L. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, pp. 124–143, 1996. *(Cited on page 11)*

[48] D. Fudenberg and J. Tirole, *Game Theory*. MIT Press, 1991. *(Cited on pages 11 and 90)*

[49] J. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D. dissertation, Massachusetts Institute of Technology, 1985. *(Cited on pages 13, 14, 91, and 95)*

[50] D. P. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Prentice Hall Inc., 1989. *(Cited on page 14)*

[51] H. Yin, U. Shanbhag, and P. Mehta, "Nash equilibrium problems with congestion costs and shared constraints," in *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, dec. 2009, pp. 4649 –4654. *(Cited on page 14)*

[52] D. P. Bertsekas, *Nonlinear Programming*. Bellmont, MA: Athena Scientific, 1999. *(Cited on pages 14 and 28)*

[53] B. T. Polyak, *Introduction to Optimization*. Optimization Software, May 1987. *(Cited on page 14)*

[54] G. Korpelevich, "The extragradient method for finding saddle points and other problems," *Matekon*, vol. 12, pp. 747–756, 1976. *(Cited on page 14)*

[55] J.-S. Pang and D. Chan, "Iterative methods for variational and complementarity problems," *Mathematical Programming*, vol. 24, no. 3, pp. 284–313, 1982. *(Cited on page 14)*

[56] M. Solodov and P. Tseng, "Modified projection-type methods for monotone variational inequalities," *SIAM Journal on Control and Optimization*, vol. 34, no. 5, pp. 1814–1830, 1996. *(Cited on page 14)*

[57] I. Konnov, "Methods for solving multivalued variational inequalities," in *At the turn of the 20th–21st centuries (Russian)*. Kazan. Mat. Obs., Kazan, 2003, pp. 68–79. *(Cited on pages 14 and 45)*

[58] I. V. Konnov, *Equilibrium models and variational inequalities*, ser. Mathematics in Science and Engineering. Elsevier B. V., Amsterdam, 2007, vol. 210. *(Cited on page 14)*

[59] Y. E. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming, Ser. A*, vol. 103, pp. 127–152, 2005. *(Cited on page 14)*

[60] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," 2008, submitted to SIAM Journal on Optimization. *(Cited on page 14)*

[61] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring," pp. 693–696, april 2009. *(Cited on page 14)*

[62] A. Beck and M. Teboulle, "Gradient-based algorithms with applications to signal-recovery problems," in *Convex Optimization in Signal Processing and Communications*, Y. Eldar and D. Palomar, Eds. Cambridge University Press, 2010, pp. 42–88. *(Cited on page 14)*

[63] A. Nemirovski, "Prox-method with rate of convergence O(1/T) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems," *SIAM Journal On Optimization*, vol. 15, pp. 229–251, 2004. *(Cited on pages 14, 54, and 56)*

[64] Y. E. Nesterov, "Dual extrapolation and its applications to solving variational inequalities and related problems," *Mathematical Programming*, vol. 109, no. 2, pp. 319–344, Mar. 2007. [Online]. Available: http://dx.doi.org/10.1007/s10107-006-0034-z *(Cited on pages 14, 54, and 56)*

[65] D. P. Bertsekas, A. Nedi'c, and A. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA: Athena Scientific, 2003. *(Cited on page 19)*

[66] H. Uzawa, "Iterative methods in concave programming," *Studies in Linear and Nonlinear Programming*, pp. 154–165, 1958. *(Cited on page 19)*

[67] R. H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM Journal Optimization*, vol. 9, no. 4, pp. 877–900 (electronic), 1999, dedicated to John E. Dennis, Jr., on his 60th birthday. *(Cited on pages 49 and 84)*

[68] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951. *(Cited on page 56)*

[69] Y. Ermoliev, "Stochastic quazigradient methods," in *Numerical Techniques for Stochastic Optimization*. Springer-Verlag, N.Y., 1988, pp. 141–186. *(Cited on page 56)*

[70] D. Bertsekas and J. Tsitsiklis, "Gradient convergence in gradient methods," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000. *(Cited on page 56)*

[71] V. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008. *(Cited on page 56)*

[72] S. Ram, A. Nedić, and V. Veeravalli, "Incremental stochastic sub-gradient algorithms for convex optimization," *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 691–717, 2009. *(Cited on page 56)*

[73] B. Polyak, *Introduction to Optimisation*. New York: Optimization Software, Inc., 1987. *(Cited on pages 56, 57, and 59)*

[74] I. Konnov, *Equilibrium Models and Variational Inequalities*. Elsevier, 2007. *(Cited on page 57)*

[75] A. Kannan and U. Shanbhag, "Distributed iterative regularization algorithms for monotone Nash games," *Proceedings of the IEEE Conference on Decision and Control (CDC)*, 2010. *(Cited on pages 57 and 94)*

[76] V. S. Borkar, "Stochastic approximation with two time scales," *Syst. Control Lett.*, vol. 29, pp. 291–294, 1997. [Online]. Available: http://portal.acm.org/citation.cfm?id=252145.252157 *(Cited on page 57)*

[77] H. D. Qi, "Tikhonov regularization methods for variational inequality problems," *Journal of Optimization Theory and Applications*, vol. 102, pp. 193–201, 1999. *(Cited on page 64)*

[78] R. Rockafellar, "Augmented lagrangians and applications of the proximal point algorithm in convex programming," *Mathematics of Operations Research*, vol. 1, no. 2, pp. 97–116, 1976. *(Cited on page 72)*

[79] F. Facchinei, A. Fischer, and V. Piccialli, "On generalized Nash games and variational inequalities," *Oper. Res. Lett.*, vol. 35, no. 2, pp. 159–164, 2007. *(Cited on page 82)*

[80] U. Ravat and U. V. Shanbhag, "On the characterization of solution sets of smooth and non-smooth convex stochastic nash games," *SIAM Journal of Optimization*, vol. 21, no. 3, pp. 1168–1199, 2011. *(Cited on page 83)*

[81] M. K. Jensen, "Aggregative games and best-reply potentials," *Econom. Theory*, vol. 43, no. 1, pp. 45–66, 2010. *(Cited on page 90)*

[82] S. S. Ram, A. Nedić, and V. Veeravalli, "Distributed stochastic subgradient projection algorithms forconvex optimization," *Journal of Optimization Theory and Applications*, vol. 147, pp. 516–545, 2010. *(Cited on pages 90 and 101)*

[83] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, 2006. *(Cited on pages 91 and 104)*

[84] J. Koshal, A. Nedić, and U. Shanbhag, "A gossip algorithm for aggregative games on graphs," accepted in Conference on Decision and Control (CDC) 2012. *(Cited on page 91)*

[85] A. Nedić and A. E. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Automat. Contr.*, vol. 54, no. 1, pp. 48–61, 2009. *(Cited on page 91)*

[86] S. S. Ram, A. Nedić, and V. Veeravalli, "Asynchronous gossip algorithms for stochastic optimization: Constant stepsize analysis," in *Recent Advances in Optimization and its Applications in Engineering*, M. Diehl, F. Glineur, E. Jarlebring, and W. Michiels, Eds. Volume of the 14th Belgian-French-German Conference on Optimization (BFG), 2010, pp. 51–60. *(Cited on page 91)*

[87] S. S. Ram, A. Nedić, and V. Veeravalli, "Asynchronous gossip algorithms for stochastic optimization," 2009, pp. 3581–3586. *(Cited on pages 91, 106, and 107)*

[88] A. Nedić, A. E. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Trans. Automat. Contr.*, vol. 55, no. 4, pp. 922–938, 2010. *(Cited on page 91)*

[89] K. Srivastava and A. Nedić, "Distributed asynchronous constrained stochastic optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 772–790, 2011. *(Cited on page 91)*

[90] A. Nedić, "Asynchronous broadcast-based convex optimization over a network," *IEEE Transactions on Automatic Control*, vol. 56, no. 6, pp. 1337–1351, 2011. *(Cited on pages 91 and 106)*

[91] I. Lobel and A. E. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Trans. Automat. Contr.*, vol. 56, no. 6, pp. 1291–1306, 2011. *(Cited on page 91)*

[92] S.-Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks," *CoRR*, vol. abs/1205.3993, 2012. *(Cited on page 91)*

[93] S. S. Ram, A. Nedić, and V. V. Veeravalli, "A new class of distributed optimization algorithms: application to regression of distributed data," *Optimization Methods and Software*, vol. 27, no. 1, pp. 71–88, 2012. *(Cited on pages 91 and 97)*

[94] S.-Y. Tu and A. H. Sayed, "On the influence of informed agents on learning and adaptation over networks," *CoRR*, vol. abs/1203.1524, 2012. *(Cited on page 91)*

[95] A. N. K. Srivastav and D. Stipanović, "Distributed bregman-distance algorithms for min-max optimization," to appear in the book Agent-Based Optimization I. Czarnowski, P. Jedrzejowicz and J. Kacprzyk (Eds.), Springer Studies in Computational Intelligence (SCI), 2012. *(Cited on page 91)*

[96] H. Boche, M. Wiczanowski, and S. Stanczak, "Unifying view on min-max fairness, max-min fairness, and utility optimization incellular networks," *EURASIP J. Wirel. Commun. Netw.*, vol. 2007, no. 1, pp. 5–5, Jan. 2007. *(Cited on page 91)*

[97] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. Tsitsiklis, "On distributed averaging algorithms and quantization effects," in *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, 2008. *(Cited on page 97)*