

© 2013 Soomin Lee

OPTIMIZATION OVER NETWORKS:
EFFICIENT ALGORITHMS AND ANALYSIS

BY

SOOMIN LEE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Associate Professor Angelia Nedić, Chair
Professor Dan Roth
Associate Professor Olgica Milenkovic
Professor Venugopal Veeravalli

ABSTRACT

A number of important problems that arise in various application domains can be formulated as a distributed convex constrained minimization problem over a multi-agent network. The problem is usually defined as a sum of convex objective functions over an intersection of convex constraint sets. The first part of this thesis is focused on the development and analysis of efficient distributed algorithms for a constrained convex optimization problem over a multi-agent network where each agent has its own objective function and constraint set. We propose gradient descent algorithms with *random projections* which use various communication protocols.

First, we present a *distributed random projection* (DRP) algorithm whereby each agent exchanges local information only with its immediate neighbors at each iteration. With reasonable assumptions, we prove that the iterates of all agents converge to the same point in the optimal set with probability 1. In addition, we consider a variant of the method that uses a mini-batch of consecutive random projections and establish its convergence. Experiments on distributed support vector machines demonstrate fast convergence of the DRP algorithm. It actually shows that the number of iterations required for convergence is much smaller than that for scanning over all training samples just once.

Second, we propose an *asynchronous gossip-based random projection* (GRP) algorithm that solves the distributed problem using gossip type communications and local computations. We analyze the convergence properties of the algorithm for an uncoordinated diminishing stepsize and a constant stepsize. For a diminishing stepsize, we prove that the iterates of all agents converge to the same optimal point with probability 1. For a constant stepsize, we establish an error bound on the expected distance from the optimal point to the iterates of the algorithm. In addition, we consider a variant of the method that uses a mini-batch of consecutive random projections and,

also, establish its convergence. Furthermore, we provide simulation results on a distributed robust model predictive control problem.

In the second part of the thesis, we discuss an efficient epoch gradient descent algorithm for obtaining fast and exact solutions of linear support vector machines (SVMs). SVMs penalized with the popular hinge-loss are strongly convex but they do not have Lipschitz continuous gradient. We find SVMs that have both strong-convexity and Lipschitz continuous gradient using a smooth approximation technique.

To my parents, for their love and support.

ACKNOWLEDGMENTS

Preparing one small step into the world, I must admit that the luckiest thing in my PhD study was to meet my advisor, Professor Angelia Nedić. I express my deep sense of gratitude for her encouragement, patience, support and guidance in the hardest time of my life. Besides that I was really impressed by the true research spirit that she has shown to me. She has been and will always be the greatest role model in my life.

I would like to thank my thesis committee members for their helpful comments during the research process. I am especially thankful to Professor Dan Roth for his wonderful courses on machine learning and fruitful discussions about my research, which helped the completion of this thesis a lot. I would also like to thank Professor Olgica Milenkovic for accepting to be my academic advisor and supporting me in time of need. I would also like to thank Professor Venu Veeravalli for serving in my thesis committee and providing me with lots of helpful comments.

I also thank Professor Yoram Bresler for being a great mentor, taking such a good care of my husband, treating us like his family and allowing me to share the office with his group members.

My gratitude also goes to Professor Seth Hutchinson, Professor Sean Meyn, and Professor Daniel Liberzon who made my TA experience so wonderful. They treated me like a colleague and not a student. While I was working with them, I learned the joy of teaching and helping others.

I greatly acknowledge my wonderful colleagues who made my stay at Urbana-Champaign so memorable and pleasant. I would like to specially thank Behrouz Touri for his friendship and fruitful research discussions. I also thank Deniz Tursun, Rasoul Etesami, Farzad Yousefian and Jayash Koshal for helping me with numerous things throughout my PhD.

Last but certainly not least, I deeply appreciate the support of my beloved family. My husband Kiryung, thank you for supporting me with endless love,

understanding and patience. It is such a blessing to have somebody who looks in the same direction. My perfect Mom and Dad, this long journey would not have been completed without your unconditional support, love and confidence you have shown in me. I dedicate this thesis to you.

TABLE OF CONTENTS

LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF ABBREVIATIONS	xi
I Random Projection Algorithms for Distributed Optimization	1
CHAPTER 1 INTRODUCTION	2
1.1 Problem Set-up	3
1.2 Previous Work and Our Contribution	4
1.3 Thesis Organization	6
CHAPTER 2 PRELIMINARIES	7
CHAPTER 3 DISTRIBUTED RANDOM PROJECTION AL- GORITHM	9
3.1 DRP Algorithm	9
3.2 Assumptions	11
3.3 Some Basic Relations	13
3.4 Convergence with Probability 1	19
3.5 Distributed Mini-batch Random Projections	25
CHAPTER 4 ASYNCHRONOUS GOSSIP-BASED RANDOM PROJECTION ALGORITHM	32
4.1 GRP Algorithm	32
4.2 Assumptions	34
4.3 Convergence Analysis	35
4.4 Error Bound	53
4.5 Gossip-based Mini-batch Random Projections	61
CHAPTER 5 APPLICATIONS	63
5.1 Distributed Support Vector Machines (DrSVM)	63
5.2 Distributed Robust Control	70

CHAPTER 6	CONCLUSIONS AND FUTURE WORK	76
6.1	Time-varying Mini-batch Random Projection	76
6.2	Random Projections for Convex Feasibility Problems	77
6.3	Distributed Optimization over Time-varying Graphs	78
 II Epoch Gradient Descent for Smoothed Hinge-loss SVMs		80
CHAPTER 7	INTRODUCTION	81
CHAPTER 8	LOSS FUNCTION WITH LIPSCHITZ CONTINUOUS GRADIENT	86
8.1	Smoothed Hinge-loss $\mathcal{L}_\mu(w)$	86
8.2	The Lipschitz Continuity of ∇f_μ	87
CHAPTER 9	EFFICIENT ALGORITHMS	91
9.1	Simple Gradient Descent with Strong-Convexity and Lipschitz Continuity	91
9.2	Epoch Gradient Descent Algorithm	93
CHAPTER 10	EXPERIMENTAL RESULTS	98
CHAPTER 11	CONCLUSIONS AND FUTURE WORK	101
REFERENCES	102

LIST OF TABLES

5.1	The statistics of three text classification data sets: n is the number of examples, d is the number of features, and s is the sparsity of data.	68
5.2	The results of DrSVM with two different graph topologies (clique and 3-regular expander graph) and three different numbers of agents ($m = 2, 6, 10$): t_{acc} is the target test accuracy; b is the number of projections per iteration. The table shows the number of iterations for all agents to reach the target test accuracy, where ‘-’ indicates that the algorithm did not converge within the 20,000 maximum iteration limit.	68
5.3	Number of agents and λ	73
10.1	The statistics of the four data sets	99
10.2	The results of SSVM and SSVM-epoch (\dagger indicates a different stopping criterion is used)	99
10.3	The results of Pegasos-batch and Pegasos-online	100

LIST OF FIGURES

5.1	$f(x)$ vs. iteration on astro-ph with 10 agents when batch size b is 1 (top) and 100 (bottom)	67
5.2	Clique (left), cycle (center) and star (right) graph used for communication topology ($m = 4$)	73
5.3	Iteration vs. $\frac{1}{m} \sum_{i=1}^m \ \mathbf{u}_i(k) - \mathbf{u}^*\ ^2$ with a diminishing step-size when $m = 4$ (top) and $m = 10$ (bottom)	74
5.4	Iteration vs. $\frac{1}{m} \sum_{i=1}^m \ \mathbf{u}_i(k) - \mathbf{u}^*\ ^2$ with a constant stepsize when $m = 4$ (top) and $m = 10$ (bottom)	75
7.1	Frequently used loss functions	83
8.1	Smoothed hinge-loss function with two different smoothing parameters $\mu = 2$ and $\mu = 5$	90

LIST OF ABBREVIATIONS

\mathbb{R}	The real number
\mathbb{R}^d	The d -dimensional Euclidean space
x	A column vector in \mathbb{R}^d
x'	Transpose of the vector x
$\ x\ $	Euclidean norm of x
$\langle x, y \rangle$	Inner product of two vectors x and y
$\mathbf{1}$	A vector whose entries are all 1
\mathcal{X}	A closed convex set contained in \mathbb{R}^d
$\text{dist}(x, \mathcal{X})$	The distance of a vector x from a set \mathcal{X} , i.e., $\min_{v \in \mathcal{X}} \ v - x\ $
$\Pi_{\mathcal{X}}[x]$	The projection of x on a set \mathcal{X} , i.e., $\arg \min_{v \in \mathcal{X}} \ v - x\ ^2$
$\Pr\{Z\}$	The probability of a random variable Z
$\mathbb{E}\{Z\}$	The expectation of a random variable Z
$w.p.1$	With probability 1
$i.i.d.$	Independent and identically distributed
DRP	Distributed Random Projection
GRP	Gossip-based Random Projection
SVM	Support Vector Machine
DrSVM	Distributed Random Projection for SVM

Part I

Random Projection Algorithms for Distributed Optimization

CHAPTER 1

INTRODUCTION

A number of important problems that arise in various application domains, including distributed control [1], large-scale machine learning [2,3], wired and wireless networks [4–7] can be formulated as a distributed convex constrained minimization problem over a multi-agent network. The problem is usually defined as a sum of convex objective functions over an intersection of convex constraint sets. The goal of the agents is to solve the problem in a distributed way, with each agent handling a component of the objective and constraint. This is useful either when the problem data are naturally collected in a distributed way or when the data are too large to be conveniently processed by a single agent.

Common to these distributed optimization problems are the following operational restrictions: 1) a component objective function and constraint set is only known to a specific network agent (the problem is fully distributed), 2) there is no central coordinator that synchronizes actions on the network or works with global information, 3) the agents usually have a limited memory, computational power and energy, and 4) communication overhead is significant due to the expensive start-up cost and network latencies. These restrictions motivate the design of computationally simple, distributed and decentralized algorithms.

The focus of this thesis is the development and analysis of efficient distributed algorithms whereby each agent exchanges local information only with its immediate neighbors at each iteration. We propose gradient descent algorithms with *random projections* which use various communication protocols.

1.1 Problem Set-up

We consider an optimization problem where the objective function and constraint sets are distributed among m agents over a network. Let a time-varying graph $G(k) = (V, E(k))$ represent the topology of the network at iteration k , with the vertex set $V = \{1, \dots, m\}$ and the edge set $E(k) \subseteq V \times V$. Let $\mathcal{N}_i(k)$ be the set of the neighbors of agent i at iteration k , i.e., $\mathcal{N}_i(k) = \{j \in V \mid \{i, j\} \in E(k)\}$. (In the thesis, we also consider time-invariant network $G = (V, E)$. In this case, we use $\mathcal{N}(i)$ to represent the set of neighbors.) The goal of the agents is to cooperatively solve the following optimization problem:

$$\min f(x) \triangleq \sum_{i=1}^m f_i(x) \quad \text{s.t. } x \in \mathcal{X} \triangleq \bigcap_{i=1}^m \mathcal{X}_i, \quad (1.1)$$

where $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is a convex function, representing the local objective of agent i , and $\mathcal{X}_i \subseteq \mathbb{R}^d$ is a closed convex set, representing the local constraint set of agent i . The function f_i and the set \mathcal{X}_i are known to agent i only. Collectively, the agents are responsible for solving problem (1.1).

We assume that problem (1.1) is feasible. Moreover, we assume each set \mathcal{X}_i is defined as the intersection of a collection of simple convex sets. That is, \mathcal{X}_i can be represented as $\mathcal{X}_i = \bigcap_{j \in I_i} \mathcal{X}_i^j$, where the superscript j is used to identify a component set and I_i is a (possibly infinite) set of indices. Each component set \mathcal{X}_i^j is assumed to be a “simple set” for the projection operation. Examples of such a simple set include a halfspace $\mathcal{X}_i^j = \{x \in \mathbb{R}^d \mid \langle a, x \rangle \leq b\}$, a box $\mathcal{X}_i^j = \{x \in \mathbb{R}^d \mid \alpha \leq x \leq \beta\}$ (the inequality is component-wise) and a ball $\mathcal{X}_i^j = \{x \in \mathbb{R}^d \mid \|x - v\| \leq r\}$, where $a, \alpha, \beta, v \in \mathbb{R}^d$ and $b, r \in \mathbb{R}$. In such cases, the projection on the whole set \mathcal{X}_i may be complex, especially when the number of components is large, while the projection on each component \mathcal{X}_i^j has a closed form expression.

In the proposed algorithms, each agent i maintains its own estimate sequence $\{x_i(k)\}$ of the decision variable x . At each iteration, each agent calculates weighted average of the received iterates (from its neighbors) and its own iterate, adjusts the iterate by using gradient information of its local objective function f_i and projects onto a constraint component that is selected randomly from its local constraint set \mathcal{X}_i . The projections are performed locally by each agent based on the random observations of the local constraint

components. In particular, agent i observes a constraint component $\mathcal{X}_i^{\Omega_i(k)}$ at time k , where $\Omega_i(k) \in I_i$ is a random variable.

Our primary interest is in the case when the whole constraint set \mathcal{X}_i for an agent i is not known in advance, but its component is revealed through random realizations $\mathcal{X}_i^{\Omega_i(k)}$. For example, consider the case when \mathcal{X}_i is given by

$$\mathcal{X}_i = \{x \in \mathbb{R}^d \mid \langle a + \xi, x \rangle \leq b\},$$

where $a \in \mathbb{R}^d$, $b \in \mathbb{R}$ are deterministic and $\xi \in \mathbb{R}^d$ is a Gaussian random noise. In such a case, a projection-based distributed algorithm cannot be directly applied to solve problem (1.1) since $|I_i|$ is infinite and the projection of a point on the uncertain set \mathcal{X}_i is impossible. However, a component \mathcal{X}_i^j can be realized from a random selection of ξ and the projection onto the realized component is always possible.

Another case of interest is when the whole constraint set \mathcal{X}_i is known in advance but it has a huge number of components. For example, in text classification problems, model parameters are trained based on a hundred thousand or more text samples and each sample constitutes a constraint component (usually a halfspace) [8]. In such a case, the projection operation on the whole constraint set \mathcal{X}_i is computationally prohibitive but the projection on a single component \mathcal{X}_i^j is simple.

1.2 Previous Work and Our Contribution

In the optimization literature, algorithms of two categories have been proposed for problem (1.1): the Markov incremental algorithm and the distributed subgradient algorithm. In the Markov incremental algorithm studied in [9, 10], the agents maintain a single estimate sequence that is sequentially updated by one agent at a time. When an agent receives the estimate, it updates the estimate using its local objective function and passes it to a randomly selected neighbor. The update order is driven by a time inhomogeneous Markov chain (as the network topology is time varying). In the distributed subgradient algorithms, each agent maintains its own estimate. It communicates the estimate with its neighbors and updates it using the local objective and constraint information. Algorithms of this type require

a consensus over all agents for convergence. However, in some distributed problems it is important that each agent maintains a good estimate at all times. For example, in a distributed online learning, each node is expected to perform in real time. Our proposed algorithms are in the distributed subgradient algorithm category.

The related distributed optimization literature includes [11–14], which are concerned with convex but unconstrained problems, and [15–17] where constrained problems are considered. The most relevant to the work in this thesis are [18–21] where the constraint set is also distributed across agents and each agent handles its own constraint set only. In [18], the convergence analysis is done for a special case when the network is completely connected. The work in [19, 20] extends the algorithm and its analysis to a more general network including the presence of noisy links, while [21] extends it to a general Markovian network model. Unlike [18] and [19], where each agent can perform projections on its entire constraint set, this thesis addresses the case when such projections are not possible or computationally prohibitive.

To the best of our knowledge, there is no previous work on asynchronous distributed optimization algorithms that utilize random projections. Finding probabilistic feasible solutions through random sampling of constraints for optimization problems with uncertain constraints have been proposed in [22, 23]. Also, the related work is the (centralized) random projection method proposed by Polyak [24] for a class of convex feasibility problems and the random projection algorithm [25] for convex set intersection problems. On a much broader scale, the work in this thesis is related to the literature on the consensus problem, where each agent starts from an initial value and ends by converging to a value common to all agents (see for example [7, 26–29]).

The contribution of this thesis is mainly in two directions. First, we propose novel distributed optimization algorithms that are based on gradient descent with random projections and local communications of agents in a network. We also propose a variant of the algorithms using a mini-batch of random projections. Second, we establish the convergence theory for these algorithms.

1.3 Thesis Organization

The rest of part I is organized as follows.

In Chapter 2, we state some results from the literature that we use in the convergence analysis of our proposed algorithms.

In Chapter 3, we introduce our distributed random projection (DRP) algorithm. We also state assumptions on the distributed problem and the network. For establishing convergence of the DRP algorithm, we first derive two important lemmas and provide the main convergence results using these lemmas. Also, we provide an extension of the algorithm that uses a mini-batch of random projections and state a convergence result for this extension.

In Chapter 4, we introduce our asynchronous gossip-based random projection (GRP) algorithm. For the convergence analysis, we provide a roadmap of the proofs and state lemmas regarding random projection errors and agent disagreements. Then, we prove the main convergence results using these lemmas.

As practical applications of our algorithms, in Chapter 5, we introduce distributed formulations of a linear support vector machine and a model predictive control. We also discuss how to apply the algorithms and present some experimental results on binary text classification tasks and robust control applications.

Chapter 6 contains concluding remarks and future directions.

CHAPTER 2

PRELIMINARIES

In this section, we state some definitions and results from the literature, which will be used in later sections.

Convexity of Euclidean norm and its square. Both the Euclidean norm and its square are convex functions; i.e., for any vectors $v_1, \dots, v_m \in \mathbb{R}^d$ and nonnegative scalars β_1, \dots, β_m such that $\sum_{i=1}^m \beta_i = 1$, we have

$$\left\| \sum_{i=1}^m \beta_i v_i \right\| \leq \sum_{i=1}^m \beta_i \|v_i\|, \quad \left\| \sum_{i=1}^m \beta_i v_i \right\|^2 \leq \sum_{i=1}^m \beta_i \|v_i\|^2. \quad (2.1)$$

Non-expansive projection property. We state a projection theorem (see [30] for its proof).

Lemma 2.1 *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty closed convex set. The function $\Pi_{\mathcal{X}} : \mathbb{R}^d \rightarrow \mathcal{X}$ is continuous and nonexpansive, i.e.,*

- (a) $\|\Pi_{\mathcal{X}}[x] - \Pi_{\mathcal{X}}[y]\| \leq \|x - y\|$ for all $x, y \in \mathbb{R}^d$.
- (b) $\|\Pi_{\mathcal{X}}[x] - y\|^2 \leq \|x - y\|^2 - \|\Pi_{\mathcal{X}}[x] - x\|^2$ for all $x \in \mathbb{R}^d$ and for all $y \in \mathcal{X}$.

Matrix convergence. Recall we defined $W(k)$ to be the matrix with (i, j) th entry equal to $w_{ij}(k)$. From Assumption 3.4, the matrix $W(k)$ is doubly stochastic. Define for all k, s with $k > s \geq 0$,

$$\Phi(k, s) = W(k)W(k-1) \cdots W(s+1)W(s), \quad (2.2)$$

with $\Phi(k, k) = W(k)$ for all $k \geq 0$. We state the convergence property of the matrix $\Phi(k, s)$ (see [13] for its proof). Let $[\Phi(k, s)]_{ij}$ denote the (i, j) th entry of the matrix $\Phi(k, s)$, and $e \in \mathbb{R}^m$ be the column vector whose all entries are equal to 1.

Lemma 2.2 *Let Assumptions 3.3 and 3.4 hold. Then,*

(a) $\lim_{k \rightarrow \infty} \Phi(k, s) = \frac{1}{m} e e^T$ for all $s \geq 0$.

(b) $|\Phi(k, s)_{ij} - \frac{1}{m}| \leq \theta \beta^{k-s}$ for all $k \geq s \geq 0$, where $\theta = (1 - \frac{\eta}{4m^2})^{-2}$ and $\beta = (1 - \frac{\eta}{4m^2})^{\frac{1}{Q}}$.

Convergence result. In our analysis of the DRP algorithm, we also make use of the following convergence result due to Robbins and Siegmund (see [31, Lemma 10-11, p. 49-50]).

Theorem 2.1 *Let $\{v_k\}$, $\{u_k\}$, $\{a_k\}$ and $\{b_k\}$ be sequences of non-negative random variables such that*

$$\mathbb{E}[v_{k+1} | \mathcal{F}_k] \leq (1 + a_k)v_k - u_k + b_k \quad \text{for all } k \geq 0 \quad \text{w.p.1,}$$

where \mathcal{F}_k denotes the collection $v_0, \dots, v_k, u_0, \dots, u_k, a_0, \dots, a_k$ and b_0, \dots, b_k . Also, let $\sum_{k=0}^{\infty} a_k < \infty$ and $\sum_{k=0}^{\infty} b_k < \infty$ w.p.1. Then, we have $\lim_{k \rightarrow \infty} v_k = v$ for a random variable $v \geq 0$ w.p.1, and $\sum_{k=0}^{\infty} u_k < \infty$ w.p.1.

The above theorem is the key in our convergence analysis. Specifically, once we show that Theorem 2.1 applies to $v_{k+1} = \sum_{i=1}^m \|x_i(k+1) - x^*\|^2$ for an optimal solution x^* , the rest of the proof just builds on the implications of the theorem.

Scalar Sequences. We also use the convergence result for scalar sequences (see Lemma 3.1 in [15] for its proof). For a scalar β and a scalar sequence $\{\gamma(k)\}$, we consider the convolution sequence $\sum_{\ell=0}^k \beta^{k-\ell} \gamma(\ell)$.

Lemma 2.3 *If $\lim_{k \rightarrow \infty} \gamma(k) = \gamma$ and $0 < \beta < 1$, then*

$$\lim_{k \rightarrow \infty} \sum_{\ell=0}^k \beta^{k-\ell} \gamma(\ell) = \frac{\gamma}{1 - \beta}.$$

We also use the following notation regarding the optimal value and optimal solutions of problem (1.1):

$$f^* = \min_{x \in \mathcal{X}} f(x), \quad \mathcal{X}^* = \{x \in \mathcal{X} \mid f(x) = f^*\}.$$

CHAPTER 3

DISTRIBUTED RANDOM PROJECTION ALGORITHM

3.1 DRP Algorithm

To solve the problem (1.1) with distributed information access, we propose an iterative gradient method with random projections. Let $x_i(k) \in \mathbb{R}^d$ denote the estimate of agent i at time k . At time k , each agent sends the estimate to its neighbors (represented by the graph $(V, E(k))$). Upon receiving the estimates $x_j(k)$ from its neighbors $j \in \mathcal{N}_i(k)$, each agent i updates according to the following two steps:

$$v_i(k) = \sum_{j \in \mathcal{N}_i(k)} w_{ij}(k) x_j(k) \quad (3.1a)$$

$$x_i(k+1) = \Pi_{\mathcal{X}_i^{\Omega_i(k)}} [v_i(k) - \alpha_k \nabla f_i(v_i(k))], \quad (3.1b)$$

where $\alpha_k > 0$ is a stepsize at time k and $x_i(0) \in \mathbb{R}^d$ is an initial estimate of agent i (which can be random).

In the above, (3.1a) is an information mixing step and (3.1b) is a local minimization and feasibility update step using a random projection. In (3.1a), the iterate $v_i(k)$ is a weighted average of agent i 's estimate and the estimates received from its neighbors $j \in \mathcal{N}_i(k)$. Specifically, $w_{ij}(k) \geq 0$ is a weight that agent i places on the estimate $x_j(k)$ received from a neighbor $j \in \mathcal{N}_i(k)$ at time k , where the total weight sum is 1, i.e., $\sum_{j \in \mathcal{N}_i(k)} w_{ij}(k) = 1$ for each agent i . The step (3.1a) can be equivalently represented as

$$v_i(k) = \sum_{j=1}^m [W(k)]_{ij} x_j(k) \quad (3.2)$$

by letting $w_{ij}(k) = 0$ for whenever $j \notin \mathcal{N}_i(k)$, and using $[W]_{ij}$ to denote the (i, j) th entry of a matrix W .

In (3.1b), agent i adjusts the average $v_i(k)$ along the negative gradient direction of its local objective f_i . At time k , agent i also observes a random realization of its local constraint component set $\mathcal{X}_i^{\Omega_i(k)}$. To reduce the feasibility violation, it projects its current estimate on this set. The random variable $\Omega_i(k)$ takes values in the index set I_i at all times k . In this way, instead of projecting onto the whole local constraint set \mathcal{X}_i , agent i projects only on a component set $\mathcal{X}_i^{\Omega_i(k)}$ which is randomly selected at time k . Note that the updated estimate $x_i(k+1)$ may not lie in \mathcal{X}_i since $\mathcal{X}_i \subset \mathcal{X}_i^{\Omega_i(k)}$.

Through the updates (3.1a) and (3.1b), agents combine their information and consider their own optimization problem of minimizing f_i over the set \mathcal{X}_i . There is neither a central node governing the whole process nor additional constraints enforcing consistency. Nevertheless, with this simple update rule, our algorithm finds the optimal solution and all agents eventually arrive at a common optimal solution (all $x_i(k)$ converge to some $x^* \in \mathcal{X}^*$, as shown in Section 4.3).

Note that algorithm (3.1a)-(3.1b) is similar to the distributed projected subgradient algorithm in [18] except for the randomization over the components of the set \mathcal{X}_i in (3.1b). At each iteration of the algorithm in [18], a projection is performed on the entire constraint set \mathcal{X}_i , which can be prohibitively expensive when \mathcal{X}_i is itself an intersection of many sets. In addition, unlike the method in [18], DRP can also handle the cases when the projection on the entire set \mathcal{X}_i is not possible since the set \mathcal{X}_i is not known in advance.

The challenges in convergence analysis of the DRP algorithm are posed mainly by its distributed nature, through the *effects of the time-varying network*, and by the *projection errors* associated with using projections on components \mathcal{X}_i^j , $j \in I_i$ of the set $\mathcal{X}_i = \bigcap_{j \in I_i} \mathcal{X}_i^j$ instead of the projection on the set \mathcal{X}_i . The fact that the DRP relies on a random component \mathcal{X}_i^j poses particular difficulties, as one needs to characterize the impact of the random projection errors, which is closely related to errors in “set-approximations.” To handle these difficulties, we make several mild assumptions. We make an assumption on the random set processes $\{\Omega_i(k)\}$, $i \in V$, that allows us to characterize the projection errors. For the network we assume that it is sufficiently connected in order to properly conduct the information among the agents. Finally, we assume that the agent weights are also properly chosen to ensure that each agent is equally influencing every other agent. These network assumptions have been typically used in distributed optimization al-

gorithms over a time-varying network (see e.g. [12, 13, 15, 32–34]). In the next subsections, we state our assumptions on the random set processes $\{\Omega_i(k)\}$, $i \in V$, the network and the weight matrices $W(k)$.

3.2 Assumptions

We use the following assumption for the functions f_i and the sets \mathcal{X}_i^j .

Assumption 3.1 *Let the following conditions hold:*

- (a) *The sets \mathcal{X}_i^j , $j \in I_i$ are closed and convex for every $i \in V$.*
- (b) *Each function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex.*
- (c) *The functions f_i , $i \in V$, are differentiable and have Lipschitz gradients with a constant L over \mathbb{R}^d ,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^d.$$

- (d) *The gradients $\nabla f_i(x)$, $i \in V$ are bounded over the set \mathcal{X} , i.e., there exists a constant G_f such that*

$$\|\nabla f_i(x)\| \leq G_f \quad \text{for all } x \in \mathcal{X} \text{ and all } i \in V.$$

When each f_i has Lipschitz gradients with a constant L_i , Assumption 3.1(c) is satisfied with $L = \max_{i \in V} L_i$. Further note that Assumption 3.1(d) is satisfied, for example, when \mathcal{X} is compact.

The next assumption is crucial in our analysis.

Assumption 3.2 *For all $i \in V$, there exists a constant $c > 0$ such that for all $x \in \mathbb{R}^d$,*

$$\text{dist}^2(x, \mathcal{X}) \leq cE \left[\text{dist}^2(x, \mathcal{X}_i^{\Omega_i(k)}) \right]. \quad (3.3)$$

Assumption 3.2 is satisfied, for example, when each set \mathcal{X}_i^j is given by either linear inequality or a linear equality, or when the intersection set \mathcal{X} has a nonempty interior. In the first case, one can verify that the assumption holds by using the results of Burke and Ferris on a set of weak sharp minima [35]. In the second case, one can use the ideas of the convergence rate analysis for

the alternating projection algorithm of Gubin, Polyak and Raik in [36]. In either case, the constant c depends on the probability distributions of $\Omega_i(k)$ and some geometric properties of the sets.

As a direct consequence of Assumption 3.2, we have the following observation. Since $\mathcal{X} \subset \mathcal{X}_i^{\Omega_i(k)}$, we have

$$\text{dist}(x, \mathcal{X}_i^{\Omega_i(k)}) \leq \text{dist}(x, \mathcal{X}) \quad \text{for all } x \in \mathbb{R}^d.$$

Thus, by Assumption 3.2 it follows that

$$\mathbb{E} \left[\text{dist}^2(x, \mathcal{X}_i^{\Omega_i(k)}) \right] \leq \text{dist}^2(x, \mathcal{X}).$$

In view of this relation and Assumption 3.2, we find that $c \geq 1$ holds always.

We rely on the graphs $(V, E(k))$, $k \geq 0$ to represent the time-varying network. We make two assumptions.

Assumption 3.3 *[Network Connectivity] There exists a scalar Q such that the graph $\left(V, \bigcup_{\ell=0, \dots, Q-1} E(k + \ell)\right)$ is strongly connected for all $k \geq 0$.*

Assumption 3.3 ensures that the agents communicate sufficiently often so that all functions and all constraints (f_i 's and \mathcal{X}_i 's) influence the iterates of all agents.

Next, we make the following assumption on the edge weights (defined below (3.2)).

Assumption 3.4 *[Doubly Stochasticity] For all $k \geq 1$,*

- (a) $[W(k)]_{ij} \geq 0$ and $[W(k)]_{ij} = 0$ when $j \notin \mathcal{N}_i(k)$,
- (b) $\sum_{j=1}^m [W(k)]_{ij} = 1$ for all $i \in V$,
- (c) There exists a scalar $\eta \in (0, 1)$ such that $[W(k)]_{ij} \geq \eta$ when $j \in \mathcal{N}_i(k)$,
- (d) $\sum_{i=1}^m [W(k)]_{ij} = 1$ for all $j \in V$.

Assumption 3.4(a) states that the weights respect the network topology at any time k . Assumption 3.4(b) means that each agent calculates a weighted average of the estimates obtained from its neighbors. Assumption 3.4(c) ensures that each agent gives sufficient weights on the information received. Assumption 3.4(d) together with Assumption 3.3 ensure that each agent is equally influential in the long run so that the agents arrive at a consensus on an optimal solution.

3.3 Some Basic Relations

Our convergence analysis is based on a critical relation that captures the decrease in values $\sum_{i=1}^m \|x_i(k+1) - x^*\|^2$ as the algorithm progresses. Such a relation is provided in Lemma 3.1, which is taken from [37] where it was developed for a centralized algorithm. This basic relation is further refined to take into account the distributed nature of the algorithm. Specifically, in Lemma 3.2, we show that the weighted averages $v_i(k)$ of the iterates approach the constraint set \mathcal{X} asymptotically. Then, in Lemma 3.4, we prove that the agents' disagreement on $v_i(k)$ is diminishing with the number k of iterations. The proof of Lemma 3.4 relies on an auxiliary result taken from [15], which is provided in Lemma 3.3.

In the analysis, we will rely on the expectation taken with respect to the past history of the algorithm, which we define as follows. Let \mathcal{F}_k be the σ -algebra generated by the entire history of the algorithm up to time $k-1$ inclusively (realizations of all the random variables but not the realizations of the indices Ω_i at time k); i.e., for all $k \geq 1$,

$$\mathcal{F}_k = \{x_i(0), i \in V\} \cup \{\Omega_i(\ell); 0 \leq \ell \leq k-1, i \in V\},$$

where $\mathcal{F}_0 = \{x_i(0), i \in V\}$. Therefore, given \mathcal{F}_k , the collection $x_i(0), \dots, x_i(k)$ and $v_i(0), \dots, v_i(k)$ generated by the algorithm (3.1a)-(3.1b) is fully determined.

3.3.1 Basic Iterate Relation

The following lemma is from the paper [37, Lemma 1], which provides relation among the iterate obtained after one step of the algorithm (3.1a), a point in the feasible set \mathcal{X} and an arbitrary point in \mathbb{R}^d .

Lemma 3.1 *Let \mathcal{Y} be a closed convex set such that $\mathcal{Y} \subseteq \mathbb{R}^d$. Let the function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and differentiable over \mathbb{R}^d with Lipschitz continuous gradients with a constant L . Let y be given by*

$$y = \Pi_{\mathcal{Y}}[x - \alpha \nabla \phi(x)] \quad \text{for some } x \in \mathbb{R}^d \text{ and } \alpha > 0.$$

Then, we have for any $\tilde{x} \in \mathcal{Y}$ and $z \in \mathbb{R}^d$,

$$\begin{aligned} \|y - \tilde{x}\|^2 &\leq (1 + A_\tau \alpha^2) \|x - \tilde{x}\|^2 - 2\alpha(\phi(z) - \phi(\tilde{x})) \\ &\quad - \frac{3}{4} \|y - x\|^2 + \left(\frac{3}{8\tau} + 2\alpha L \right) \|x - z\|^2 + B_\tau \alpha^2 \|\nabla \phi(\tilde{x})\|^2, \end{aligned} \quad (3.4)$$

where $A_\tau = 8L^2 + 16\tau L^2$, $B_\tau = 8\tau + 8$ and $\tau > 0$ is arbitrary.

Lemma 3.1 provides a measure of progress toward an optimal point of the function ϕ when moving from a point x in the direction opposite of the gradient $\nabla \phi(x)$. Specifically, if x^* is a minimizer of $\phi(x)$ over \mathcal{Y} , the lemma (with $\tilde{x} = x^*$) will provide us with a relation between the distances $\|y - x^*\|$ and $\|x - x^*\|$, where the point y is resulting from a projected-gradient step away from the point x . The lemma provides a relation that helps us measure the progress of a gradient-based algorithm for minimizing ϕ . Lemma 3.1, with a specific identification of the terms, will be a starting point for our convergence proof.

3.3.2 Projection Error Estimate

In the next lemma, we show that the sequences $\{v_i(k)\}$, $i \in V$, approach the constraint set \mathcal{X} . The result does not say that these sequences necessarily have accumulation points in \mathcal{X} , but rather that the distance between $v_i(k)$ and the set \mathcal{X} tends to 0, as $k \rightarrow \infty$, for all i . Furthermore, these distances converge to 0 rather fast, as the sum of all squared distances over time is finite, which is a critical relation in our analysis.

Lemma 3.2 *Let Assumption 1 hold. Let each $W(k)$ be doubly stochastic, and let $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. Then,*

$$\sum_{k=0}^{\infty} \text{dist}^2(v_i(k), \mathcal{X}) < \infty \quad \text{for all } i \in V \text{ w.p.1.}$$

Proof In Lemma 3.1, let $y = x_i(k+1)$, $x = v_i(k)$, $\mathcal{Y} = \mathcal{X}_i^{\Omega_i(k)}$, $\alpha = \alpha_k$, $\phi = f_i$ and $\tau = c$ where c is the constant from Assumption 3. Then, for any

$\tilde{x} \in \mathcal{X}$ (also in $\mathcal{X}_i^{\Omega_i(k)}$, since $\mathcal{X} \subseteq \mathcal{X}_i^{\Omega_i(k)}$) and any $z \in \mathbb{R}^d$, we obtain

$$\begin{aligned} \|x_i(k+1) - \tilde{x}\|^2 &\leq (1 + A\alpha_k^2)\|v_i(k) - \tilde{x}\|^2 - 2\alpha_k(f_i(z) - f_i(\tilde{x})) \\ &\quad - \frac{3}{4}\|x_i(k+1) - v_i(k)\|^2 + \left(\frac{3}{8c} + 2\alpha_k L\right)\|v_i(k) - z\|^2 + B\alpha_k^2 G_f^2, \end{aligned}$$

where $A = 8L^2 + 16cL^2$ and $B = 8c + 8$. Here, we have also used Assumption 3.1(d), according to which the gradients $\nabla f_i(x)$ are bounded on the set \mathcal{X} , i.e., $\|\nabla f_i(\Pi_{\mathcal{X}}[v_i(k)])\| \leq G_f$ for all k and i .

Letting $\tilde{x} = z = \Pi_{\mathcal{X}}[v_i(k)]$ in the preceding relation, we find

$$\begin{aligned} \|x_i(k+1) - \Pi_{\mathcal{X}}[v_i(k)]\|^2 &\leq (1 + A\alpha_k^2)\text{dist}^2(v_i(k), \mathcal{X}) - \frac{3}{4}\|x_i(k+1) - v_i(k)\|^2 \\ &\quad + \left(\frac{3}{8c} + 2\alpha_k L\right)\text{dist}^2(v_i(k), \mathcal{X}) + B\alpha_k^2 G_f^2. \end{aligned} \quad (3.5)$$

By the definition of the projection, we have

$$\text{dist}(x_i(k+1), \mathcal{X}) = \|x_i(k+1) - \Pi_{\mathcal{X}}[x_i(k+1)]\| \leq \|x_i(k+1) - \Pi_{\mathcal{X}}[v_i(k)]\|,$$

$$\|x_i(k+1) - v_i(k)\| \geq \left\| \Pi_{\mathcal{X}_i^{\Omega_i(k)}}[v_i(k)] - v_i(k) \right\| = \text{dist}(v_i(k), \mathcal{X}_i^{\Omega_i(k)}).$$

Upon substituting these estimates in (3.5), we obtain

$$\begin{aligned} \text{dist}^2(x_i(k+1), \mathcal{X}) &\leq (1 + A\alpha_k^2)\text{dist}^2(v_i(k), \mathcal{X}) - \frac{3}{4}\text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \\ &\quad + \left(\frac{3}{8c} + 2\alpha_k L\right)\text{dist}^2(v_i(k), \mathcal{X}) + B\alpha_k^2 G_f^2. \end{aligned} \quad (3.6)$$

Taking the expectation in (3.6) conditioned on \mathcal{F}_k , and using

$$\mathbb{E} \left[\text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \mid \mathcal{F}_k \right] \geq \frac{1}{c}\text{dist}^2(v_i(k), \mathcal{X}),$$

which follows by Assumption 3, we find that with probability 1

$$\begin{aligned} \mathbb{E} \left[\text{dist}^2(x_i(k+1), \mathcal{X}) \mid \mathcal{F}_k \right] &\leq (1 + A\alpha_k^2)\text{dist}^2(v_i(k), \mathcal{X}) \\ &\quad - \left(\frac{3}{8c} - 2\alpha_k L \right) \text{dist}^2(v_i(k), \mathcal{X}) + B\alpha_k^2 G_f^2. \end{aligned} \quad (3.7)$$

By using the definition of $v_i(k)$ (as a convex combination of $x_j(k)$ in (3.2)) and the convexity of the distance function $x \mapsto \text{dist}^2(x, \mathcal{X})$ (see [30, p. 88]),

we find that

$$\text{dist}^2(v_i(k), \mathcal{X}) \leq \sum_{j=1}^m [W(k)]_{ij} \text{dist}^2(x_j(k), \mathcal{X}).$$

The preceding relation and (3.7) imply that with probability 1 for all $k \geq 0$,

$$\begin{aligned} \mathbb{E} \left[\text{dist}^2(x_i(k+1), \mathcal{X}) \mid \mathcal{F}_k \right] &\leq (1 + A\alpha_k^2) \sum_{j=1}^m [W(k)]_{ij} \text{dist}^2(x_j(k), \mathcal{X}) \\ &\quad - \left(\frac{3}{8c} - 2\alpha_k L \right) \text{dist}^2(v_i(k), \mathcal{X}) + B\alpha_k^2 G_f^2. \end{aligned}$$

Finally, by summing over all i and using the fact that each $W(k)$ has column sums equal to 1, we arrive at the following relation: with probability 1 for all $k \geq 0$,

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^m \text{dist}^2(x_i(k+1), \mathcal{X}) \mid \mathcal{F}_k \right] &\leq (1 + A\alpha_k^2) \sum_{j=1}^m \text{dist}^2(x_j(k), \mathcal{X}) \\ &\quad - \left(\frac{3}{8c} - 2\alpha_k L \right) \sum_{i=1}^m \text{dist}^2(v_i(k), \mathcal{X}) + mB\alpha_k^2 G_f^2. \end{aligned}$$

Since $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, it follows that $\alpha_k \rightarrow 0$, implying that there exists \bar{k} such that $\frac{3}{8c} - 2\alpha_k L > 0$ for all $k \geq \bar{k}$. Therefore, for all $k \geq \bar{k}$, all the conditions of the convergence theorem are satisfied (Theorem 2.1). By applying the convergence theorem (to a time-delayed process from \bar{k} onward) we conclude that

$$\sum_{k=0}^{\infty} \text{dist}^2(v_i(k), \mathcal{X}) < \infty \quad \text{for all } i \in V \quad \text{w.p.1.} \quad \blacksquare$$

Lemma 3.2 shows that the points $v_i(k)$ are getting close to the set \mathcal{X} relatively fast, as $k \rightarrow \infty$. If the set \mathcal{X} was compact, this would imply that all accumulation points of $\{v_i(k)\}$ would lie in the set \mathcal{X} . However, there would be no guarantee that the accumulation points of any two sequences $\{v_i(k)\}$ and $\{v_j(k)\}$ would be the same. Furthermore, Lemma 3.2 would give no information about optimality of any of the accumulation points. In the next section, we provide a result that helps us claim later on that any two sequences $\{v_i(k)\}$ and $\{v_j(k)\}$ have the same accumulation points.

3.3.3 Disagreement Estimate

We now quantify the agent disagreements in time. We measure the disagreements by using the norm $\|v_i(k) - \bar{v}(k)\|$ of the differences between the estimates $v_i(k)$ generated by different agents according the algorithm (3.1a)-(3.1b) and their instantaneous average $\bar{v}(k) = \frac{1}{m} \sum_{\ell=1}^m v_\ell(k)$. The proof of our result relies on a lemma (adopted from [16, Theorem 4.2]), which states that the iterates generated by a “perturbed” consensus protocol are guaranteed to arrive at a consensus when the perturbations are small in some sense. This lemma is provided next.

Lemma 3.3 *Let Assumptions 3.3 and 3.4 hold. Consider the iterates generated by*

$$\theta_i(k+1) = \sum_{j=1}^m [W(k)]_{ij} \theta_j(k) + \epsilon_i(k) \text{ for all } i \in V. \quad (3.8)$$

Suppose there exists a non-negative non-increasing scalar sequence $\{\alpha_k\}$ such that $\sum_{k=0}^{\infty} \alpha_k \|\epsilon_i(k)\| < \infty$ for all $i \in V$. Then, for all $i, j \in V$,

$$\sum_{k=0}^{\infty} \alpha_k \|\theta_i(k) - \theta_j(k)\| < \infty.$$

Using Lemma 3.3, we prove the following disagreement results that will be important in our analysis later.

Lemma 3.4 *Let Assumptions 3.1, 3.3 and 3.4 hold. Also, assume that the stepsize sequence $\{\alpha_k\}$ is non-increasing and such that $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. Define*

$$e_i(k) = x_i(k+1) - v_i(k) \quad \text{for all } i \in V \text{ and } k \geq 0.$$

Then, we have with probability 1

$$\sum_{k=0}^{\infty} \|e_i(k)\|^2 < \infty \quad \text{for all } i \in V, \quad (3.9)$$

$$\sum_{k=0}^{\infty} \alpha_k \|v_i(k) - \bar{v}(k)\| < \infty \quad \text{for all } i \in V, \quad (3.10)$$

where $\bar{v}(k) = \frac{1}{m} \sum_{\ell=1}^m v_\ell(k)$.

Proof Define $z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$. Consider $\|e_i(k)\|$, for which we can write

$$\begin{aligned}\|e_i(k)\| &\leq \|x_i(k+1) - z_i(k)\| + \|z_i(k) - v_i(k)\| \\ &= \left\| \Pi_{\mathcal{X}_i^{\Omega_i(k)}}[v_i(k) - \alpha_k \nabla f_i(v_i(k))] - z_i(k) \right\| + \|z_i(k) - v_i(k)\|.\end{aligned}$$

Since $\mathcal{X} \subseteq \mathcal{X}_i^{\Omega_i(k)}$ and $z_i(k) \in \mathcal{X}$, we have $z_i(k) \in \mathcal{X}_i^{\Omega_i(k)}$. Using the projection theorem (Lemma 2.1), we obtain

$$\begin{aligned}\|e_i(k)\| &\leq \|v_i(k) - \alpha_k \nabla f_i(v_i(k)) - z_i(k)\| + \|z_i(k) - v_i(k)\| \\ &\leq 2\|v_i(k) - z_i(k)\| + \alpha_k \|\nabla f_i(v_i(k))\| \\ &\leq 2\|v_i(k) - z_i(k)\| + \alpha_k \|\nabla f_i(z_i(k))\| + \alpha_k \|\nabla f_i(v_i(k)) - \nabla f_i(z_i(k))\| \\ &\leq (2 + \alpha_0 L)\|v_i(k) - z_i(k)\| + \alpha_k G_f,\end{aligned}\tag{3.11}$$

where the last inequality follows by using $\alpha_k \leq \alpha_0$, the Lipschitz gradient property of f_i and the gradient boundedness property (Assumptions 3.1(c) and 3.1(d)). Therefore, applying $(a+b)^2 \leq 2a^2 + 2b^2$ in inequality (3.11), we have for all $i \in V$ and $k \geq 0$,

$$\|e_i(k)\|^2 \leq 2(2 + \alpha_0 L)^2 \|v_i(k) - z_i(k)\|^2 + 2\alpha_k^2 G_f^2.\tag{3.12}$$

Recall that we defined $z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$, so we have $\|v_i(k) - z_i(k)\| = \text{dist}(v_i(k), \mathcal{X})$. In the light of Lemma 3.2, we also have $\sum_{k=0}^{\infty} \|v_i(k) - z_i(k)\|^2 < \infty$ with probability 1. Since $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, we conclude that

$$\sum_{k=0}^{\infty} \|e_i(k)\|^2 < \infty \quad \text{for all } i \in V \text{ w.p.1.}$$

By applying the inequality $2ab \leq a^2 + b^2$ to each term $\alpha_k \|e_i(k)\|$, we see that for all $i \in V$

$$\sum_{k=0}^{\infty} \alpha_k \|e_i(k)\| \leq \frac{1}{2} \sum_{k=0}^{\infty} \alpha_k^2 + \frac{1}{2} \sum_{k=0}^{\infty} \|e_i(k)\|^2 < \infty.$$

Now, we note that $x_i(k+1) = v_i(k) + e_i(k)$ with $v_i(k) = \sum_{j=1}^m [W(k)]_{ij} x_j(k)$ and the error $e_i(k)$ satisfying $\sum_{k=0}^{\infty} \alpha_k \|e_i(k)\| < \infty$ with probability 1. There-

fore, by Lemma 3.3, it follows that

$$\sum_{k=0}^{\infty} \alpha_k \|x_i(k) - x_j(k)\| < \infty \text{ for all } i \text{ and } j \text{ w.p.1.} \quad (3.13)$$

Next, we consider $\|v_i(k) - \bar{v}(k)\|$. Recalling that $v_i(k) = \sum_{j=1}^m [W(k)]_{ij} x_j(k)$ (see (3.2)) and $W(k)$ is stochastic (Assumption 3.4), and by using the convexity of the norm, we obtain

$$\|v_i(k) - \bar{v}(k)\| \leq \sum_{j=1}^m w_{ij}(k) \|x_j(k) - \bar{v}(k)\| \leq \sum_{j=1}^m \left\| x_j(k) - \frac{1}{m} \sum_{\ell=1}^m x_{\ell}(k) \right\|,$$

where in the last equality we use $0 \leq [W(k)]_{ij} \leq 1$ and $\bar{v}(k) = \frac{1}{m} \sum_{\ell=1}^m x_{\ell}(k)$, which holds since $v_i(k) = \sum_{j=1}^m [W(k)]_{ij} x_j(k)$ and each $W(k)$ is doubly stochastic. Therefore, by using the convexity of the norm again, we see

$$\sum_{j=1}^m \left\| x_j(k) - \frac{1}{m} \sum_{\ell=1}^m x_{\ell}(k) \right\| \leq \frac{1}{m} \sum_{j=1}^m \sum_{\ell=1}^m \|x_j(k) - x_{\ell}(k)\|.$$

We thus have

$$\alpha_k \|v_i(k) - \bar{v}(k)\| \leq \frac{\alpha_k}{m} \sum_{j=1}^m \sum_{\ell=1}^m \|x_j(k) - x_{\ell}(k)\|,$$

and by using the relation in (3.13), we conclude that

$$\sum_{k=0}^{\infty} \alpha_k \|v_i(k) - \bar{v}(k)\| < \infty \quad \text{for all } i \in V \text{ w.p.1.} \quad \blacksquare$$

3.4 Convergence with Probability 1

We are now ready to assert the convergence of the method (3.1a)-(3.1b) using the lemmas established in Section 3.3. To outline the rough idea of the proof, let us note that Lemma 3.2 allows us to infer that $v_i(k)$ approaches the set \mathcal{X} . Lemma 3.4 will allow us to claim that any two sequences $\{v_i(k)\}$ and $\{v_j(k)\}$ have the same accumulation with probability 1, under some mild assumptions on the stepsize. To claim the convergence of the iterates to an optimal solution, it remains to relate the accumulation points of $\{v_i(k)\}$ to

the optimal solutions of problem (1.1). This last piece is provided by the iterate relation of Lemma 3.1, supported by the convergence Theorem 2.1.

We have the following convergence result.

Proposition 3.1 *Let Assumptions 3.1-3.4 hold. Let the stepsize be such that $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. Assume that problem (1.1) has a nonempty optimal set \mathcal{X}^* . Then, the iterates $\{x_i(k)\}$, $i \in V$, generated by the method (3.1a)-(3.1b) converge with probability 1 to some random point in the optimal set \mathcal{X}^* , i.e., for some random vector $x^* \in \mathcal{X}^*$,*

$$\lim_{k \rightarrow \infty} x_i(k) = x^* \quad \text{for all } i \in V \text{ w.p.1.}$$

Proof We use the definition of the iterate $x_i(k)$ in (3.1a)-(3.1b) and lemma 3.1 with the following identification: $\mathcal{Y} = \mathcal{X}_i^{\Omega_i(k)}$, $y = x_i(k+1)$, $x = v_i(k)$, $z = z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$, $\alpha = \alpha_k$ and $\tau = c$ where c is the constant from the relation (3.3). Thus, for any $\tilde{x} \in \mathcal{X}$, $k \geq 0$ and $i \in V$, we have

$$\begin{aligned} \|x_i(k+1) - \tilde{x}\|^2 &\leq (1 + A\alpha_k^2) \|v_i(k) - \tilde{x}\|^2 - 2\alpha_k (f_i(z_i(k)) - f_i(\tilde{x})) \\ &\quad - \frac{3}{4} \|x_i(k+1) - v_i(k)\|^2 + \left(\frac{3}{8c} + 2\alpha_k L \right) \|v_i(k) - z_i(k)\|^2 + B\alpha_k^2 \|\nabla f_i(\tilde{x})\|^2, \end{aligned}$$

with $A = 8L^2 + 16cL^2$ and $B = 8c + 8$. We next sum the preceding relations over $i = 1, \dots, m$. Also, we use the convexity of the squared-norm (cf. (2.1)) and the doubly stochasticity of the weights to obtain the following relation:

$$\begin{aligned} \sum_{i=1}^m \|v_i(k) - \tilde{x}\|^2 &\leq \sum_{i=1}^m \sum_{j=1}^m [W(k)]_{ij} \|x_j(k) - \tilde{x}\|^2 \\ &= \sum_{j=1}^m \left(\sum_{i=1}^m [W(k)]_{ij} \right) \|x_j(k) - \tilde{x}\|^2 = \sum_{j=1}^m \|x_j(k) - \tilde{x}\|^2. \end{aligned}$$

By doing so, and taking into account that the gradients $\|\nabla f_i(\tilde{x})\|$ are bounded over \mathcal{X} by a scalar G_f (Assumption 3.1(d)), we obtain for any $\tilde{x} \in \mathcal{X}$ and $k \geq 0$,

$$\begin{aligned} \sum_{i=1}^m \|x_i(k+1) - \tilde{x}\|^2 &\leq (1 + A\alpha_k^2) \sum_{i=1}^m \|x_i(k) - \tilde{x}\|^2 \\ &\quad - 2\alpha_k \sum_{i=1}^m (f_i(z_i(k)) - f_i(\tilde{x})) - \frac{3}{4} \sum_{i=1}^m \|x_i(k+1) - v_i(k)\|^2 \end{aligned}$$

$$+ \left(\frac{3}{8c} + 2\alpha_k L \right) \sum_{i=1}^m \|v_i(k) - z_i(k)\|^2 + mB\alpha_k^2 G_f^2. \quad (3.14)$$

Let $\bar{z}(k) \triangleq \frac{1}{m} \sum_{\ell=1}^m z_\ell(k)$ and recall that $f(x) = \sum_{i=1}^m f_i(x)$. Using $\bar{z}(k)$ and f , we can rewrite the second term on the right hand side in (3.14) as follows:

$$\begin{aligned} \sum_{i=1}^m (f_i(z_i(k)) - f_i(\check{x})) &= \sum_{i=1}^m (f_i(z_i(k)) - f_i(\bar{z}(k))) \\ &\quad + (f(\bar{z}(k)) - f(\check{x})). \end{aligned} \quad (3.15)$$

We estimate the first term on the right hand side of the above equation as follows. Using the convexity of each function f_i , we obtain

$$\begin{aligned} \sum_{i=1}^m (f_i(z_i(k)) - f_i(\bar{z}(k))) &\geq \sum_{i=1}^m \langle \nabla f_i(\bar{z}(k)), z_i(k) - \bar{z}(k) \rangle \\ &\geq - \sum_{i=1}^m \|\nabla f_i(\bar{z}(k))\| \|z_i(k) - \bar{z}(k)\|. \end{aligned}$$

Since $\bar{z}(k)$ is a convex combination of points $z_i(k) \in \mathcal{X}$, it follows that $\bar{z}(k) \in \mathcal{X}$. This observation and Assumption 3.1(d), stating that the gradients $\nabla f_i(x)$ are uniformly bounded for $x \in \mathcal{X}$, yield

$$\sum_{i=1}^m (f_i(z_i(k)) - f_i(\bar{z}(k))) \geq -G_f \sum_{i=1}^m \|z_i(k) - \bar{z}(k)\|. \quad (3.16)$$

We next consider the term $\|z_i(k) - \bar{z}(k)\|$, for which by using $\bar{z}(k) \triangleq \frac{1}{m} \sum_{\ell=1}^m z_\ell(k)$ we have

$$\begin{aligned} \|z_i(k) - \bar{z}(k)\| &= \left\| \frac{1}{m} \sum_{\ell=1}^m (z_i(k) - z_\ell(k)) \right\| \\ &\leq \frac{1}{m} \sum_{\ell=1}^m \|z_i(k) - z_\ell(k)\| \leq \frac{1}{m} \sum_{\ell=1}^m \|v_i(k) - v_\ell(k)\|, \end{aligned}$$

where the first inequality is obtained by the convexity of the norm (see (2.1)) and the last inequality follows by the non-expansive projection property (Lemma 2.1). Furthermore, by using $\|v_i(k) - v_\ell(k)\| \leq \|v_i(k) - \bar{v}(k)\| +$

$\|v_\ell(k) - \bar{v}(k)\|$, we obtain for every $i \in V$,

$$\|z_i(k) - \bar{z}(k)\| \leq \|v_i(k) - \bar{v}(k)\| + \frac{1}{m} \sum_{\ell=1}^m \|v_\ell(k) - \bar{v}(k)\|.$$

Upon summing over $i \in V$, we find that

$$\sum_{i=1}^m \|z_i(k) - \bar{z}(k)\| \leq 2 \sum_{i=1}^m \|v_i(k) - \bar{v}(k)\|. \quad (3.17)$$

Combining relations (3.17) and (3.16), and substituting the resulting relation in equation (3.15), we find that

$$\sum_{i=1}^m (f_i(z_i(k)) - f_i(\check{x})) \geq -2G_f \sum_{i=1}^m \|v_i(k) - \bar{v}(k)\| + (f(\bar{z}(k)) - f(\check{x})).$$

Finally, by using the preceding estimate in inequality (3.14), we obtain for any $\check{x} \in \mathcal{X}$ and $k \geq 0$,

$$\begin{aligned} \sum_{i=1}^m \|x_i(k+1) - \check{x}\|^2 &\leq (1 + A\alpha_k^2) \sum_{i=1}^m \|x_i(k) - \check{x}\|^2 - 2\alpha_k(f(\bar{z}(k)) - f(\check{x})) \\ &\quad - \frac{3}{4} \sum_{i=1}^m \|x_i(k+1) - v_i(k)\|^2 + \left(\frac{3}{8c} + 2\alpha_k L\right) \sum_{i=1}^m \|v_i(k) - z_i(k)\|^2 \\ &\quad + 4\alpha_k G_f \sum_{i=1}^m \|v_i(k) - \bar{v}(k)\| + mB\alpha_k^2 G_f^2. \end{aligned} \quad (3.18)$$

By the definition of $x_i(k+1)$, we have $x_i(k+1) \in \mathcal{X}_i^{\Omega_i(k)}$, which implies $\|x_i(k+1) - v_i(k)\| \geq \text{dist}(v_i(k), \mathcal{X}_i^{\Omega_i(k)})$ for $i \in V$. Also, from the definition of $z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$, we have $\|v_i(k) - z_i(k)\| = \text{dist}(v_i(k), \mathcal{X})$ for $i \in V$. Using these relations and letting $\check{x} = x^*$ for an arbitrary $x^* \in \mathcal{X}^*$, from (3.18) we obtain for all $k \geq 0$,

$$\begin{aligned} \sum_{i=1}^m \|x_i(k+1) - x^*\|^2 &\leq (1 + A\alpha_k^2) \sum_{i=1}^m \|x_i(k) - x^*\|^2 - 2\alpha_k(f(\bar{z}(k)) - f^*) \\ &\quad - \frac{3}{4} \sum_{i=1}^m \text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) + \left(\frac{3}{8c} + 2\alpha_k L\right) \sum_{i=1}^m \text{dist}^2(v_i(k), \mathcal{X}) \\ &\quad + 4\alpha_k G_f \sum_{i=1}^m \|v_i(k) - \bar{v}(k)\| + mB\alpha_k^2 G_f^2. \end{aligned}$$

By taking the expectation conditioned on \mathcal{F}_k , and noting that $x_i(k)$, $v_i(k)$, $\bar{v}(k)$, and $\bar{z}(k)$ are fully determined by \mathcal{F}_k , we have with probability 1 for all $x^* \in \mathcal{X}$ and $k \geq 0$,

$$\begin{aligned}
& \mathbb{E} \left[\sum_{i=1}^m \|x_i(k+1) - x^*\|^2 \mid \mathcal{F}_k \right] \\
& \leq (1 + A\alpha_k^2) \sum_{i=1}^m \|x_i(k) - x^*\|^2 - 2\alpha_k(f(\bar{z}(k)) - f^*) \\
& \quad - \frac{3}{4} \mathbb{E} \left[\sum_{i=1}^m \text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \mid \mathcal{F}_k \right] + \left(\frac{3}{8c} + 2\alpha_k L \right) \sum_{i=1}^m \text{dist}^2(v_i(k), \mathcal{X}) \\
& \quad + 4\alpha_k G_f \sum_{i=1}^m \|v_i(k) - \bar{v}(k)\| + mB\alpha_k^2 G_f^2.
\end{aligned}$$

By Assumption 3.2, we have $\text{dist}^2(x, \mathcal{X}) \leq c \mathbb{E} [\text{dist}^2(x, \mathcal{X}_i^{\Omega_i(k)}) \mid \mathcal{F}_k]$ for all $x \in \mathcal{X}$ and all $i \in V$. Furthermore, since $\alpha_k \rightarrow 0$, by choosing \bar{k} large enough so that $2\alpha_k L \leq \frac{3}{8c}$, we have for all $k \geq \bar{k}$,

$$-\frac{3}{4} \mathbb{E} \left[\sum_{i=1}^m \text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \mid \mathcal{F}_k \right] + \left(\frac{3}{8c} + 2\alpha_k L \right) \sum_{i=1}^m \text{dist}^2(v_i(k), \mathcal{X}) \leq 0.$$

Thus, we obtain with probability 1 for all $k \geq \bar{k}$ and $x^* \in \mathcal{X}^*$,

$$\begin{aligned}
& \mathbb{E} \left[\sum_{i=1}^m \|x_i(k+1) - x^*\|^2 \mid \mathcal{F}_k \right] \leq (1 + A\alpha_k^2) \sum_{i=1}^m \|x_i(k) - x^*\|^2 \\
& \quad - 2\alpha_k(f(\bar{z}(k)) - f^*) + 4\alpha_k G_f \sum_{i=1}^m \|v_i(k) - \bar{v}(k)\| + mB\alpha_k^2 G_f^2. \quad (3.19)
\end{aligned}$$

Since $\bar{z}(k) \in \mathcal{X}$, we have $f(\bar{z}(k)) - f^* \geq 0$. Thus, under the assumption $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ and Lemma 3.4, relation (3.19) satisfies all the conditions of the convergence Theorem 2.1. Hence, the sequence $\{\|x_i(k) - x^*\|^2\}$ is convergent with probability 1 for any $i \in V$ and $x^* \in \mathcal{X}^*$, and

$$\sum_{k=0}^{\infty} \alpha_k (f(\bar{z}(k)) - f(x^*)) < \infty \quad \text{w.p.1.}$$

The preceding relation and the condition $\sum_{k=0}^{\infty} \alpha_k = \infty$ imply that

$$\liminf_{k \rightarrow \infty} (f(\bar{z}(k)) - f(x^*)) = 0 \quad \text{w.p.1.} \quad (3.20)$$

By Lemma 3.2, noting that $z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$, we have $\sum_{k=1}^{\infty} \sum_{i=1}^m \|v_i(k) - z_i(k)\|^2 < \infty$ with probability 1, implying

$$\lim_{k \rightarrow \infty} \|v_i(k) - z_i(k)\| = 0 \quad \text{for all } i \in V \text{ w.p.1.} \quad (3.21)$$

Recall that the sequence $\{\|x_i(k) - x^*\|\}$ is convergent with probability 1 for all $i \in V$ and every $x^* \in \mathcal{X}^*$. Then, in view of relation (3.1a), we have that the sequence $\{\|v_i(k) - x^*\|\}$ is also convergent with probability 1 for all $i \in V$ and $x^* \in \mathcal{X}^*$. By relation (3.21) it follows that $\{\|z_i(k) - x^*\|\}$ is also convergent with probability 1 for all $i \in V$ and $x^* \in \mathcal{X}^*$. Since $\|\bar{v}_i(k) - x^*\| \leq \frac{1}{m} \sum_{i=1}^m \|v_i(k) - x^*\|$ and the sequence $\{\|v_i(k) - x^*\|\}$ is convergent with probability 1 for all $i \in V$ and $x^* \in \mathcal{X}^*$, it follows that $\{\|\bar{v}(k) - x^*\|\}$ is convergent with probability 1 for all $x^* \in \mathcal{X}^*$. Using a similar argument, we can conclude that $\{\|\bar{z}(k) - x^*\|\}$ is convergent with probability 1 for all $x^* \in \mathcal{X}^*$. As a particular consequence, it follows that the sequences $\{\bar{v}(k)\}$ and $\{\bar{z}(k)\}$ are bounded with probability 1 and, hence, they have accumulation points. From relation (3.20) and the continuity of f , it follows that the sequence $\{\bar{z}(k)\}$ must have one accumulation point in the set \mathcal{X}^* with probability 1. This and the fact that $\{\|\bar{z}(k) - x^*\|\}$ is convergent with probability 1 for every $x^* \in \mathcal{X}^*$ imply that for a random point $x^* \in \mathcal{X}^*$,

$$\lim_{k \rightarrow \infty} \bar{z}(k) = x^* \quad \text{w.p.1.} \quad (3.22)$$

Now, from $\bar{z}(k) = \frac{1}{m} \sum_{\ell=1}^m z_{\ell}(k)$ and $\bar{v}(k) = \frac{1}{m} \sum_{i=\ell}^m v_{\ell}(k)$, using relation (3.21) and the convexity of the norm (cf. (2.1)), we obtain with probability 1

$$\lim_{k \rightarrow \infty} \|\bar{v}(k) - \bar{z}(k)\| \leq \frac{1}{m} \sum_{\ell=1}^m \lim_{k \rightarrow \infty} \|v_{\ell}(k) - z_{\ell}(k)\| = 0.$$

In view of relation (3.22), it follows that

$$\lim_{k \rightarrow \infty} \bar{v}(k) = x^* \quad \text{w.p.1.} \quad (3.23)$$

By relation (3.10) in Lemma 3.4, we have

$$\liminf_{k \rightarrow \infty} \|v_i(k) - \bar{v}(k)\| = 0 \quad \text{for all } i \in V \text{ w.p.1.} \quad (3.24)$$

The fact that $\{\|v_i(k) - x^*\|\}$ is convergent with probability 1 for all i , together with (3.23) and (3.24) implies that

$$\lim_{k \rightarrow \infty} \|v_i(k) - x^*\| = 0 \quad \text{for } i \in V \text{ w.p.1.} \quad (3.25)$$

Finally, from relation (3.9) in Lemma 3.4, we have $\lim_{k \rightarrow \infty} \|x_i(k+1) - v_i(k)\| = 0$ for all $i \in V$ with probability 1, which together with the limit in (3.25) yields $\lim_{k \rightarrow \infty} x_i(k) = x^*$ for all $i \in V$ with probability 1. ■

3.5 Distributed Mini-batch Random Projections

As an extension of the algorithm in (3.1a)–(3.1b), one may consider an algorithm where the agents use several random projections at each iteration. Namely, after generating $v_i(k)$ each agent may take (or nature may reveal them) several random samples $\Omega_i^1(k), \dots, \Omega_i^b(k)$, where each $\Omega_i^r(k) \in I_i$ and $b \geq 1$ is the batch-size. Each collection $\Omega_i^1(k), \dots, \Omega_i^b(k)$ consists of mutually independent random variables and is independent of the past realizations. More specifically, we have b random independent samples of the *i.i.d.* random variable $\Omega_i(k)$ (taking values in I_i). Using the compact form (3.2) for the update in (3.1a), in the mini-batch version of the algorithm, each agent $i \in V$, performs the following steps:

$$v_i(k) = \sum_{j=1}^m [W(k)]_{ij} x_j(k), \quad (3.26a)$$

$$\psi_i^0(k) = v_i(k) - \alpha_k \nabla f_i(v_i(k)), \quad (3.26b)$$

$$\psi_i^r(k) = \Pi_{\mathcal{X}_i^{\Omega_i^r(k)}}[\psi_i^{r-1}(k)] \text{ for } r = 1, \dots, b, \quad (3.26c)$$

$$x_i(k+1) = \psi_i^b(k), \quad (3.26d)$$

where $\alpha_k > 0$ is a stepsize at time k and $x_i(0) \in \mathbb{R}^d$ is an initial estimate of agent i (which can be random). The steps in (3.26b)–(3.26d) are the successive (random) projections on the sets $\mathcal{X}_i^{\Omega_i^1(k)}, \dots, \mathcal{X}_i^{\Omega_i^b(k)}$ of the point

$$v_i(k) - \alpha_k \nabla f_i(v_i(k)).$$

The algorithm using mini-batches for random projections is of interest when the set I_i is large, i.e., the number of constraint set components \mathcal{X}_i^j , $j \in I_i$, of the set $\mathcal{X}_i = \bigcap_{j \in I_i} \mathcal{X}_i^j$ is large. In such cases, taking several projection steps is beneficial for reducing the infeasibility of the iterates $x_i(k)$ with respect to the set \mathcal{X}_i . More concretely, if each set \mathcal{X}_i is the intersection of about 10^4 simpler sets, then one sample of these sets will render a poor approximation of the true set \mathcal{X}_i , whereas 100 samples will provide a better approximation of the set. Let \tilde{x} be a point in the feasible set \mathcal{X} . If just one sample is considered at each iteration, by the non-expansive projection property (Lemma 2.1), the distance between the next iterate and a point in \mathcal{X} can be estimated as:

$$\|x_i(k+1) - \tilde{x}\| = \|\psi_i^1(k) - \tilde{x}\| \leq \|\psi_i^0(k) - \tilde{x}\|,$$

whereas if 100 samples are considered for projections,

$$\|x_i(k+1) - \tilde{x}\| = \|\psi_i^{100}(k) - \tilde{x}\| \leq \dots \leq \|\psi_i^1(k) - \tilde{x}\| \leq \|\psi_i^0(k) - \tilde{x}\|,$$

which may yield a larger infeasibility reduction.

The convergence proof of this algorithm is similar to that of Proposition 3.1. We construct the proof by adjusting the result of Lemma 3.1, and by verifying that Lemma 3.2 and Lemma 3.4 apply to the mini-batch variant of the DRP method. The basic insight that guides the proof is that the operation of successive projections on components \mathcal{X}_i^j of the set $\mathcal{X}_i = \bigcap_{j \in I_i} \mathcal{X}_i^j$ remains a non-expansive operation with respect to points that belong to the set \mathcal{X}_i , as well as with respect to the points in the intersection set $\mathcal{X} = \bigcap_{i=1}^m \mathcal{X}_i$.

Basic Iterate Relation for Mini-Batch Algorithm

For the iterates generated by the mini-batch random projection algorithm in (3.26a)–(3.26d), we have the following basic result.

Lemma 3.5 *Let Assumption 3.1 hold. Then, for any $\tilde{x} \in \mathcal{X}$, and for all*

$i \in V$ and all $k \geq 0$,

$$\begin{aligned} \|x_i(k+1) - \tilde{x}\|^2 &\leq (1 + A_\tau \alpha_k^2) \|v_i(k) - \tilde{x}\|^2 - 2\alpha_k(f_i(z) - f_i(\tilde{x})) \\ &\quad - \frac{3}{4} \|\psi_i^1(k) - v_i(k)\|^2 + \left(\frac{3}{8\tau} + 2\alpha_k L\right) \|v_i(k) - z\|^2 + B_\tau \alpha_k^2 G_f^2, \end{aligned}$$

where $A_\tau = 8L^2 + 16\tau L^2$, $B_\tau = 8\tau + 8$ and $\tau > 0$ is arbitrary.

Proof By using the non-expansiveness property of projection operation (Lemma 2.1(a)), we have for arbitrary $\tilde{x} \in \mathcal{X}$ (since $\mathcal{X} \subseteq \mathcal{X}_i^j$ for all $j \in I_i$), and for all $i \in V$ and $k \geq 0$,

$$\|x_i(k+1) - \tilde{x}\| \leq \|\psi_i^{b-1}(k) - \tilde{x}\| \leq \dots \leq \|\psi_i^1(k) - \tilde{x}\|. \quad (3.27)$$

The intermediate iterate $\psi_i^1(k)$ is just obtained after one projection step,

$$\psi_i^1(k) = \Pi_{\mathcal{X}_i^{\Omega_i^1(k)}}[v_i(k) - \alpha_k \nabla f_i(v_i(k))],$$

so it satisfies Lemma 3.1 with $y = \psi_i^1(k)$, $\mathcal{Y} = \mathcal{X}_i^{\Omega_i^1(k)}$, $x = v_i(k)$, $\alpha = \alpha_k$, and $\phi = f_i$. Thus, we have for any $\tilde{x} \in \mathcal{X}$ and $z \in \mathbb{R}^d$,

$$\begin{aligned} \|\psi_i^1(k) - \tilde{x}\|^2 &\leq (1 + A_\tau \alpha_k^2) \|v_i(k) - \tilde{x}\|^2 - 2\alpha_k(f_i(z) - f_i(\tilde{x})) \\ &\quad - \frac{3}{4} \|\psi_i^1(k) - v_i(k)\|^2 + \left(\frac{3}{8\tau} + 2\alpha_k L\right) \|v_i(k) - z\|^2 + B_\tau \alpha_k^2 \|\nabla f_i(\tilde{x})\|^2. \end{aligned} \quad (3.28)$$

From (3.27) and (3.28), by using the gradient boundedness property of Assumption 3.1(d), we obtain the stated relation. \blacksquare

Conditional Expectation Relation for Mini-Batch Algorithm

The convergence proof of Proposition 3.2 requires a relation for the iterates involving expectations with respect to the past history of the method. For this, we need to define a relevant σ -algebra. We let $\tilde{\mathcal{F}}_k$ be the σ -algebra generated by the entire history of the algorithm up to time $k-1$ inclusively. Thus, $\tilde{\mathcal{F}}_k$ includes the realizations of all the random variables but not the realizations of the indices $\Omega_i^1(k), \dots, \Omega_i^b(k)$ at time k . Specifically, it is given

by for all $k \geq 1$,

$$\tilde{\mathcal{F}}_k = \{x_i(0), i \in V\} \cup \{\Omega_i^r(\ell); 0 \leq \ell \leq k-1, 1 \leq r \leq b, i \in V\},$$

where $\tilde{\mathcal{F}}_0 = \{x_i(0), i \in V\}$.

Now, with this definition of the σ -algebra, we have the following result.

Lemma 3.6 *Let Assumptions 3.1 and 3.2 hold. Then, with probability 1 for any $\tilde{x} \in \mathcal{X}$, and for all $i \in V$ and all $k \geq 0$,*

$$\begin{aligned} \mathbb{E} \left[\|x_i(k+1) - \tilde{x}\|^2 \mid \tilde{\mathcal{F}}_k \right] &\leq (1 + A\alpha_k^2) \|v_i(k) - \tilde{x}\|^2 - 2\alpha_k (f_i(z_i(k)) - f_i(\tilde{x})) \\ &\quad - \left(\frac{3}{8c} - 2\alpha_k L \right) \text{dist}^2(v_i(k), \mathcal{X}) + B\alpha_k^2 G_f^2, \end{aligned}$$

where $z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$, $A = 8L^2 + 16cL^2$, $B = 8c + 8$, and c is from Assumption 3.2.

Proof By letting $z = z_i(k)$ and $\tau = c$ in Lemma 3.5, we obtain

$$\begin{aligned} \mathbb{E} \left[\|x_i(k+1) - \tilde{x}\|^2 \mid \tilde{\mathcal{F}}_k \right] &\leq (1 + A\alpha_k^2) \|v_i(k) - \tilde{x}\|^2 - 2\alpha_k (f_i(z_i(k)) - f_i(\tilde{x})) \\ &\quad - \frac{3}{4} \mathbb{E} \left[\|\psi_i^1(k) - v_i(k)\|^2 \mid \tilde{\mathcal{F}}_k \right] + \left(\frac{3}{8c} + 2\alpha_k L \right) \|v_i(k) - z_i(k)\|^2 + B\alpha_k^2 G_f^2, \end{aligned}$$

where $A = 8L^2 + 16cL^2$ and $B = 8c + 8$.

Since $\psi_i^1(k) \in \mathcal{X}_i^{\Omega_i^1(k)}$, by the projection property we have $\|\psi_i^1(k) - v_i(k)\|^2 \geq \|\Pi_{\mathcal{X}_i^{\Omega_i^1(k)}}[v_i(k)] - v_i(k)\|^2$. Then,

$$\begin{aligned} \mathbb{E} \left[\|\psi_i^1(k) - v_i(k)\|^2 \mid \tilde{\mathcal{F}}_k \right] &\geq \mathbb{E} \left[\|\Pi_{\mathcal{X}_i^{\Omega_i^1(k)}}[v_i(k)] - v_i(k)\|^2 \mid \tilde{\mathcal{F}}_k \right] \\ &= \mathbb{E} \left[\|\Pi_{\mathcal{X}_i^{\Omega_i^1(k)}}[v_i(k)] - v_i(k)\|^2 \mid v_i(k) \right]. \end{aligned}$$

Furthermore, by Assumption 3.2 we have

$$\mathbb{E} \left[\|\Pi_{\mathcal{X}_i^{\Omega_i^1(k)}}[v_i(k)] - v_i(k)\|^2 \mid v_i(k) \right] \geq \frac{1}{c} \text{dist}^2(v_i(k), \mathcal{X}).$$

The preceding relations and $\text{dist}(v_i(k), \mathcal{X}) = \|v_i(k) - z_i(k)\|$ yield the desired relation. \blacksquare

Lemma 3.2 and Lemma 3.4 hold

Using Lemma 3.6, we argue that the results of Lemma 3.2 and Lemma 3.4 apply to the mini-batch random projection algorithm.

Claim 1 *Lemma 3.2 holds for the iterates generated by method (3.26a)–(3.26d).*

Proof By letting $\tilde{x} = \Pi_{\mathcal{X}}[v_i(k)]$ in Lemma 3.6, and noting that $\|x_i(k+1) - \Pi_{\mathcal{X}}[v_i(k)]\| \geq \text{dist}(x_i(k+1), \mathcal{X})$ and $\|v_i(k) - \Pi_{\mathcal{X}}[v_i(k)]\| = \text{dist}(v_i(k), \mathcal{X})$, we obtain

$$\begin{aligned} \mathbb{E} \left[\text{dist}^2(x_i(k+1), \mathcal{X}) \mid \tilde{\mathcal{F}}_k \right] &\leq (1 + A\alpha^2) \text{dist}^2(v_i(k), \mathcal{X}) \\ &\quad - \left(\frac{3}{8c} - 2\alpha_k L \right) \text{dist}^2(v_i(k), \mathcal{X}) + B\alpha_k^2 G_f^2, \end{aligned}$$

which is the same as relation (3.7) within the proof of Lemma 3.2. The rest of the proof of Lemma 3.2 holds exactly as given, and the result of Lemma 3.2 remains valid. ■

Claim 2 *Lemma 3.4 holds for the iterates generated by method (3.26a)–(3.26d).*

Proof Define $e_i(k) = x_i(k+1) - v_i(k)$ and $z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$. Now, consider $\|e_i(k)\|$ for which we have

$$\|e_i(k)\| \leq \|x_i(k+1) - z_i(k)\| + \|z_i(k) - v_i(k)\|.$$

The non-expansiveness projection property and the fact $z_i(k) \in \mathcal{X} \subset \mathcal{X}_i^{\Omega_i^r(k)}$, for all $r = 1, \dots, b$, and any realization of these sets imply

$$\begin{aligned} \|x_i(k+1) - z_i(k)\| &\leq \|\psi_i^{b-1}(k) - z_i(k)\| \leq \dots \leq \|\psi_i^1(k) - z_i(k)\| \\ &\leq \|v_i(k) - \alpha_k \nabla f_i(v_i(k)) - z_i(k)\|. \end{aligned}$$

Therefore

$$\|e_i(k)\| \leq \|v_i(k) - \alpha_k \nabla f_i(v_i(k)) - z_i(k)\| + \|z_i(k) - v_i(k)\|,$$

which is the same as the first inequality in (3.11) within the proof of Lemma 3.4. The rest of the proof of that lemma holds in verbatim, and the result follows. ■

We now connect the preceding results and provide the proof of the following proposition.

Proposition 3.2 *Let Assumptions 3.1-3.4 hold, and let the stepsize satisfy $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. Assume that problem (1.1) has a nonempty optimal set \mathcal{X}^* . Then, the iterates $\{x_i(k)\}$, $i \in V$, produced by the method (3.26a)-(3.26d) converge to some random point in the optimal set \mathcal{X}^* with probability 1, i.e., for some random vector $x^* \in \mathcal{X}^*$,*

$$\lim_{k \rightarrow \infty} x_i(k) = x^* \quad \text{for all } i \in V \text{ w.p.1.}$$

Proof Starting from the relation in Lemma 3.6, after summing over all $i \in V$, we can see that with probability 1 for all $\tilde{x} \in \mathcal{X}$ and all $k \geq 0$,

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^m \|x_i(k+1) - \tilde{x}\|^2 \mid \tilde{\mathcal{F}}_k \right] &\leq (1 + A\alpha_k^2) \sum_{i=1}^m \|v_i(k) - \tilde{x}\|^2 \\ &\quad - 2\alpha_k \sum_{i=1}^m (f_i(z_i(k)) - f_i(\tilde{x})) - \left(\frac{3}{8c} - 2\alpha_k L \right) \sum_{i=1}^m \|v_i(k) - z_i(k)\|^2 + mB\alpha_k^2 G_f^2, \end{aligned}$$

where $z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$.

Now, the same as in the proof of Proposition 3.1, using the properties of the matrices $W(k)$ and the convexity of the squared-norm function (see (2.1)), we can show that

$$\sum_{i=1}^m \|v_i(k) - \tilde{x}\|^2 \leq \sum_{j=1}^m \|x_j(k) - \tilde{x}\|^2.$$

Also, using verbatim arguments, we can show that

$$\sum_{i=1}^m (f_i(z_i(k)) - f_i(\tilde{x})) \geq -2G_f \sum_{i=1}^m \|v_i(k) - \bar{v}(k)\| + (f(\bar{z}(k)) - f(\tilde{x})),$$

where $\bar{z}(k) = \frac{1}{m} \sum_{\ell=1}^m z_{\ell}(k)$ and $\bar{v}(k) = \frac{1}{m} \sum_{\ell=1}^m v_{\ell}(k)$. Under the conditions of Proposition 3.2, we have $\alpha_k \rightarrow 0$. Choosing \bar{k} large enough so that $2\alpha_k L \leq \frac{3}{8c}$ for all $k \geq \bar{k}$, we have

$$- \left(\frac{3}{8c} - 2\alpha_k L \right) \sum_{i=1}^m \|v_i(k) - z_i(k)\|^2 \leq 0.$$

By combining all the preceding relations, we obtain with probability 1 for all $\tilde{x} \in \mathcal{X}$ and all $k \geq \bar{k}$,

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^m \|x_i(k+1) - \tilde{x}\|^2 \mid \tilde{\mathcal{F}}_k \right] &\leq (1 + A\alpha_k^2) \sum_{i=1}^m \|x_i(k) - \tilde{x}\|^2 \\ &\quad - 2\alpha_k(f(\bar{z}(k)) - f(\tilde{x})) + 4\alpha_k G_f \sum_{i=1}^m \|v_i(k) - \bar{v}(k)\| + mB\alpha_k^2 G_f^2. \end{aligned}$$

Letting $\tilde{x} = x^*$ for an arbitrary optimal solution $x^* \in \mathcal{X}^*$, from the preceding relation we arrive at relation (3.19) in the proof of Proposition 3.1. From relation (3.19) onward, the proof of Proposition 3.1 holds verbatim, and the stated convergence with probability 1 of the mini-batch method follows. ■

CHAPTER 4

ASYNCHRONOUS GOSSIP-BASED RANDOM PROJECTION ALGORITHM

4.1 GRP Algorithm

We propose a variant distributed optimization algorithm for problem (1.1) that is based on the random projections and the gossip communication protocol. Gossip algorithms robustly achieve consensus through sparse communications in the network. That is, only one edge $\{i, j\}$ in the network is randomly selected for communication at each iteration and agents i and j simply average their values. From now on, we refer to our algorithm as *Gossip-based Random Projection* (GRP).

GRP uses an asynchronous time model as in [38]. The time model assumes that each agent has a random clock. The notion of randomness is imposed to desynchronize actions on the network. More specifically, each agent has a local clock that ticks at a Poisson rate of 1. The setting can be visualized as having a single virtual clock that ticks whenever any local Poisson clock ticks. Thus, the ticking of the virtual clock is a Poisson random process with rate m . Let Z_k be the absolute time of the k th tick of the virtual clock. The time is discretized according to the intervals $[Z_{k-1}, Z_k)$ and this time slot corresponds to our discrete time k . Let I_k denote the index of the agent that wakes up at time k and J_k denote the index of agent I_k 's neighbor that is selected for communication. We assume that only one agent wakes up at a time.

The distribution by which J_k is selected is characterized by a nonnegative stochastic $m \times m$ matrix $[\Pi]_{ij} = \pi_{ij}$ that conforms with the graph topology $G = (V, E)$, i.e., $\pi_{ij} > 0$ only if $\{i, j\} \in E$. At iteration k , agent I_k wakes up and contacts one of its neighbors J_k with probability $\pi_{I_k J_k}$.

Let $x_i(k)$ denote the estimate of agent i at time k . GRP updates these estimates according to the following rule. Each agent starts with some initial

vector $x_i(0)$, which can be randomly selected. For $k \geq 1$, agents other than I_k and J_k do not update:

$$x_i(k) = x_i(k-1) \quad \text{for all } i \notin \{I_k, J_k\}. \quad (4.1)$$

Agents I_k and J_k calculate the average of their estimates, and adjust the average by using their local gradient information and by projecting onto a randomly selected component of their local constraint sets; i.e., for $i \in \{I_k, J_k\}$:

$$\begin{aligned} v_i(k) &= (x_{I_k}(k-1) + x_{J_k}(k-1))/2, \\ x_i(k) &= \Pi_{\mathcal{X}_i^{\Omega_i(k)}} [v_i(k) - \alpha_i(k) \nabla f_i(v_i(k))], \end{aligned} \quad (4.2)$$

where $\alpha_i(k)$ is a stepsize of agent i , and $\Omega_i(k)$ is a random variable drawn from the set I_i . The key difference between the work in [39–41] and this thesis is the random projection step. Instead of projecting on the whole constraint set \mathcal{X}_i , a component set $\mathcal{X}_i^{\Omega_i(k)}$ is selected (or revealed by nature) and the projection is made on that set, which reduces the required computations per iteration.

For an alternative representation of GRP we define a nonnegative matrix $W(k)$ as follows:

$$W(k) = I - \frac{1}{2}(e_{I_k} - e_{J_k})(e_{I_k} - e_{J_k})' \quad \text{for } k \geq 1,$$

where I is the m -dimensional identity matrix, $e_i \in \mathbb{R}^m$ is a vector whose i th entry is equal to 1 and all other entries are equal to 0. Each $W(k)$ is doubly stochastic by construction, implying that $\mathbb{E}[W(k)]$ is also doubly stochastic. Using $W(k)$, algorithm (4.1)–(4.2) can be equivalently represented as

$$v_i(k) = \sum_{j=1}^m [W(k)]_{ij} x_j(k-1), \quad (4.3a)$$

$$p_i(k) = \Pi_{\mathcal{X}_i^{\Omega_i(k)}} [v_i(k) - \alpha_i(k) \nabla f(v_i(k))] - v_i(k), \quad (4.3b)$$

$$x_i(k) = v_i(k) + p_i(k) \chi_{\{i \in \{I_k, J_k\}\}}, \quad (4.3c)$$

where $\chi_{\mathcal{E}}$ is the characteristic function of an event \mathcal{E} , that is, $\chi_{\mathcal{E}} = 1$ if \mathcal{E} happens, and $\chi_{\mathcal{E}} = 0$ otherwise.

From here onward, we will shorten $\mathbb{E}[W(k)] = \bar{W}$ since the matrices $W(k)$ are identically distributed. Let λ denote to the second largest eigenvalue of \bar{W} . If the underlying graph G is connected, the incidence graph associated with the positive entries in the matrix \bar{W} is connected with a self-loop at each node. Hence, we have $\lambda < 1$.

In the convergence analysis of the algorithm (4.3a)-(4.3c), we use two different choices of stepsize for asynchronous algorithms. For a diminishing stepsize, we use $\alpha_i(k) = \frac{1}{\Gamma_i(k)}$ where $\Gamma_i(k)$ denotes the number of updates that agent i has performed until time k . Since every agent $i \in V$ has access to a locally defined quantity $\Gamma_i(k)$, the stepsize of agent i is independent of every other agent and no coordination is needed for its update. We provide a convergence proof for this random diminishing stepsize. Another choice is a constant deterministic stepsize $\alpha_i(k) = \alpha_i > 0$. For the constant stepsize, we provide an error bound.

4.2 Assumptions

We next discuss our assumptions, the first of which ensures that the information of each agent influences every other agent.

Assumption 4.1 *The underlying graph $G = (V, E)$ is connected. Furthermore, the neighbor selection process is i.i.d., whereby at any time agent i is chosen by its neighbor $j \in \mathcal{N}(i)$ with probability $\pi_{ji} > 0$ ($\pi_{ji} = 0$ if $j \notin \mathcal{N}(i)$) independent of the other agents in the network.*

We use the following assumption for the functions f_i and the sets \mathcal{X}_i^j .

Assumption 4.2 *Let the following conditions hold:*

- (a) *The sets \mathcal{X}_i^j , $j \in I_i$, are closed and convex for every $i \in V$.*
- (b) *Each function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex over \mathbb{R}^d .*
- (c) *Each function f_i is differentiable and has Lipschitz gradients with a constant L_i over \mathbb{R}^d ,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^d.$$

(d) The gradients $\nabla f_i(x)$, $i \in V$, are bounded over the set \mathcal{X} , i.e., there exists a constant G_f such that

$$\|\nabla f_i(x)\| \leq G_f \quad \text{for all } x \in \mathcal{X} \text{ and all } i \in V.$$

Assumption 4.2(d) is satisfied, for example, when the constraint set \mathcal{X} is compact.

The next assumption states set regularity, which is crucial in our convergence analysis.

Assumption 4.3 *There exists a constant $c > 0$ such that for all $i \in V$ and $x \in \mathbb{R}^d$,*

$$\text{dist}^2(x, \mathcal{X}) \leq c \mathbb{E} \left[\text{dist}^2(x, \mathcal{X}_i^{\Omega_i(k)}) | \Omega_\ell(t), t \in [1, k], \ell \in V \right].$$

Assumption 4.3 is satisfied, for example, if each set \mathcal{X}_i^j is an affine set, or the constraint set \mathcal{X} has a nonempty interior.

4.3 Convergence Analysis

In this section, we assert the convergence of the GRP method when the uncoordinated random diminishing stepsize $\alpha_i(k) = \frac{1}{\Gamma_i(k)}$ is used in algorithm (4.3a)-(4.3c). We start by providing roadmap and lemmas that will help us with the convergence proof.

4.3.1 Roadmap of the Proof

The asynchronous GRP algorithm with a diminishing stepsize has three random factors, namely, random gossip communication, random stepsizes and random projections. Each of these will be efficiently handled as follows.

Random Gossip Communication: At each iteration of the algorithm, an agent and one of its neighbors is randomly selected and the gossip communication matrix $W(k)$ is realized. In the analysis, we can work with \bar{W} instead of $W(k)$ due to the following properties of the matrices $W(k)$:

- (i) Each $W(k)$ is a symmetric projection matrix. Therefore, $\bar{W}'\bar{W} = \bar{W}$ and also $(\bar{W} - \frac{1}{m}\mathbf{1}\mathbf{1}')^2 = \bar{W} - \frac{1}{m}\mathbf{1}\mathbf{1}'$.

- (ii) Since \bar{W} is doubly stochastic, the largest eigenvalue of \bar{W} is 1. Therefore, the largest eigenvalue of the matrix $\bar{W} - \frac{1}{m}\mathbf{1}\mathbf{1}'$ is the same as λ (the second largest eigenvalue of \bar{W}).

The properties (i) and (ii) immediately yield the following relation for any $y \in \mathbb{R}^m$:

$$\left\| \left(\bar{W} - \frac{1}{m}\mathbf{1}\mathbf{1}' \right) y \right\|^2 \leq \lambda \|y\|^2. \quad (4.4)$$

Furthermore, in view of the connectivity of the underlying graph (Assumption 4.1), we have $\lambda < 1$.

Random Stepsizes: We examine a long-term behavior of the stepsizes in Lemma 4.2. The random diminishing stepsize $\alpha_i(k) = \frac{1}{\Gamma_i(k)}$ exhibits the same behavior as the deterministic stepsize $1/k$ in a long run. It enables us to handle the cross dependencies of the random stepsizes and the other randomness in the GRP method.

Random Projections: A projection error is incurred at each iteration of the algorithm since we select only one component from the constraint set and project the current point onto that single component. In Lemma 4.4, we characterize this projection error by showing that the intermediate iterates $\{v_i(k)\}$ approach the feasible set \mathcal{X} . We also show that the error between the two iterates $\{v_i(k)\}$ and $\{x_i(k)\}$ goes to zero with probability 1.

In addition to those, we need to show that i) the agents' estimates $x_i(k)$ eventually arrive at a consensus and ii) the consensus point lies in the optimal set. For part i), we quantify the agents' disagreements on the estimates $v_i(k)$ in Lemma 4.6, and show that the disagreements accumulate to zero. For part ii), we use Lemma 2.1 by letting $v_k = \sum_{i=1}^m \|x_i(k) - x^*\|^2$ for some optimal point x^* . For this, we first state Lemma 4.3 which provides a relation similar to that of Lemma 2.1 for GRP algorithm.

4.3.2 Basic Iterate Relation for GRP

We provide a relation among the iterates obtained after one step of the algorithm (4.3a)-(4.3c) and a point in the feasible set \mathcal{X} in Lemma 4.3.

For this, we begin with an auxiliary lemma which provides some basic relations among a point $\tilde{x} \in \mathcal{X}$, an arbitrary point $z \in \mathbb{R}^d$, and two consecutive

iterates x and y of a projected-gradient algorithm. The point z is used to accommodate the iterations $v_i(k)$ of the GRP method which may not lie in the constraint set \mathcal{X} , while \tilde{x} will most often be a suitably chosen point in \mathcal{X} .

Lemma 4.1 *Let $\mathcal{Y} \subseteq \mathbb{R}^d$ be a closed convex set. Let the function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and differentiable over \mathbb{R}^d with Lipschitz continuous gradients with a constant L . Let y be given by*

$$y = \Pi_{\mathcal{Y}}[x - \alpha \nabla \phi(x)] \quad \text{for some } x \in \mathbb{R}^d \text{ and } \alpha > 0.$$

Then, we have:

(a) *For any $\tilde{x} \in \mathcal{Y}$ and $z \in \mathbb{R}^d$,*

$$\begin{aligned} \|y - \tilde{x}\|^2 &\leq (1 + 8\alpha^2 L^2) \|x - \tilde{x}\|^2 - 2\alpha (\phi(z) - \phi(\tilde{x})) - \frac{3}{4} \|y - x\|^2 \\ &\quad + (8 + \tau_2) \alpha^2 \|\nabla \phi(\tilde{x})\|^2 + \tau_1 \alpha^2 L^2 \|z - \tilde{x}\|^2 + \left(\frac{1}{\tau_1} + \frac{1}{\tau_2} \right) \|x - z\|^2, \end{aligned}$$

where $\tau_1, \tau_2 > 0$ are arbitrary.

(b) *In addition, if ϕ is strongly-convex with a constant $\sigma > 0$, then for any $\tilde{x} \in \mathcal{Y}$, $z \in \mathbb{R}^d$, and $\tau_1, \tau_2 > 0$,*

$$\begin{aligned} \|y - \tilde{x}\|^2 &\leq (1 - \alpha\sigma + 8\alpha^2 L^2) \|x - \tilde{x}\|^2 - 2\alpha \langle \nabla \phi(\tilde{x}), z - \tilde{x} \rangle - \frac{3}{4} \|y - x\|^2 \\ &\quad + (8 + \tau_2) \alpha^2 \|\nabla \phi(\tilde{x})\|^2 + \tau_1 \alpha^2 L^2 \|z - \tilde{x}\|^2 + \left(\frac{1}{\tau_1} + \frac{1}{\tau_2} \right) \|x - z\|^2. \end{aligned}$$

Proof We start with the proof for part (a). From the relation of x and y and the strictly non-expansive projection property in Lemma 2.1(b), we obtain for any $\tilde{x} \in \mathcal{Y}$,

$$\begin{aligned} \|y - \tilde{x}\|^2 &\leq \|x - \tilde{x}\|^2 - 2\alpha \langle \nabla \phi(x), x - \tilde{x} \rangle \\ &\quad - \|y - x\|^2 + 2\alpha \langle \nabla \phi(x), x - y \rangle. \end{aligned} \tag{4.5}$$

We next estimate the term $2\alpha \langle \nabla \phi(x), x - y \rangle$. By using Cauchy-Swartz inequality we obtain $2\alpha \langle \nabla \phi(x), x - y \rangle \leq 2\alpha \|\nabla \phi(x)\| \|x - y\|$. By writing

$2\alpha\|\nabla\phi(x)\|\|x-y\| = 2(2\alpha\|\nabla\phi(x)\|)(\|x-y\|/2)$, we find that

$$2\alpha\langle\nabla\phi(x), x-y\rangle \leq 4\alpha^2\|\nabla\phi(x)\|^2 + \frac{1}{4}\|x-y\|^2. \quad (4.6)$$

Furthermore, we have $\|\nabla\phi(x)\|^2 \leq \|(\nabla\phi(x) - \nabla\phi(\tilde{x})) + \nabla\phi(\tilde{x})\|^2$, which by the square property $(a+b)^2 \leq 2(a^2+b^2)$ yields $\|\nabla\phi(x)\|^2 \leq 2\|\nabla\phi(x) - \nabla\phi(\tilde{x})\|^2 + 2\|\nabla\phi(\tilde{x})\|^2$. The preceding relation and the Lipschitz gradient property of ϕ imply

$$\|\nabla\phi(x)\|^2 \leq 2L\|x - \tilde{x}\|^2 + 2\|\nabla\phi(\tilde{x})\|^2. \quad (4.7)$$

Therefore, from (4.5)–(4.7) we obtain

$$\begin{aligned} \|y - \tilde{x}\|^2 &\leq (1 + 8\alpha^2L^2)\|x - \tilde{x}\|^2 - 2\alpha\langle\nabla\phi(x), x - \tilde{x}\rangle \\ &\quad - \frac{3}{4}\|y - x\|^2 + 8\alpha^2\|\nabla\phi(\tilde{x})\|^2. \end{aligned} \quad (4.8)$$

Next, we estimate the term $2\alpha\langle\nabla\phi(x), x - \tilde{x}\rangle$ using the convexity of ϕ ,

$$\langle\nabla\phi(x), x - \tilde{x}\rangle \geq \phi(x) - \phi(\tilde{x}) = (\phi(x) - \phi(z)) + (\phi(z) - \phi(\tilde{x})), \quad (4.9)$$

where $z \in \mathbb{R}^d$ is some given point. It remains to bound the term $\phi(x) - \phi(z)$, for which by convexity of ϕ we further have

$$\phi(x) - \phi(z) \geq \langle\nabla\phi(z), x - z\rangle \geq -\|\nabla\phi(z)\|\|x - z\|.$$

By writing $\|\nabla\phi(z)\| \leq \|\nabla\phi(z) - \nabla\phi(\tilde{x})\| + \|\nabla\phi(\tilde{x})\|$ and using the Lipschitz-gradient property of ϕ , we obtain

$$\phi(x) - \phi(z) \geq -L\|z - \tilde{x}\|\|x - z\| - \|\nabla\phi(\tilde{x})\|\|x - z\|.$$

Multiplying the preceding relation with 2α and using $2\alpha L\|z - \tilde{x}\|\|x - z\| = 2(\alpha\sqrt{\tau_1}L\|z - \tilde{x}\|)(\|x - z\|/\sqrt{\tau_1}) \leq \tau_1\alpha^2L^2\|z - \tilde{x}\|^2 + \|x - z\|^2/\tau_1$, $2\alpha\|\nabla\phi(\tilde{x})\|\|x - z\| = 2(\alpha\sqrt{\tau_2}\|\nabla\phi(\tilde{x})\|)(\|x - z\|/\sqrt{\tau_2}) \leq \tau_2\alpha^2\|\nabla\phi(\tilde{x})\|^2 + \|x - z\|^2/\tau_2$ for some $\tau_1, \tau_2 > 0$, we obtain

$$2\alpha(\phi(x) - \phi(z)) \geq -\tau_1\alpha^2L^2\|z - \tilde{x}\|^2 - \tau_2\alpha^2\|\nabla\phi(\tilde{x})\|^2 - \left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right)\|x - z\|^2. \quad (4.10)$$

Thus, from Eqs. (4.8)–(4.10) it follows that

$$\begin{aligned} \|y - \check{x}\|^2 &\leq (1 + 8\alpha^2 L^2) \|x - \check{x}\|^2 - 2\alpha (\phi(z) - \phi(\check{x})) - \frac{3}{4} \|y - x\|^2 \\ &\quad + (8 + \tau_2) \alpha^2 \|\nabla \phi(\check{x})\|^2 + \tau_1 \alpha^2 L^2 \|z - \check{x}\|^2 + \left(\frac{1}{\tau_1} + \frac{1}{\tau_2} \right) \|x - z\|^2, \end{aligned} \quad (4.11)$$

thus proving the relation in part (a).

The relation in part (b) follows similarly by using the strong convexity of ϕ in Eq. (4.9), according to which we have $\langle \nabla \phi(x), x - \check{x} \rangle \geq \phi(x) - \phi(\check{x}) + \frac{\sigma}{2} \|x - \check{x}\|^2$. ■

The proof of Lemma 4.3 relies on Lemma 4.1(a) and the fact that the event $E_i(k) = \{i \in \{I_k, J_k\}\}$ that agent i updates at any time is independent of the past. Let γ_i be the probability of the event $E_i(k)$. Then, $\gamma_i = \frac{1}{m} + \frac{1}{m} \sum_{j \in \mathcal{N}(i)} \pi_{ji}$ for all $i \in V$, where $\pi_{ji} > 0$ is the probability that agent i is chosen by its neighbor j to communicate.

The long term estimates for the stepsize $\alpha_i(k) = \frac{1}{\Gamma_i(k)}$ in terms of the probability γ_i that agent i updates are given in the following lemma.

Lemma 4.2 (see [42]) *Let $\alpha_i(k) = 1/\Gamma_i(k)$ for all $k \geq 1$ and $i \in V$. Let $\pi_{\min} = \min_{\{i,j\} \in E} \pi_{ij}$. Also, let q be a constant such that $0 < q < \frac{1}{2}$. Then, there exists a large enough \tilde{k} (which depends on q and m) such that with probability 1 for all $k \geq \tilde{k}$ and $i \in V$,*

$$\begin{aligned} (a) \quad \alpha_i(k) &\leq \frac{2}{k\gamma_i}, \quad (b) \quad \alpha_i^2(k) \leq \frac{4m^2}{k^2(1 + \pi_{\min})^2}, \\ (c) \quad \left| \alpha_i(k) - \frac{1}{k\gamma_i} \right| &\leq \frac{2}{k^{\frac{3}{2}-q}(1 + \pi_{\min})^2}. \end{aligned}$$

According to this lemma, the stepsizes $\alpha_i(k)$ exhibit the same behavior as the deterministic stepsize $1/k$ in a long run. The result is critical for dealing with the cross dependencies of the random stepsizes and the other randomness in the GRP method.

Finally, we prove Lemma 4.3 using Lemma 4.1(a) and Lemma 4.2. For this, we define the history of the algorithm as follows. Let \mathcal{F}_k be the σ -algebra generated by the entire history of the algorithm up to time k inclusively; i.e., for all $k \geq 1$,

$$\mathcal{F}_k = \{x_i(0); i \in V\} \cup \{I_\ell, J_\ell, \Omega_i(\ell); i \in \{I_\ell, J_\ell\}, 1 \leq \ell \leq k\},$$

and $\mathcal{F}_0 = \{x_i(0); i \in V\}$.

The following relations will be frequently used in the analysis. By the definition of $v_i(k)$ in method (4.3a), the convexity of the norm square function and the doubly stochasticity of the weights, we have

$$\begin{aligned} \sum_{i=1}^m \mathbb{E}[\|v_i(k) - x^*\|^2 \mid \mathcal{F}_{k-1}] &\leq \sum_{i=1}^m \sum_{j=1}^m \bar{W}_{ij} \|x_j(k-1) - x^*\|^2 \\ &= \sum_{j=1}^m \|x_j(k-1) - x^*\|^2. \end{aligned} \quad (4.12)$$

Also, by the convexity of the distance function $x \mapsto \text{dist}^2(x, \mathcal{X})$ (see [30, p. 88]), we have

$$\begin{aligned} \sum_{i=1}^m \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] &\leq \sum_{i=1}^m \sum_{j=1}^m \bar{W}_{ij} \text{dist}^2(x_j(k-1), \mathcal{X}) \\ &= \sum_{j=1}^m \text{dist}^2(x_j(k-1), \mathcal{X}). \end{aligned} \quad (4.13)$$

Lemma 4.3 [*Basic Iterate Relation*] *Let Assumptions 4.2-4.3 hold. Let $\{x_i(k)\}$ be the iterates generated by the algorithm (4.3a)-(4.3c). Then, for any $q \in (0, 1/2)$ there is a sufficiently large \hat{k} , such that with probability 1, for all $\check{x} \in \mathcal{X}$, $k \geq \hat{k}$ and $i \in V$,*

$$\begin{aligned} \mathbb{E}[\|x_i(k) - \check{x}\|^2 \mid \mathcal{F}_{k-1}] &\leq \left(1 + \frac{a_1}{k^2}\right) \mathbb{E}[\|v_i(k) - \check{x}\|^2 \mid \mathcal{F}_{k-1}] \\ &\quad - \frac{2}{k} \mathbb{E}[f_i(z_i(k)) - f_i(\check{x}) \mid \mathcal{F}_{k-1}] - \frac{\gamma_i}{4c} \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] \\ &\quad + \frac{a_2}{k^2} + \frac{a_4}{k^{\frac{3}{2}-q}} + \left(\frac{a_3}{k^2} + \frac{a_5}{k^{\frac{3}{2}-q}}\right) \mathbb{E}[\|z_i(k) - \check{x}\|^2 \mid \mathcal{F}_{k-1}], \end{aligned}$$

where $z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$, $a_j > 0$ are some constants, c is the scalar from Assumption 4.3 and γ_i is the probability that agent i updates.

Proof First, we fix $i \in \{I_k, J_k\}$ and use Lemma 4.1(a) with the following identification: $\mathcal{Y} = \mathcal{X}_i^{\Omega_i(k)}$ and $\check{x} \in \mathcal{X}$, $y = x_i(k)$, $x = v_i(k)$, $z = z_i(k) \triangleq$

$\Pi_{\mathcal{X}}[v_i(k)]$, $\phi = f_i$, and $\alpha = \alpha_i(k)$. Then, for any $\tilde{x} \in \mathcal{X}$, $k \geq 1$ and $i \in \{I_k, J_k\}$

$$\begin{aligned} \|x_i(k) - \tilde{x}\|^2 &\leq (1 + 8\alpha_i^2(k)L_i^2)\|v_i(k) - \tilde{x}\|^2 - 2\alpha_i(k)(f_i(z_i(k)) - f_i(\tilde{x})) \\ &\quad - \frac{3}{4}\|x_i(k) - v_i(k)\|^2 + (8 + \tau_2)\alpha_i^2(k)\|\nabla f_i(\tilde{x})\|^2 \\ &\quad + \tau_1\alpha_i^2(k)L_i^2\|z_i(k) - \tilde{x}\|^2 + \left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right)\|v_i(k) - z_i(k)\|^2. \end{aligned}$$

By Assumption 4.2(d), we have $\|\nabla f_i(\tilde{x})\| \leq G_f$. Further, we let $\tau_1 = \tau_2 = 4\eta$ for some $\eta > 0$, and by using Lemma 4.2(b) we find that w.p.1 for k large enough

$$\begin{aligned} \|x_i(k) - \tilde{x}\|^2 &\leq \left(1 + \frac{a_1}{k^2}\right)\|v_i(k) - \tilde{x}\|^2 - 2\alpha_i(k)(f_i(z_i(k)) - f_i(\tilde{x})) \\ &\quad - \frac{3}{4}\|x_i(k) - v_i(k)\|^2 + \frac{a_2}{k^2} + \frac{a_3}{k^2}\|z_i(k) - \tilde{x}\|^2 + \frac{1}{2\eta}\|v_i(k) - z_i(k)\|^2, \end{aligned} \quad (4.14)$$

where $a_1 = \frac{32m^2\bar{L}^2}{(1+\pi_{\min})^2}$, $a_2 = \frac{4(8+4\eta)m^2G_f^2}{(1+\pi_{\min})^2}$ and $a_3 = \frac{16\eta m^2\bar{L}^2}{(1+\pi_{\min})^2}$. We next estimate $2\alpha_i(k)(f_i(z_i(k)) - f_i(\tilde{x}))$, for which we can write

$$\begin{aligned} 2\alpha_i(k)(f_i(z_i(k)) - f_i(\tilde{x})) &\geq \frac{2}{k\gamma_i}(f_i(z_i(k)) - f_i(\tilde{x})) \\ &\quad - 2\left|\alpha_i(k) - \frac{1}{k\gamma_i}\right||f_i(z_i(k)) - f_i(\tilde{x})|. \end{aligned}$$

Since f_i has bounded gradients over the set \mathcal{X} , it is Lipschitz continuous over \mathcal{X} . Thus, since $z_i(k), \tilde{x} \in \mathcal{X}$, it follows that $|f_i(z_i(k)) - f_i(\tilde{x})| \geq G_f\|z_i(k) - \tilde{x}\|$. This and Lemma 4.2(c) imply

$$\begin{aligned} &2\alpha_i(k)(f_i(z_i(k)) - f_i(\tilde{x})) \\ &\geq \frac{2}{k\gamma_i}(f_i(z_i(k)) - f_i(\tilde{x})) - 2\frac{2}{k^{\frac{3}{2}-q}(1+\pi_{\min})^2}G_f\|z_i(k) - \tilde{x}\| \\ &\geq \frac{2}{k\gamma_i}(f_i(z_i(k)) - f_i(\tilde{x})) - \frac{2}{k^{\frac{3}{2}-q}(1+\pi_{\min})^2}(G_f^2 + \|z_i(k) - \tilde{x}\|^2), \end{aligned}$$

where the last inequality follows by the Cauchy-Schwarz inequality. Combining the preceding relation with Eq. (4.14), we obtain w.p.1 for k large

enough

$$\begin{aligned} \|x_i(k) - \tilde{x}\|^2 &\leq \left(1 + \frac{a_1}{k^2}\right) \|v_i(k) - \tilde{x}\|^2 - \frac{2}{\gamma_i k} (f_i(z_i(k)) - f_i(\tilde{x})) \\ &\quad - \frac{3}{4} \|x_i(k) - v_i(k)\|^2 + \frac{a_2}{k^2} + \frac{a_4}{k^{\frac{3}{2}-q}} \\ &\quad + \left(\frac{a_3}{k^2} + \frac{a_5}{k^{\frac{3}{2}-q}}\right) \|z_i(k) - \tilde{x}\|^2 + \frac{1}{2\eta} \|v_i(k) - z_i(k)\|^2, \end{aligned} \quad (4.15)$$

where $a_4 = \frac{2}{(1+\pi_{\min})^2} G_f^2$ and $a_5 = \frac{2}{(1+\pi_{\min})^2}$.

By the definition of the projection, we have $\|v_i(k) - z_i(k)\| = \text{dist}(v_i(k), \mathcal{X})$ and

$$\|x_i(k) - v_i(k)\| \geq \left\| \Pi_{\mathcal{X}_i^{\Omega_i(k)}}[v_i(k)] - v_i(k) \right\| = \text{dist}(v_i(k), \mathcal{X}_i^{\Omega_i(k)}).$$

Taking the expectation in (4.15) conditioned on \mathcal{F}_{k-1}, I_k and J_k jointly we obtain for any $\tilde{x} \in \mathcal{X}$, $i \in \{I_k, J_k\}$ w.p.1 for all k large enough

$$\begin{aligned} \mathbb{E}[\|x_i(k) - \tilde{x}\|^2 \mid \mathcal{F}_{k-1}, I_k, J_k] &\leq \left(1 + \frac{a_1}{k^2}\right) \|v_i(k) - \tilde{x}\|^2 \\ &\quad - \frac{2}{\gamma_i k} (f_i(z_i(k)) - f_i(\tilde{x})) - \frac{3}{4} \mathbb{E} \left[\text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \mid v_i(k) \right] \\ &\quad + \frac{a_2}{k^2} + \frac{a_4}{k^{\frac{3}{2}-q}} + \left(\frac{a_3}{k^2} + \frac{a_5}{k^{\frac{3}{2}-q}}\right) \|z_i(k) - \tilde{x}\|^2 + \frac{1}{2\eta} \text{dist}^2(v_i(k), \mathcal{X}). \end{aligned}$$

Using Assumption 4.3, we have

$$\mathbb{E} \left[\text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \mid v_i(k) \right] \geq \frac{1}{c} \text{dist}^2(v_i(k), \mathcal{X}).$$

Thus, by letting $\eta = c$, from the preceding two relations we have w.p.1 for all k large enough

$$\begin{aligned} \mathbb{E}[\|x_i(k) - \tilde{x}\|^2 \mid \mathcal{F}_{k-1}, I_k, J_k] &\leq \left(1 + \frac{a_1}{k^2}\right) \|v_i(k) - \tilde{x}\|^2 \\ &\quad - \frac{2}{\gamma_i k} (f_i(z_i(k)) - f_i(\tilde{x})) - \frac{1}{4c} \text{dist}^2(v_i(k), \mathcal{X}) \\ &\quad + \frac{a_2}{k^2} + \frac{a_4}{k^{\frac{3}{2}-q}} + \left(\frac{a_3}{k^2} + \frac{a_5}{k^{\frac{3}{2}-q}}\right) \|z_i(k) - \tilde{x}\|^2. \end{aligned}$$

Now we use the fact that the preceding inequality holds with probability γ_i (when agent i updates), and $x_i(k) = v_i(k)$ with probability $1 - \gamma_i$ (when agent i does not update), and we obtain w.p.1 for any $\tilde{x} \in \mathcal{X}$, all $i \in V$, and

all k large enough

$$\begin{aligned} \mathbb{E}[\|x_i(k) - \tilde{x}\|^2 \mid \mathcal{F}_{k-1}] &\leq \left(1 + \frac{\gamma_i a_1}{k^2}\right) \mathbb{E}[\|v_i(k) - \tilde{x}\|^2 \mid \mathcal{F}_{k-1}] \\ &\quad - \frac{2}{k} \mathbb{E}[f_i(z_i(k)) - f_i(\tilde{x}) \mid \mathcal{F}_{k-1}] - \frac{\gamma_i}{4c} \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] \\ &\quad + \frac{\gamma_i a_2}{k^2} + \frac{\gamma_i a_4}{k^{\frac{3}{2}-q}} + \gamma_i \left(\frac{a_3}{k^2} + \frac{a_5}{k^{\frac{3}{2}-q}}\right) \mathbb{E}[\|z_i(k) - \tilde{x}\|^2 \mid \mathcal{F}_{k-1}]. \end{aligned}$$

Since $\gamma_i \leq 1$, the relation of Lemma 4.3 follows. \blacksquare

4.3.3 Projection Error Estimate

In the next lemma, we show that the distance between the estimates $\{v_i(k)\}$, $i \in V$, and the constraint set \mathcal{X} goes to zero as $k \rightarrow \infty$ with probability 1. We also show that the error sequence $e_i(k) = x_i(k) - v_i(k)$, $i \in V$, converges to zero with probability 1.

Lemma 4.4 [*Projection Error*] *Let Assumptions 4.2-4.3 hold. Then, with probability 1, we have*

- (a) $\sum_{k=1}^{\infty} \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] < \infty$ and $\lim_{k \rightarrow \infty} \text{dist}(v_i(k), \mathcal{X}) = 0$ for all $i \in V$.
- (b) $\sum_{k=1}^{\infty} \mathbb{E}[\|e_i(k)\|^2 \mid \mathcal{F}_{k-1}] < \infty$ and $\lim_{k \rightarrow \infty} \|e_i(k)\| = 0$ for all $i \in V$, where $e_i(k) = x_i(k) - v_i(k)$ for all $i \in V$ and $k \geq 1$.

Proof For the proof of part (a) we start with Lemma 4.3, where we let $\tilde{x} = z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$. Then, for all k large enough and all $i \in V$, we obtain w.p.1,

$$\begin{aligned} \mathbb{E}[\|x_i(k) - \Pi_{\mathcal{X}}[v_i(k)]\|^2 \mid \mathcal{F}_{k-1}] &\leq \left(1 + \frac{a_1}{k^2}\right) \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] \\ &\quad - \frac{\gamma_i}{4c} \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] + \frac{a_2}{k^2} + \frac{a_4}{k^{\frac{3}{2}-q}}. \end{aligned}$$

By the definition of the projection, we have $\text{dist}(x_i(k), \mathcal{X}) \leq \|x_i(k) - \Pi_{\mathcal{X}}[v_i(k)]\|$. Using this and relation (4.13), we have w.p.1 for all k large enough

and all $i \in V$,

$$\begin{aligned} \mathbb{E}[\text{dist}^2(x_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] &\leq \left(1 + \frac{a_1}{k^2}\right) \sum_{j=1}^m \bar{W}_{ij} \text{dist}^2(x_j(k-1), \mathcal{X}) \\ &\quad - \frac{\gamma_i}{4c} \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] + \frac{a_2}{k^2} + \frac{a_4}{k^{\frac{3}{2}-q}}. \end{aligned}$$

By summing over all i and using the fact that each $W(k)$ has column sums equal to 1, we conclude that, after taking the total expectation, for all k large enough,

$$\begin{aligned} \sum_{i=1}^m \mathbb{E}[\text{dist}^2(x_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] &\leq \left(1 + \frac{a_1}{k^2}\right) \sum_{j=1}^m \text{dist}^2(x_j(k-1), \mathcal{X}) \\ &\quad - \frac{\underline{\gamma}}{4c} \sum_{i=1}^m \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] + \frac{a_2 m}{k^2} + \frac{a_4 m}{k^{\frac{3}{2}-q}}, \end{aligned} \quad (4.16)$$

where $\underline{\gamma} = \min_i \gamma_i$. Therefore, for all k large enough, all the conditions of Lemma 2.1 are satisfied (for a time-delayed sequence), so we conclude that $\sum_{k=1}^{\infty} \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] < \infty$ for all $i \in V$. Taking the total expectation in relation (4.16), it also follows that $\sum_{k=1}^{\infty} \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X})] < \infty$ for all $i \in V$, which by the Monotone Convergence Theorem [43] implies $\lim_{k \rightarrow \infty} \text{dist}(v_i(k), \mathcal{X}) = 0$ for all i w.p.1, showing the result in part (a).

For part (b), note that for $\|e_i(k)\|$, using $z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$, we can write for $i \in \{I_k, J_k\}$,

$$\begin{aligned} \|e_i(k)\| &\leq \|x_i(k) - z_i(k)\| + \|z_i(k) - v_i(k)\| \\ &= \left\| \Pi_{\mathcal{X}_i^{\Omega_i(k)}}[v_i(k) - \alpha_i(k) \nabla f_i(v_i(k))] - z_i(k) \right\| + \|z_i(k) - v_i(k)\|. \end{aligned}$$

Since $\mathcal{X} \subseteq \mathcal{X}_i^{\Omega_i(k)}$ and $z_i(k) \in \mathcal{X}$, we have $z_i(k) \in \mathcal{X}_i^{\Omega_i(k)}$. Using the projection non expansiveness property of Lemma 2.1(a), we obtain

$$\begin{aligned} \|e_i(k)\| &\leq \|v_i(k) - \alpha_i(k) \nabla f_i(v_i(k)) - z_i(k)\| + \|z_i(k) - v_i(k)\| \\ &\leq 2\|v_i(k) - z_i(k)\| + \alpha_i(k) \|\nabla f_i(v_i(k))\| \\ &\leq 2\|v_i(k) - z_i(k)\| + \alpha_i(k) (\|\nabla f_i(v_i(k)) - \nabla f_i(z_i(k))\| + \|\nabla f_i(z_i(k))\|) \\ &\leq (2 + \alpha_i(1)L_i)\|v_i(k) - z_i(k)\| + \alpha_i(k)G_f, \end{aligned} \quad (4.17)$$

where the last inequality follows by using $\alpha_i(k) \leq \alpha_i(1)$, the Lipschitz

gradient property of f_i and the gradient boundedness property (Assumptions 4.2(c) and 4.2(d)). Using the Cauchy-Schwartz inequality and Lemma 4.2(a) (i.e. $\alpha_i(k) \leq 2/(k\gamma_i)$), we have for all $i \in \{I_k, J_k\}$ and $k \geq \tilde{k}$,

$$\|e_i(k)\|^2 \leq 2(2 + \alpha_i(1)L_i)^2 \|v_i(k) - z_i(k)\|^2 + \frac{8m^2}{k^2} G_f^2, \quad (4.18)$$

where we also use $\gamma_i \geq \frac{1}{m}$. Taking the expectation in (4.18) conditioned on $\mathcal{F}_{k-1}, I_k, J_k$ and noting that the preceding inequality holds with probability γ_i , and $x_i(k) = v_i(k)$ with probability $1 - \gamma_i$, we obtain with probability 1 for all $k \geq \tilde{k}$ and $i \in V$,

$$\mathbb{E}[\|e_i(k)\|^2 \mid \mathcal{F}_{k-1}] \leq 2\gamma_i(2 + \alpha_i(1)L_i)^2 \mathbb{E}[\|v_i(k) - z_i(k)\|^2 \mid \mathcal{F}_{k-1}] + \frac{8\gamma_i m^2}{k^2} G_f^2.$$

Since $z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$, it follows that $\|v_i(k) - z_i(k)\| = \text{dist}(v_i(k), \mathcal{X})$. By part (a) of this lemma, we have $\sum_{k=1}^{\infty} \mathbb{E}[\|v_i(k) - z_i(k)\|^2 \mid \mathcal{F}_{k-1}] < \infty$ w.p.1 for all i . As $\sum_{k=1}^{\infty} \frac{1}{k^2} < \infty$, we conclude that $\sum_{k=1}^{\infty} \mathbb{E}[\|e_i(k)\|^2 \mid \mathcal{F}_{k-1}] < \infty$ for all $i \in V$ w.p.1. Furthermore, by relation (4.18) and part (a) of the lemma we find that $\lim_{k \rightarrow \infty} \|e_i(k)\| = 0$ for all i w.p.1. ■

Lemma 4.4(a) together with Lemma 4.4(b) ensures that $\lim_{k \rightarrow \infty} \text{dist}^2(x_i(k), \mathcal{X}) = 0$ with probability 1 for all $i \in V$.

4.3.4 Disagreement Estimate

We provide a relation for the agent disagreements on the vectors $v_i(k)$ in Lemma 4.6. The proof of this Lemma makes use of an additional result given below.

Lemma 4.5 *Let $\{W(k)\}$ be an i.i.d. sequence of $m \times m$ symmetric and stochastic matrices. Consider a sequence $\{\theta(k)\} \subset \mathbb{R}^m$ generated by the following dynamics:*

$$\theta(k) = W(k)\theta(k-1) + \epsilon(k) \quad \text{for } k \geq 1. \quad (4.19)$$

Then, we have with probability 1 for all $k \geq 1$,

$$\mathbb{E}[\|\Delta(k)\| \mid \mathcal{F}_{k-1}] \leq \sqrt{\lambda} \|\Delta(k-1)\| + \mathbb{E}[\|\epsilon(k)\| \mid \mathcal{F}_{k-1}],$$

where $\Delta(k) \triangleq \theta(k) - \frac{1}{m} \mathbf{1} \mathbf{1}^T \theta(k)$ and $\lambda < 1$ is the second largest eigenvalue of $\bar{W} = \mathbb{E}[W(k)]$.

Proof Define the sequences of averaged coordinate values as $\theta_k^{\text{ave}} \triangleq \frac{1}{m} \sum_{i=1}^m \theta_i(k)$ and $\epsilon_k^{\text{ave}} \triangleq \frac{1}{m} \sum_{i=1}^m \epsilon_i(k)$. From relation (4.19), by taking averages over the coordinates, we have $\theta_k^{\text{ave}} = \theta_{k-1}^{\text{ave}} + \epsilon_k^{\text{ave}}$. Using $\mathbf{1} \in \mathbb{R}^m$, a vector with all its elements 1, we can write $\theta_k^{\text{ave}} \mathbf{1} = \theta_{k-1}^{\text{ave}} \mathbf{1} + \epsilon_k^{\text{ave}} \mathbf{1}$, or equivalently,

$$\frac{1}{m} \mathbf{1} \mathbf{1}^T \theta(k) = \frac{1}{m} \mathbf{1} \mathbf{1}^T \theta(k-1) + \frac{1}{m} \mathbf{1} \mathbf{1}^T \epsilon(k). \quad (4.20)$$

From equations (4.19) and (4.20), we have:

$$\theta(k) - \frac{1}{m} \mathbf{1} \mathbf{1}^T \theta(k) = \left(W(k) - \frac{1}{m} \mathbf{1} \mathbf{1}^T \right) \theta(k-1) + \left(I - \frac{1}{m} \mathbf{1} \mathbf{1}^T \right) \epsilon(k).$$

Since $(W(k) - \frac{1}{m} \mathbf{1} \mathbf{1}^T) \frac{1}{m} \mathbf{1} \mathbf{1}^T \theta(k-1) = (W(k) - \frac{1}{m} \mathbf{1} \mathbf{1}^T) \theta_{k-1}^{\text{ave}} \mathbf{1} = 0$, it follows that

$$\begin{aligned} \theta(k) - \frac{1}{m} \mathbf{1} \mathbf{1}^T \theta(k) &= \left(W(k) - \frac{1}{m} \mathbf{1} \mathbf{1}^T \right) \left(\theta(k-1) - \frac{1}{m} \mathbf{1} \mathbf{1}^T \theta(k-1) \right) \\ &\quad + \left(I - \frac{1}{m} \mathbf{1} \mathbf{1}^T \right) \epsilon(k). \end{aligned}$$

Let $\Delta(k) \triangleq \theta(k) - \frac{1}{m} \mathbf{1} \mathbf{1}^T \theta(k)$, $D_k \triangleq W(k) - \frac{1}{m} \mathbf{1} \mathbf{1}^T$ and $M \triangleq I - \frac{1}{m} \mathbf{1} \mathbf{1}^T$. Then, for all $k \geq 1$,

$$\Delta(k) = D_k \Delta(k-1) + M \epsilon(k).$$

By taking the norm and the expectation conditioned on the past history, from the preceding relation we have w.p.1 for $k \geq 1$,

$$\mathbb{E}[\|\Delta(k)\| \mid \mathcal{F}_{k-1}] = \mathbb{E}[\|D_k \Delta(k-1)\| \mid \mathcal{F}_{k-1}] + \mathbb{E}[\|M \epsilon(k)\| \mid \mathcal{F}_{k-1}]. \quad (4.21)$$

From Eq. (4.4) and the fact that $W(k)$ is independent of the past \mathcal{F}_{k-1} , we obtain $\mathbb{E}[\|D_k \Delta(k-1)\|^2 \mid \mathcal{F}_{k-1}] \leq \lambda \|\Delta(k-1)\|^2$, where λ is the second largest eigenvalue of the matrix \bar{W} . Using $\mathbb{E}[\|x\|] \leq \sqrt{\mathbb{E}[\|x\|^2]}$, we obtain for all $k \geq 1$, $\mathbb{E}[\|D_k \Delta(k-1)\| \mid \mathcal{F}_{k-1}] \leq \sqrt{\lambda} \|\Delta(k-1)\|$. For the second term on the right-hand side of (4.21), we have $\mathbb{E}[\|M \epsilon(k)\| \mid \mathcal{F}_{k-1}] = \mathbb{E}[\|\epsilon(k)\| \mid \mathcal{F}_{k-1}]$, where the equality is from the fact that $M = I - \frac{1}{m} \mathbf{1} \mathbf{1}^T$ is a projection matrix

and thus $\|M\| = 1$. Upon substituting the above two relations in (4.21), we obtain the desired result. ■

Lemma 4.6 *[Disagreement] Let Assumptions 4.1-4.2 hold. Let $\{v_i(k)\}$ be generated by the method (4.3a)-(4.3c) with $\alpha_i(k) = 1/\Gamma_i(k)$ and $\Gamma_i(k)$ being the number of updates that agent i has performed until time k . Then, for $\bar{v}(k) = \frac{1}{m} \sum_{i=1}^m v_i(k)$ we have $\sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}[\|v_i(k) - \bar{v}(k)\| \mid \mathcal{F}_{k-1}] < \infty$ with probability 1 for all $i \in V$.*

Proof We consider coordinate-wise relations by defining the vector $y_\ell(k) \in \mathbb{R}^m$ for $\ell = 1, \dots, d$ such that $[y_\ell(k)]_i = [x_i(k)]_\ell$ for all i . From algorithm (4.3a)-(4.3c), we have

$$y_\ell(k) = W(k)y_\ell(k-1) + \delta_\ell(k) \quad \text{for } k \geq 1,$$

where $\delta_\ell(k) \in \mathbb{R}^m$ is a vector defined as

$$[\delta_\ell(k)]_i = \begin{cases} \left[\Pi_{\mathcal{X}_i^{\Omega_i(k)}}[v_i(k) - \alpha_i(k)\nabla f(v_i(k))] - v_i(k) \right]_\ell & \text{if } i \in \{I_k, J_k\}, \\ 0 & \text{otherwise.} \end{cases} \quad (4.22)$$

Since the matrices $W(k)$ are doubly stochastic for all $k \geq 1$, from Lemma 4.5 we obtain

$$\mathbb{E}[\|y_\ell(k) - [\bar{x}(k)]_\ell \mathbf{1}\| \mid \mathcal{F}_{k-1}] \leq \sqrt{\lambda} \|y_\ell(k-1) - [\bar{x}(k-1)]_\ell \mathbf{1}\| + \mathbb{E}[\|\delta_\ell(k)\| \mid \mathcal{F}_{k-1}], \quad (4.23)$$

where $[\bar{x}(k)]_\ell = \frac{1}{m} \mathbf{1}^T y_\ell(k)$ and $\lambda < 1$ by Assumption 4.1.

We next consider $\delta_\ell(k)$ as given by (4.22), for which we have for all $k \geq 1$,

$$\|\delta_\ell(k)\|^2 \leq \sum_{i=1}^m \left\| \Pi_{\mathcal{X}_i^{\Omega_i(k)}}[v_i(k) - \alpha_i(k)(\nabla f_i(v_i(k)))] - v_i(k) \right\|^2.$$

Letting $z_i(k) \triangleq \Pi_{\mathcal{X}}[v_i(k)]$, observing that $z_i(k) \in \mathcal{X}_i^{\Omega_i(k)}$ and using the projection property in Lemma 2.1(a), we obtain

$$\begin{aligned} \|\delta_\ell(k)\|^2 &\leq \sum_{i=1}^m \left(\left\| \Pi_{\mathcal{X}_i^{\Omega_i(k)}}[v_i(k) - \alpha_i(k)\nabla f_i(v_i(k))] - z_i(k) \right\| + \|z_i(k) - v_i(k)\| \right)^2 \\ &\leq \sum_{i=1}^m (\alpha_i(k) \|\nabla f_i(v_i(k))\| + 2\|z_i(k) - v_i(k)\|)^2. \end{aligned}$$

Applying the Cauchy-Schwartz inequality, we can obtain

$$\|\delta_\ell(k)\|^2 \leq \sum_{i=1}^m (2\alpha_i^2(k) \|\nabla f_i(v_i(k))\|^2 + 4\|z_i(k) - v_i(k)\|^2).$$

The term $\|\nabla f_i(v_i(k))\|^2$ can be further evaluated by using the Lipschitz property and the bounded gradient assumption (Assumption 4.2(d)),

$$\begin{aligned} \|\nabla f_i(v_i(k))\|^2 &\leq 2\|\nabla f_i(v_i(k)) - \nabla f_i(z_i(k))\|^2 + 2\|\nabla f_i(z_i(k))\|^2 \\ &\leq 2L^2\|v_i(k) - z_i(k)\|^2 + 2G_f^2. \end{aligned}$$

From Lemma 4.2(b), there exists a large enough \tilde{k} such that $\alpha_i^2(k) \leq 4m^2/k^2 \leq 4m^2/\tilde{k}^2$ w.p.1 for all $k \geq \tilde{k}$. Therefore, noting that $\|z_i(k) - v_i(k)\| = \text{dist}(v_i(k), \mathcal{X})$, we obtain for all $k \geq \tilde{k}$ with probability 1,

$$\|\delta_\ell(k)\|^2 \leq \left(4 + \frac{16m^2}{\tilde{k}^2}L^2\right) \sum_{i=1}^m \text{dist}^2(v_i(k), \mathcal{X}) + \frac{16m^2}{k^2}G_f^2.$$

Taking the expectation conditioned on \mathcal{F}_{k-1} and using $\mathbf{E}[\|x\|] \leq \sqrt{\mathbf{E}[\|x\|^2]}$, we obtain

$$\mathbf{E}[\|\delta_\ell(k)\| \mid \mathcal{F}_{k-1}] \leq b_k, \quad (4.24)$$

where

$$b_k = \sqrt{\left(4 + \frac{16m^2}{\tilde{k}^2}L^2\right) \sum_{i=1}^m \mathbf{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] + \frac{16m^2}{k^2}G_f^2}.$$

By relations (4.23) and (4.24), we obtain for all $k \geq \tilde{k}$ with probability 1,

$$\begin{aligned} \frac{1}{k} \mathbf{E}[\|y_\ell(k) - [\bar{x}(k)]_\ell \mathbf{1}\| \mid \mathcal{F}_{k-1}] &\leq \frac{1}{k-1} \|y_\ell(k-1) - [\bar{x}(k-1)]_\ell \mathbf{1}\| \\ &\quad - \frac{1 - \sqrt{\lambda}}{k} \|y_\ell(k-1) - [\bar{x}(k-1)]_\ell \mathbf{1}\| + \frac{1}{k} b_k. \end{aligned} \quad (4.25)$$

Noting that $\frac{1}{k}b_k \leq (1/k^2 + b_k^2)/2$, and that $\sum_{k=1}^\infty b_k^2 < \infty$ by Lemma 4.4(a), the term $\frac{1}{k}b_k$ is summable. From this and the fact that $1 - \sqrt{\lambda} > 0$, relation (4.25) satisfies all the conditions in Lemma 2.1. It follows that $\sum_{k=1}^\infty \frac{1}{k} \mathbf{E}[\|y_\ell(k) - [\bar{x}(k)]_\ell \mathbf{1}\| \mid \mathcal{F}_{k-1}] < \infty$ with probability 1 for any $\ell = 1, \dots, d$. This and the

definition of $y_\ell(k)$ imply that with probability 1

$$\sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}[\|x_i(k) - \bar{x}(k)\| \mid \mathcal{F}_{k-1}] < \infty \text{ for all } i \in V, \quad (4.26)$$

where $\bar{x}(k) = \sum_{j=1}^m x_j(k)$. Next, consider $\|v_i(k) - \bar{v}(k)\|$. Since $v_i(k) = \sum_{j=1}^m [W(k)]_{ij} x_j(k-1)$ (see (4.3a)) and $W(k)$ is doubly stochastic, by using the convexity of the norm, for $\bar{v}(k) = \frac{1}{m} \sum_{j=1}^m v_j(k)$ we can see that $\sum_{i=1}^m \|v_i(k) - \bar{v}(k)\| \leq \sum_{j=1}^m \|x_j(k-1) - \bar{x}(k-1)\|$. By using relation (4.26), we conclude that $\sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}[\|v_i(k) - \bar{v}(k)\| \mid \mathcal{F}_{k-1}] < \infty$ for all $i \in V$ w.p.1. \blacksquare

4.3.5 Convergence of GRP

We assert the convergence of the method (4.3a)-(4.3c) using the lemmas established in Section 4.3.2-Section 4.3.4. Note that Lemma 4.4 allows us to infer that $v_i(k)$ approaches the set \mathcal{X} , while Lemma 4.6 allows us to claim that any two sequences $\{v_i(k)\}$ and $\{v_j(k)\}$ have the same limit points with probability 1. To claim the convergence of the iterates to an optimal solution, it remains to relate the limit points of $\{v_i(k)\}$ and the solutions of problem (1.1). This connection is provided by the iterate relation of Lemma 4.3, supported by the convergence result in Lemma 2.1.

Proposition 4.1 (Convergence w.p.1) *Let Assumptions 4.2-4.3 hold. Let us assume that the problem (1.1) has a nonempty optimal set \mathcal{X}^* and the iterates $\{x_i(k)\}$ are generated by the algorithm (4.3a)-(4.3c) with $\alpha_i(k) = 1/\Gamma_i(k)$. Then, the sequences $\{x_i(k)\}$, for $i \in V$, converge to some random point x^* in the optimal set \mathcal{X}^* with probability 1. i.e.,*

$$\lim_{k \rightarrow \infty} x_i(k) = x^* \quad \text{w.p.1 for all } i \in V.$$

Proof We start the proof by invoking Lemma 4.3 stating that for all $\tilde{x} \in \mathcal{X}$

and all $k \geq \hat{k}$,

$$\begin{aligned} \mathbb{E}[\|x_i(k) - \check{x}\|^2 \mid \mathcal{F}_{k-1}] &\leq \left(1 + \frac{a_1}{k^2}\right) \mathbb{E}[\|v_i(k) - \check{x}\|^2 \mid \mathcal{F}_{k-1}] \\ &\quad - \frac{2}{k} \mathbb{E}[f_i(z_i(k)) - f_i(\check{x}) \mid \mathcal{F}_{k-1}] - \frac{\gamma_i}{4c} \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] \\ &\quad + \frac{a_2}{k^2} + \frac{a_4}{k^{\frac{3}{2}-q}} + \left(\frac{a_3}{k^2} + \frac{a_5}{k^{\frac{3}{2}-q}}\right) \mathbb{E}[\|z_i(k) - \check{x}\|^2 \mid \mathcal{F}_{k-1}]. \end{aligned}$$

Since $\|z_i(k) - \check{x}\|^2 \leq 2\|z_i(k) - v_i(k)\| + 2\|v_i(k) - \check{x}\|^2$ and $\|z_i(k) - v_i(k)\| = \text{dist}(v_i(k), \mathcal{X})$, we obtain

$$\begin{aligned} \mathbb{E}[\|x_i(k) - \check{x}\|^2 \mid \mathcal{F}_{k-1}] &\leq \left(1 + \frac{2a_5}{k^{\frac{3}{2}-q}} + \frac{a_6}{k^2}\right) \mathbb{E}[\|v_i(k) - \check{x}\|^2 \mid \mathcal{F}_{k-1}] \\ &\quad - \frac{2}{k} \mathbb{E}[f_i(z_i(k)) - f_i(\check{x}) \mid \mathcal{F}_{k-1}] \\ &\quad - \left(\frac{\gamma_i}{4c} - \frac{2a_3}{k^2} - \frac{2a_5}{k^{\frac{3}{2}-q}}\right) \mathbb{E}[\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] + \frac{a_2}{k^2} + \frac{a_4}{k^{\frac{3}{2}-q}}, \end{aligned}$$

where $a_6 = a_1 + 2a_3$. Note that we can choose \hat{k} large enough so that $\left(\frac{\gamma_i}{4c} - \frac{2a_3}{k^2} - \frac{2a_5}{k^{\frac{3}{2}-q}}\right) \geq 0$ for all i . Then, by summing the preceding relations over i and using relation (4.12), we find that w.p.1 for all $\check{x} \in \mathcal{X}$ and all $k \geq \hat{k}$,

$$\begin{aligned} \sum_{i=1}^m \mathbb{E}[\|x_i(k) - \check{x}\|^2 \mid \mathcal{F}_{k-1}] &\leq \left(1 + \frac{2a_5}{k^{\frac{3}{2}-q}} + \frac{a_6}{k^2}\right) \sum_{j=1}^m \|x_j(k-1) - \check{x}\|^2 \\ &\quad - \frac{2}{k} \sum_{i=1}^m \mathbb{E}[f_i(z_i(k)) - f_i(\check{x}) \mid \mathcal{F}_{k-1}] + \frac{a_2 m}{k^2} + \frac{a_4 m}{k^{\frac{3}{2}-q}}. \quad (4.27) \end{aligned}$$

Recall that $f(x) = \sum_{i=1}^m f_i(x)$. Let $\bar{z}(k) \triangleq \frac{1}{m} \sum_{\ell=1}^m z_\ell(k)$. Using $\bar{z}(k)$ and f , we can rewrite the term $f_i(z_i(k)) - f_i(\check{x})$ as follows:

$$\begin{aligned} \sum_{i=1}^m (f_i(z_i(k)) - f_i(\check{x})) &= \sum_{i=1}^m (f_i(z_i(k)) - f_i(\bar{z}(k))) \\ &\quad + (f(\bar{z}(k)) - f(\check{x})). \quad (4.28) \end{aligned}$$

Furthermore, using the convexity of each function f_i , we obtain

$$\begin{aligned} \sum_{i=1}^m (f_i(z_i(k)) - f_i(\bar{z}(k))) &\geq \sum_{i=1}^m \langle \nabla f_i(\bar{z}(k)), z_i(k) - \bar{z}(k) \rangle \\ &\geq - \sum_{i=1}^m \|\nabla f_i(\bar{z}(k))\| \|z_i(k) - \bar{z}(k)\|. \end{aligned}$$

Since $\bar{z}(k)$ is a convex combination of points $z_i(k) \in \mathcal{X}$, it follows that $\bar{z}(k) \in \mathcal{X}$. This observation and Assumption 4.2(d), stating that the gradients $\nabla f_i(x)$ are uniformly bounded for $x \in \mathcal{X}$, yield

$$\sum_{i=1}^m (f_i(z_i(k)) - f_i(\bar{z}(k))) \geq -G_f \sum_{i=1}^m \|z_i(k) - \bar{z}(k)\|. \quad (4.29)$$

We next consider the term $\|z_i(k) - \bar{z}(k)\|$, for which by using $\bar{z}(k) \triangleq \frac{1}{m} \sum_{\ell=1}^m z_\ell(k)$ we have

$$\begin{aligned} \|z_i(k) - \bar{z}(k)\| &= \left\| \frac{1}{m} \sum_{\ell=1}^m (z_i(k) - z_\ell(k)) \right\| \\ &\leq \frac{1}{m} \sum_{\ell=1}^m \|z_i(k) - z_\ell(k)\| \leq \frac{1}{m} \sum_{\ell=1}^m \|v_i(k) - v_\ell(k)\|, \end{aligned}$$

where the first inequality is obtained by the convexity of the norm and the last inequality follows by the non-expansive projection property in Lemma 2.1(a). Furthermore, by letting $\bar{v}(k) = \frac{1}{m} \sum_{j=1}^m v_j(k)$ and using $\|v_i(k) - v_\ell(k)\| \leq \|v_i(k) - \bar{v}(k)\| + \|v_\ell(k) - \bar{v}(k)\|$, we obtain $\|z_i(k) - \bar{z}(k)\| \leq \|v_i(k) - \bar{v}(k)\| + \frac{1}{m} \sum_{\ell=1}^m \|v_\ell(k) - \bar{v}(k)\|$ for every $i \in V$. Upon summing these relations over $i \in V$, we find

$$\sum_{i=1}^m \|z_i(k) - \bar{z}(k)\| \leq 2 \sum_{i=1}^m \|v_i(k) - \bar{v}(k)\|. \quad (4.30)$$

Combining relations (4.30) and (4.29), and substituting the resulting relation

in Eq. (4.28), we obtain

$$\begin{aligned} \sum_{i=1}^m (f_i(z_i(k)) - f_i(\tilde{x})) &\geq -2G_f \sum_{i=1}^m \|v_i(k) - \bar{v}(k)\| \\ &\quad + (f(\bar{z}(k)) - f(\tilde{x})). \end{aligned} \quad (4.31)$$

Finally, by using the preceding estimate in inequality (4.27) and letting $\tilde{x} = x^*$ for an arbitrary $x^* \in \mathcal{X}^*$, we have w.p.1 for any $x^* \in \mathcal{X}^*$ and $k \geq \bar{k}$,

$$\begin{aligned} \sum_{i=1}^m \mathbb{E}[\|x_i(k) - x^*\|^2 \mid \mathcal{F}_{k-1}] &\leq \left(1 + \frac{2a_5}{k^{\frac{3}{2}-q}} + \frac{a_6}{k^2}\right) \sum_{j=1}^m \|x_i(k-1) - x^*\|^2 \\ &\quad - \frac{2}{k} \mathbb{E}[f(\bar{z}(k)) - f^* \mid \mathcal{F}_{k-1}] \\ &\quad + \frac{4G_f}{k} \sum_{i=1}^m \mathbb{E}[\|v_i(k) - \bar{v}(k)\| \mid \mathcal{F}_{k-1}] + \frac{a_2 m}{k^2} + \frac{a_4 m}{k^{\frac{3}{2}-q}}. \end{aligned} \quad (4.32)$$

Since $\bar{z}(k) \in \mathcal{X}$, we have $f(\bar{z}(k)) - f^* \geq 0$. Thus, in light of Lemma 4.6, relation (4.32) satisfies all the conditions of Lemma 2.1. Hence, the sequence $\{\|x_i(k) - x^*\|^2\}$ is convergent with probability 1 for any $i \in V$ and $x^* \in \mathcal{X}^*$, and $\sum_{k=0}^{\infty} \frac{1}{k} (f(\bar{z}(k)) - f^*) < \infty$ w.p.1. Since $\sum_{k=0}^{\infty} \frac{1}{k} = \infty$, it follows that

$$\liminf_{k \rightarrow \infty} (f(\bar{z}(k)) - f^*) = 0 \quad \text{w.p.1.} \quad (4.33)$$

By Lemma 4.4(a), noting that $z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$, we have

$$\lim_{k \rightarrow \infty} \|v_i(k) - z_i(k)\| = 0 \quad \text{for all } i \in V \text{ w.p.1.} \quad (4.34)$$

Since the sequence $\{\|x_i(k) - x^*\|\}$ is convergent with probability 1 for any $i \in V$ and every $x^* \in \mathcal{X}^*$, in view of the relations (4.3a) and (4.34), respectively, so are the sequences $\{\|v_i(k) - x^*\|\}$ and $\{\|z_i(k) - x^*\|\}$, as well as their average sequences $\{\|\bar{v}(k) - x^*\|\}$ and $\{\|\bar{z}(k) - x^*\|\}$. Therefore, the sequences $\{\bar{v}(k)\}$ and $\{\bar{z}(k)\}$ are bounded with probability 1, and they have accumulation points. From relation (4.33) and the continuity of f , the sequence $\{\bar{z}(k)\}$ must have one accumulation point in \mathcal{X}^* with probability 1. This and the fact that $\{\|\bar{z}(k) - x^*\|\}$ is convergent with probability 1 for every $x^* \in \mathcal{X}^*$

imply that for a random point $x^* \in \mathcal{X}^*$,

$$\lim_{k \rightarrow \infty} \bar{z}(k) = x^* \quad \text{w.p.1.} \quad (4.35)$$

Now, from $\bar{z}(k) = \frac{1}{m} \sum_{\ell=1}^m z_\ell(k)$ and $\bar{v}(k) = \frac{1}{m} \sum_{i=\ell}^m v_\ell(k)$, using relation (4.34) and the convexity of the norm, we obtain $\lim_{k \rightarrow \infty} \|\bar{v}(k) - \bar{z}(k)\| \leq \frac{1}{m} \sum_{\ell=1}^m \lim_{k \rightarrow \infty} \|v_\ell(k) - z_\ell(k)\| = 0$ w.p.1. In view of relation (4.35), it follows that

$$\lim_{k \rightarrow \infty} \bar{v}(k) = x^* \quad \text{w.p.1.} \quad (4.36)$$

By Lemma 4.6, we have

$$\liminf_{k \rightarrow \infty} \|v_i(k) - \bar{v}(k)\| = 0 \quad \text{for all } i \in V \text{ w.p.1.} \quad (4.37)$$

The fact that $\{\|v_i(k) - x^*\|\}$ is convergent with probability 1 for all i and any $x^* \in \mathcal{X}^*$, together with (4.36) and (4.37), implies that

$$\lim_{k \rightarrow \infty} \|v_i(k) - x^*\| = 0 \quad \text{for all } i \in V \text{ w.p.1.} \quad (4.38)$$

Finally, by Lemma 4.4(b), we have $\lim_{k \rightarrow \infty} \|x_i(k) - v_i(k)\| = 0$ for all $i \in V$ w.p.1, which together with the limit in (4.38) yields $\lim_{k \rightarrow \infty} x_i(k) = x^*$ for all $i \in V$ with probability 1. ■

Proposition 4.1 states that the agents asymptotically arrive at a consensus and the consensus point is in the optimal set \mathcal{X}^* .

4.4 Error Bound

We now focus on a constant stepsize $\alpha_i(k) = \alpha_i > 0$ for $i \in V$ and establish a limiting error bound assuming that each f_i is strongly convex over the set \mathcal{X} with constant σ_i .

4.4.1 Basic Results for GRP

We start by providing some lemmas that are valid for a constant stepsize. The first result shows a basic iterate relation.

Lemma 4.7 *Let Assumptions 4.2-4.3 hold, where Assumption 4.2(b) is replaced with a condition that each function f_i be strongly convex with a constant σ_i over \mathbb{R}^d . Let the stepsize in method (4.3a)-(4.3c) be such that $\alpha_i(k) = \alpha_i > 0$, $8\alpha_i^2 L_i^2 - \frac{1}{2c} \leq 0$, and $\rho_i = 1 - \alpha_i \sigma_i + 8(1+c)\alpha_i^2 L_i^2 \in (0, 1)$, where c is the constant from Assumption 4.3. Then, for the solution x^* of problem (1.1) we have w.p.1 for all $k \geq 1$ and $i \in V$,*

$$\begin{aligned} \mathbb{E}[\|x_i(k) - x^*\|^2 \mid \mathcal{F}_{k-1}, I_k, J_k] \\ \leq \rho_i \|v_i(k) - x^*\|^2 - 2\alpha_i \langle \nabla f_i(x^*), z_i(k) - x^* \rangle + 8(1+c)\alpha_i^2 G_f^2. \end{aligned}$$

Proof The function f is strongly convex with a constant $\sigma = \sum_{i=1}^m \sigma_i$, and therefore problem (1.1) has a unique optimal solution x^* . We use the definition of the iterate $x_i(k)$ in (4.3a)-(4.3c) and Lemma 4.1(b) with the following identification: $\mathcal{Y} = \mathcal{X}$, $y = x_i(k)$, $x = v_i(k)$, $z = z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$, $\alpha = \alpha_i$, $\tilde{x} = x^*$, $f = f_i$, $L = L_i$ and $\tau_1 = \tau_2 = 8c$. Thus, we obtain for the optimal point x^* , $k \geq 1$ and $i \in \{I_k, J_k\}$,

$$\begin{aligned} \|x_i(k) - x^*\|^2 &\leq (1 - \sigma_i \alpha_i + 8\alpha_i^2 L_i^2) \|v_i(k) - x^*\|^2 \\ &\quad - 2\alpha_i \langle \nabla f_i(x^*), z_i(k) - x^* \rangle - \frac{3}{4} \|x_i(k) - v_i(k)\|^2 \\ &\quad + 8(1+c)\alpha_i^2 \|\nabla f_i(x^*)\|^2 + 4c\alpha_i^2 L_i^2 \|z_i(k) - x^*\|^2 + \frac{1}{4c} \|v_i(k) - z_i(k)\|^2. \end{aligned}$$

By Assumption 4.2(d), we have $\|\nabla f_i(x^*)\| \leq G_f$. Furthermore, $\|z_i(k) - x^*\|^2 \leq 2\|z_i(k) - v_i(k)\|^2 + 2\|v_i(k) - x^*\|^2$. Since $z_i(k) = \Pi_{\mathcal{X}}[v_i(k)]$, we have $\|v_i(k) - z_i(k)\| = \text{dist}(v_i(k), \mathcal{X})$. Therefore,

$$\begin{aligned} \|x_i(k) - x^*\|^2 &\leq \rho_i \|v_i(k) - x^*\|^2 - 2\alpha_i \langle \nabla f_i(x^*), z_i(k) - x^* \rangle \\ &\quad - \frac{3}{4} \|x_i(k) - v_i(k)\|^2 + 8(1+c)\alpha_i^2 G_f^2 + \left(8\alpha_i^2 L_i^2 + \frac{1}{4c}\right) \text{dist}^2(v_i(k), \mathcal{X}), \end{aligned}$$

with $\rho_i = 1 - \sigma_i \alpha_i + 8(1+c)\alpha_i^2 L_i^2$. By the definition of $x_i(k)$, we have $x_i(k) \in \mathcal{X}_i^{\Omega_i(k)}$, which implies

$$\mathbb{E}[\|v_i(k) - x_i(k)\| \mid \mathcal{F}_{k-1}, I_k, J_k] \geq \mathbb{E}[\text{dist}(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \mid \mathcal{F}_{k-1}, I_k, J_k].$$

By Assumption 4.3 it follows

$$\text{dist}^2(v_i(k), \mathcal{X}) \leq c\mathbb{E} \left[\text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \mid \mathcal{F}_{k-1}, I_k, J_k \right] \text{ for all } i.$$

Therefore, we have w.p.1 for all k and i ,

$$\begin{aligned} \mathbb{E}[\|x_i(k) - x^*\|^2 \mid \mathcal{F}_{k-1}, I_k, J_k] &\leq \rho_i \|v_i(k) - x^*\|^2 - 2\alpha_i \langle \nabla f_i(x^*), z_i(k) - x^* \rangle \\ &\quad + 8(1+c)\alpha_i^2 G_f^2 + \left(8\alpha_i^2 L_i^2 - \frac{1}{2c} \right) \text{dist}^2(v_i(k), \mathcal{X}), \end{aligned}$$

from which the result follows by $8\alpha_i^2 L_i^2 - \frac{1}{2c} \leq 0$. \blacksquare

In the next lemma we provide an asymptotic upper bound for the distance between the iterates $x_i(k)$ and the set \mathcal{X} .

Lemma 4.8 *Let Assumptions 4.2-4.3 hold, where Assumption 4.2(b) is replaced with a condition that each f_i is strongly convex with a scalar $\sigma_i > 0$ over \mathbb{R}^d . Assume that $\mu_i = 1 - \gamma_i \alpha_i \sigma_i + 8\gamma_i \alpha_i^2 L_i^2 \in (0, 1)$. Then, we have with probability 1*

$$\limsup_{k \rightarrow \infty} \sum_{i=1}^m \mathbb{E}[\text{dist}^2(x_i(k), \mathcal{X})] \leq C_{pe} m \bar{\gamma} \bar{\alpha}^2 G_f^2,$$

where $C_{pe} = \frac{8(1+c)}{1-q_{pe}}$, $q_{pe} = \max_i \mu_i$, $\bar{\gamma} = \max_i \gamma_i$, and $\bar{\alpha} = \max_i \alpha_i$.

Proof We fix $i \in \{I_k, J_k\}$ and use Lemma 4.1(b) with the following identification: $\mathcal{Y} = \mathcal{X}_i^{\Omega_i(k)}$, $y = x_i(k)$, $x = v_i(k)$, $\tilde{x} = z = \Pi_{\mathcal{X}}[v_i(k)]$, $\sigma = \sigma_i$, $\alpha = \alpha_i > 0$ and $\phi = f_i$, $L = L_i$, and obtain for $i \in \{I_k, J_k\}$ and $k \geq 1$,

$$\begin{aligned} \|x_i(k) - \Pi_{\mathcal{X}}[v_i(k)]\|^2 &\leq (1 - \alpha_i \sigma_i + 8\alpha_i^2 L_i^2) \|v_i(k) - \Pi_{\mathcal{X}}[v_i(k)]\|^2 \\ &\quad - \frac{3}{4} \|x_i(k) - v_i(k)\|^2 + (8 + \tau_2) \alpha_i^2 \|\nabla f_i(\Pi_{\mathcal{X}}[v_i(k)])\|^2 \\ &\quad + \left(\frac{1}{\tau_1} + \frac{1}{\tau_2} \right) \|v_i(k) - \Pi_{\mathcal{X}}[v_i(k)]\|^2. \end{aligned}$$

By Assumption 4.2(d), we have $\|\nabla f_i(\Pi_{\mathcal{X}}[v_i(k)])\| \leq G_f$. Thus, by letting $\tau_1 = \tau_2 = 8c$ (where c is from Assumption 4.3), we obtain for $i \in \{I_k, J_k\}$

and $k \geq 1$ with probability 1

$$\begin{aligned} \mathbb{E} [\|x_i(k) - \Pi_{\mathcal{X}}[v_i(k)]\|^2 \mid \mathcal{F}_{k-1}, I_k, J_k] &\leq (1 - \alpha_i \sigma_i + 8\alpha_i^2 L_i^2) \text{dist}^2(v_i(k), \mathcal{X}) \\ &- \frac{3}{4} \mathbb{E} [\|x_i(k) - v_i(k)\|^2 \mid \mathcal{F}_{k-1}, I_k, J_k] + 8(1 + c)\alpha_i^2 G_f^2 + \frac{1}{4c} \text{dist}(v_i(k), \mathcal{X}). \end{aligned} \quad (4.39)$$

We also make use of the projection properties, which yield $\text{dist}(x_i(k), \mathcal{X}) \leq \|x_i(k) - \Pi_{\mathcal{X}}[v_i(k)]\|$ and $\|v_i(k) - x_i(k)\| \geq \text{dist}(v_i(k), \mathcal{X}_i^{\Omega_i(k)})$. Furthermore, using Assumption 4.3, we have

$$\mathbb{E} [\text{dist}^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \mid \mathcal{F}_{k-1}, I_k, J_k] \geq \frac{1}{c} \text{dist}^2(v_i(k), \mathcal{X}).$$

Substituting these estimates in relation (4.39), ignoring the term with the factor $-1/2c$, and rearranging the other terms accordingly, we have w.p.1 for $i \in \{I_k, J_k\}$ and $k \geq 1$,

$$\begin{aligned} \mathbb{E}[\text{dist}^2(x_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}, I_k, J_k] \\ \leq (1 - \alpha_i \sigma_i + 8\alpha_i^2 L_i^2) \text{dist}^2(v_i(k), \mathcal{X}) + 8(1 + c)\alpha_i^2 G_f^2. \end{aligned}$$

The preceding relation holds with probability γ_i , and $x_i(k) = v_i(k)$ with probability $1 - \gamma_i$. Thus, with probability 1 for all $k \geq 1$ and $i \in V$,

$$\mathbb{E}[\text{dist}^2(x_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] \leq \mu_i \mathbb{E} [\text{dist}^2(v_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] + 8(1 + c)\gamma_i \alpha_i^2 G_f^2,$$

where $\mu_i = 1 - \gamma_i \alpha_i \sigma_i + 8\gamma_i \alpha_i^2 L_i^2$. By summing over i and using relation (4.13), we obtain

$$\sum_{i=1}^m \mathbb{E}[\text{dist}^2(x_i(k), \mathcal{X}) \mid \mathcal{F}_{k-1}] \leq q_{pe} \sum_{j=1}^m \text{dist}^2(x_j(k-1), \mathcal{X}) + 8(1 + c)m\bar{\gamma}\bar{\alpha}^2 G_f^2,$$

where $q_{pe} = \max_i \mu_i$, $\bar{\gamma} = \max_i \gamma_i$, and $\bar{\alpha} = \max_i \alpha_i$. Taking the total expectation, it follows that $\sum_{i=1}^m \mathbb{E}[\text{dist}^2(x_i(k), \mathcal{X})] \leq q_{pe} \sum_{i=1}^m \mathbb{E}[\text{dist}^2(x_i(k-1), \mathcal{X})] + 8(1 + c)m\bar{\gamma}\bar{\alpha}^2 G_f^2$. Assuming that each agent selects the stepsize so that $\mu_i = 1 - \gamma_i \alpha_i \sigma_i + 8\gamma_i \alpha_i^2 L_i^2 \in (0, 1)$, the desired relation follows. \blacksquare

The bound shows the asymptotic distance in terms of the number of agents, the maximum stepsize, the properties of the objective function and the agent-update probability γ_i .

In the next lemma, we provide an estimate for the disagreement among the agents.

Lemma 4.9 *Let Assumptions 4.1-4.3 hold and f_i be strongly convex over \mathbb{R}^d . Let the stepsizes in method (4.3a)-(4.3c) be such that $\mu_i = 1 - \gamma_i \alpha_i \sigma_i + 8\gamma_i \alpha_i^2 L_i^2 \in (0, 1)$ for all $i \in V$. Let $\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i(k)$ for all k . Then, for the iterates $\{x_i(k)\}$ generated by method (4.3a)-(4.3c), we have with probability 1*

$$\limsup_{k \rightarrow \infty} \sum_{i=1}^m \mathbb{E}[\|x_i(k) - \bar{x}(k)\|^2] \leq C_{de} \frac{m \bar{\alpha}^2 G_f^2}{(1 - \sqrt{\lambda})^2},$$

where $C_{de} = 4 \left(\frac{8\bar{\gamma}(1+\bar{\alpha}^2 \bar{L}^2)(1+c)}{1-q_{pe}} + 1 \right)$, $q_{pe} = \max_i \mu_i$, $\bar{\gamma} = \max_i \gamma_i$, $\bar{\alpha} = \max_i \alpha_i$, and $\bar{L} = \max L_i$.

Proof We consider coordinate-wise relations similar to the proof of Lemma 4.6. Since the matrices $W(k)$ are doubly stochastic for all $k \geq 1$, from relation (4.21) with $\|M\| = 1$ and Hölder's inequality, we obtain

$$\begin{aligned} & \sum_{\ell=1}^d \mathbb{E}[\|y_\ell(k) - [\bar{x}(k)]_\ell \mathbf{1}\|^2] \\ & \leq \left(\sqrt{\sum_{\ell=1}^d \mathbb{E}[\|D_k(y_\ell(k-1) - [\bar{x}(k-1)]_\ell \mathbf{1})\|^2]} + \sqrt{\sum_{\ell=1}^d \mathbb{E}[\|\delta_\ell(k)\|^2]} \right)^2, \end{aligned} \quad (4.40)$$

where $[\bar{x}(k)]_\ell = \frac{1}{m} \mathbf{1}^T y_\ell(k)$, $D_k = W(k) - \frac{1}{m} \mathbf{1} \mathbf{1}^T$, and $\lambda < 1$ is the second largest eigenvalue of \bar{W} . From relation (4.4), we know that

$$\begin{aligned} & \sum_{\ell=1}^d \mathbb{E}[\|D_k(y_\ell(k-1) - [\bar{x}(k-1)]_\ell \mathbf{1})\|^2] \\ & \leq \lambda \sum_{\ell=1}^d \mathbb{E}[\|y_\ell(k-1) - [\bar{x}(k-1)]_\ell \mathbf{1}\|^2]. \end{aligned} \quad (4.41)$$

The second term in (4.40) is evaluated similar to that in Lemma 4.6. Hence, we obtain for all $k \geq 1$ with probability 1

$$\sqrt{\sum_{\ell=1}^d \mathbb{E}[\|\delta_\ell(k)\|^2]} \leq \beta_k, \quad (4.42)$$

where $\beta_k = \sqrt{(4 + 4\bar{\alpha}^2\bar{L}^2) \sum_{i=1}^m \mathbb{E} [\text{dist}^2(v_i(k), \mathcal{X})] + 4m\bar{\alpha}^2 G_f^2}$, $\bar{\alpha} = \max_i \alpha_i$ and $\bar{L} = \max_i L_i$.

Letting $u_k = \sqrt{\sum_{\ell=1}^d \mathbb{E} [\|y_\ell(k) - [\bar{x}(k)]_\ell \mathbf{1}\|^2]}$ in (4.40) and using relations (4.41) and (4.42), we have for all $k \geq 1$

$$u_k \leq \sqrt{\lambda} u_{k-1} + \beta_k.$$

Since $\lambda < 1$, by Lemma 2.3, it follows that

$$\limsup_{k \rightarrow \infty} u_k \leq \frac{1}{1 - \sqrt{\lambda}} \limsup_{k \rightarrow \infty} \beta_k. \quad (4.43)$$

By the definition of $y_\ell(k)$, we have $u_k = \sqrt{\sum_{i=1}^m \mathbb{E} [\|x_i(k) - \bar{x}(k)\|^2]}$. The convexity of the norm function, the doubly stochastic \bar{W} , and the definition of $v_i(k)$ imply $\sum_{i=1}^m \mathbb{E} [\text{dist}^2(v_i(k), \mathcal{X})] \leq \sum_{j=1}^m \mathbb{E} [\text{dist}^2(x_j(k-1), \mathcal{X})]$. Therefore, we obtain

$$\limsup_{k \rightarrow \infty} \beta_k^2 \leq (4 + 4\bar{\alpha}^2\bar{L}^2) \limsup_{k \rightarrow \infty} \sum_{j=1}^m \mathbb{E} [\text{dist}^2(x_j(k-1), \mathcal{X})] + 4m\bar{\alpha}^2 G_f^2. \quad (4.44)$$

The desired relation follows from Eqs. (4.43) and (4.44) and Lemma 4.8. \blacksquare

The bound in Lemma 4.9 captures the variance of the estimates $x_i(k)$ and their average in terms of the number of agents, the maximum stepsize and the spectral gap $1 - \sqrt{\lambda}$ of the matrix \bar{W} .

4.4.2 Error Bound of GRP

We now establish the error bound of the method (4.3a)-(4.3c) with a constant stepsize using the lemmas provided in Section 4.4.1.

Proposition 4.2 (Error bound) *Let Assumptions 4.2-4.3 hold, where Assumption 4.2(b) is replaced with a condition that each function f_i be strongly convex with a constant σ_i over \mathbb{R}^d . Let λ be the second largest eigenvalue of the matrix \bar{W} and $\Delta_{\gamma\alpha} = \max_i \gamma_i \alpha_i - \min_i \gamma_i \alpha_i$. Let $\{x_i(k)\}$, $i \in V$ be the iterates generated by the algorithm (4.3a)-(4.3c) with a constant step size $\alpha_i(k) = \alpha_i > 0$ and an agent selection probability γ_i that satisfies*

$q = \max_i \{1 - \gamma_i \alpha_i \sigma_i + 8(1+c)\gamma_i \alpha_i^2 L_i^2 + \frac{\Delta_{\gamma\alpha}}{m}\} \in (0, 1)$. Then, we have

$$\limsup_{k \rightarrow \infty} \frac{1}{m} \sum_{i=1}^m \mathbb{E}[\|x_i(k) - x^*\|^2] \leq \left(\frac{C_1}{1 - \sqrt{\lambda}} + C_2 \right) \bar{\gamma} \bar{\alpha}^2 G_f^2 + \Delta_{\gamma\alpha} C_3,$$

where $C_1 = \frac{2\sqrt{C_{de}}}{1-q}$, $C_2 = \frac{8(1+c)}{1-q}$ and $C_3 = \frac{G_f^2}{1-q}$, $\bar{\gamma} = \max_i \gamma_i$, $\bar{\alpha} = \max_i \alpha_i$ and C_{de} is the constant from Lemma 4.9.

Proof The proof starts with the relation of Lemma 4.7. Define $\bar{z}(k) = \frac{1}{m} \sum_{i=1}^m z_i(k)$, so that $\bar{z}(k) \in \mathcal{X}$. We have $\langle \nabla f_i(x^*), z_i(k) - x^* \rangle = \langle \nabla f_i(x^*), \bar{z}(k) - x^* \rangle + \langle \nabla f_i(x^*), z_i(k) - \bar{z}(k) \rangle$, which in view of the gradient boundedness (Assumption 4.2(d)) implies that

$$\langle \nabla f_i(x^*), z_i(k) - x^* \rangle \geq \langle \nabla f_i(x^*), \bar{z}(k) - x^* \rangle - G_f \|z_i(k) - \bar{z}(k)\|.$$

Substituting the above relation in the relation of Lemma 4.7, we have for all $k \geq 1$ w.p.1,

$$\begin{aligned} \mathbb{E}[\|x_i(k) - x^*\|^2 \mid \mathcal{F}_{k-1}, I_k, J_k] &\leq \rho_i \|v_i(k) - x^*\|^2 - 2\alpha_i \langle \nabla f_i(x^*), \bar{z}(k) - x^* \rangle \\ &\quad + 2\alpha_i G_f \|z_i(k) - \bar{z}(k)\| + 8(1+c)\alpha_i^2 G_f^2. \end{aligned}$$

Taking the expectation with respect to \mathcal{F}_{k-1} and using the fact that the preceding inequality holds with probability γ_i , and $x_i(k) = v_i(k)$ with probability $1 - \gamma_i$, we obtain w.p.1 for all $k \geq 1$ and $i \in V$,

$$\begin{aligned} \mathbb{E}[\|x_i(k) - x^*\|^2 \mid \mathcal{F}_{k-1}] &\leq \nu_i \mathbb{E}[\|v_i(k) - x^*\|^2 \mid \mathcal{F}_{k-1}] \\ &\quad - 2\gamma_i \alpha_i \mathbb{E}[\langle \nabla f_i(x^*), \bar{z}(k) - x^* \rangle \mid \mathcal{F}_{k-1}] \\ &\quad + 2\gamma_i \alpha_i G_f \mathbb{E}[\|z_i(k) - \bar{z}(k)\| \mid \mathcal{F}_{k-1}] + 8(1+c)\gamma_i \alpha_i^2 G_f^2, \end{aligned}$$

where $\nu_i = 1 - \gamma_i \alpha_i \sigma_i + 8(1+c)\gamma_i \alpha_i^2 L_i^2$. Let $\underline{\alpha} = \min_i \alpha_i$, $\underline{\gamma} = \min_i \gamma_i$, $\bar{\alpha} = \max_i \alpha_i$ and $\bar{\gamma} = \max_i \gamma_i$. By adding and subtracting $2\gamma_i \alpha_i \mathbb{E}[\langle \nabla f_i(x^*), \bar{z}(k) -$

$x^*\rangle \mid \mathcal{F}_{k-1}]$, we find that

$$\begin{aligned} \mathbb{E}[\|x_i(k) - x^*\|^2 \mid \mathcal{F}_{k-1}] &\leq \nu_i \mathbb{E}[\|v_i(k) - x^*\|^2 \mid \mathcal{F}_{k-1}] \\ &\quad - 2\underline{\gamma}\underline{\alpha} \mathbb{E}[\langle \nabla f_i(x^*), \bar{z}(k) - x^* \rangle \mid \mathcal{F}_{k-1}] \\ &\quad + 2\Delta_{\gamma\alpha} \mathbb{E}[\|\nabla f_i(x^*)\| \|\bar{z}(k) - x^*\| \mid \mathcal{F}_{k-1}] \\ &\quad + 2\bar{\gamma}\bar{\alpha}G_f \mathbb{E}[\|z_i(k) - \bar{z}(k)\| \mid \mathcal{F}_{k-1}] + 8(1+c)\bar{\gamma}\bar{\alpha}^2G_f^2, \end{aligned} \quad (4.45)$$

where $\Delta_{\gamma\alpha} = \max_i \gamma_i \alpha_i - \min_i \gamma_i \alpha_i$. We can further estimate

$$\|\nabla f_i(x^*)\| \|\bar{z}(k) - x^*\| \leq \frac{G_f}{m} \sum_{i=1}^m \|\Pi_{\mathcal{X}}[v_i(k)] - x^*\| \leq \frac{G_f}{m} \sum_{i=1}^m \|v_i(k) - x^*\|,$$

where the first inequality follows by Assumption 4.2(d), $\bar{z}(k) = \frac{1}{m} \sum_{i=1}^m \Pi_{\mathcal{X}}[v_i(k)]$ and the convexity of the norm function and the second inequality follows by the projection property in Lemma 2.1(a). Also, from relation the square property $ab \leq \frac{1}{2}(a^2 + b^2)$ and Hölder's inequality, we have

$$\|\nabla f_i(x^*)\| \|\bar{z}(k) - x^*\| \leq \frac{1}{2} \left(G_f^2 + \frac{1}{m} \sum_{i=1}^m \|v_i(k) - x^*\|^2 \right). \quad (4.46)$$

Summing relation (4.45) over $i = 1, \dots, m$, using estimate (4.12), (4.46) and $\sum_{i=1}^m \langle \nabla f_i(x^*), \bar{z}(k) - x^* \rangle \geq f(\bar{z}(k)) - f(x^*) \geq 0$, we have

$$\begin{aligned} \sum_{i=1}^m \mathbb{E}[\|x_j(k-1) - x^*\|^2 \mid \mathcal{F}_{k-1}] &\leq q \sum_{i=1}^m \mathbb{E}[\|v_i(k) - x^*\|^2 \mid \mathcal{F}_{k-1}] + \Delta_{\gamma\alpha} m G_f^2 \\ &\quad + 2\bar{\gamma}\bar{\alpha}G_f \sum_{i=1}^m \mathbb{E}[\|z_i(k) - \bar{z}(k)\| \mid \mathcal{F}_{k-1}] + 8(1+c)m\bar{\gamma}\bar{\alpha}^2G_f^2, \end{aligned}$$

where $q = \max_i \{\nu_i + \frac{\Delta_{\gamma\alpha}}{m}\}$. If α_i and γ_i are chosen such that that $q \in (0, 1)$, we obtain

$$\begin{aligned} \limsup_{k \rightarrow \infty} \sum_{i=1}^m \mathbb{E}[\|x_i(k) - x^*\|^2] &\leq \frac{2\bar{\gamma}\bar{\alpha}G_f}{1-q} \limsup_{k \rightarrow \infty} \sum_{i=1}^m \mathbb{E}[\|z_i(k) - \bar{z}(k)\|] \\ &\quad + \frac{\Delta_{\gamma\alpha} m G_f^2}{1-q} + \frac{8(1+c)m\bar{\gamma}\bar{\alpha}^2G_f^2}{1-q}. \end{aligned} \quad (4.47)$$

From the projection property in Lemma 2.1(a), it follows that

$$\begin{aligned} \sum_{i=1}^m \mathbb{E}[\|z_i(k) - \bar{z}(k)\|] &\leq \sum_{i=1}^m \mathbb{E}[\|z_i(k) - \Pi_{\mathcal{X}}[\bar{v}(k)]\|] \\ &\leq \sum_{i=1}^m \mathbb{E}[\|v_i(k) - \bar{v}(k)\|]. \end{aligned} \quad (4.48)$$

Furthermore, using Hölder's inequality, we have

$$\begin{aligned} \sum_{i=1}^m \mathbb{E}[\|v_i(k) - \bar{v}(k)\|] &\leq \sqrt{m \mathbb{E} \left[\sum_{i=1}^m \|v_i(k) - \bar{v}(k)\|^2 \right]} \\ &\leq \sqrt{m \mathbb{E} \left[\sum_{i=1}^m \|x_i(k) - \bar{x}(k)\|^2 \right]}, \end{aligned} \quad (4.49)$$

where the last inequality follows by the convexity of the norm-squared and the definition of $v_i(k)$. The result follows from (4.47)–(4.49) and Lemma 4.9. \blacksquare

Proposition 4.2 provides an asymptotic error bound for the average of the expected distances between the optimal solution x^* and the iterates of GRP algorithm. The bound shows that the convergence of the algorithm is upper bounded by a sum of three terms. The first term is a penalty incurred due to the distributed nature of the problem, which is controlled by the spectral gap $1 - \sqrt{\lambda}$ of the matrix \bar{W} . The second term is an optimization error term due to a non-diminishing stepsize that is common to gradient descent algorithms. The third term is an error term due to the different agent selection probability γ_i . To minimize the error, it shows that agents who update less frequently should be more aggressive and choose a larger stepsize than those who update more frequently.

4.5 Gossip-based Mini-batch Random Projections

As an extension of the algorithm in (4.3a)–(4.3c), one may consider an algorithm where the agents use several random projections at each iteration. Namely, after generating $v_i(k)$ agent $i \in \{I_k, J_k\}$ may take (or nature may reveal them) several random samples $\Omega_i^1(k), \dots, \Omega_i^b(k)$, where each $\Omega_i^r(k) \in I_i$

and $b \geq 1$ is the batch-size. Each collection $\Omega_i^1(k), \dots, \Omega_i^b(k)$ consists of mutually independent random variables and is independent of the past realizations. More specifically, we have b random independent samples of the *i.i.d.* random variable $\Omega_i(k)$ (taking values in I_i). In the mini-batch version of the algorithm, agent $i \in \{I_k, J_k\}$, performs the following steps:

$$v_i(k) = \frac{1}{2}(x_{I_k}(k-1) + x_{J_k}(k-1)), \quad (4.50a)$$

$$\psi_i^0(k) = v_i(k) - \alpha_i(k) \nabla f(v_i(k)), \quad (4.50b)$$

$$\psi_i^r(k) = \Pi_{\mathcal{X}_i^{\Omega_i^r(k)}}[\psi_i^{r-1}(k)] \quad \text{for } r = 1, \dots, b, \quad (4.50c)$$

$$x_i(k) = \frac{1}{b} \sum_{j=1}^b \psi_i^j(k), \quad (4.50d)$$

where $\alpha_i(k) > 0$ is a stepsize at time k and $x_i(0) \in \mathbb{R}^d$ is an initial estimate of agent i (which can be random). The steps in (4.50a)–(4.50c) are the successive (random) projections on the sets $\mathcal{X}^{\Omega_i^1(k)}, \dots, \mathcal{X}^{\Omega_i^b(k)}$ of the point $v_i(k) - \alpha_k \nabla f_i(v_i(k))$ whereas the step (4.50d) calculates the average of the projected points. Agents other than I_k and J_k do not update.

For the algorithm using random mini-batch projections, we have the following convergence result.

Proposition 4.3 *Let Assumptions 4.2–4.3 hold. Assume that problem (1.1) has a nonempty optimal set \mathcal{X}^* . Then, the iterates $\{x_i(k)\}$, $i \in V$, produced by the method (4.50a)–(4.50d) converge to some random point in the optimal set \mathcal{X}^* with probability 1, i.e., for some random vector $x^* \in \mathcal{X}^*$,*

$$\lim_{k \rightarrow \infty} x_i(k) = x^* \quad \text{for all } i \in V \text{ w.p.1.}$$

Proof The proof of this result is similar to that of Proposition 4.1. It requires some adjustments of Lemma 4.4 and Lemma 4.6. The proof is omitted.

CHAPTER 5

APPLICATIONS

5.1 Distributed Support Vector Machines (DrSVM)

In this section, we apply our DRP algorithm and its mini-batch variant to Support vector machines (SVMs). We provide a brief introduction to SVMs in Section 5.1.1. In Section 5.1.2, we derive mathematical formulae for the projection of a point onto the intersection of two halfspaces. The formulae will be needed to apply our DRP in SVM applications. In Section 5.1.3, we report our numerical results on some text classification data sets.

5.1.1 Support Vector Machines

Support vector machines (SVMs) are popular classification tools with a strong theoretical background. Given a set of n example-label pairs $\{(a_j, b_j)\}_{j=1}^n$, $a_j \in \mathbb{R}^d$ and $b_j \in \{+1, -1\}$, we need to find a vector $x = [y^T \ \boldsymbol{\xi}^T]^T \in \mathbb{R}^{d+n}$ that solves the following optimization problem (a bias term is included in y for convenience):

$$\begin{aligned} \min_{y, \boldsymbol{\xi}} f(y, \boldsymbol{\xi}) &= \frac{1}{2} \|y\|^2 + C \sum_{j=1}^n \xi_j \\ \text{s.t.} \quad & b_j \langle y, a_j \rangle \geq 1 - \xi_j, \quad \xi_j \geq 0, \quad \text{for all } j \in \{1, \dots, n\}. \end{aligned} \tag{5.1}$$

Here, we use slack variables ξ_j , for $j = 1, \dots, n$, to consider linearly non-separable cases as well. If the optimal solution $(y^*, \boldsymbol{\xi}^*)$ to this problem exists, the solution y^* is the maximum-margin separating hyperplane [44].

For applying DRP to problem (5.1), we can define f_i and \mathcal{X}_i , as follows:

$$f_i(x) = \frac{1}{2m} \|y\|^2 + C \sum_{j \in I_i} \xi_j,$$

$$\mathcal{X}_i = \{x \in \mathbb{R}^{d+n} \mid b_j \langle y, a_j \rangle \geq 1 - \xi_j, \xi_j \geq 0, \forall j \in I_i\},$$

where I_i is a set of indices such that $\cup_{i=1}^m I_i = \{1, \dots, n\}$, $I_i \cap I_j = \emptyset$ for $i \neq j$ and $j \in I_i$ if and only if \mathcal{X}_i contains inequalities associated with the data (a_j, b_j) . Note that each set $\mathcal{X}_i^j = \{x \in \mathbb{R}^{d+n} \mid b_j \langle y, a_j \rangle \geq 1 - \xi_j, \xi_j \geq 0\}$ is the intersection of two halfspaces, the projection onto which can be computed in a few steps (see the next subsection).

5.1.2 Projection onto the Intersection of Two Halfspaces

Given $v \in \mathbb{R}^d$, we are interested in solving the following optimization problem:

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \quad & \frac{1}{2} \|w - v\|^2 \\ \text{s.t.} \quad & \langle a, w \rangle \leq b, \quad w_i \geq 0, \end{aligned} \tag{5.2}$$

where $a \in \mathbb{R}^d$, $b \in \mathbb{R}$ and w_i is the i -th component of the vector w .

The two half-spaces divide the \mathbb{R}^d space into four parts. Therefore, there are only four cases to consider:

1. $\langle a, v \rangle \leq b$ and $v_i \geq 0$.

In this case, v is already in the intersection and $w = v$.

2. $\langle a, v \rangle > b$ and $v_i < 0$.

In this case, v is projected onto the intersection of the two hyperplanes $\{w \mid \langle a, w \rangle = b\}$ and $\{w \mid w_i = 0\}$. Finding such a projection is equivalent to solving the following optimization problem:

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \quad & \frac{1}{2} \|w - v\|^2 \\ \text{s.t.} \quad & \langle a, w \rangle = b, \quad w_i = 0. \end{aligned} \tag{5.3}$$

The Lagrangian of the problem (5.3) is

$$\mathcal{L}(w, \theta, \zeta) = \frac{1}{2} \|w - v\|^2 + \theta \left(\sum_{j=1}^d a_j w_j - b \right) + \zeta w_i,$$

where $\theta, \zeta \in \mathbb{R}$ are Lagrange multipliers. Differentiating the Lagrangian

and setting it to zero gives the optimality condition,

$$w_i^* - v_i + a_i \theta^* + \zeta^* = 0,$$

$$w_j^* - v_j + a_j \theta^* = 0, \quad \text{for } j \neq i.$$

From the primal feasibility, we have the following relations:

$$w_i^* = 0 \implies \zeta^* = v_i - a_i \theta^*,$$

$$\sum_{j=1}^n a_j w_j^* = \sum_{j \neq i} a_j w_j^* = \sum_{j \neq i} a_j (v_j - a_j \theta^*) = b \implies \theta^* = \frac{\sum_{j \neq i} a_j v_j - b}{\sum_{j \neq i} a_j^2}.$$

Therefore, the projection is given by

$$w_j^* = \begin{cases} 0 & \text{if } j = i, \\ v_j - a_j \theta^* & \text{otherwise.} \end{cases}$$

Let $w^* = [w_1^*, \dots, w_d^*]^T$.

3. $\langle a, v \rangle > b$ and $v_i \geq 0$.

In this case, v will be projected either onto the hyperplane $\{w \mid \langle a, w \rangle = b\}$ or onto the intersection of the two hyperplanes $\{w \mid \langle a, w \rangle = b\}$ and $\{w \mid w_i = 0\}$. Let \hat{w} be the projection of v onto $\{w \mid \langle a, w \rangle = b\}$, i.e.,

$$\hat{w} = v - \left(\frac{\langle a, v \rangle - b}{\|a\|^2} \right) a.$$

The projection of v in this case is given by

$$w = \begin{cases} \hat{w} & \text{if } \hat{w}_i \geq 0, \\ w^* & \text{otherwise.} \end{cases}$$

4. $\langle a, v \rangle \leq b$ and $v_i < 0$.

Let \hat{w} be the projection of v onto the hyperplane $\{w \mid w_i = 0\}$, i.e.,

$$\hat{w} = v - (v_i - b) e_i,$$

where $e_i \in \mathbb{R}^d$ is the vector whose i -th component is one and all the

other components are zero. Then, the projection of v is given by

$$w = \begin{cases} \hat{w} & \text{if } \langle a, \hat{w} \rangle \leq b, \\ w^* & \text{otherwise.} \end{cases}$$

5.1.3 Simulations

In the section, we perform some experiments with our DRP algorithm. We refer to our DRP algorithm applied on SVMs as **DrSVM**. The purpose of the experiments is to verify the convergence and to show in how many iterations the proposed method can actually arrive at consensus in distributed settings. We use the DRP algorithm in (3.1a)-(3.1b) and its variant in (3.26a)-(3.26d) with the stepsize $\alpha_k = \frac{1}{k+1}$ for $k \geq 0$. We vary the number of batches b as 1, 100 or 1000 to observe the different convergence speed, where $b = 1$ corresponds to the algorithm in (3.1a)-(3.1b). To show the effect of connectivity, we compare two different time-invariant network topologies, i) a completely connected graph (clique) and ii) a 3-regular expander graph. The 3-regular expander graph is a sparse graph that has strong connectivity with every node having degree 3.

We use 3 text classification data sets for our experiments. The data sets were kindly provided by Thorsten Joachims (see [8] for their descriptions). Table 5.1 lists the statistics of the data sets. All of the data sets are from binary document classification. Since the data sets used here are very unlikely separable, we use the formulation (5.1) with $C = 1$. In each experiment the number n of constraints is divided among the agents equally (if n is not divisible by m , the m -th agent gets the remainder). To estimate the generalization (or testing) performance, we split the data and use 80% for training and 20% for testing.

DrSVM is implemented with C/C++ and all experiments were performed on a 64-bit machine running Fedora 16 with an Intel Core 2 Quad Processor Q9400 and 8G of RAM. The experiments are not performed on a real networked environment so we do not consider delays and node/link failures that may exist in networks.

For stopping criteria, we first run a centralized random incremental projection [37] on the 80% training set with $b = 1$ until the relative error of

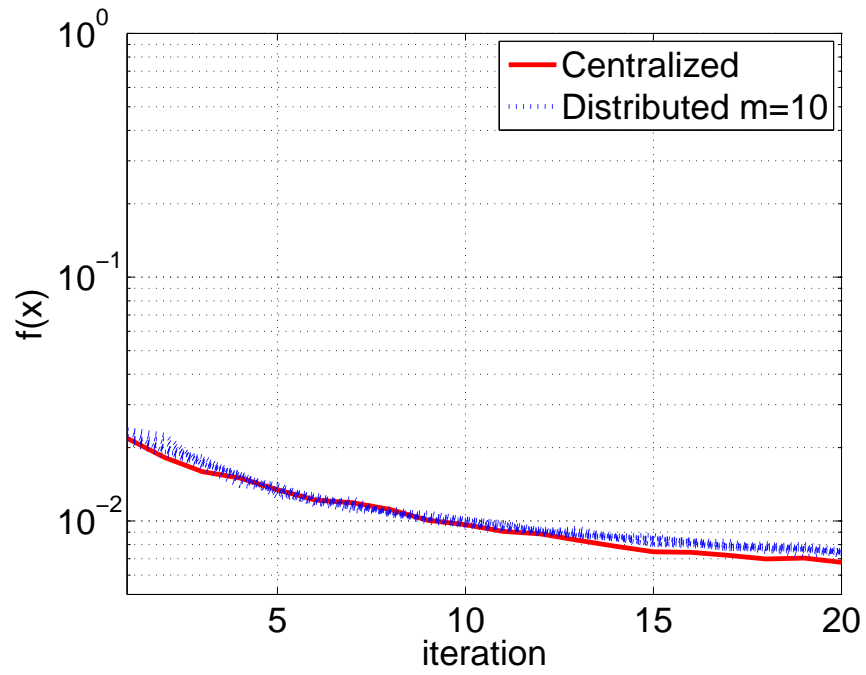
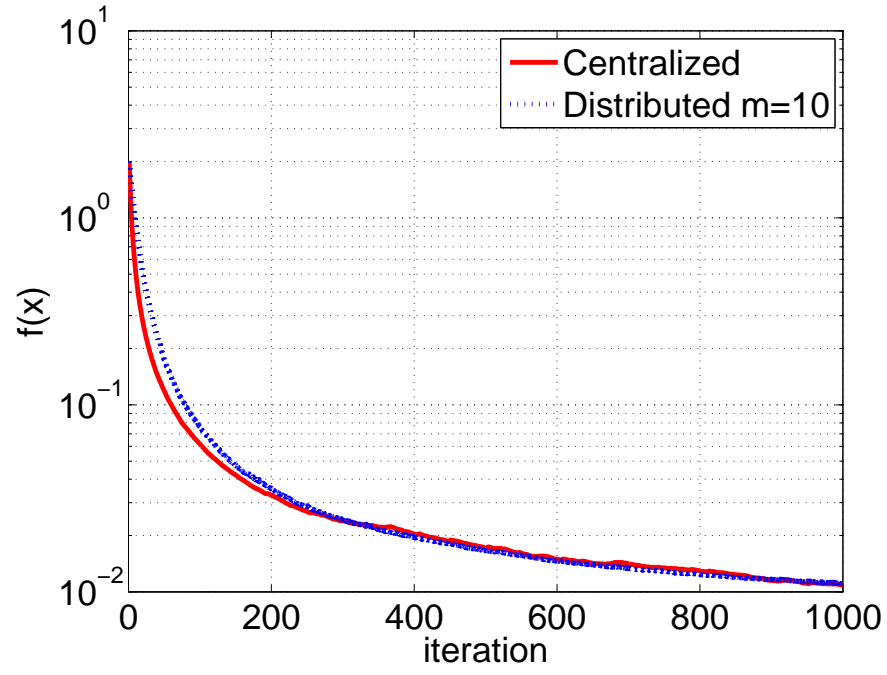


Figure 5.1: $f(x)$ vs. iteration on `astro-ph` with 10 agents when batch size b is 1 (top) and 100 (bottom)

Table 5.1: The statistics of three text classification data sets: n is the number of examples, d is the number of features, and s is the sparsity of data.

Data set	Statistics		
	n	d	s
astro-ph	62,369	99,757	0.08%
CCAT	804,414	47,236	0.16%
C11	804,414	47,236	0.16%

Table 5.2: The results of **DrSVM** with two different graph topologies (clique and 3-regular expander graph) and three different numbers of agents ($m = 2, 6, 10$): t_{acc} is the target test accuracy; b is the number of projections per iteration. The table shows the number of iterations for all agents to reach the target test accuracy, where ‘-’ indicates that the algorithm did not converge within the 20,000 maximum iteration limit.

Data set	t_{acc}	b	$m = 2$	Clique		3-regular expander	
				$m = 6$	$m = 10$	$m = 6$	$m = 10$
astro-ph	0.95	1	1,055	695	697	695	-
		100	11	8	11	11	11
		1000	2	2	2	2	2
CCAT	0.91	1	752	511	362	517	-
		100	11	10	8	10	8
		1000	2	3	2	3	3
C11	0.97	1	1,511	1,255	799	1,226	-
		100	16	17	12	17	15
		1000	2	2	2	2	2

objective values in two consecutive iterations is less than 0.001; i.e.,

$$|f(x(k)) - f(x(k+1))|/f(x(k)) < 0.001.$$

We then measure the test accuracy of the final solution on the remaining 20% test set, which will become the target test accuracy t_{acc} . For experiments in the distributed setting, we measure the test accuracy of every agent’s solution at the end of every iteration. If every solution at a certain iteration satisfies the target value t_{acc} , we conclude that the agents arrived at a consensus and the algorithm converged. The maximum number of iterations in each simulation is limited to 20,000.

Table 5.2 shows the results. As we do more projections per iteration, the total number of iterations required for convergence is less, regardless of the number of agents. For the given stopping criteria, it seems that fewer iterations are needed for **DrSVM** to converge as the number of agents increases. We can also observe the effect of network connectivity. When all the other parameters (m and b) are the same, for most of the cases, the number of iterations required for the 3-regular expander graph to converge is greater than or equal to that for the clique.

The table reports the number of iterations required for all the agents to achieve the target test accuracy. Therefore, the total number of projections is *at most* the number of iterations times m times b . This is because no projection is required if the current estimate is already in the selected constraint component. For example, the total number of projections for **astro-ph** with $m = 6$ and $b = 100$ is at most $4,800 (= 8 \times 6 \times 100)$.

The runtime (or the number of calculations) of the algorithm is not only proportional to the number of projections, but also to the number of gradient updates. For example, for **astro-ph** with $m = 6$ and $b = 1$, the total number of projections is $4,170 (= 695 \times 6 \times 1)$, while the total number of gradient updates is $4,170 (= 695 \times 6)$. For the same example with $m = 6$ and $b = 100$, the total number of projections is $4,800 (= 8 \times 6 \times 100)$, but the total number of gradient updates is only $48 (= 8 \times 6)$. In any case, the numbers are much smaller than the number 62,369 of the training data points. This shows that **DrSVM** can quickly find a good quality solution before examining the training samples even once.

To show the convergence (and consensus) of the algorithm, we plot in

Figure 5.1 the objective value $f(x)$ of centralized random projection (CRP) and DRP with 10 agents for example **astro-ph**. Note that we plot the convergence of the objective value instead of the solution. This is because CRP and DRP may converge to different optimal points as the problem (5.1) may not have a unique optimal solution. For Figure 5.1(a) and 5.1(b), we applied the random projection once and 100 times per iteration, respectively. From the figures, we can observe that the objective values of CRP and the 10 agents in DRP are almost identical. The final objective of Figure 5.1(b) seems smaller than that of Figure 5.1(a). This is because the stepsize at iteration 1000 is too small.

5.2 Distributed Robust Control

In this section, we apply our GRP algorithm to a distributed robust model predictive control (MPC) example. The purpose of this experiment is to verify the error bound obtained in Section 4.4 and to show in how many iterations the proposed method actually arrives at almost-consensus in various distributed settings.

A linear, time-invariant, discrete-time system is given by the following state equation for $t = 1, \dots, T$:

$$x(t) = Ax(t-1) + Bu(t), \quad (5.4)$$

where

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix},$$

with initial state $x(0) = [7, 0]'$.

The goal of the agents on the network is to find an optimal control $\mathbf{u} \triangleq [u(1), \dots, u(T)]'$ of the system (5.4) over time $t = 1, \dots, T$ with some random terminal constraints. The distributed optimization problem is given as the following:

$$\min_{\mathbf{u}} f(\mathbf{u}) = \sum_{i=1}^m f_i(\mathbf{u}) \quad \text{s.t. } \mathbf{u} \in \mathcal{X}. \quad (5.5)$$

Here,

$$f_i(\mathbf{u}) = \sum_{t=1}^T \|x(t) - z_i\|^2 + ru(t), \text{ for } i = 1, \dots, m,$$

is the local objective of agent i and $r > 0$ is a control parameter. Hence, the agents on the network jointly find a control \mathbf{u} that generates a trajectory $x(t)$, for $t = 1, \dots, T$ such that the trajectory minimizes the deviations from the points $z_i \in \mathbb{R}^2$ together with the control effort. The information about the points z_i for $i = 1, \dots, m$ are private and only agent i knows where the z_i is located.

The constraint set \mathcal{X} is a set of control inputs that satisfies the following constraints:

$$\|u(t)\|_\infty \leq 2, \quad \text{for } t = 1, \dots, T, \quad (5.6a)$$

$$x(t) = Ax(t-1) + Bu(t), \quad \text{for } t = 1, \dots, T, \quad (5.6b)$$

$$x(0) = [7, 0]', \quad (5.6c)$$

$$\max_{\ell=1,2,3,4} \{(a_\ell + \delta_\ell)'x(T) - b_\ell\} \leq 0. \quad (5.6d)$$

The constraint (5.6a) states that the control inputs are constrained so that $\|u(t)\|_\infty \leq 2$, for $t = 1, \dots, T$. The constraints (5.6b)-(5.6c) describe the system dynamics. (5.6d) refers to the random terminal constraints given by the linear inequalities $(a_\ell + \delta_\ell)'x(T) \leq b_\ell$ and the perturbations δ_ℓ are uniform random vectors in boxes $\|\delta_\ell\|_\infty \leq \beta_\ell$. Note that $u(t)$, for $t = 1, \dots, T$, are the only variables here since $x(t)$, for $t = 1, \dots, T$, are fully determined by the state equation (5.6b)-(5.6c) once $u(t)$, for $t = 1, \dots, T$, are given. For this problem, $\mathcal{X} = \mathcal{X}_i$.

The constraint set \mathcal{X} is uncertain and not exactly known in advance since the perturbations are uniform random vectors in boxes. To apply the GRP algorithm (4.3a)-(4.3c) in solving this robust optimal control problem, at iteration k , each agent I_k and J_k draws a realization of one of the linear inequality terminal constraints, and each of them projects its current iterate on the selected constraint. Subsequently, they perform their projections onto the box constraint (5.6a).

Since the uncertainty exists in a box, the problem (5.5) has an equivalent quadratic programming (QP) formulation. Note that the following represen-

tations are all equivalent:

$$(a_\ell + \delta_\ell)'x(T) \leq b_\ell, \quad \forall(\delta_\ell : \|\delta_\ell\|_\infty \leq \beta_\ell) \quad (5.7a)$$

$$\Leftrightarrow \max_{\|\delta_\ell\|_\infty \leq \beta_\ell} \delta_\ell'x(T) \leq b_\ell - a_\ell'x(T) \quad (5.7b)$$

$$\Leftrightarrow a_\ell'x(T) + \beta_\ell|[x(T)]_1| + \beta_\ell|[x(T)]_2| \leq b_\ell. \quad (5.7c)$$

Therefore, the inequality (5.6d) admits an equivalent representation of (5.7c) by a system of linear inequalities with additional variables t_1 and t_2 :

$$-t_j \leq [x(T)]_j \leq t_j, \quad \text{for } j = 1, 2, \quad (5.8a)$$

$$\max_{\ell=1,2,3,4} \{a_\ell'x(T) + \beta_\ell t_1 + \beta_\ell t_2 - b_\ell\} \leq 0. \quad (5.8b)$$

This alternative representation is only available since we are considering simple box uncertainty sets for the sake of comparison. Note that our GRP algorithm is applicable not just to box uncertainty but to more complicated perturbations such as Gaussian or other distributions.

In the experiment, we use $m = 4$ and $m = 10$ agents with $T = 10$ and $r = 0.1$. We solve the problem on three different network topologies, namely, clique, cycle and star (see Figure 5.2). For the agent selection probability, we use uniform distribution. That is, at each iteration, one of the m agents is uniformly selected and the selected agent uniformly picks one of its neighbors. Table 5.3 shows the second largest eigenvalue λ of \bar{W} for the three network topologies when $m = 4$ and $m = 10$. When m is larger, we can see that λ is very close to one for all of the three cases.

We evaluate the algorithm performance by carrying out 100 Monte-Carlo runs, each with 40,000 iterations for $m = 4$ and 100,000 iterations for $m = 10$. For the stepsize, we use either a diminishing one ($1/\Gamma_i(k)$) or a constant $\alpha_i = 10^{-5}$ for $m = 4$ and $\alpha_i = 10^{-6}$ for $m = 10$.

In Figures 5.3 and 5.4, we depict $\frac{1}{m} \sum_{i=1}^m \|\mathbf{u}_i(k) - \mathbf{u}^*\|^2$ over 40,000 and 100,000 iterations when the diminishing and constant stepsize are used, respectively. The optimal solution \mathbf{u}^* was obtained by solving the equivalent QP problem (i.e., problem (5.5) with constraints (5.6a)-(5.6c) and (5.8a)-(5.8b)) using a commercial QP solver.

We can observe for both cases that the errors go down quickly. An interesting observation is that the network topology does not affect the algorithm

Table 5.3: Number of agents and λ

m	Clique	Cycle	star
4	0.6667	0.7500	0.8333
10	0.8889	0.9809	0.9444

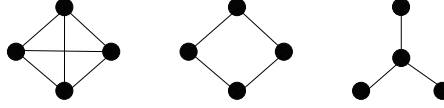


Figure 5.2: Clique (left), cycle (center) and star (right) graph used for communication topology ($m = 4$)

performance when the diminishing stepsize is used. When the constant stepsize is used for the $m = 4$ case, star network converges much slower than the other two networks. This is because the agent selection probability γ_i is different for the center node and the peripheral nodes. As the bound in Proposition 4.2 captures, a more aggressive stepsize α_i should have been used for the peripheral nodes. For the $m = 10$ case, however, the difference is not as clearly visible as in the $m = 4$ case. This can be explained by almost the same spectral gap $1 - \sqrt{\lambda}$ (as shown in Proposition 4.2 and Table 5.3).

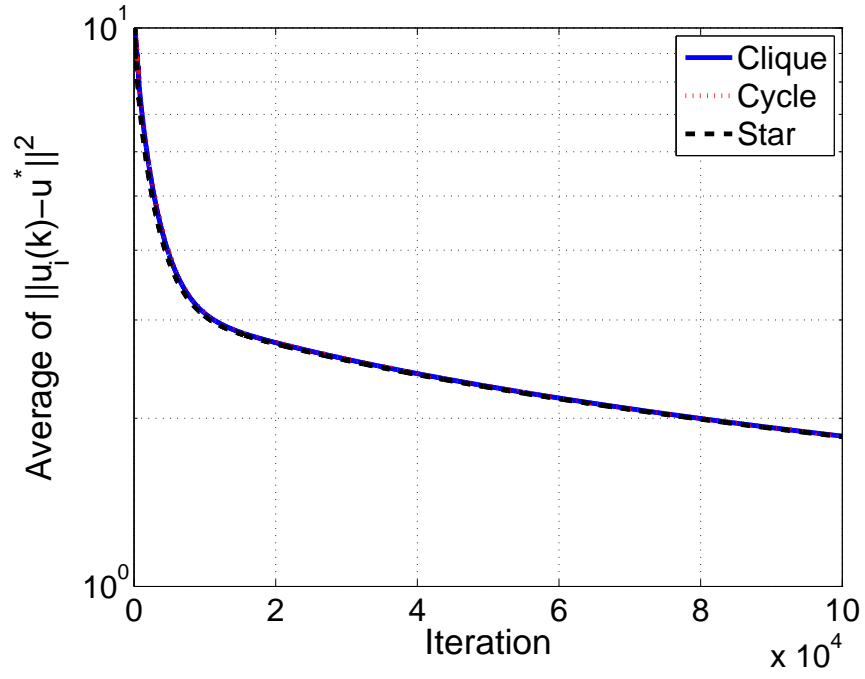
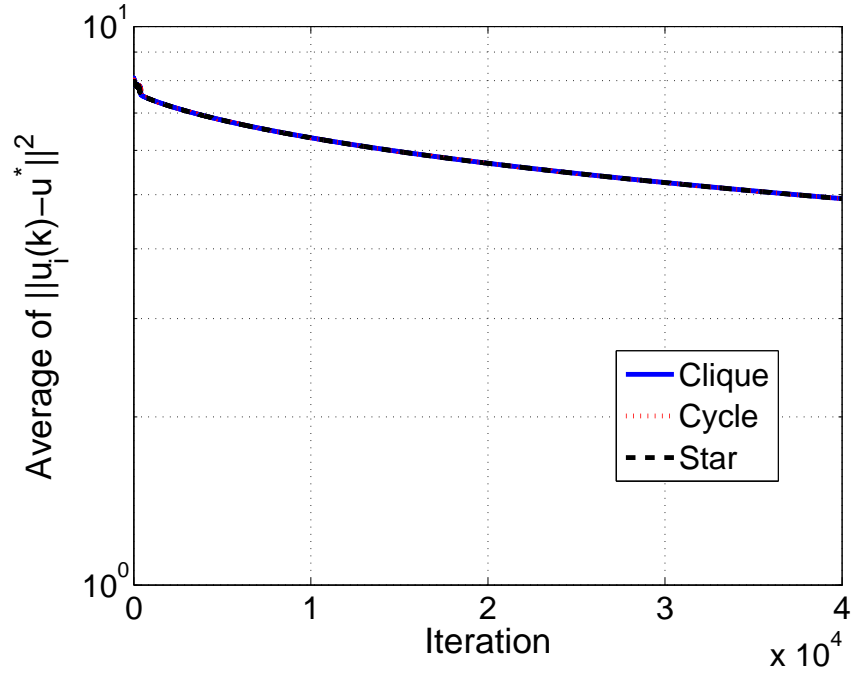


Figure 5.3: Iteration vs. $\frac{1}{m} \sum_{i=1}^m \|u_i(k) - u^*\|^2$ with a diminishing stepsize when $m = 4$ (top) and $m = 10$ (bottom)

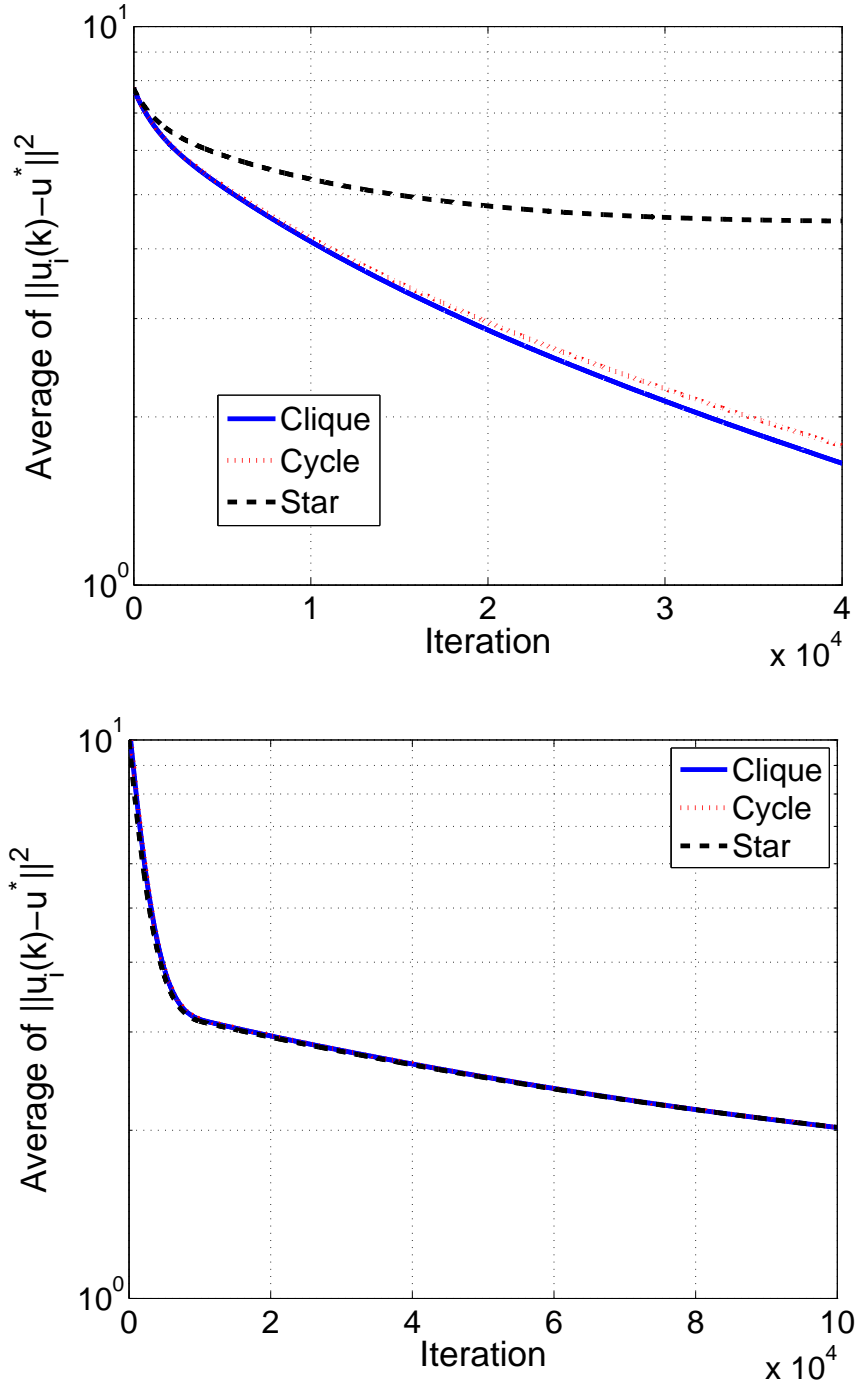


Figure 5.4: Iteration vs. $\frac{1}{m} \sum_{i=1}^m \|\mathbf{u}_i(k) - \mathbf{u}^*\|^2$ with a constant stepsize when $m = 4$ (top) and $m = 10$ (bottom)

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In this part of the thesis, we have proposed and analyzed distributed gradient algorithms with random incremental projections for a network of agents. We considered the most general cases, where each agent has a unique and different objective and constraint. The proposed algorithms are applicable to problems where the whole constraint set is not known in advance but its component is revealed in time, or where the projection onto the whole set is computationally prohibitive. We have established convergence with probability 1 for the algorithms when the objective is convex under typical assumptions. For the GRP algorithm, we have additionally established an error bound when constant stepsizes are used. Also, we have provided variants of the algorithms using a mini-batch of consecutive projections and established its convergence with probability 1. Experiments on three text classification benchmarks using SVMs were performed to verify the performance of the proposed algorithm. We have also provided a simulation result for a distributed robust model predictive control problem.

There are a few directions to extend this research.

6.1 Time-varying Mini-batch Random Projection

As a variant of the mini-batch random projection algorithms in (3.26a)-(3.26d) and (4.50a)-(4.50d), we can think of a time-varying mini-batch random projection algorithm. Instead of fixing the batch size as b , we use a sequence b_k which can vary from 1 to the number of components $|I_i|$ in the set \mathcal{X}_i .

When $b_k = 1$, each iteration of the algorithm is inexpensive as we sample and project on just one constraint component. The setting can make great progress initially, but is often slow as it approaches a solution. In contrast,

when $b_k = |I_i|$, the algorithm achieves steady convergence at the expense of full projections.

Therefore, we can think of a hybrid method of random projections and full projections that exploits the benefits of both approaches.

6.2 Random Projections for Convex Feasibility Problems

In the convex feasibility problem, we are interested in solving the following optimization problem:

$$\begin{aligned} \min_x \quad & 0 \\ \text{s.t. } \quad & x \in \mathcal{X}, \quad \mathcal{X} \triangleq \bigcap_{i=1}^m \mathcal{X}_i. \end{aligned}$$

If the problem is feasible, any point in the set \mathcal{X} is considered as an optimal solution. The feasible set \mathcal{X} is the intersection of a number of components, where each component is represented as a convex inequality.

The objective is to investigate several different random projection algorithms that would converge in some sense. For the three algorithms below, ω_k is defined as a random variable that takes values from the set $\{1, \dots, m\}$. The idea of time-varying batch can also be used here.

1. Average of subsequent projections

$$\begin{aligned} \psi_k^0 &= x_{k-1} \\ \psi_k^i &= \Pi_{\mathcal{X}_{\omega_k}}[\psi_k^{i-1}] \quad \text{for } i = 1, \dots, b_k \\ x_k &= \frac{1}{b_k} \sum_{j=1}^{b_k} \psi_k^j \end{aligned}$$

In this approach, the previous point ψ_k^{i-1} is used for projection onto the new selected component set \mathcal{X}_{ω_k} . At each iteration, we repeat this for b_k times and set the next iterate x_k as the average of the projected points.

2. Average of independent projections

$$\begin{aligned}\psi_k^0 &= x_{k-1} \\ \psi_k^i &= \Pi_{\mathcal{X}_{\omega_k}}[\psi_k^0] \quad \text{for } i = 1, \dots, b_k \\ x_k &= \frac{1}{b_k} \sum_{j=1}^{b_k} \psi_k^j\end{aligned}$$

In this approach, we keep selecting a new component but the projection is made from the previous iterate x_{k-1} . Again, we repeat this for b_k times and set the next iterate x_k as the average of the projected points.

3. Projection onto the intersection of subsets

$$\begin{aligned}\psi_k^0 &= x_{k-1} \\ \mathcal{X}_k &= \bigcap_{i=1}^{b_k} \mathcal{X}_{\omega_k} \\ x_k &= \Pi_{\mathcal{X}_k}[\psi_k^0]\end{aligned}$$

In this approach, we first select b_k components and define a subset \mathcal{X}_k which is the intersection of the selected component sets. Then, we consider the projection of the previous iterate x_{k-1} on \mathcal{X}_k .

6.3 Distributed Optimization over Time-varying Graphs

As a variant of the Distributed Random Projection algorithm in (3.1a)-(3.1b), we can think of distributed optimization over time-varying graphs with much weaker assumptions on the weight matrices $W(k)$. The objective is to show that the algorithm converges without the doubly stochasticity assumption on $W(k)$.

We define $N_i^{\text{in}}(k)$ and $N_i^{\text{out}}(k)$ for the in and out neighborhoods of agent i at time k . The neighborhoods include the agent i itself. Formally, we define

$$\begin{aligned}N_i^{\text{in}}(k) &= \{j \mid (j, i) \in E(k)\} \cup \{i\}, \\ N_i^{\text{out}}(k) &= \{j \mid (i, j) \in E(k)\} \cup \{i\},\end{aligned}$$

and $d_i(t) = |N_i^{\text{out}}(k)|$.

Every node i maintains vector estimate sequences $x_i(t)$, $w_i(t)$ in \mathbb{R}^d and a scalar estimate sequence $y_i(t)$ in \mathbb{R} . With $y_i(0) = 1$ for all $i \in V$, these quantities are updated by the following rules:

$$\begin{aligned} w_i(k+1) &= \sum_{N_i^{\text{out}}(k)} \frac{x_j(k)}{d_j(k)}, \\ y_i(k+1) &= \sum_{N_i^{\text{out}}(k)} \frac{y_j(k)}{d_j(k)}, \\ z_i(k+1) &= \frac{w_i(k+1)}{y_i(k+1)}, \\ x_i(k+1) &= \Pi_{\mathcal{X}_i^{\Omega_i(k)}}[w_i(k+1) - \alpha(k+1)g_i(k+1)], \end{aligned}$$

where $g_i(k+1)$ is a subgradient of the function f_i at $z_i(k+1)$ and $\alpha(k+1)$ is a nonincreasing stepsize at time $k+1$. We refer to this algorithm as subgradient-push random projection method. This algorithm follows from the paper [45], where the subgradient-push method is developed for unconstrained distributed optimization.

At each iteration, each agent i simply broadcasts $\frac{x_j(t)}{d_j(t)}$ and $\frac{y_j(t)}{d_j(t)}$ to its out neighborhood and no other communication is needed. The stepsize $\alpha(k+1)$ needs to satisfy the following conditions:

$$\sum_{k=1}^{\infty} \alpha(k) = \infty, \quad \sum_{k=1}^{\infty} \alpha^2(k) < \infty.$$

The Q-strong connectivity of $G(k)$ (Assumption 3.3) still needs to be held. This is a typical assumption for the control of multi-agent systems.

Part II

Epoch Gradient Descent for Smoothed Hinge-loss SVMs

CHAPTER 7

INTRODUCTION

Support vector machines (SVMs) are popular classification tools. Given a set of example-label pairs $\{(x_i, y_i)\}_{i=1}^m$, $x_i \in \mathbb{R}^n$ and $y_i \in \{+1, -1\}$, SVMs find $w \in \mathbb{R}^n$ that solves the following optimization problem. (A bias term is included in w for convenience.)

$$\begin{aligned} \min f(w) &= \frac{1}{2} \|w\|^2 \\ \text{s.t. } y_i w^T x_i &\geq 1, \quad \forall i \in 1, \dots, m. \end{aligned} \tag{7.1}$$

If an optimal solution to this problem exists, it is a maximum-margin hyperplane separating the two classes.

SVMs have strong theoretical backgrounds. It is shown in [44] that SVMs directly minimize VC dimensions; therefore, the solutions obtained usually have a good generalization performance. Many researchers have developed a number of SVM algorithms for the primal form (7.1) and its dual form. Especially, it has been shown in [46] that SVMs can also be efficiently trained in the primal space. Our focus in this part of the thesis is to address the primal problem (7.1) with a simple first order method.

If a function f constrained on a closed convex set \mathcal{C} is λ -strongly convex and its gradient is Lipschitz continuous with a constant L , we can obtain its optimal solution with a linear convergence rate by just using a simple projected gradient descent method. That is, let $w_0 \in \mathbb{R}^n$ and w^* be the minimizer of the function f . Then, we have

$$f(w_k) - f(w^*) \leq q^k [f(w_0) - f(w^*)], \tag{7.2}$$

where $q = 1 - \lambda/L$ if the iterate $\{w_k\}_{k \geq 1}$ is updated via the following rule:

$$w_{k+1} = \Pi_{\mathcal{C}} \left[w_k - \frac{1}{L} \nabla f(w_k) \right], \tag{7.3}$$

where $\Pi_{\mathcal{C}}[x]$ is the projection of x on the set \mathcal{C} , i.e. $\Pi_{\mathcal{C}}[x] = \arg \min_{v \in \mathcal{C}} \|v - x\|^2$. The convergence rate (7.2) is determined by q . If q is smaller (or the condition number L/λ is smaller), we have a faster convergence.

In fact, the formulation (7.1) is not so suitable for obtaining this linear convergence. It has a nice quadratic objective function but the projection onto the constraint set is too complicated. The problem (7.1) can be reformulated to the following unconstrained optimization problem with a penalty parameter $\gamma > 0$:

$$\min f(w) = \frac{1}{2}\|w\|^2 + \gamma \sum_{i=1}^m \mathcal{L}(w; (x_i, y_i)), \quad (7.4)$$

where $\mathcal{L}^i(w)$ is a non-negative loss function representing the violation of the inequality constraint associated with the data point (x_i, y_i) . Note that this penalized form can also handle the cases when the given data sets are nonseparable.

We divide the function (7.4) by γm and let $\lambda = \frac{1}{\gamma m}$. This does not change the solution of (7.4).

$$\min f(w) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{m} \sum_{i=1}^m \mathcal{L}(w; (x_i, y_i)). \quad (7.5)$$

This objective function is λ -strongly convex with respect to the Euclidean norm $\|\cdot\|$ due to the quadratic term $(\lambda/2)\|w\|^2$. Therefore, the minimizer of f always exists and is unique.

Additional properties of the function would be determined by the loss term $\mathcal{L}(\cdot)$. Frequently used loss functions are:

- Hinge-loss: $\max\{0, 1 - y_i \langle w, x_i \rangle\}$
- Squared hinge-loss: $\max\{0, 1 - y_i \langle w, x_i \rangle\}^2$
- Logistic-loss: $(1/\log 2) \log(1 + \exp\{-y_i \langle w, x_i \rangle\})$

These three loss functions are depicted in Figure 7.1.

Hinge-loss is an exact penalty function. That is, there exists a constant $\gamma > 0$ such that the minimum of (7.5) with hinge-loss coincides with the minimum of (7.1). However, the function is non-smooth and its gradient is not Lipschitz continuous. Squared hinge-loss has Lipschitz continuous

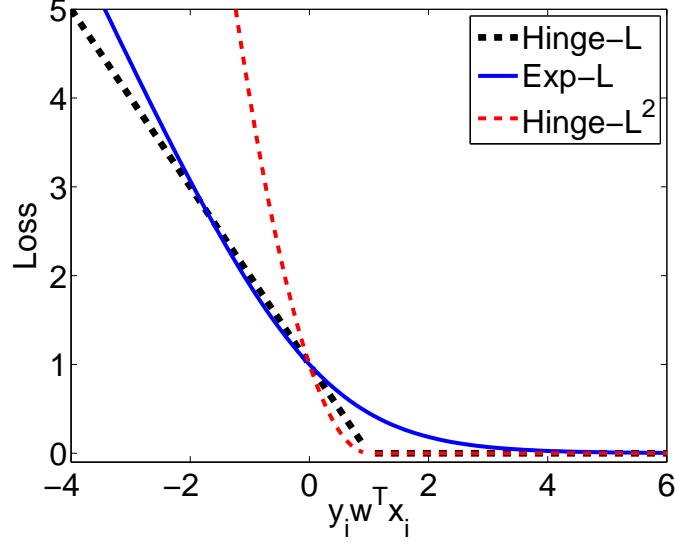


Figure 7.1: Frequently used loss functions

gradient and logistic-loss is a smooth surrogate of hinge-loss. However, these loss functions are usually not exact. That is, there is usually no choice of $\gamma < \infty$ for which the penalized function yields the optimal solution.

The trade-off here is an exact solution versus a linear convergence rate. If we use hinge-loss, we will obtain an exact solution but the convergence rate of the simple gradient descent method will not be linear. If we use the other loss functions with Lipschitz continuous gradient, the convergence rate will be linear but the solution will not be exact. To obtain an exact solution as well as a fast convergence, we propose a hybrid approach, *epoch gradient descent*. For this purpose, we use hinge-loss and Nesterov's smooth approximation technique [47].

In epoch gradient descent, we start from solving a very smooth problem (obtained by the technique [47]) to obtain a quick and inexact solution with the linear convergence rate. To obtain an exact solution, this solution is fed to the original non-smooth problem. There would be no gain in the convergence rate, but given a very good quality starting point, the total *time* required for achieving the same quality solution would be less.

Using this approach naturally arises the following two questions:

- (1) How can we choose a proper smoothing parameter?
- (2) Up to what accuracy level does the smooth approximated problem have

to be solved?

To answer question (1), we solve a sequence of smooth problems in our epoch gradient descent. That is, we start from a very smooth problem, and gradually decrease the amount of smoothing. The solution to the current epoch is fed to the problem of the next epoch. This reduces the risk of choosing an inappropriate smoothing parameter. For question (2), we analyze the algorithm’s error bound at each iteration. This helps us determine the number of iterations required for the smoothed problem at the current epoch.

Thesis Organization

The second part of this thesis is organized as follows. In Chapter 8, we find a smoothed hinge-loss function using the Nesterov’s smooth approximation technique. We verify the smoothed function is a uniform smooth approximation of a hinge-loss and it also has Lipschitz continuous gradient. In Chapter 9, we present our simple/epoch gradient descent algorithms on the smooth SVM and provide results on the error bounds. In Chapter 10, we provide experimental results of our proposed methods on some text classification benchmarks. Chapter 11 contains conclusions and some future directions.

Previous Work

We only review some primal gradient-descent based methods which are the most relevant to our work. To the best of our knowledge, no gradient descent methods on SVMs have exploited both strong-convexity and gradient Lipschitz continuity.

NORMA [48] and an algorithm by [49] are stochastic gradient descent methods for solving linear/nonlinear SVMs. It minimizes hinge-loss penalized SVM objectives to find some f in a Hilbert space. The algorithm requires $\mathcal{O}(1/\epsilon^2)$ iterations to converge to an ϵ -accurate solution. **Pegasos** [50] is also a stochastic gradient descent algorithm on hinge-loss penalized SVMs. It operates on a mini-batch A of the training data, with a batch size $|A|$ varies from 1 to m . It converges to an ϵ -accurate solution in $\mathcal{O}(d/\lambda\epsilon)$ iterations, where d is the maximum number of nonzero features in each example.

Even though online methods require more iterations to obtain an ϵ -accurate solution, the computation overhead per iteration is less ($\mathcal{O}(n)$ while batch methods require $\mathcal{O}(mn)$). Therefore, the total training time of online

methods might be shorter than that of batch methods. The problem of on-line methods is that they converge fast initially but slow as approaching a solution. On the other hand, batch methods converge steadily at the expense of full gradient evaluations.

We would like to point out that there is no reason to always prefer online methods to batch methods although researchers have preferred online methods due to their simplicity. In fact, we can never achieve the fastest $\mathcal{O}(\log_q \epsilon)$ convergence rate if we update only with partial information on the gradients. Moreover, in some situations (for example, high-dimensional training problems with only a small set of training examples is available), batch methods would be more advantageous.

There are some other works which proposed to use smooth loss functions for the sake of differentiability [51–53]. However, none of the smooth formulations are appropriate for our purpose.

CHAPTER 8

LOSS FUNCTION WITH LIPSCHITZ CONTINUOUS GRADIENT

In this chapter, we use Nesterov's smoothing technique [47] to find a proper loss function for our purpose and prove that the smoothed loss function has Lipschitz continuous gradient.

8.1 Smoothed Hinge-loss $\mathcal{L}_\mu(w)$

Let $\mathcal{L}_\mu(w)$ be the smoothed loss function with parameter μ and $\mathcal{L}(w)$ be the hinge loss function.

Theorem 8.1 *We define a smoothed loss function \mathcal{L}_μ associated with a data point (x, y) as*

$$\mathcal{L}_\mu(w) = \begin{cases} \frac{1}{4\mu}(1 + \mu - yw^T x)^2, & \text{if } |1 - yw^T x| \leq \mu \\ 1 - yw^T x, & \text{if } yw^T x < 1 - \mu \\ 0, & \text{otherwise,} \end{cases}$$

where $\mu > 0$ is a smoothing parameter. Then, \mathcal{L}_μ is a uniform smooth approximation of hinge loss \mathcal{L} . That is, for any $w \in \mathbb{R}^n$, we have

$$\mathcal{L}(w) \leq \mathcal{L}_\mu(w) \leq \mathcal{L}(w) + \frac{\mu}{2}.$$

Proof Note that the hinge-loss function can be equivalently represented like the following:

$$\mathcal{L}(w) = \max\{0, 1 - y\langle w, x \rangle\} = \max_{\alpha \in \mathbb{S}_2} \{\alpha^T A w + \alpha^T b\},$$

where α is constrained on a 2-dimensional probability simplex $\mathbb{S}_2 = \{\alpha \in \mathbb{R}^2 \mid \alpha(1) + \alpha(2) = 1, \alpha(1) \geq 0, \alpha(2) \geq 0\}$; $A^T = [0 \ -yx] \in \mathbb{R}^{n \times 2}$, and $b = [0 \ 1]^T \in \mathbb{R}^2$.

Given this representation, we use the smoothing technique in [47]. We introduce a smoothing parameter $\mu \in \mathbb{R}_+$ and a strongly-convex function $\frac{\mu}{2}\|\alpha\|^2$. Then,

$$\mathcal{L}_\mu(w) = \max_{\alpha \in \mathbb{S}_2} \left\{ \alpha^T A w + \alpha^T b + (1 - \|\alpha\|^2) \frac{\mu}{2} \right\} \quad (8.1)$$

is the smooth approximation of the hinge-loss. Note that the constant term $\frac{\mu}{2}$ is added so that $\mathcal{L}_\mu(w) = 0$ when there is no constraint violation. The closed form solution to the optimization problem (8.1) gives the desired result. (See [54] for details.)

Since α is constrained in a 2-dimensional probability simplex, $0 \leq \|\alpha\|^2 \leq 1$. Therefore, for any $w \in \mathbb{R}^n$, we have

$$\mathcal{L}(w) \leq \mathcal{L}_\mu(w) \leq \mathcal{L}(w) + \frac{\mu}{2}. \quad (8.2)$$

■

8.2 The Lipschitz Continuity of ∇f_μ

We are interested in finding a solution w_μ^* for the following smooth SVM problem:

$$\min_{w \in \mathbb{R}^n} \left\{ f_\mu(w) = \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^m \mathcal{L}_\mu^i(w) \right\}, \quad (8.3)$$

where $\mathcal{L}_\mu^i(w)$ is the smoothed hinge-loss function associated with the data point (x_i, y_i) . Let w_μ^* be an optimal solution of the function f_μ and w^* be the optimal solution of the original function f .

Theorem 8.2 *The gradient of the function f_μ is Lipschitz continuous with constant L_μ , where*

$$L_\mu = \lambda + \frac{\sum_{i=1}^m \|x_i\|^2}{m\mu}. \quad (8.4)$$

Proof Let $\mathcal{L}_\mu(w)$ attain a maximum value at $\alpha(w)$, i.e.

$$\alpha(w) = \arg \max_{\alpha \in \mathbb{S}_2} \left\{ \alpha^T A_i w + \alpha^T b_i + (1 - \|\alpha\|^2) \frac{\mu}{2} \right\}.$$

From the optimality condition of a convex function [30] *et al.*, for any

$w_1, w_2 \in \mathbb{R}^n$, we have

$$\langle A_i w_1 + b_i - \mu \alpha(w_1), \alpha - \alpha(w_1) \rangle \leq 0, \quad \forall \alpha \in \mathbb{S}_2,$$

$$\langle A_i w_2 + b_i - \mu \alpha(w_2), \alpha - \alpha(w_2) \rangle \leq 0, \quad \forall \alpha \in \mathbb{S}_2.$$

Using the above inequalities with $\alpha := \alpha(w_2)$ and $\alpha := \alpha(w_1)$ respectively, we have

$$\langle A_i w_1 + b_i - \mu \alpha(w_1), \alpha(w_2) - \alpha(w_1) \rangle \leq 0,$$

$$\langle A_i w_2 + b_i - \mu \alpha(w_2), \alpha(w_1) - \alpha(w_2) \rangle \leq 0.$$

By adding the above two inequalities and arranging the terms appropriately, we have

$$\begin{aligned} \mu \|\alpha(w_1) - \alpha(w_2)\|^2 &\leq \langle \alpha(w_1) - \alpha(w_2), A_i(w_1 - w_2) \rangle \\ &\leq \|\alpha(w_1) - \alpha(w_2)\| \|A_i\| \|w_1 - w_2\|, \end{aligned}$$

and therefore,

$$\|\alpha(w_1) - \alpha(w_2)\| \leq \frac{\|A_i\|}{\mu} \|w_1 - w_2\|. \quad (8.5)$$

Using the inequality (8.5), the following relation holds for any $w_1, w_2 \in \mathbb{R}^n$:

$$\begin{aligned} &\|\nabla \mathcal{L}_\mu(w_1) - \nabla \mathcal{L}_\mu(w_2)\| \\ &= \|A_i^T(\alpha(w_1) - \alpha(w_2))\| \leq \|A_i\| \|\alpha(w_1) - \alpha(w_2)\| \\ &\leq \frac{\|A_i\|^2}{\mu} \|w_1 - w_2\| = \frac{\|x_i\|^2}{\mu} \|w_1 - w_2\|, \end{aligned}$$

where the last equality is from the fact that $\|A_i\| = \|x_i\|$. This shows that $\nabla \mathcal{L}_\mu(w)$ is Lipschitz continuous with constant $\frac{\|x_i\|^2}{\mu}$.

Also, we know that the gradient of the function $\frac{\lambda}{2} \|w\|^2$ is Lipschitz continuous with constant λ . Therefore, $\nabla f_\mu(w)$ is Lipschitz continuous with constant L_μ , where

$$L_\mu = \lambda + \frac{1}{m} \sum_{i=1}^m \frac{\|x_i\|^2}{\mu}.$$

■

It is straightforward from the relation (8.2) that the following also holds for any $w \in \mathbb{R}^n$:

$$f(w) \leq f_\mu(w) \leq f(w) + \frac{\mu}{2}. \quad (8.6)$$

Figure 8.1 shows $y_i w^T x_i$ versus the smoothed hinge-loss function \mathcal{L}_μ with two different smoothing parameters $\mu = 2$ and $\mu = 5$. Note that we also penalize the constraints when $1 \leq y_i w^T x_i \leq 1 + \mu$.

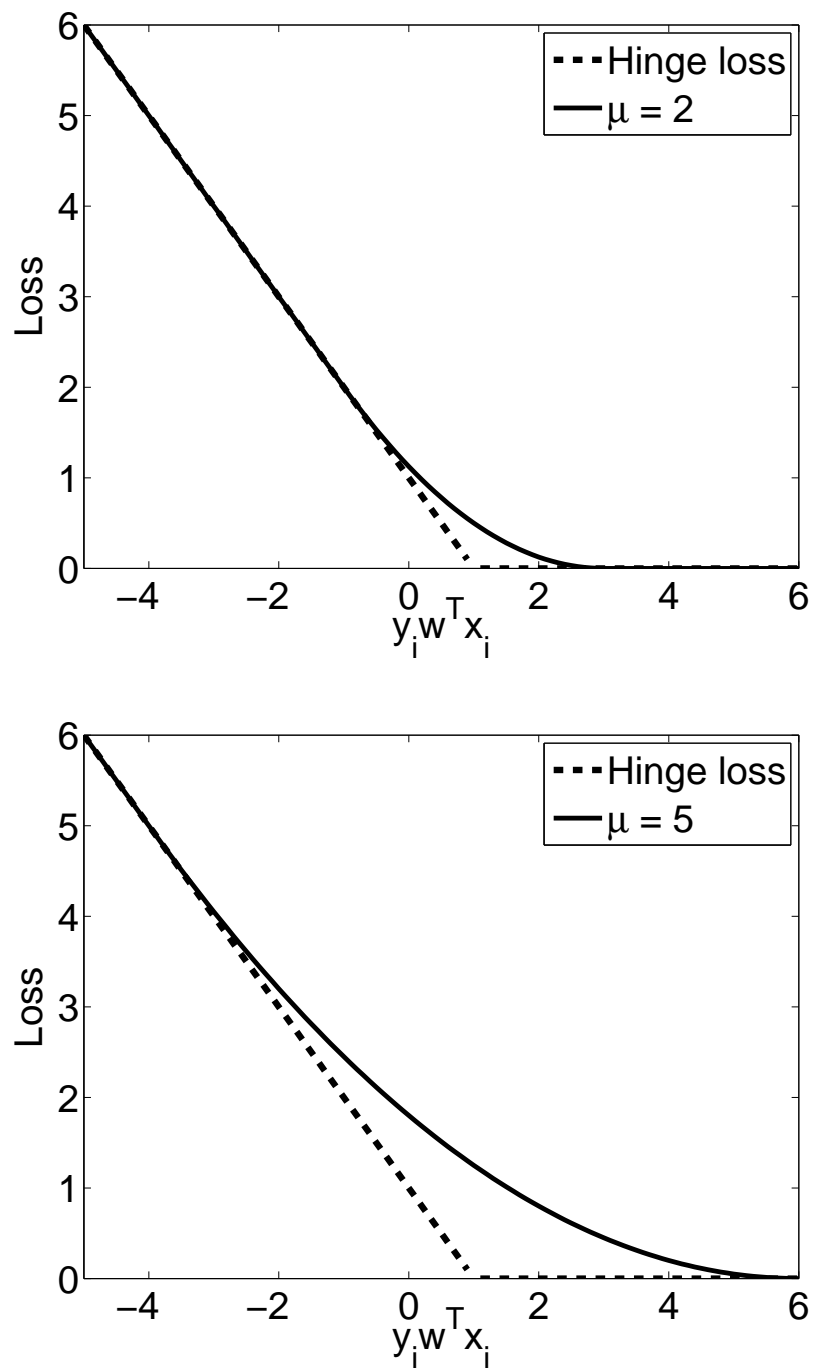


Figure 8.1: Smoothed hinge-loss function with two different smoothing parameters $\mu = 2$ and $\mu = 5$

CHAPTER 9

EFFICIENT ALGORITHMS

In this chapter, we present our simple and epoch gradient descent algorithms applied on the smooth SVM problems and provide results on the error bounds generated by these algorithms.

9.1 Simple Gradient Descent with Strong-Convexity and Lipschitz Continuity

In the most basic version of gradient descent algorithms that we consider, each iterate is updated via the following rule:

$$w_{k+1} = w_k - \alpha \nabla f_\mu(w_k), \quad (9.1)$$

where the step size $\alpha = \frac{1}{L_\mu}$ can be used since the function f_μ has a Lipschitz continuous gradient with a constant L_μ . The initial iterate w_0 is selected randomly from \mathbb{R}^n .

In Theorem 9.1, we state the error bound of this simple gradient descent algorithm applied on the problem (8.3).

Theorem 9.1 *An optimal solution to the problem (8.3) always exists and is unique. Using the gradient descent algorithm (9.1) on f_μ for K iterations yields the following bound:*

$$f_\mu(w_K) - f_\mu(w_\mu^*) \leq q(\mu)^K [f_\mu(w_0) - f_\mu(w_\mu^*)], \quad (9.2)$$

where $q(\mu) = 1 - \lambda/L_\mu$ and w_μ^* is the optimal solution to the problem (8.3). Therefore, the number of iterations required for an ϵ -accurate solution is $\mathcal{O}(\log_q \epsilon)$.

Proof See [31].

In Lemma 9.1, we analyze the difference between the iterates generated by the simple gradient algorithm on the problem (8.3) and the optimal solution w^* of the hinge-loss penalized problem.

Lemma 9.1 *Let w^* be the optimal solution of the hinge-loss penalized problem. After K iterations of the simple gradient descent algorithm (9.1), the following bound holds:*

$$f_\mu(w_K) - f(w^*) \leq q(\mu)^K [f_\mu(w_0) - f(w^*)] + (1 - q(\mu)^K) \frac{\mu}{2}, \quad (9.3)$$

where $q(\mu) = 1 - \lambda/L_\mu$.

Proof Let w_μ^* be the optimal solution of f_μ . Applying the inequality (9.2), we have:

$$\begin{aligned} f_\mu(w_K) - f(w^*) &= f_\mu(w_K) - f_\mu(w_\mu^*) + f_\mu(w_\mu^*) - f(w^*) \\ &\leq q(\mu)^K [f_\mu(w_0) - f_\mu(w_\mu^*)] + f_\mu(w_\mu^*) - f(w^*) \\ &= q(\mu)^K [f_\mu(w_0) - f(w^*)] + (1 - q(\mu)^K) [f_\mu(w_\mu^*) - f(w^*)]. \end{aligned} \quad (9.4)$$

We have $f_\mu(w_\mu^*) \leq f_\mu(w^*)$ and from (8.6), $f(w^*) \leq f_\mu(w^*) \leq f(w^*) + \frac{\mu}{2}$. Therefore,

$$\begin{aligned} f_\mu(w_\mu^*) - f(w^*) &= f_\mu(w_\mu^*) - f_\mu(w^*) + f_\mu(w^*) - f(w^*) \\ &\leq f_\mu(w^*) - f(w^*) \leq \frac{\mu}{2}. \end{aligned} \quad (9.5)$$

Combining the inequalities (9.4) and (9.5) gives the desired result. ■

Note that the value $f_\mu(w_K) - f(w^*)$ in the inequality (9.3) is bounded by the sum of two error terms. Since $q(\mu) < 1$, the first term $q(\mu)^K [f_\mu(w_0) - f(w^*)]$ decay to zero as $K \rightarrow \infty$. This ensures that the first term can be reduced to an arbitrary small level. However, the second term $(1 - q(\mu)^K) \frac{\mu}{2}$ does not decay to zero as $K \rightarrow \infty$. In fact, this term upper bounds the difference between $f_\mu(w_\mu^*)$ and $f(w^*)$.

The lemma also illustrates a trade-off between the convergence rate and the accuracy of the solution w_μ^* compared to w^* . For a bigger μ (or a smaller L_μ), the gradient descent method will converge to w_μ^* faster as the ratio $q(\mu)$

is smaller. However, the solution w_μ^* gets far from the original solution w^* . Therefore, if we need an exact solution, the reduction of μ is necessary. This leads to an epoch gradient descent algorithm in the subsequent section.

9.2 Epoch Gradient Descent Algorithm

From the fact that there is a trade-off between the convergence rate and the accuracy of the solution, we develop an epoch gradient descent algorithm. We will start from a very large smoothing parameter μ_1 and gradually decrease it. The solution will converge very close to the optimal of the corresponding hinge-loss SVM since the sequence of the smoothing parameter $\{\mu_t\}$ eventually converges to zero. At the same time, we can benefit from the fast convergence of very smooth problems at initial stages.

The pseudo code of this algorithm is represented in Algorithm 1. A sequence of non-increasing smoothing parameters $\{\mu_t\}_{t \geq 1}$ and the corresponding sequence of smooth problems $\{f_{\mu_t}\}_{t \geq 1}$ is generated. At first, we pick a random point $w_0^1 \in \mathbb{R}^n$ and set μ_1 to some value depending on problems. Then, we decide the duration K_t of an epoch t . At each epoch t , we apply the simple gradient descent K_t times on problem f_{μ_t} . Then, we decrease the smoothing parameter from μ_t to μ_{t+1} . The last iterate $w_{K_t}^t$ of epoch t is used as an initial point w_0^{t+1} of epoch $t+1$. We continue this with a new problem $f_{\mu_{t+1}}$. The algorithm terminates when the error gets smaller than some predefined ϵ .

This algorithm basically requires to specify a rule for deciding at what iteration to decrease the smoothing parameter and to what extent it should be decreased. Let us consider (9.3) at epoch t with $\mu := \mu_t$.

$$f_{\mu_t}(w_K^t) - f(w^*) \leq q(\mu_t)^K [f_{\mu_t}(w_0^t) - f(w^*)] + (1 - q(\mu_t)^K) \frac{\mu_t}{2}. \quad (9.6)$$

Suppose at the beginning of epoch t (with $K = 0$),

$$f_{\mu_t}(w_0^t) - f(w^*) > \frac{\mu_t}{2}. \quad (9.7)$$

Since $0 < q(\mu_t) < 1$, $f_{\mu_t}(w_0^t) - f(w^*)$ decreases when multiplied with $q(\mu_t)^K$ for $K \geq 0$. Therefore, there exists $K > 0$ for which $q(\mu_t)^K [f_{\mu_t}(w_0^t) - f(w^*)]$ drops below $(1 - q(\mu_t)^K) \frac{\mu_t}{2}$. We take this K as our K_t , i.e., the smallest K

Algorithm 1 EPOCH GRADIENT DESCENT

Pick $w_0^1 \in \mathbb{R}^n$, $\epsilon, \epsilon_f \in \mathbb{R}_+$ and $\mu_1 \in \mathbb{R}_+$.
Set $t := 1$
while $\mu_t > \epsilon$ **do**
 Set $L_{\mu_t} = \lambda + \sum_{i=1}^m \frac{\|x_i\|^2}{m\mu_t}$.
 Set $K_t = \lceil \frac{1}{-\log_2 q(\mu_t)} \rceil$.
 for $k = 0$ to K_t **do**
 $w_{k+1}^t = w_k^t - \frac{1}{L_{\mu_t}} \nabla f_{\mu_t}(w_k^t)$
 if $|f_{\mu_t}(w_{k+1}^t) - f_{\mu_t}(w_k^t)| / f_{\mu_t}(w_k^t) < \epsilon_f$ **then**
 $K_t := k$
 Break **for** loop
 end if
 end for
 Set $\mu_{t+1} := 0.5\mu_t$ and $w_0^{t+1} := w_{K_t}^t$
 Set $t := t + 1$
end while

for which the following inequality holds:

$$q(\mu_t)^K [f_{\mu_t}(w_0^t) - f(w^*)] < (1 - q(\mu_t)^K) \frac{\mu_t}{2}. \quad (9.8)$$

As we do not have knowledge on the value of $f(w^*)$, we assume that $f(w^*) \approx 0$. This assumption is reasonable since $\lambda(\leq 1/m)$ is generally very small and the loss term evaluated at the optimal point w^* is close to zero. Setting $f(w^*) = 0$ at (9.7) and (9.8), we have,

$$q(\mu_t)^K \frac{\mu_t}{2} < q(\mu_t)^K f_{\mu_t}(w_0^t) < (1 - q(\mu_t)^K) \frac{\mu_t}{2} < \frac{\mu_t}{2}. \quad (9.9)$$

Solving this inequality with respect to K gives the duration K_t :

$$K_t = \left\lceil -\frac{1}{\log_2 q(\mu_t)} \right\rceil. \quad (9.10)$$

In practice, K_t generated by this equation is usually larger than necessary (due to the assumption $f(w^*) = 0$). We stop updating if the relative error in function values does not change more than ϵ_f .

At the end of epoch t (at $K = K_t$), our error $f_{\mu_t}(w_{K_t}^t) - f(w^*)$ is guaranteed to be smaller than μ_t . This is clear from the relations (9.6) and (9.9).

Next, we establish the error bound of our epoch gradient descent algorithm in the following theorem using Lemma 9.1 and calculate the total number of

gradient updates required to obtain an ϵ -accurate solution.

Theorem 9.2 *At each epoch t of the EPOCH GRADIENT DESCENT ALGORITHM, the following error bound holds:*

$$\begin{aligned} f_{\mu_t}(w_{K_t}^t) - f(w^*) &\leq \prod_{i=1}^t q(\mu_i)^{K_i} [f_{\mu_1}(w_0^1) - f(w^*)] \\ &\quad + \sum_{i=1}^{t-1} \prod_{j=i+1}^t q(\mu_j)^{K_j} (1 - q(\mu_i)^{K_i}) \frac{\mu_i}{2} \\ &\quad + (1 - q(\mu_t)^{K_t}) \frac{\mu_t}{2}, \end{aligned} \quad (9.11)$$

where $q(\mu_i) = 1 - \lambda/L_{\mu_i}$. The total number of gradient update for an ϵ -accurate solution is $\mathcal{O}(\frac{1}{\epsilon})$.

Proof From the algorithm, we have $w_0^{t+1} = w_{K_t}^t$.

$$f_{\mu_{t+1}}(w_0^{t+1}) - f(w^*) = f_{\mu_{t+1}}(w_{K_t}^t) - f(w^*) \leq f_{\mu_t}(w_{K_t}^t) - f(w^*). \quad (9.12)$$

Also, using the inequality (9.3) with $\mu := \mu_t$ and $K := K_t$, we have

$$f_{\mu_t}(w_{K_t}^t) - f(w^*) \leq q(\mu_t)^{K_t} [f_{\mu_t}(w_0^t) - f(w^*)] + (1 - q(\mu_t)^{K_t}) \frac{\mu_t}{2}, \quad (9.13)$$

where $q(\mu_t) = 1 - \lambda/L_{\mu_t}$.

We use induction on t to prove our results. Note that the inequality (9.11) holds trivially for $t = 1$ from (9.13). Next, assume that the inequality (9.11) holds for some t . From the inequality (9.13), consider the case for $t + 1$.

$$\begin{aligned} f_{\mu_{t+1}}(w_{K_{t+1}}^{t+1}) - f(w^*) &\leq q(\mu_{t+1})^{T_{k+1}} [f_{\mu_{t+1}}(w_0^{t+1}) - f(w^*)] \\ &\quad + (1 - q(\mu_{t+1})^{K_{t+1}}) \frac{\mu_{t+1}}{2}. \end{aligned}$$

Applying the induction hypothesis to the right-hand side of the above

inequality together with (9.12),

$$\begin{aligned}
& f_{\mu_{t+1}}(w_{K_{t+1}}^{t+1}) - f(w^*) \\
& \leq q(\mu_{t+1})^{K_{t+1}} \left[\prod_{i=1}^t q(\mu_i)^{K_i} [f_{\mu_1}(w_0^1) - f(w^*)] \right. \\
& \quad + \sum_{i=1}^{t-1} \prod_{j=i+1}^t q(\mu_j)^{K_j} (1 - q(\mu_i)^{K_i}) \frac{\mu_i}{2} \\
& \quad \left. + (1 - q(\mu_t)^{K_t}) \frac{\mu_t}{2} \right] + (1 - q(\mu_{t+1})^{K_{t+1}}) \frac{\mu_{t+1}}{2} \\
& = \prod_{i=1}^{t+1} q(\mu_i)^{K_i} [f_{\mu_1}(w_0^1) - f(w^*)] \\
& \quad + \sum_{i=1}^t \prod_{j=i+1}^{t+1} q(\mu_j)^{K_j} (1 - q(\mu_i)^{K_i}) \frac{\mu_i}{2} \\
& \quad + (1 - q(\mu_{t+1})^{K_{t+1}}) \frac{\mu_{t+1}}{2}.
\end{aligned}$$

Given that the error at the end of epoch t is bounded by $\mu_t = \mu_1(0.5)^t$, to obtain an ϵ -accurate solution we need $N := \lceil \log_2 \frac{\mu_1}{\epsilon} \rceil$ epochs. Let $S := \sum_{i=1}^m \|x_i\|^2$. Using the equation (8.4),

$$\sum_{t=1}^N K_t = \sum_{t=1}^N \frac{1}{\log_2 q(\mu_t)} = \sum_{t=1}^N \frac{1}{\log_2 \left(1 + \frac{m\lambda\mu_1}{S} (0.5)^t\right)}.$$

Note if $\{\alpha_t\}_{t \geq 1}$ is a decreasing sequence, the following relation holds:

$$\log_2(1 + \alpha_t) \geq c\alpha_t,$$

where $c = \frac{\log_2(1+\alpha_1)}{\alpha_1}$. Using the above relation with $\alpha_t := \frac{m\lambda\mu_1}{S}(0.5)^t$ and $c := \log_2(1 + \frac{m\lambda\mu_1}{2S}) / \frac{m\lambda\mu_1}{2S}$, we have

$$\begin{aligned}
\sum_{t=1}^N K_t & \leq \sum_{t=1}^N \frac{1}{c \frac{m\lambda\mu_1}{S} (0.5)^t} = \frac{1}{c} \frac{S}{m\lambda\mu_1} \sum_{t=1}^N 2^t \\
& = \frac{1}{\log_2 \left(1 + \frac{m\lambda\mu_1}{2S}\right)} \left(\frac{\mu_1}{\epsilon} - 1 \right), \tag{9.14}
\end{aligned}$$

which is $\mathcal{O}(\frac{1}{\epsilon})$. ■

It is clear from this Theorem that all of the three terms on the right hand side of the inequality (9.11) decay to zero as $t \rightarrow 0$. Also, the total number of iterations $\mathcal{O}(\frac{1}{\epsilon})$ is optimal. We cannot do any better than this (in terms of the order, not the running time in practice) if the original problem has only strong-convexity [55]. Note that the inequality (9.14) is a theoretical bound. In practice, we need far fewer iterations in general.

CHAPTER 10

EXPERIMENTAL RESULTS

In the chapter, we analyze the performance of our gradient descent and epoch-gradient descent algorithms on primal SVMs penalized with a smooth hinge-loss. We perform experiments on the most commonly used text classification benchmarks.

We compare our implementation with an SVM solver **Pegasos** [50]. **Pegasos** uses a stochastic incremental subgradient or subgradient method on primal SVMs penalized with a hinge-loss. It is one of the state-of-the-art solvers and one of the most relevant methods with ours.

Table 10.1 lists the statistics of data sets. We use 4 data sets for our experiments. The data sets were kindly provided by Thorsten Joachims (see [8] for their descriptions). All of the data sets are from binary document classification. Note that the training vectors in three of the four data sets are really sparse.

To estimate the generalization (or testing) performance, we split the data and use 80% for training and 20% for testing. We use $\lambda = \frac{1}{m}$ for all the experiments. Given that $\lambda \leq \frac{1}{m\bar{\gamma}}$ with $\bar{\gamma}$ not known, $\lambda = \frac{1}{m}$ seems to be a reasonable choice.

We compare our two different implementations of linear smooth SVM solvers with **Pegasos**.

- **SSVM**: The simple gradient descent method applied on primal SVMs penalized with a μ -smoothed hinge-loss. A smoothing parameter μ is chosen in advance and fixed throughout the simulation. In the experiment, μ is set such that the ratio $q(\mu) = 1 - \frac{\lambda}{L_\mu}$ is 0.5.
- **SSVM-epoch**: The epoch gradient descent method on primal SVMs penalized with a μ_t -smoothed hinge-loss. At epoch t , the hinge-loss is smoothed with a different smoothing parameter μ_t . In the experiments, we use $\mu_1 = \sum_{i=1}^n \|x_i\|^2/9$ and $\mu_t = \mu_1(0.5)^t$.

Table 10.1: The statistics of the four data sets

Data set	Statistics		
	m	n	s
astro-ph	62,369	99,757	0.08%
cov1	522,911	54	22.22%
CCAT	804,414	47,236	0.16%
C11	804,414	47,236	0.16%

Table 10.2: The results of **SSVM** and **SSVM-epoch** (\dagger indicates a different stopping criterion is used)

Data set	SSVM			SSVM-epoch \dagger		
	n_i	t	e_{gen}	n_i	t	e_{gen}
astro-ph	3	0.08	0.20	529	11.16	0.04
cov1	5	0.15	0.29	526	15.49	0.23
CCAT	2	0.65	0.16	276	69.99	0.09
C11	5	1.64	0.03	736	162.49	0.02

- **Pegasos-batch**: The subgradient descent method on primal SVMs penalized with the ordinary hinge-loss.
- **Pegasos-online**: The stochastic incremental subgradient descent method on primal SVMs penalized with the ordinary hinge-loss. At each iteration, only one example is randomly selected and used for update.

All the above methods (except **Pegasos** which is downloadable from the author's homepage¹) are implemented with C/C++ and all experiments were performed on a 64-bit machine running Fedora 16 with an Intel Core 2 Quad Processor Q9400 and 8G of RAM.

For a stopping criterion, we use a relative error of objective values in two consecutive iterations. We stop either if the error is reduced to 0.1% (i.e., $|f(w_k) - f(w_{k+1})|/f(w_k) < 0.001$) or if the maximum number of iterations (500) is reached. Note that for **SSVM-epoch**, we use a different stopping criterion. **SSVM-epoch** stops if μ_t (the error upper bound) is reduced to $\epsilon = 0.01$. We use $\epsilon_f = 0.001$.

¹<http://ttic.uchicago.edu/~shai/code>

Table 10.3: The results of **Pegasos-batch** and **Pegasos-online**

Data set	Pegasos-batch			Pegasos-online		
	n_i	t	e_{gen}	n_i	t	e_{gen}
astro-ph	52	2.01	0.03	898110	0.66	0.03
cov1	499	90.44	0.31	2091645	0.98	0.23
CCAT	322	175.31	0.05	10296496	8.64	0.05
C11	285	138.86	0.03	9652965	7.65	0.03

Table 10.2 and 10.3 shows the results. Note that **SSVM**, **SSVM-epoch** and **Pegasos-batch** are batch methods but **Pegasos-online** is an online method. We can observe that **SSVM** is the fastest one of the four in terms of both time and the number of iterations. Its generalization performance is not very good compared to the other three methods. This is because **SSVM** finds solutions to approximated problems.

If we compare the timing results of the two batch methods, **SSVM-epoch** and **Pegasos-batch**, **SSVM-epoch** is comparable or faster in 3 out of 4 data sets than the non-smooth implementation **Pegasos-batch** even if a stricter stopping criterion is used for **SSVM-epoch**. This indicates that solving a sequence of very smooth problems at initial stages is advantageous. However, if we compare **SSVM-epoch** and **Pegasos-online**, **Pegasos-online** is faster. This is because of the extreme sparsity of the data set. In the future, we will test our algorithms on non-sparse data sets as well.

Also, the generalization performance of **SSVM-epoch** is comparable with the two hinge-loss SVM implementations **Pegasos-batch** and **Pegasos-online**. This indicates that the solutions of **SSVM-epoch** indeed converge to the optimal of hinge-loss penalized SVMs.

CHAPTER 11

CONCLUSIONS AND FUTURE WORK

In this part of the thesis, we studied a primal SVM penalized with smooth loss functions. We showed that the simple gradient descent method applied on this approximated SVM converges linearly due to the strong-convexity and Lipschitz property. Observing that there is a trade-off between the convergence rate and the amount of error, we developed an epoch based gradient descent algorithm to achieve a fast convergence and an accurate solution at the same time. We also showed the convergence behavior of this algorithm and the number of iterations required to obtain an ϵ accurate solution is $\mathcal{O}(\frac{1}{\epsilon})$. Experiments on four text classification benchmarks were performed on the smooth SVM.

There are two directions to extend this research. First, we can extend this idea to nonlinear SVMs and see how much improvement we can get from the smooth loss functions. Second, though we showed batch methods are still promising, it is computationally too heavy to consider all the examples at every iteration if training vectors are sparse and the number is huge. Therefore, we can also consider online methods with the smooth loss functions. We will not obtain the linear convergence rate if gradients are evaluated with only a subset of training examples but total training time may be reduced depending on problems.

REFERENCES

- [1] E. Camponogara, D. Jia, B. Krogh, and S. Talukdar, “Distributed model predictive control,” *IEEE Control Systems*, vol. 22, no. 1, pp. 44–52, February 2002.
- [2] R. Bekkerman, M. Bilenko, and J. Langford, *Scaling Up Machine Learning: Parallel and Distributed Approaches*. Cambridge University Press, 2011.
- [3] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- [4] B. Johansson, “On distributed optimization in networked systems,” Ph.D. dissertation, Royal Institute of Technology (KTH), tRITA-EE 2008:065, 2008.
- [5] S. S. Ram, V. V. Veeravalli, and A. Nedić, “Distributed non-autonomous power control through distributed convex optimization,” in *IEEE INFOCOM*, 2009, pp. 3001–3005.
- [6] M. Rabbat and R. D. Nowak, “Distributed optimization in sensor networks,” in *IPSN*, 2004, pp. 20–27.
- [7] S. Kar and J. Moura, “Distributed consensus algorithms in sensor networks: Quantized data and random link failures,” *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1383–1400, March 2010.
- [8] T. Joachims, “Training linear SVMs in linear time,” in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 217–226.
- [9] B. Johansson, M. Rabi, and M. Johansson, “A simple peer-to-peer algorithm for distributed optimization in sensor networks,” in *Proceedings of the 46th IEEE Conference on Decision and Control*, Dec. 2007, pp. 4705–4710.
- [10] S. S. Ram, A. Nedić, and V. V. Veeravalli, “Incremental stochastic sub-gradient algorithms for convex optimization,” *SIAM J. on Optimization*, vol. 20, no. 2, pp. 691–717, June 2009.

- [11] A. Nedić and A. Ozdaglar, “On the rate of convergence of distributed asynchronous subgradient methods for multi-agent optimization,” in *Proceedings of the 46th IEEE Conference on Decision and Control*, New Orleans, USA, 2007, pp. 4711–4716.
- [12] A. Nedić and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.
- [13] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. Tsitsiklis, “Distributed subgradient methods and quantization effects,” in *Proceedings of 47th IEEE Conference on Decision and Control*, December 2008, pp. 4177–4184.
- [14] I. Lobel and A. Ozdaglar, “Distributed subgradient methods for convex optimization over random networks,” *IEEE Transactions on Automatic Control*, vol. 56, no. 6, pp. 1291–1306, June 2011.
- [15] S. S. Ram, A. Nedić, and V. V. Veeravalli, “Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization,” *Journal of Optimization Theory and Applications*, vol. 147, pp. 516–545, 2010.
- [16] 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.
- [17] J. Duchi, A. Agarwal, and M. Wainwright, “Dual averaging for distributed optimization: Convergence analysis and network scaling,” *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, March 2012.
- [18] A. Nedić, A. Ozdaglar, and P. Parrilo, “Constrained consensus and optimization in multi-agent networks,” *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, April 2010.
- [19] K. Srivastava, A. Nedić, and D. Stipanović, “Distributed constrained optimization over noisy networks,” in *Proceedings of the 49th IEEE Conference on Decision and Control (CDC)*, Dec. 2010, pp. 1945–1950.
- [20] K. Srivastava and A. Nedić, “Distributed asynchronous constrained stochastic optimization,” *IEEE J. Sel. Topics. Signal Process.*, vol. 5, no. 4, pp. 772–790, 2011.
- [21] I. Lobel, A. Ozdaglar, and D. Feijer, “Distributed multi-agent optimization with state-dependent communication,” *Mathematical Programming*, vol. 129, no. 2, pp. 255–284, 2011.

- [22] T. Alamo, R. Tempo, and E. Camacho, “Randomized strategies for probabilistic solutions of uncertain feasibility and optimization problems,” *IEEE Trans. Autom. Control*, vol. 54, no. 11, pp. 2545–2559, Nov 2009.
- [23] G. C. Calafiore, “Random convex programs,” *SIAM J. Optimiz.*, vol. 20, no. 6, pp. 3427–3464, Dec. 2010.
- [24] B. Polyak, “Random algorithms for solving convex inequalities,” in *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, ser. Studies in Computational Mathematics, Y. C. Dan Butnariu and S. Reich, Eds. Elsevier, 2001, vol. 8, pp. 409 – 422.
- [25] A. Nedić, “Random projection algorithms for convex set intersection problems,” in *Proc. of the 49th IEEE Conference on Decision and Control*, 2010, pp. 7655–7660.
- [26] A. Jadbabaie, J. Lin, and A. Morse, “Coordination of groups of mobile autonomous agents using nearest neighbor rules,” *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988 – 1001, June 2003.
- [27] R. Olfati-Saber and R. Murray, “Consensus problems in networks of agents with switching topology and time-delays,” *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520 – 1533, Sep. 2004.
- [28] L. Xiao, S. P. Boyd, and S.-J. Kim, “Distributed average consensus with least-mean-square deviation,” *J. Parallel Distrib. Comput.*, vol. 67, no. 1, pp. 33–46, 2007.
- [29] A. Olshevsky and J. N. Tsitsiklis, “Convergence speed in distributed consensus and averaging,” *SIAM J. Control Optim.*, vol. 48, no. 1, pp. 33–55, Feb. 2009.
- [30] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Athena Scientific, 2003.
- [31] B. Polyak, *Introduction to Optimization*. Optimization software, Inc., Publications division, New York, 1987.
- [32] J. N. Tsitsiklis, “Problems in decentralized decision making and computation,” Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, 1984.
- [33] J. Tsitsiklis, D. Bertsekas, and M. Athans, “Distributed asynchronous deterministic and stochastic gradient optimization algorithms,” *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803 – 812, Sep. 1986.

- [34] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, Belmont, 1997.
- [35] J. V. Burke and M. C. Ferris, “Weak sharp minima in mathematical programming,” *SIAM Journal on Control and Optimization*, vol. 31, pp. 1340–1359, 1993.
- [36] L. Gubin, B. Polyak, and E. Raik, “The method of projections for finding the common point of convex sets,” *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 6, pp. 1 – 24, 1967.
- [37] A. Nedić, “Random algorithms for convex minimization problems,” *Mathematical Programming - B*, vol. 129, pp. 225–253, 2011.
- [38] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, “Randomized gossip algorithms,” *IEEE Trans. Inform. Theory*, vol. 52, no. 6, pp. 2508–2530, June 2006.
- [39] A. Nedić and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.
- [40] A. Nedić, A. Ozdaglar, and A. Parrilo, “Constrained consensus and optimization in multi-agent networks,” *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.
- [41] 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.
- [42] A. Nedić, “Asynchronous broadcast-based convex optimization over a network,” *IEEE Trans. Automat. Contr.*, vol. 56, no. 6, pp. 1337–1351, 2011.
- [43] H. L. Royden, *Real Analysis*, 3rd ed. Prentice Hall, 1998.
- [44] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995.
- [45] A. Nedić and A. Olshevsky, “Distributed optimization over time-varying directed graphs,” 2013, <http://arxiv.org/abs/1303.2289>.
- [46] O. Chapelle, “Training a support vector machine in the primal,” *Neural Comput.*, vol. 19, pp. 1155–1178, May 2007.
- [47] Y. Nesterov, “Smooth minimization of non-smooth functions,” *Math. Program.*, vol. 103, pp. 127–152, May 2005.

- [48] J. Kivinen, A. J. Smola, and R. C. Williamson, “Online learning with kernels,” *IEEE Transactions on Signal Processing*, vol. 52, pp. 2165–2176, 2004.
- [49] T. Zhang, “Solving large scale linear prediction problems using stochastic gradient descent algorithms,” in *Proceedings of the 21st ICML*, 2004, pp. 116–123.
- [50] Y. Singer and N. Srebro, “Pegasos: Primal estimated sub-gradient solver for SVM,” in *ICML*, 2007, pp. 807–814.
- [51] Y. J. Lee and O. L. Mangasarian, “SSVM: A smooth support vector machine for classification,” *Computational Optimization and Applications*, vol. 20, no. 1, pp. 5–22, 2001.
- [52] J. Zhang, R. Jin, Y. Yang, and A. G. Hauptmann, “Modified logistic regression: An approximation to SVM and its applications in large-scale text categorization,” in *Proceedings of the 20th ICML*, Menlo Park, AAAI, 2003, pp. 888–895.
- [53] J. Zhu and T. Hastie, “Kernel logistic regression and the import vector machine,” *Journal of Computational and Graphical Statistics*, vol. 14, no. 1, pp. 185–205, 2005.
- [54] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, “Efficient projections onto the l_1 -ball for learning in high dimensions,” in *Proceedings of the 25th ICML*, 2008, pp. 272–279.
- [55] A. Nemirovsky and D. Yudin, *Informational Complexity and Efficient Methods for Solution of Convex Extremal Problems*. J. Wiley & Sons, New York, 1983.