SELF-TUNED, BLOCK-COORDINATE, AND INCREMENTAL MIRROR DESCENT METHODS WITH APPLICATIONS IN MACHINE LEARNING AND WIRELESS COMMUNICATIONS

By

NAHIDSADAT MAJLESINASAB

Bachelor of Science in Industrial Engineering
Golpayegan College of Engineering
Golpayegan, Iran
2010

Master of Science in Industrial Engineering
Isfahan University of Technology
Isfahan, Iran
2013

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2020

SELF-TUNED, BLOCK-COORDINATE, AND INCREMENTAL MIRROR DESCENT METHODS WITH APPLICATIONS IN MACHINE LEARNING AND WIRELESS COMMUNICATIONS

Dissertation Approved:

Dr. Farzad Yousefian

Dissertation Advisor

Dr. Balabhaskar Balasundaram

Dr. Manjunath Kamath

Dr. Mahdi Asgari

Dedicated to my

beloved mother, father, and husband.

The dedication reflects the views of the author and are not endorsed by committee members or Oklahoma State University.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Farzad Yousefian, for the continuous support to my Ph.D study and for his motivation and immense knowledge. My dissertation could not have been completed without his guidance. I really appreciate his thoughtful comments and recommendations.

I would like to express my special appreciation and thanks to my committee members, Dr. Balabhaskar Balasundaram, Dr. Manjunath Kamath, and Dr. Mahdi Asgari for their brilliant comments and suggestions during my Ph.D. study. I am also very grateful to Dr. Chaoyue Zhao and Dr. Yuanxiong Guo for serving as my committee members before leaving Oklahoma State University.

I am very much indebted and grateful to Dr. Arash Pourhabib, my former advisor, who has helped me a lot during the past five years. I will be forever thankful for his support and encouragement. Additionally, I thank Dr. Austin Buchanan for his generous help during my job search.

My heartfelt thanks to Dr. Sunderesh Heragu for all his support during my PhD and for giving me the opportunity to teach the IEM Engineering Economics course which was one of the most wonderful experiences in my life.

Very special thanks to Dr. Balabhaskar Balasundaram for his endless help and support through which I found an exceptional opportunity to work for the MODE Transportation Company as an intern for two consecutive semesters and thereafter.

Acknowledgments reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

I am extremely grateful to Dr. MohammadJavad Feizollahi for his assistance and constructive suggestions which helped improve my work.

I would like to sincerely thank the staff of the department of Industrial Engineering and Management at Oklahoma State University, in particular Ms. Laura Brown, Ms. Valerie Quirey, Mr. Matt Taylor and Ms. Megan R. Hughes.

I am very thankful to all Oklahoma State University faculties who shared their knowledge with me and I learned a lot from including Dr. Mahdi Asgari, Dr. Balabhaskar Balasundaram, Dr. Austin Buchanan, Dr. Terry Collins, Dr. Manjunath Kamath, Dr. Tieming Liu., Dr. Jiahong Wu, and Dr. Farzad Yousefian.

I owe so much thanks to my parents for all the sacrifices they made for me. Without them, I would never been the person I am today. I would also like to thank my brother and sisters for their help and encouragement.

Last but not least, my deepest gratitude to my caring, loving, and supportive husband, Khosro Sasan, who always encouraged and supported me when the times got rough. Without his unconditional support and patience, it would not have been possible to successfully complete this journey.

Acknowledgments reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: NAHIDSADAT MAJLESINASAB

Date of Degree: July, 2020

Title of Study: SELF-TUNED, BLOCK-COORDINATE, AND INCREMENTAL MIRROR DESCENT METHODS WITH APPLICATIONS IN MACHINE LEARNING AND WIRELESS COMMUNICATIONS

Major Field: INDUSTRIAL ENGINEERING AND MANAGEMENT

Abstract: Uncertainty, high-dimensionality, and matrix structure of the decision variables are among the main challenges that may arise in addressing a wide range of stochastic optimization and equilibrium problems in machine learning and signal processing. Accordingly, the main goal of this dissertation lies in the development of suitable computational methods that can cope with the aforementioned challenges. To this end, we consider the stochastic mirror descent (SMD) methods that are among the popular avenues in solving stochastic optimization and variational inequality problems. Despite the significant advances in the convergence and complexity analysis of the SMD methods in the past two decades, there seems to be much to learn about ways to reduce the sensitivity of the performance of these methods with respect to the choice of the step-size rule. Motivated by this research gap, in the first part of this dissertation, we develop a unifying self-tuned randomized block-coordinate SMD method for solving high-dimensional stochastic optimization problems. The proposed method is unifying in the sense that it addresses both smooth and nonsmooth regimes. We establish an almost sure convergence for the generated iterate by the scheme. Importantly, we show that a mean-squared error of the method is minimized resulting in a faster convergence compared to the standard SMD methods. The numerical experiments on training support vector machines display that the self-tuned schemes are robust with respect to the choice of problem parameters and data sets. The second part of this dissertation is focused on multi-user optimization problems over semidefinite matrix spaces. The motivation arises in wireless communication networks composed of transmitters and receivers that generate and detect the signals, respectively. The competition among the transmitters in the network can be characterized as a non-cooperative Nash game with positive semidefinite matrix variables. To compute the equilibrium, we develop an SMD method equipped with a convergence rate statement. In addressing cooperative regimes, we develop an incremental mirror descent method where the users communicate with their adjacent user over the network. We establish the convergence for the proposed algorithm and also derive a non-asymptotic convergence rate statement.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

The mirror descent (MD) method was first proposed by Yudin and Nemirovski [1983] for solving convex optimization problems. MD is shown to be successful in solving high-dimensional deterministic optimization problems arising in reconstructing medical images [Ben-Tal et al., 2001] and stochastic optimization problems arising in network planning and power market [Nemirovski et al., 2009; Nedić and Lee, 2014]. The complexity of MD method is moderately dependent on the dimension of decision variables [Beck and Teboulle, 2003]. Consider the following minimization problem,

$$\min_{\beta \in \mathcal{B}} F(\beta), \tag{P1}$$

where $\mathcal{B} \subset \mathbb{R}^n$ is a closed convex set and $F : \mathcal{B} \to \mathbb{R}$ is a nonsmooth convex function. Let $\mathbf{g}_t \in \nabla F(\beta_t)$ denote the gradient of function F at point $\beta_t \in \mathscr{B}$. Let $\omega : \mathcal{B} \to \mathbb{R}$, called the distance generating function, be a continuously differentiable and strongly convex function on $\mathcal{B}$ with strong convexity parameter $\mu_\omega > 0$. The outline of MD method is as follows:

Algorithm 1 Mirror descent method

```
1: initialization: pick $\beta_0 \in \mathscr{B}$ arbitrarily and set $y_1 = \nabla\omega^*(\beta_0)$.
2: General step: for any $t = 1, 2, \ldots$ do the following:
  (a) $\beta_t = \nabla\omega^*(y_t)$,
  (b) $y_{t+1} = \nabla\omega(\beta_t) - \eta_t \mathbf{g}_t$,
```

where the conjugate of ω is defined by $\omega^\star(y) = \max_{\beta \in \mathcal{B}} \{\langle \beta, y \rangle - \omega(\beta)\}$ and $\{\eta_t\}$ denotes the stepsize sequence.

Algorithm 1 can be viewed as a generalization of the standard projected subgradient method as well. The subgradient projection method generates iterates, starting with an initial point $\beta_0 \in \mathcal{B}$, according to the following update rule:

$$\beta_{t+1} := \operatorname*{argmin}_{\beta \in \mathscr{B}} \|\beta_t - \eta_t \mathbf{g}_t - \beta\|_2 \quad \text{for all} \quad t \geq 0. \tag{1.1}$$

An iterative scheme such as subgradient method that uses the subdifferential/gradient information of the objective function to generate each iterate is called a first-order method. In the past few decades, first-order methods have proved to be very successful in addressing the optimization problem (P1) in stochastic, distributed, and large-scale regimes. In particular, their asymptotic convergence and non-asymptotic convergence rates can often be characterized in such regimes. It is for these important reasons that the first-order methods have been more favorable compared to their interior point-based counterparts.

The Bregman divergence function associated with ω is defined as $D_\omega : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and is given as

$$D_\omega(\beta_1, \beta_2) = \omega(\beta_2) - \omega(\beta_1) - \langle \nabla\omega(\beta_1), \beta_2 - \beta_1 \rangle,$$

for all $\beta_1, \beta_2 \in \mathscr{B}$. Beck and Teboulle [2003] showed that the MD method can be written equivalently as the following nonlinear projected subgradient type method in which a general distance function ω is used,

$$\beta_{t+1} := \operatorname*{argmin}_{\beta \in \mathscr{B}} \{\eta_t \langle \mathbf{g}_t, \beta - \beta_t \rangle + D_\omega(\beta_t, \beta)\}. \tag{1.2}$$

In other words, if we use the distance generating function $\omega(\beta) := \frac{1}{2}\|\beta\|_2^2$ in the update rule of the scheme (1.2), this method will be equivalent to the scheme (1.1).

Later, the stochastic variants of the mirror descent method including stochastic gradient mirror descent (SGMD) and stochastic subgradient mirror descent (SSMD) [Nemirovski et al., 2009; Nedić and Lee, 2014] were developed to solve the following canonical stochastic optimization problem,

$$\begin{aligned} &\text{minimize} \quad F(\beta) := \mathbb{E}[f(\beta,\xi)] \\ &\text{subject to} \quad \beta \in \mathscr{B}, \end{aligned} \tag{StochOpt}$$

where $f : \mathscr{B} \times \Omega \to \mathbb{R}$ is a stochastic function, and the vector $\xi : \Omega \to \mathbb{R}^d$ is a random vector associated with a probability space represented by $(\Omega, \mathcal{F}, \mathbb{P})$. Problem (StochOpt) is challenging because: (i) in statistical learning problems, usually the distribution of ξ is unknown; (ii) if dimension of ξ is more than 5, the expectation cannot be efficiently computed; (iii) when the dimensionality of solution space is huge, the first-order methods become impractical. In the update rule of SSMD method, the true value of subgradient $\mathbf{g}_t \in \partial F(\beta_t)$ is substituted by $\tilde{g}_t$, the noisy subgradient of $f(\beta, \xi_t)$ at $\beta := \beta_t$.

1.1 Multi-agent Optimization Problems

In the past two decades, there has been much interest in development of distributed and decentralized algorithms for multi-agent optimization problems in vector spaces [Nedić and Ozdaglar, 2009; Lobel and Ozdaglar, 2011; Shi et al., 2015] where the goal is to optimize a sum of convex component functions corresponding to m agents (also called users) as follows:

$$\underset{x \in \mathcal{R}^n}{\text{minimize}} \sum_{i=1}^{m} f_i(x). \tag{1.3}$$

Problems of the form (1.3) have been widely found in sensor network information processing, multi-agent control and coordination, and distributed machine learning. In these applications, agents refer to sensors, processors, etc. The notion of distributed algorithms refers to the algorithms that can be distributed across many agents. While in centralized algorithms, it is assumed that there is a centralized coordinator connected with all other agents that aggregates the information (e.g., gradients) computed from other agents and updates the model parameter, in decentralized algorithms, all agents can only communicate with their neighbors and there does not exist a central agent. In this line of research, incremental gradient/subgradient methods and their accelerated aggregated variants [Nedić and Bertsekas, 2001; Ram et al., 2009a; Gürbuzbalaban et al., 2017] have been developed where a local gradient/subgradient is taken at each step of an iteration and is followed by communicating with adjacent agents. More recently, Boţ and Böhm [2019] proposed an incremental mirror descent method with a stochastic sweeping of the component functions.

Although the agents would like to cooperate, it might not be practical or possible to communicate and exchange the information in some applications. Also, there might be a competition among the agents and it is to their benefit to optimize their local objective. In these cases, the distributed optimization techniques discussed above cannot be applied. However, the competition among the agents can be characterized as a non-cooperative Nash game. In a Nash game, N agents (users) with conflicting interests compete to minimize their own payoff function or maximize their utility function. Suppose each player controls a variable $x_i \in X_i$ where $X_i \subset \mathcal{R}^n$ denotes the set of all possible actions of player i. We let $x_{-i} :\triangleq (x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)$ denote the possible actions of other players and $f_i(x_i, x_{-i})$ denote the payoff function of player i. Therefore, the following Nash game needs to be solved:

$$\underset{x_i \in X_i}{\text{minimize}} \quad f_i(x_i, x_{-i}), \quad \text{for all } i = 1, \cdots, N, \tag{G1}$$

which includes N optimization problems. A solution $x^* = (x_1^*, \ldots, x_N^*)$ to this game called a Nash equilibrium is a feasible action profile such that $f_i(x_i^*, x_{-i}^*) \leq f_i(x_i, x_{-i}^*)$, for all $x_i \in X_i$, $i = 1, \ldots, N$. It can be proved that the optimality conditions of Nash game (G1) can be formulated as a Cartesian stochastic VI(X, F) where $X :\triangleq \{X | X = \text{diag}(x_1, \cdots, x_N),\ x_i \in X_i,\ \text{for all } i = 1, \ldots, N\}$ and $F(X) :\triangleq \text{diag}(\nabla_{x_1} f_1(x), \cdots, \nabla_{x_N} f_N(x))$.

1.2 Motivation

In this section, we motivate our research and explain the main research questions. Much of the interest in the literature of stochastic mirror descent (SMD) methods has focused on convergence and rate analysis in terms of magnitude of the error bounds. Yet, the finite-time performance of this class of methods can be significantly sensitive with respect to problem parameters, algorithm settings (e.g., stepsize choice), and the uncertainty (e.g., induced by the data). For instance, selecting a large step-size may result in divergence and choosing a small step-size may cause a very slow convergence. Therefore, the performance of the algorithm depends on the selection of a step-size as much as it depends on the selection of a search direction. In the development of efficient stepsize rules for stochastic approximation schemes, it is well-known that when the stepsize diminishes not too fast ($\sum_{t=0}^{\infty} \eta_t = \infty$) and not too slow ($\sum_{t=0}^{\infty} \eta_t^2 < \infty$), the method converges to the solution of problem (StochOpt) almost surely [Polyak, 1987]. For example, Spall [2005] discusses a harmonic stepsize of the form $\eta_t = \frac{a}{(t+1+b)^{\alpha}}$ where $a > 0$ is a tuning parameter, $b \geq 0$ is the stability constant and $0.5 < \alpha \leq 1$. It is recommended selecting a positive b that guarantees the stable behavior of the algorithm in a sense that it is not running slow in early or later iterations. It can be seen that problem parameters do not play a role in this choice of stepsize. Moreover, the performance of SMD method is not robust with respect to parameters a and b in practice. In the following example, we explain the drawback of harmonic stepsizes.

Example 1. [Support vector machines] Consider the following support vector machine problem,

$$\min \quad F(\beta) \triangleq \frac{1}{m}\sum_{i=1}^{m} L(\langle \beta, \mathbf{x}_i \rangle, y_i) + \frac{\lambda}{2}\|\beta\|_2^2 \ ,$$

where $L(\langle \beta, \mathbf{x}_i \rangle, y_i) \triangleq max\{0, 1 - y_i \langle \beta, \mathbf{x}_i \rangle\}$ is the hinge-loss function and $\lambda > 0$ is a regularization parameter. In this example, $\xi_i = (x_i, y_i)$ is drawn from a certain, but unknown distribution. We apply the SMD method (scheme (1.2)) using harmonic stepsizes of the form $\frac{a}{t+b}$ to solve the support vector machine problem which is discussed in detail later. Figure 1.1 illustrates the performance of the SMD method with harmonic stepsizes for two different data sets including RCV [Lewis et al., 2004] and Magic [Bock et al., 2004] data sets. The Reuters Corpus Volume (RCV) data set is a collection of newswire stories produced by Reuters journalists from 1996-1997. The articles are categorized into four different classes including Corporate/Industrial, Economics, Government/Social, and Markets. In this data sets, the samples are documents and the features represent the existence or nonexistence of a given word with 1 or 0 values. We chose a part of a data with 199,328 samples and 138,921 features. The goal is to predict whether an article belongs to Markets class or not and as a result, we have labels $y_i = \pm 1$. The other data sets, Magic, is from UCI Machine Learning Repository. The Magic data set includes some features to distinguish high-energy gamma particles from hadron particles using a gamma telescope and it includes 19,020 samples and 10 features. The vertical axis of each plot represents the logarithm of the objective function while the horizontal axis corresponds to iteration number. Parameters a and b are tuned and chosen according to Table 1.1. There are 6 different settings of these two parameters for each data set. From Figure 1.1, it can be seen that the SMD method with harmonic stepsizes are very sensitive to different choices of parameters a and b. For RCV data set, the stepsize with larger values of a for a fixed b performs better while for the Magic one, the smaller values

Data	$b_1 = 1000$	$b_2 = 2000$
RCV	$a_1 = 900$	$a_1 = 1800$
	$a_2 = 100000$	$a_2 = 200000$
	$a_3 = 250000$	$a_3 = 500000$
Magic	$a_1 = 25$	$a_1 = 50$
	$a_2 = 50$	$a_2 = 100$
	$a_3 = 100$	$a_3 = 200$

Table 1.1: Choice of parameters a and b

works better. Motivated by this example, our first goal, in this dissertation, is to develop self-tuned SMD schemes that are characterized in terms of problem parameters as well as algorithm settings and are robust with respect to the choice of problem parameters and data sets. We aim to develop such schemes for smooth, nonsmooth, and high-dimensional optimization problems.

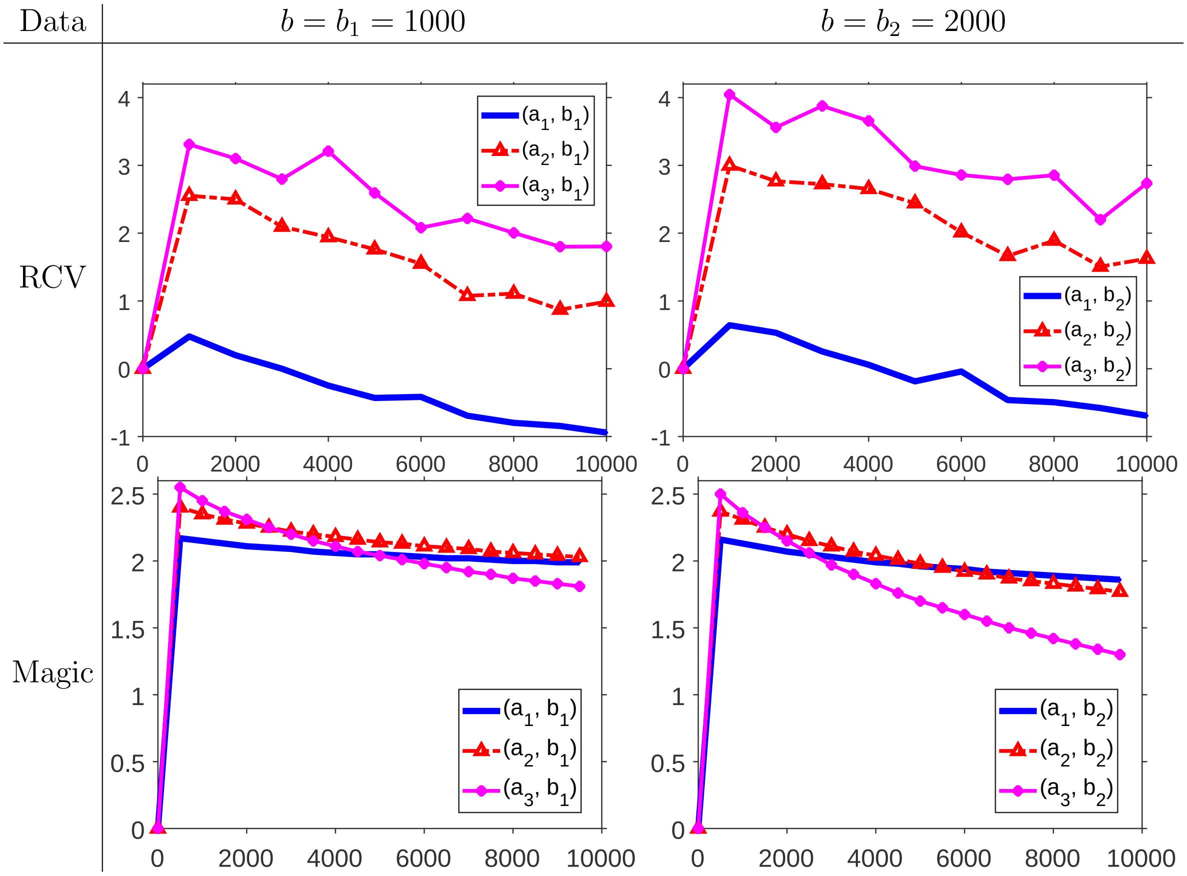

Figure 1.1: Comparison of SMD method applying harmonic stepsize $\frac{a}{t+b}$ for two data sets

The second research motivation arises from the need for addressing multi-user optimization problems on semidefinite matrix spaces. This includes cooperative multi-agent problems and non-cooperative Nash games. First, we consider the following multi-agent finite-sum optimization problem which involves a network of multiple agents who optimize a global objective,

$$\underset{X\in\mathcal{B}}{\text{minimize}}\ \sum_{i=1}^{m} f_i(X) \tag{1.4}$$

where $\mathcal{B} = \{X \in \mathbb{S}_n : X \succeq 0 \text{ and } \mathrm{tr}(X) = 1\}$, and $f_i : \mathcal{B} \to \mathbb{R}$ is a convex function. Note that each agent i is associated with the local objective $f_i(X)$ and all agents cooperatively minimize the network objective $\sum_{i=1}^{m} f_i(X)$. Assume that the allocation of all the objective components at one node is not possible due to memory or computational power constraints. Hence, for solving this problem, a distributed incremental algorithm is needed where the agents (players) should communicate with their adjacent agents to spread the distributed information over the network.

The sparse covariance inverse estimation explained in Example 2 is a specific application of finite-sum problem which sets a certain number of coefficients in the inverse covariance to zero to improve the stability of covariance matrix estimation [Price, 1972]. The goal is to find a sparse representation of the sample data and to highlight independence relationships between the sample variables.

Example 2. [Distributed sparse estimation of covariance inverse] Given a set of samples $\{z_i^j\}_{j=1}^{n_i}$ associated with agent i, where $z_i \sim \mathcal{N}(\mu, \Sigma)$, n_i is the sample size of the ith agent, $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ are the mean and covariance matrix of a multivariate Gaussian distribution, respectively. To estimate μ and Σ, consider the maximum likelihood estimators (MLE) given by

$$\hat{\mu}, \hat{\Sigma} = \underset{\mu,\Sigma}{\text{argmax}} \prod_{i=1}^{m}\prod_{j=1}^{n} \frac{1}{\sqrt{(2\pi)^{n_i}\det(\Sigma)}} \exp\left(-\frac{1}{2}(z_i^j-\mu)^T\Sigma^{-1}(z_i^j-\mu)\right).$$

This equation can then be cast as a distributed inverse covariance estimation problem

$$\min_{\Sigma^{-1}\succ 0} -\sum_{i=1}^{m} \log\left(\det\Sigma^{-1}\right) + \sum_{i=1}^{m} \mathrm{tr}\left(S_i\Sigma^{-1}\right),$$

where $S_i \triangleq \frac{1}{n_i}\sum_{j=1}^{n_i} -\frac{1}{2}(z_i^j - \hat{\mu}_i)^T(z_i^j - \hat{\mu}_i)$ with $\hat{\mu}_i \triangleq \frac{1}{n_i}\sum_{j=1}^{n_i} z_i^j$. To have a sparse solution, we consider the addition of a lasso penalty $P * \Sigma^{-1}$ to the likelihood as follows

$$\min_{\Sigma^{-1}\succ 0} -\sum_{i=1}^{m} \log\left(\det\Sigma^{-1}\right) + \sum_{i=1}^{m} \mathrm{tr}\left(S_i\Sigma^{-1}\right) + \lambda\|P * \Sigma^{-1}\|_1, \tag{1.5}$$

where P is an arbitrary matrix with nonnegative elements, $\lambda > 0$ is the regularization parameter, and $*$ denotes element-wise multiplication. For a matrix A, we define $\|A\|_1 = \sum_{i,j} |[A]_{ij}|$. Two common choices for P would be the matrix of all ones or this matrix with zeros on the diagonal to avoid shrinking diagonal elements of Σ [Bien and Tibshirani, 2011]. Problem (1.5) can be viewed as an instance of the Problem (1.4), where we define $f_i(\Sigma^{-1}) = -\log\left(\det\Sigma^{-1}\right) + \mathrm{tr}(S_i\Sigma^{-1}) + \frac{\lambda}{m}\|P * \Sigma^{-1}\|_1$.

Motivated by the above example, one of our research goals in this dissertation is to develop a matrix mirror descent incremental subgradient (M-MDIS) method to solve problem (1.4).

As mentioned previously, in this dissertation, we address multi-user optimization problems on semidfinite matrix spaces including cooperative multi-agent problems and non-cooperative Nash games. We already talked about cooperative optimization problem of interest and now we would like to talk about the non-cooperative Nash game. Here, we refer to the game (G1) introduced in the section 1.1. Assume there are N players competing to minimize their own payoff function. Suppose each player controls a positive semidefinite matrix variable $X_i \in \mathcal{X}_i$ where $\mathcal{X}_i$ denotes the set of all possible actions of player i. We let $X_{-i} :\triangleq (X_1, ..., X_{i-1}, X_{i+1}, ..., X_N)$ denote the possible actions of other players and $f_i(X_i, X_{-i})$

denote the payoff function of player i. Therefore, the following Nash game needs to be solved:

$$\underset{X_i \in \mathcal{X}_i}{\text{minimize}} \quad f_i(X_i, X_{-i}), \quad \text{for all } i = 1, \cdots, N, \tag{G2}$$

which includes N semidefinite optimization problem. A solution $X^* = (X_1^*, \ldots, X_N^*)$ to this game called a Nash equilibrium is a feasible action profile such that $f_i(X_i^*, X_{-i}^*) \leq f_i(X_i, X_{-i}^*)$, for all $X_i \in \mathcal{X}_i = \{X_i | X_i \in \mathbb{S}_{n_i}^+, \mathrm{tr}(X_i) = 1\}$, $i = 1, \ldots, N$ where $\mathbb{S}_{n_i}^+$ denotes the cone of all $n_i \times n_i$ positive semidefinite matrices. The next example discusses one of the applications of problem (G2) in wireless communication network.

Example 3. [Wireless communication networks] A wireless network is composed of transmitters and receivers that generate and detect radio signals respectively. An antenna enables a transmitter to send signals into space and a receiver to pick up signals from space. In a multiple-input multiple-output (MIMO) wireless transmission system, multiple antennas is applied in transmitters and receivers in order to improve its performance. In some MIMO systems such as MIMO broadcast channels and MIMO multiple access channels, there are multiple users which mutually interfere. In recent years, MIMO systems under uncertainty have been studied where the state channel information is subject to noise, delays and other imperfections [Mertikopoulos et al., 2017]. Here, the problem of interest is the throughput maximization in multi-user MIMO networks under feedback errors. In this network, N MIMO links (users) compete where each link i represents a pair of transmitter-receiver with m_i antennas at the transmitter and n_i antennas at the receiver. Let $\mathbf{x}_i \in \mathbb{C}^{n_i}$ and $\mathbf{y}_i \in \mathbb{C}^{m_i}$ denote the signal transmitted from and received by the ith link, respectively. The signal model can be described by $\mathbf{y}_i = H_{ii}\mathbf{x}_i + \sum_{j \neq i} H_{ji}\mathbf{x}_j + \epsilon_i$, where $H_{ii} \in \mathbb{C}^{m_i \times n_i}$ is the direct-channel matrix of link i, $H_{ji} \in \mathbb{C}^{m_i \times n_j}$ is the cross-channel matrix between transmitter j and receiver i, and $\epsilon_i \in \mathbb{C}^{m_i}$ is a zero-mean circularly symmetric complex Gaussian noise vector with the covariance matrix $\mathbf{I}_{m_i}$ [Mertikopoulos and Moustakas, 2016]. Each transmitter

i tries to improve its performance by transmitting at its maximum power level. Hence, The action for each player is the transmit power. However, doing so would result in a conflict in the system since the overall interference increases and affects the capability of all involved transmitters. Here, we consider the interference generated by other users as an additive noise. Therefore, $\sum_{j\neq i} H_{ji}\mathbf{x}_j$ represents the multi-user interference (MUI) received by ith player and generated by other users. Assuming the random vector $\mathbf{x}_i$ follows a complex Guassian distribution, transmitter i controls its input signal covariance matrix $X_i :\triangleq \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\dagger]$ subject to two constraints: first the signal covariance matrix is positive semidefinite and second each transmitter's maximum transmit power is bounded by a positive scalar p. Under these assumptions, each user's transmission throughput for a given set of users' covariance matrices $X_1, \ldots, X_N$ is given by

$$R_i(X_i, X_{-i}) = \log\det\left(\mathbf{I}_{m_i} + \sum\nolimits_{j=1}^{N} H_{ji}X_jH_{ji}^\dagger\right) - \log\det(W_{-i}), \tag{1.6}$$

where $W_{-i} = \mathbf{I}_{m_i} + \sum_{j\neq i} H_{ji}X_jH_{ji}^\dagger$ is the MUI-plus-noise covariance matrix at receiver i [Telatar, 1999]. Let $\mathcal{X}_i = \{X_i \in \mathbb{C}^{n_i\times n_i} : X_i \succeq 0,\ \mathrm{tr}(X_i) \leq p\}$. The goal is to solve

$$\underset{X_i\in\mathcal{X}_i}{\text{maximize}} \quad R_i(X_i, X_{-i}), \quad \text{for all } i = 1, \ldots, N. \tag{1.7}$$

Later, we prove that the optimality conditions of Nash game (G2) can be formulated as a Cartesian stochastic VI$(\mathcal{X}, F)$ where $\mathcal{X} :\triangleq \{X | X = \mathrm{diag}(X_1, \cdots, X_N),\ X_i \in \mathcal{X}_i,\ \text{for all } i = 1, \ldots, N\}$ and $F(X) :\triangleq \mathrm{diag}(\nabla_{X_1} f_1(X),\ \cdots, \nabla_{X_N} f_N(X))$. There are several challenges in solving CSVIs on semidefinite matrix spaces including presence of uncertainty, the semidefinite solution space and the Cartesian product structure. Much of the interest in the theory of variational inequality (VI) has focused on addressing VIs on vector spaces. There are a few methods addressing VIs on matrix spaces. Some of these methods require a two-loop framework where at each iteration, a projection problem, i.e., a semidefinite optimization problem needs to be solved. Others rely on assumptions that either does not hold in applications, or it is hard to verify. Motivated by this gap, our goal is to develop a single-loop first-order method under the assumption that the mapping is merely monotone.

1.3 Research Contributions

In this section, we discuss the key contributions of our work. In Chapter II, motivated by big data applications, we consider stochastic mirror descent (SMD) methods for solving stochastic optimization problems with strongly convex objective functions. Our goal is to develop SMD schemes that achieve a rate of convergence with a minimum constant factor with respect to the choice of the stepsize sequence. To this end, we consider three variants of SMD methods namely (a) subgradient SMD methods addressing nonsmooth problems, (b) gradient SMD methods addressing smooth problems, and (c) randomized block coordinate SMD methods addressing high-dimensional problems. For each scheme, we develop self-tuned stepsize rules that are characterized in terms of problem parameters and algorithm settings. Using self-tuned stepsize rules, we show that the non-averaging iterate generated by the underlying SMD method converges to the optimal solution both in an almost sure and a mean sense. For each scheme, we derive error bounds and show that using the corresponding self-tuned stepsizes, such an error bound is minimized. Moreover, in the case where problem parameters are unknown, we develop a unifying self-tuned update rule that can be applied in both smooth and nonsmooth settings. We show that for any arbitrary and small enough initial stepsize, a suitably defined error bound is minimized. Finally, We provide constant factor comparisons with standard SMD methods. We also investigate the robustness of self-tuned SMD schemes with respect to the choice of data set, problem parameters, and initial stepsize.

In Chapter III, we consider multi-user optimization problems on semidefinite matrix spaces. We develop mirror descent methods where we choose the distance generating function to be defined as the quantum entropy. These methods are single-loop first-order methods in the sense that they only require a gradient-type of update at each iteration. In the first part of the chapter, we propose a mirror descent incremental subgradient method for minimizing a convex function that consists of sum of component functions. This type of minimization over semidefinite matrix spaces arises in cooperative multi-agent problems such as sparse estimation of a covariance matrix. We show that the iterate generated by the algorithm converges asymptotically to the optimal solution and derive a non-asymptotic convergence rate. Motivated by non-cooperative Nash games in stochastic regimes, in the second part of the chapter, we consider Cartesian stochastic variational inequality (CSVI) problems where the variables are positive semidefinite matrices. We develop a stochastic mirror descent method that require monotonicity assumption which holds in many applications. The originality of this work lies in the convergence analysis. Employing an auxiliary sequence of stochastic matrices and averaging techniques, we show that the iterate generated by the algorithm converges to a weak solution of the CSVI. Then, we derive a rate of convergence in terms of the expected value of a suitably defined gap function. We also implement the proposed method for solving a multiple-input multiple-output multi-cell cellular wireless network composed of seven hexagonal cells. We investigate the robustness of our scheme with respect to problem parameters and uncertainty. Finally, in chapter IV, we conclude this research.

1.4 Notations and Definitions

In this section, first, we introduce some basic notations which are used in this dissertation. Then, we recall some definitions.

Throughout the first and second chapter, we use $\langle \beta_1, \beta_2 \rangle$ to denote the inner product of two vectors $\beta_1, \beta_2 \in \mathbb{R}^n$. It is assumed that $\mathbb{R}^n$ is equipped with some norm $\|\cdot\|$ and $\|\cdot\|_*$ denotes its dual norm. We use $Prob\ (Z)$ and $\mathbb{E}[z]$ to denote the probability of an event Z, and the expectation of a random variable z, respectively. We let $\beta^i \in \mathbb{R}^{n_i}$ denote the ith block coordinate of vector $\beta \in \mathbb{R}^n$, and the subscript i represent the ith block of a mapping in $\mathbb{R}^n$. For any $i = 1, \ldots, l$, we use $\|\cdot\|_i$ to denote the general norm on $\mathbb{R}^{n_i}$ and $\|\cdot\|_{*i}$ to denote its dual norm. The inner product of vectors $u, v \in \mathbb{R}^n$ is defined by $\langle u, v\rangle :\triangleq \sum_{i=1}^{l} \langle u^i, v^i\rangle$. We define norm $\|\cdot\|$ as $\|x\|^2 :\triangleq \sum_{i=1}^{d} \|x^i\|_i^2$ for any $x \in \mathbb{R}^n$, and denote its dual norm by$\|\cdot\|_*$. Throughout, p_i denotes the probability associated with choosing the ith block coordinate. We use the notation $p_\wedge :\triangleq \min\limits_{1\leq i\leq l} p_i$, $p_\vee :\triangleq \max\limits_{1\leq i\leq l} p_i$, $\mathfrak{L}_{max} :\triangleq \max\limits_{1\leq i\leq l} \mathfrak{L}_{\omega_i}$, and $\mu_{min} :\triangleq \min\limits_{1\leq i\leq l} \mu_{\omega_i}$.

Throughout the third chapter, we let $\mathbb{S}_n$ denote the set of all $n \times n$ symmetric matrices and $\mathbb{S}_n^+$ the cone of all positive semidefinite matrices. The set of solutions to $\mathrm{VI}(\mathcal{X}, F)$ is denoted by $\mathrm{SOL}(\mathcal{X}, F)$. We define the set $\mathscr{X} := \{X \in \mathbb{S}_n^+ : \mathrm{tr}(X) \leq 1\}$. We let $[A]_{uv}$ denote the components of matrix A and $\mathbb{C}$ the set of complex numbers. The spectral norm of a matrix A being the largest singular value of A is denoted by the norm $\|A\|_2$. The trace norm of a matrix A being the sum of singular values of the matrix is denoted by $\mathrm{tr}(A)$. Note that spectral and trace norms are dual to each other [Fazel et al., 2001]. We let $\mathbf{A}^\dagger$ denote the conjugate transpose of matrix $\mathbf{A}$. A square matrix A that is equal to its conjugate transpose is called Hermitian.

Next, we recall some definitions that will be referred to in Chapters 2 and 3.

Definition 1 (subgradient of function F)**.** Consider a set $\mathcal{B} \in \mathbb{R}^n$ and a function $F : \mathcal{B} \to \mathbb{R}^n$. $\mathbf{g} \in \partial F(\beta_1)$ is called a subgradient of function F at point $\beta_1 \in \mathscr{B}$, if a vector $\mathbf{g}$ exists such that

$$F(\beta_1) + \langle \mathbf{g}, \beta_2 - \beta_1\rangle \leq F(\beta_2), \quad \text{for all } \beta_2 \in \mathscr{B}.$$

Definition 2 (Types of convexity)**.** Consider a convex set $\mathcal{B} \in \mathbb{R}^n$ and a function $F : \mathcal{B} \to \mathbb{R}^n$.

(a) F is called a convex function if for any $\beta_1, \beta_2 \in \mathscr{B}$ and $\mathbf{g} \in \partial F(\beta_2)$

$$F(\beta_1) \geq F(\beta_2) + \langle \mathbf{g}, \beta_1 - \beta_2 \rangle.$$

(b) F is called a strictly convex function if for any $\beta_1, \beta_2 \in \mathscr{B}$ and $\mathbf{g} \in \partial F(\beta_2)$

$$F(\beta_1) > F(\beta_2) + \langle \mathbf{g}, \beta_1 - \beta_2 \rangle.$$

(c) F is called a strongly convex function with parameter $\mu_F > 0$ with respect to the underlying norm $\| \cdot \|$ if for any $\beta_1, \beta_2 \in \mathscr{B}$ and $\mathbf{g} \in \partial F(\beta_2)$

$$F(\beta_1) \geq F(\beta_2) + \langle \mathbf{g}, \beta_1 - \beta_2 \rangle + \frac{\mu_F}{2} \|\beta_1 - \beta_2\|^2. \tag{1.8}$$

Definition 3 (Types of monotonicity)**.** Consider a set $\mathcal{X} \in \mathbb{R}^{n \times n}$ and a mapping $F : \mathcal{X} \to \mathbb{R}^{n \times n}$.

(a) F is called a monotone mapping if for any $X, Y \in \mathcal{X}$, we have $\operatorname{tr}\big((X - Y)^T (F(X) - F(Y))\big) \geq 0$.

(b) F is called a λ-strongly monotone mapping if there is $\lambda > 0$ such that for any $X, Y \in \mathcal{X}$, we have $\operatorname{tr}\big((X - Y)^T (F(X) - F(Y))\big) \geq \lambda D(X, Y)$.

(c) F is called a pseudo-monotone mapping if for any $X, Y \in \mathcal{X}$, $\operatorname{tr}\big((X - Y)^T F(Y)\big) \geq 0$, implies that $\operatorname{tr}\big((X - Y)^T F(X)\big) \geq 0$.

(d) F is called a λ-strongly pseudo-monotone mapping if for any $X, Y \in \mathcal{X}$, $\operatorname{tr}\big((X - Y)^T F(Y)\big) \geq 0$, implies that $\operatorname{tr}\big((X - Y)^T F(X)\big) \geq \lambda D(X, Y)$.

Definition 4 (Almost sure convergence)**.** Let $\{x_n\}$ be a sequence of random variables defined on a sample space Ω. We say that $\{x_n\}$ is almost surely convergent (a.s. convergent) to a random variable x defined on Ω if and only if the sequence of real numbers $\{x_n\}$ converges to x almost surely, i.e., if and only if there exists a zero-probability event E such that

$$\{\omega \in \Omega : x_n(\omega) \text{ does not converge to } x(\omega)\} \subseteq E.$$

x is called the almost sure limit of the sequence and convergence is indicated by $x_n \xrightarrow{a.s.} x$.

CHAPTER II

SELF-TUNED STOCHASTIC MIRROR DESCENT METHODS FOR STOCHASTIC OPTIMIZATION

In this chapter, motivated by big data applications, we consider stochastic mirror descent (SMD) methods for solving stochastic optimization problems with strongly convex objective functions. A significant part of the literature for developing SMD techniques has concentrated on convergence and rate analysis as far as greatness of the error bounds. However, the finite-time execution of this class of methods is tied to the selection of stepsize sequence. As such, our goal is to develop SMD schemes that achieve a rate of convergence with a minimum constant factor with respect to the choice of the stepsize sequence. To this end, we consider three variations of SMD techniques to be specific (a) subgradient SMD methods addressing nonsmooth problems, (b) gradient SMD methods addressing smooth problems, and (c) randomized block coordinate SMD methods addressing high-dimensional problems. For each scheme, we develop self-tuned stepsize rules that are characterized in terms of problem parameters and algorithm settings. Our main contributions are as follows: (i) utilizing self-tuned stepsize rules, we show that the non-averaging iterate generated by the underlying SMD method converges to the optimal solution both in an almost sure and a mean sense; (ii) for each scheme, we derive error bounds and show that this error bound is minimized using the corresponding self-tuned step sizes; (iii) to address the cases that some problem parameters are not known, we develop a unifying self-tuned update rule that can be utilized in both smooth and nonsmooth settings. We show that for any arbitrary and small enough initial stepsize, a suitably defined error bound is minimized; (iv) We provide constant factor comparisons with standard SMD methods.

2.1 Problem Formulation and Background

In this chapter, we consider the canonical stochastic optimization problem given by

$$\begin{aligned} &\text{minimize} \quad F(\beta) := \mathbb{E}[f(\beta, \xi)] \\ &\text{subject to} \quad \beta \in \mathscr{B}, \end{aligned} \tag{StochOpt}$$

where $\mathscr{B} \subset \mathbb{R}^n$ is a nonempty, closed, and convex set and $f : \mathscr{B} \times \Omega \to \mathbb{R}$ is a stochastic function. The vector $\xi : \Omega \to \mathbb{R}^d$ is a random vector associated with a probability space represented by $(\Omega, \mathcal{F}, \mathbb{P})$. A wide range of problems in machine learning and signal processing can be formulated as problem (StochOpt). In these applications, given a set of training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ of size m, where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ are the ith input and output objects, respectively, the goal lies in learning a function $h : \mathbb{R}^n \times \mathscr{B} \to \mathbb{R}$ by solving an empirical risk minimization (ERM) problem given as follows:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{m} \sum_{i=1}^m L(h(\beta, \mathbf{x}_i), y_i) + \lambda R(\beta) \\ &\text{subject to} \quad \beta \in \mathscr{B}, \end{aligned} \tag{ERM}$$

where $L : \mathbb{R}^2 \to \mathbb{R}$ is a loss function, $R : \mathbb{R}^n \to \mathbb{R}$ is a regularizer, constant $\lambda > 0$ is the regularization parameter. In addressing problem (StochOpt), challenges arise in the development of efficient solution methods mainly due to the following reasons: (i) presence of uncertainty: in many applications arising in statistical learning, the probability distribution $\mathbb{P}$ is unknown. In such cases, the sample average approximation (SAA) scheme can be applied. However, the efficiency of SAA scheme deteriorates as the sample size increases (cf. Nemirovski et al. [2009]). Even when the probability distribution $\mathbb{P}$ is known, the evaluation of the expectation of function f becomes costly, specially when $d > 5$; (ii) high-dimensionality: another difficulty in addressing problem (StochOpt) arises when the dimensionality of the solution space, i.e., n is huge. In such applications, the computational complexity per iteration of the first-order methods (e.g., deterministic and stochastic gradient method) increases significantly, making such methods impractical for large values of n (e.g., 10^{12} or more). In addressing uncertainty, stochastic approximation (SA) method was first developed by Robbins and Monro [1951]. Since then, SA method and its variants have been vastly employed to solve stochastic optimization [Nevelśon and Hasḿinskii, 1973; Ermoliev, 1983; Ruszczyński and Syski, 1986; Kushner and Yin, 2003] and equilibrium problems [Juditsky et al., 2011; Jiang and Xu, 2008; Wang and Bertsekas, 2015]. Averaging techniques first introduced by Polyak and Juditsky [1992] proved successful in increasing the robustness of SA method. In vector spaces equipped with non-Euclidean norms, prox generalizations of deterministic gradient method [Yudin and Nemirovski, 1983; Beck and Teboulle, 2003] were introduced and applied in smooth and nonsmooth settings. Also, in stochastic regime, Nemirovski et al. [2009] developed the stochastic mirror descent (SMD) method for solving problem (StochOpt) when the objective function F is nonsmooth and merely convex. In this method, a weighted averaging sequence is computed that is characterized by the stepsize sequence and the previously generated iterates. Under a window-based averaging scheme, the rate of $\mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$ is established. Nedić and Lee [2014] showed that under a different set of weights and employing a full-window averaging scheme, the convergence rate $\mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$ can be established for the subgradient SMD method. Generalizations of this optimal averaging technique was developed for SA schemes in [Yousefian et al., 2017], and more recently for stochastic mirror prox methods in [Yousefian et al., 2018] for addressing stochastic variational inequalities with merely monotone mappings. When the dimensionality of the solution space n is huge, the SA and SMD schemes become inefficient as they require arithmetic operations of the order n at each iteration. To reduce this computational burden, block coordinate descent (CD) methods have been developed in the recent decades. While Ortega and Rheinboldt [2000] appear amongst the first to study such a concept, Luo and Tseng [1992, 1993]; Tseng and Yun [2009], Bertsekas and Tsitsiklis [2000], Nesterov [2010] and others [Mareček et al., 2015; Richtárik and Takáč, 2014; Xu and Yin, 2013] studied the convergence and complexity analysis of the CD schemes. Recently, Dang and Lan [2015b] developed randomized block coordinate SMD methods for solving problem (StochOpt) when the objective function is nonsmooth and the set $\mathscr{B}$ is given as the Cartesian product of l component sets. They showed that using averaging techniques, the convergence rate of $\mathcal{O}\left(\frac{l}{\sqrt{t}}\right)$ and $\mathcal{O}\left(\frac{l}{t}\right)$ can be established for the case when F is merely convex, and strongly convex, respectively. While these non-asymptotic convergence orders are known to be optimal for the SMD method, the performance of this method can be significantly sensitive with respect to problem parameters, algorithm settings (e.g., stepsize choice), and the uncertainty (e.g., induced by the data). Much of the interest in the literature has focused on establishing the optimal convergence rates, and there is little guidance on development of stepsize update rules for the SMD method in order to minimize the constant factor of the associated error bounds. Motivated by this gap, our goal in this chapter lies in improvement of the finite-time behavior of the SMD methods through development of self-tuned stepsizes. Several efforts have been done in development of efficient stepsize rules for SA schemes. Of these, Kesten et al. [1958] proposed a stepsize rule in which the stepsize is decreased by one when the errors in successive iterations have opposite signs. Saridis [1970] extended Kesten's rule and suggested the stepsize should also increase when error estimates in successive iterations have the same sign. Spall [2005] discusses a harmonic stepsize of the form $\eta_t = \frac{a}{(t+1+A)^\alpha}$ where $a > 0$ is a tuning parameter, and $A \geq 0$ is the stability constant. George and Powell [2006] propose a class of harmonic stepsizes which minimizes the mean squared estimation error. Other works include but are not limited to [Benveniste et al., 1990], [Pflug, 1988], Kalman filter [Stengel, 2012] and the "search then converge" algorithm [Darken and Moody, 1992]. Self-tuned stepsizes were first introduced in

[Yousefian et al., 2012] where a recursive update rule is developed for the stochastic gradient and subgradient methods. It is shown that using such update rules, the mean squared error of the method is minimized w.r.t. the choice of the stepsize. In this work, we consider problem (StochOpt) where the objective function F is strongly convex with parameter $\mu_F > 0$. We consider three cases where (i) function F is nondifferentiable, (ii) function F is differentiable and has Lipschitz gradients, and (iii) the dimensionality of the problem, i.e., n, is huge. For case (i) and (ii), the subgradient SMD and gradient SMD method are considered, respectively. For case (iii), we consider the randomized block coordinate variant of the SMD method. While the SMD methods developed in the literature (cf. [Nemirovski et al., 2009; Dang and Lan, 2015b]) employ averaging, our goal lies in developing non-averaging schemes. Our main contributions are as follows:

(1) Convergence and complexity analysis: For each variant of the aforementioned SMD methods, we develop new recursive error bounds in terms of the prox function. These error bounds are given by Lemmas 2, 5, and 6, 7 for cases (i), (ii) and (iii), respectively. In each case, we then develop self-tuned stepiszes that are characterized in terms of problem parameters and algorithm settings. We show that under such update rules, the error function of the underlying SMD method converges to zero in an almost sure and a mean sense. Importantly, we show that the expected value of the error is minimized under the self-tuned stepize rules within a specified range. We also derive bounds on the probability of error of the SMD schemes in terms of problem parameters, algorithm settings, and iteration number. The convergence and rate results are provided by Propositions 1, 2, and 3-4 for cases (i), (ii) and (iii), respectively. Our results in this chapter extend the previous findings on self-tuned stepsizes in [Yousefian et al., 2012, 2016] to a broader class of algorithms i.e., SMD methods. Moreover, our approach in addressing nonsmoothness is different than that considered in [Yousefian et al., 2012, 2016]. Here we develop subgradient variants of SMD method allowing us to prove convergence to an exact optimal solution to problem (StochOpt), while in [Yousefian et al., 2012] and [Yousefian

et al., 2016] a smoothing scheme is applied and convergence is established to an approximate optimal solution.

(2) Unifying self-tuned stepsizes: When some of problem parameters are unavailable, we develop a generalized class of stepsize rules namely *unifying self-tuned stepsizes* and prove convergence in both an almost sure and a mean sense. Importantly, we show that for an arbitrary and small enough initial stepsize, a suitably defined error bound of the SMD scheme is minimized. (see Theorem 1). This indeed implies robustness of the proposed schemes w.r.t. the choice of initial stepize and addresses a common challenge associated with the harmonic choice of stepsizes.

(3) Constant factor comparison: While we prove the superiority of the constant factor of the error bounds associated with SMD schemes under the developed self-tuned stepsizes versus any arbitrary choice of stepsizes, we also provide two sets of comparisons: (i) with a widely used harmonic stepsizes (e.g., in [Nemirovski et al., 2009; Spall, 2005]), and also (ii) with an averaging SMD scheme developed in [Dang and Lan, 2015b]. In case (ii), our comparison implies the constant factor for the class of stochastic subgradient methods can be improved up to four times under non-averaging schemes versus using the averaging scheme in [Dang and Lan, 2015b].

(4) Implementation results: We present the performance of the unifying self-tuned stepsizes applied on SVM models under three different data sets. Our results indicate the robustness of the developed schemes with respect to problem parameters, uncertainty, and the initial stepsize.

2.2 Self-tuned SMD Methods

In this section, we first start with the case where the objective function is non-differentiable. Later, in Section 2.2.2, we discuss the case of differentiable objective functions with Lipschitz gradients. In Section 2.2.3, we provide unifying self-tuned update rules addressing both cases in absence of problem parameters.

2.2.1 Self-tuned Stochastic Subgradient Mirror Descent Methods

Consider problem (StochOpt) where we assume F is a non-differentiable convex function of β. Throughout, for $t = 0, 1, \ldots$, we let $\mathbf{g}_t \in \partial F(\beta_t)$ denote a subgradient of function F at point $\beta_t \in \mathscr{B}$. Similarly, for any $\xi \in \Omega$, we let $\tilde{g}_t \in \partial f(\beta_t, \xi)$ denote a subgradient of function $f(\cdot, \xi)$ at point β_t. Throughout this section, we assume that F is strongly convex with parameter $\mu_F > 0$ over the set $\mathscr{B}$ with respect to the underlying norm $\|\cdot\|$.

In our analysis, we make use of the following result.

Lemma 1. Consider problem (StochOpt). Let F be strongly convex with parameter $\mu_F > 0$. Then, there exists a unique optimal solution $\beta^* \in \mathscr{B}$. Moreover, we have

$$F(\beta) - F(\beta^*) \geq \frac{\mu_F}{2}\|\beta - \beta^*\|^2, \quad \text{for all } \beta \in \mathscr{B}.$$

Proof. The existence and uniqueness of β^* follows by the assumption that $\mathscr{B}$ is non-empty, closed and convex, and that F is strongly convex (see Theorem 2.2.3 in [Facchinei and Pang, 2003]). By the first-order optimality conditions, for all $\beta \in \mathscr{B}$, we have $\langle \mathbf{g}^*, \beta - \beta^* \rangle \geq 0$ where $\mathbf{g}^* \in \partial F(\beta^*)$. Using this inequality and invoking relation (1.8) for $\beta_1 := \beta$ and $\beta_2 := \beta^*$, we obtain the desired inequality. □

To address problem (StochOpt), the method of interest in this section is the stochastic subgradient mirror descent method. The convergence and rate analysis of the deterministic and stochastic variants of this method have been studied in [Nemirovski et al., 2009; Beck and Teboulle, 2003; Nedić and Lee, 2014] under averaging schemes. Our focus in this section pertains to a non-averaging variant of this method. To describe the method, we first provide the settings and notations associated with the method. Let $\omega : \mathbb{R}^n \to \mathbb{R}$, called the distance generating function, be a continuously differentiable and strongly convex function with constant μ_ω, i.e.,

$$\omega(\beta_2) \geq \omega(\beta_1) + \langle \nabla\omega(\beta_1), \beta_2 - \beta_1 \rangle + \frac{\mu_\omega}{2}\|\beta_2 - \beta_1\|^2,$$

for all $\beta_1, \beta_2 \in \mathscr{B}$. For example, under Euclidean norm $\|.\|_2$, function $\omega(\beta) := \frac{1}{2}\|\beta\|_2^2$ meets these requirements with $\mu_\omega = 1$. The Bregman divergence function associated with ω is defined as $D_\omega : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and is given as

$$D_\omega(\beta_1, \beta_2) = \omega(\beta_2) - \omega(\beta_1) - \langle \nabla\omega(\beta_1), \beta_2 - \beta_1 \rangle,$$

for all $\beta_1, \beta_2 \in \mathscr{B}$. Given an arbitrary $\beta_0 \in \mathscr{B}$, the stochastic subgradient mirror descent method is given by the following update rule:

$$\beta_{t+1} := \underset{\beta \in \mathscr{B}}{\operatorname{argmin}}\{\eta_t \langle \tilde{g}_t, \beta - \beta_t \rangle + D_\omega(\beta_t, \beta)\}, \tag{SSMD}$$

for all $t \geq 0$, where η_t is the stepsize, and $\tilde{g}_t$ is the noisy subgradient of $f(\beta, \xi_t)$ at $\beta = \beta_t$. Note that in the deterministic variant of this scheme, the stochastic subgradient $\tilde{g}_t$ is substituted by the true value of subgradient $\mathbf{g}_t \in \partial F(\beta_t)$. Before we proceed with the analysis, we recall some of the properties of the Bregman divergence function. Note that $D_\omega(\beta_1, \beta_2)$ is differentiable with respect to the variable β_2. Let $\nabla_{\beta_2} D_\omega(\cdot, \cdot)$ denote the partial derivative of $D_\omega(\beta_1, \beta_2)$ with respect to β_2. Then we have for all $\beta_1, \beta_2 \in \mathscr{B}$

$$\nabla_{\beta_2} D_\omega(\beta_1, \beta_2) = \nabla\omega(\beta_2) - \nabla\omega(\beta_1). \tag{2.1}$$

Based on the definition, the Bregman divergence function has the following property

$$D_\omega(\beta_1, \beta_2) - D_\omega(\beta_3, \beta_2) = D_\omega(\beta_1, \beta_3) + \langle \nabla\omega(\beta_3) - \nabla\omega(\beta_1), \beta_2 - \beta_3 \rangle, \tag{2.2}$$

for all $\beta_1, \beta_2, \beta_3 \in \mathscr{B}$ and by the strong convexity of function ω, we have

$$D_\omega(\beta_1, \beta_2) \geq \frac{\mu_\omega}{2}\|\beta_2 - \beta_1\|^2, \quad \text{for all } \beta_1, \beta_2 \in \mathscr{B}. \tag{2.3}$$

Throughout, we assume the distance generating function ω has Lipschitz gradients with parameter $\mathfrak{L}_\omega$, i.e., for all $\beta_1, \beta_2, \beta_3 \in \mathscr{B}$

$$\omega(\beta_2) \leq \omega(\beta_1) + \langle \nabla\omega(\beta_1), \beta_2 - \beta_1 \rangle + \frac{\mathfrak{L}_\omega}{2}\|\beta_2 - \beta_1\|^2.$$

From the preceding inequality, the definition of D_ω implies that for all $\beta_1, \beta_2 \in \mathscr{B}$

$$D_\omega(\beta_1, \beta_2) \leq \frac{\mathfrak{L}_\omega}{2}\|\beta_2 - \beta_1\|^2. \tag{2.4}$$

Next, we state some standard assumptions on the stochastic subgradients that will be used in the convergence analysis.

Assumption 1. [First and second moment of stochastic subgradients] Let the stochastic subgradient $\tilde{g}(\beta) \in \partial f(\beta, \xi)$ be such that a.s. for all $\beta \in \mathscr{B}$, we have $\mathbb{E}[\tilde{g}(\beta)|\beta] = \mathbf{g}(\beta) \in \partial F(\beta)$. Moreover, there exists a scalar $C > 0$ such that

$$\mathbb{E}\left[\|\tilde{g}(\beta)\|_*^2 | \beta\right] \leq C^2, \quad \text{for all } \beta \in \mathscr{B}. \tag{2.5}$$

Throughout, we let $\mathcal{F}_t$ be the history of the algorithm up to time t, i.e, $\mathcal{F}_t = \{\beta_0, \xi_0, \xi_1, \ldots, \xi_{t-1}\}$ for $t \geq 1$, with $\mathcal{F}_0 = \{\beta_0\}$. To begin the analysis, in the following, we develop a recursive inequality in terms of the error of the (SSMD) scheme. Such a recursive inequality will be employed in the following sections to develop a self-tuned stepsize rule.

Lemma 2. [A recursive error bound for the (SSMD) scheme] Let Assumption (1) hold. Then, for all $t \geq 0$ we have a.s.

$$\mathbb{E}[D_\omega(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t] \leq \left(1 - \frac{2\mu_F}{\mathfrak{L}_\omega}\eta_t\right) D_\omega(\beta_t, \beta^*) + \frac{C^2\eta_t^2}{2\mu_\omega}, \tag{2.6}$$

where β^* is the unique optimal solution to problem (StochOpt).

Proof. Consider the update rule (SSMD). Using the first-order optimality conditions, we have for all $\beta \in \mathscr{B}$

$$\langle \eta_t \tilde{g}_t + \nabla_{\beta_{t+1}} D_\omega(\beta_t, \beta_{t+1}), \beta - \beta_{t+1} \rangle \geq 0,$$

Using equality (2.1), from the preceding inequality we obtain

$$\langle \eta_t \tilde{g}_t + \nabla\omega(\beta_{t+1}) - \nabla\omega(\beta_t), \beta - \beta_{t+1} \rangle \geq 0,$$

for all $\beta \in \mathscr{B}$ which is equivalent to

$$\langle \nabla\omega(\beta_{t+1}) - \nabla\omega(\beta_t), \beta - \beta_{t+1} \rangle \geq \eta_t \langle \tilde{g}_t, \beta_{t+1} - \beta \rangle, \tag{2.7}$$

for all $\beta \in \mathscr{B}$. Invoking relation (2.2), from the preceding relation we can write

$$D_\omega(\beta_t, \beta) - D_\omega(\beta_{t+1}, \beta) - D_\omega(\beta_t, \beta_{t+1}) \geq \eta_t \langle \tilde{g}_t, \beta_{t+1} - \beta \rangle,$$

for all $\beta \in \mathscr{B}$. From the strong convexity of $\omega(\beta)$ and relation (2.3), we have

$$D_\omega(\beta_t, \beta) - D_\omega(\beta_{t+1}, \beta) - \frac{\mu_\omega}{2}\|\beta_t - \beta_{t+1}\|^2 \geq \eta_t \langle \tilde{g}_t, \beta_{t+1} - \beta \rangle. \tag{2.8}$$

Next, we find a lower bound on the term $\eta_t \langle \tilde{g}_t, \beta_{t+1} - \beta \rangle$. By adding and subtracting $\langle \eta_t \tilde{g}_t, \beta_t \rangle$, we get

$$\begin{aligned}\eta_t\langle \tilde{g}_t, \beta_{t+1}-\beta\rangle &= \eta_t\langle \tilde{g}_t, \beta_{t+1}-\beta_t\rangle + \eta_t\langle \tilde{g}_t, \beta_t-\beta\rangle \\ &\geq -\left|\langle \frac{\eta_t}{\sqrt{\mu_\omega}}\tilde{g}_t, \sqrt{\mu_\omega}(\beta_{t+1}-\beta_t)\rangle\right| + \eta_t\langle \tilde{g}_t, \beta_t-\beta\rangle \\ &\geq -\frac{\eta_t^2}{2\mu_\omega}\|\tilde{g}_t\|_*^2 - \frac{\mu_\omega}{2}\|\beta_{t+1}-\beta_t\|^2 + \eta_t\langle \tilde{g}_t, \beta_t-\beta\rangle, \qquad (2.9)\end{aligned}$$

where the last inequality follows from Fenchel's inequality, i.e., $|\langle x, y\rangle| \leq \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|_*^2$. Combining (2.8) and (2.9) yields

$$\eta_t\langle \tilde{g}_t, \beta_t-\beta\rangle + D_\omega(\beta_{t+1}, \beta) \leq D_\omega(\beta_t, \beta) + \frac{\eta_t^2}{2\mu_\omega}\|\tilde{g}_t\|_*^2,$$

for all $\beta \in \mathscr{B}$. By taking the conditional expectation on $\mathcal{F}_t$ from both sides of the preceding relation and setting $\beta := \beta^*$, we have for all $\beta \in \mathscr{B}$

$$\eta_t\langle \mathbf{g}_t, \beta_t-\beta^*\rangle + \mathbb{E}[D_\omega(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t] \leq D_\omega(\beta_t, \beta^*) + \frac{\eta_t^2 C^2}{2\mu_\omega}, \qquad (2.10)$$

where we used $\mathbb{E}[\tilde{g}_t \mid \mathcal{F}_t] = \mathbf{g}_t$ and $\mathbb{E}[\|\tilde{g}_t\|_*^2 \mid \mathcal{F}_t] \leq C^2$ from Assumption 1. Using the strong convexity of function F in (1.8), we can write

$$\begin{aligned}\eta_t\langle \mathbf{g}_t, \beta_t-\beta^*\rangle &\geq \eta_t(F(\beta_t)-F(\beta^*)) + \frac{\mu_F\eta_t}{2}\|\beta_t-\beta^*\|^2 \\ &\geq \eta_t(F(\beta_t)-F(\beta^*)) + \frac{\mu_F\eta_t}{\mathfrak{L}_\omega}D_\omega(\beta_t, \beta^*)\end{aligned}$$

where the last inequality follows by relation (2.4). From the preceding relation and inequality (2.10), we obtain for all $t \geq 0$

$$\mathbb{E}[D_\omega(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t] + \eta_t(F(\beta_t)-F(\beta^*)) \leq \left(1-\frac{\mu_F\eta_t}{\mathfrak{L}_\omega}\right)D_\omega(\beta_t, \beta^*) + \frac{\eta_t^2 C^2}{2\mu_\omega}. \qquad (2.11)$$

Using Lemma 1 and relation (2.4), we have

$$F(\beta_t) - F(\beta^*) \geq \frac{\mu_F}{2}\|\beta_t - \beta^*\|^2 \geq \frac{\mu_F}{\mathfrak{L}_\omega} D_\omega(\beta_t, \beta^*). \tag{2.12}$$

Combining relations (2.11) and (2.12) yields the desired relation. □

The inequality (2.6) provides a recursive relation that can be used to derive an upper bound for the term $\mathbb{E}[D_\omega(\beta_t, \beta^*)]$. This term can be seen as the expected error of the (SSMD) method that quantifies the deviation between β_t and the optimal solution β^* in the mean sense. Note that using Lemma 2, the bound on this error term is characterized by problem parameters such as μ_F and C, by algorithm settings such as μ_ω, $\mathfrak{L}_\omega$, and also by the stepsize η_t. To develop an update formula for η_t, our main objective is to analyze the recursive relation (2.6). To this end, we make use of the following lemma. This lemma provides a general recursive sequence, called self-tuned sequences, that can be used for minimizing the recursive error bounds of the form in Lemma 2. We summarize some important properties of the self-tuned sequences. Some of these properties can be found in [Yousefian et al., 2012, 2016].

Lemma 3. [Self-tuned sequences] Let θ and δ be positive scalars, and $\{er_t\}$ be a non-negative sequence for $t \geq 0$, such that the following equality holds for an arbitrary non-negative sequence $\{\eta_t\}$:

$$er_{t+1} := (1 - \theta\eta_t)er_t + \delta\eta_t^2, \quad \text{for all } t \geq 1. \tag{2.13}$$

Let $er_0 \leq \frac{2\delta}{\theta^2}$ and let the self-tuned sequence $\{\eta_t^*\}$ be given by $\eta_t^* = \eta_{t-1}^*\left(1 - \frac{\theta}{2}\eta_{t-1}^*\right)$ for any $t \geq 1$, where $\eta_0^* = \frac{\theta}{2\delta}er_0$. Then the following properties hold:

(a) For any fixed $t \geq 1$, the vector $(\eta_0^*, \eta_1^*, \ldots, \eta_{t-1}^*)$ minimizes the function $er_t(\eta_0, \eta_1, \ldots,$

$\eta_{t-1})$ over the set

$$\mathbb{U}_t \triangleq \left\{\gamma \in \mathbb{R}^t : 0 < \gamma_j \leq \frac{1}{\theta} \ \text{ for } j = 1, \ldots, t\right\}.$$

More precisely, for any $t \geq 1$, and any $(\eta_0, \eta_1, \ldots, \eta_{t-1}) \in \mathbb{U}_t$, we have

$$er_t(\eta_0, \eta_1, \ldots, \eta_{t-1}) - er_t(\eta_0^*, \eta_1^*, \ldots, \eta_{t-1}^*) \geq \delta(\eta_{t-1} - \eta_{t-1}^*)^2.$$

(b) For all $t \geq 1$, we have $\eta_t^* < \frac{2}{\theta}\left(\frac{1}{t}\right)$. Moreover, under the choice of $\eta_t := \eta_t^*$, the term er_t is bounded by $\mathcal{O}(1/t)$, i.e.,

$$er_t(\eta_0^*, \eta_1^*, \ldots, \eta_{t-1}^*) \leq \frac{4\delta}{\theta^2}\left(\frac{1}{t}\right), \quad \text{for all } t \geq 1. \tag{2.14}$$

(c) We have $\sum_{t=0}^{\infty} \eta_t^* = \infty$ and $\sum_{t=0}^{\infty} \eta_t^{*2} < \infty$.

Proof. (a) To show part (a), we first use induction on t to show that er_t satisfies

$$er_t(\eta_0^*, \eta_1^*, \ldots, \eta_{t-1}^*) = \frac{2\delta}{\theta}\eta_t^*, \quad \text{for all } t \geq 0. \tag{2.15}$$

Note that it holds for $t = 0$ from the definition $\eta_0^* = \frac{\theta}{2\delta} er_0$. Next, let us assume (2.15) holds for t. From this and relation (2.13), we have

$$\begin{aligned} er_{t+1}(\eta_0^*, \eta_1^*, \ldots, \eta_t^*) &= (1 - \theta\eta_t^*) er_t(\eta_0^*, \eta_1^*, \ldots, \eta_{t-1}^*) + \delta\eta_t^{*2} = (1 - \theta\eta_t^*)\frac{2\delta}{\theta}\eta_t^* + \delta\eta_t^{*2} \\ &= \frac{2\delta}{\theta}\eta_t^*\Big(1 - \theta\eta_t^* + \frac{\theta\eta_t^*}{2}\Big) = \frac{2\delta}{\theta}\eta_t^*\Big(1 - \frac{\theta\eta_t^*}{2}\Big) = \frac{2\delta}{\theta}\eta_{t+1}^*, \end{aligned}$$

where in the last equation, we used the definition of η_{t+1}^*. This implies that relation (2.15) holds for $t + 1$ and therefore, for any $t \geq 0$. We now use induction on t to prove that $(\eta_0^*, \eta_1^*, \ldots, \eta_{t-1}^*)$ minimizes er_t for all $t \geq 1$. By the definition of er_1 and the relation

$er_1(\eta_0^*) = \frac{2\delta}{\theta}\eta_1^*$ shown previously, we have

$$er_1(\eta_0) - er_1(\eta_0^*) = (1 - \theta\eta_0)er_0 + \delta\eta_0^2 - \frac{2\delta}{\theta}\eta_1^*.$$

Therefore, using relation $\eta_1^* = \eta_0^*\left(1 - \frac{\theta}{2}\eta_0^*\right)$ and $\eta_0^* = \frac{\theta}{2\delta}er_0$, we can write

$$er_1(\eta_0) - er_1(\eta_0^*) = (1 - \theta\eta_0)\frac{2\delta}{\theta}\eta_0^* + \delta\eta_0^2 - \frac{2\delta}{\theta}\eta_0^*\Big(1 - \frac{\theta}{2}\eta_0^*\Big) = \delta(\eta_0 - \eta_0^*)^2.$$

This implies that part (a) holds for $t = 1$. In the rest of the proof, for the sake of simplicity, we use er_{t+1} for an arbitrary vector $(\eta_0, \eta_1, \ldots, \eta_t) \in \mathbb{U}_{t+1}$ and er_{t+1}^* for er_{t+1} evaluated at $(\eta_0^*, \eta_1^*, \ldots, \eta_t^*)$. Now suppose part (a) holds for some $t \geq 1$ implying that $er_t \geq er_t^*$ holds for any $(\eta_0, \eta_1, \ldots, \eta_{t-1}) \in \mathbb{U}_t$. Using (2.13) and (2.15), we have

$$er_{t+1} - er_{t+1}^* = (1 - \theta\eta_t)er_t + \delta\eta_t^2 - \frac{2\delta}{\theta}\eta_{t+1}^*.$$

Using $er_t \geq er_t^*$, relation (2.15), the definition of η_{t+1}^* and that $\eta_t \leq \frac{1}{\theta}$, we get

$$er_{t+1} - er_{t+1}^* \geq (1 - \theta\eta_t)\frac{2\delta}{\theta}\eta_t^* + \delta\eta_t^2 - \frac{2\delta}{\theta}\eta_t^*\Big(1 - \frac{\theta}{2}\eta_t^*\Big) = \delta(\eta_t - \eta_t^*)^2.$$

Therefore, part (a) holds for $t + 1$. We conclude that the result of part (a) is true for any $t \geq 1$.

(b) Using the recursive relation $\eta_{t+1}^* = \eta_t^*\left(1 - \frac{\theta}{2}\eta_t^*\right)$, we have

$$\frac{1}{\eta_{t+1}} = \frac{1}{\eta_t\left(1 - \frac{\theta}{2}\eta_t\right)} = \frac{1}{\eta_t} + \frac{\frac{\theta}{2}}{1 - \frac{\theta}{2}\eta_t}, \quad \text{for all } t \geq 0.$$

Summing up from $t = 0$ to k and canceling the common terms from both sides, we obtain

$$\frac{1}{\eta_{k+1}} = \frac{1}{\eta_0} + \frac{\theta}{2}\sum_{t=0}^{k} \frac{1}{1-\frac{\theta}{2}\eta_t} > \frac{\theta}{2}\sum_{t=0}^{k} \frac{1}{1-\frac{\theta}{2}\eta_t}. \tag{2.16}$$

Note that from the definition of η_0^* and er_0, we have $0 < \eta_0^* \leq \frac{1}{\theta}$. From relation $\eta_t^* = \eta_{t-1}^*\left(1-\frac{\theta}{2}\eta_{t-1}^*\right)$ we have $0 < \eta_t^* \leq \frac{1}{\theta}$ for all $t \geq 0$. Consequently, the term $1-\frac{\theta}{2}\eta_t^*$ is a number between zero and one. Therefore, $\left(1-\frac{\theta}{2}\eta_t^*\right)^{-1} > 1$ which implies that $\sum_{t=0}^{k}\left(1-\frac{\theta}{2}\eta_t^*\right)^{-1} > k+1$. Therefore, using relation (2.16), for all $k \geq 1$ we have $\eta_k^* < \dfrac{2}{\theta k}$. Combining inequality (2.15) and the preceding inequality, we obtain the desired result.

(c) First, we show $\sum_{t=0}^{\infty}\eta_t^* = \infty$. From $\eta_t^* = \eta_{t-1}^*\left(1-\frac{\theta}{2}\eta_{t-1}^*\right)$ for all $t \geq 0$, we obtain

$$\eta_{t+1}^* = \eta_0^*\prod_{i=0}^{t}\left(1-\frac{\theta}{2}\eta_i^*\right). \tag{2.17}$$

Note that since $\eta_0^* \in \left(0, \frac{1}{\theta}\right]$, from $\eta_t^* = \eta_{t-1}^*\left(1-\frac{\theta}{2}\eta_{t-1}^*\right)$ it follows that $\{\eta_t^*\}$ is positive non-increasing sequence. Therefore, the limit $\lim_{t\to\infty}\eta_t^*$ exists and it is less than $\frac{2}{\theta}$. Thus, by taking the limits from both sides in $\eta_t^* = \eta_{t-1}^*\left(1-\frac{\theta}{2}\eta_{t-1}^*\right)$, we obtain $\lim_{t\to\infty}\eta_t^* = 0$. Then, by taking limits in (2.17), we further obtain

$$\lim_{t\to\infty}\prod_{i=0}^{t}\left(1-\frac{\theta}{2}\eta_i^*\right) = 0.$$

To arrive at a contradiction, suppose that $\sum_{i=0}^{\infty}\eta_i^* < \infty$. Then, there is an $\epsilon \in (0,1)$ such that for j sufficiently large, we have $\frac{\theta}{2}\sum_{i=j}^{t}\eta_i^* \leq \epsilon$, for all $t \geq j$. Since $\prod_{i=j}^{t}\left(1-\frac{\theta}{2}\eta_i^*\right) \geq 1-\frac{\theta}{2}\sum_{i=j}^{t}\eta_i^*$ for all $j < t$, by letting $t \to \infty$, we obtain for all j sufficiently large,

$$\prod_{i=j}^{\infty}\left(1-\frac{\theta}{2}\eta_i^*\right) \geq 1-\frac{\theta}{2}\sum_{i=j}^{\infty}\eta_i^* \geq 1-\epsilon > 0.$$

This contradicts the statement $\lim_{t\to\infty}\prod_{i=0}^{t}\left(1-\frac{\theta}{2}\eta_i^*\right) = 0$. Hence, we conclude that

$\sum_{t=0}^{\infty} \eta_t^* = \infty$. Next, we show that $\sum_{t=0}^{\infty} \eta_t^{*2} < \infty$. From $\eta_t^* = \eta_{t-1}^* \left(1 - \frac{\theta}{2}\eta_{t-1}^*\right)$ we have

$$\eta_i^* = \eta_{i-1}^* - \frac{\theta}{2}{\eta_{i-1}^*}^2, \quad \text{for all } i \geq 1.$$

Summing the preceding relation from $i = 0$ to t and canceling the common terms, we obtain

$$\eta_t^* = \eta_0^* - \frac{\theta}{2}\sum_{i=0}^{t-1} \eta_i^{*2}, \quad \text{for all } t \geq 1.$$

By taking limits and recalling that $\lim_{t\to\infty} \eta_t^* = 0$, we obtain the desired result. □

Before we proceed with presenting the main result in this section, we revisit the following lemma (see Polyak [1987], page 50) that will be used in the analysis of the (SSMD) method.

Lemma 4. Let $\{v_t\}$ be a sequence of non-negative random variables where $\mathbb{E}[v_0] < \infty$, let $\{\alpha_t\}$ and $\{\lambda_t\}$ be deterministic scalar sequences such that:

$$\begin{aligned} &\mathbb{E}[v_{t+1}|v_0, \dots, v_t] \leq (1-\alpha_t)v_t + \lambda_t, \quad \text{a.s. for all } t \geq 0, \\ &0 \leq \alpha_t \leq 1, \quad \lambda_t \geq 0, \quad \sum_{t=0}^{\infty} \alpha_t = \infty, \quad \sum_{t=0}^{\infty} \lambda_t < \infty, \quad \lim_{t\to\infty} \frac{\lambda_t}{\alpha_t} = 0. \end{aligned}$$

Then, $v_t \longrightarrow 0$ a.s., $\lim_{t\to\infty} \mathbb{E}[v_t] = 0$, and for any $\epsilon > 0$ and for all $t > 0$

$$\text{Prob}(v_j \leq \epsilon \text{ for all } j \geq t) \geq 1 - \frac{1}{\epsilon}\left(\mathbb{E}[v_t] + \sum_{i=t}^{\infty} \lambda_i\right).$$

In the following, we present the self-tuned stepsizes for the (SSMD) method and discuss their properties.

Preposition 1. [Self-tuned stepsizes for (SSMD) method] Let $\{\beta_t\}$ be generated by the (SSMD) method. Let the function F be strongly convex with modulus μ_F and the set $\mathscr{B}$ be convex, closed, and bounded such that $\|\beta\| \leq M$ for all $\beta \in \mathscr{B}$ and some $M > 0$. Let Assumption 1 hold for some C large enough such that $C^2\mathfrak{L}_\omega \geq 8M^2\mu_\omega\mu_F^2$. Let the stepsize η_t be given by

$$\eta_0^* := \frac{4\mu_\omega\mu_F M^2}{C^2}, \qquad \eta_t^* := \eta_{t-1}^*\left(1 - \frac{\mu_F}{\mathfrak{L}_\omega}\eta_{t-1}^*\right), \quad \text{for all } t \geq 1.$$

Then, the following hold:

(a) The sequence $\{\beta_t\}$ generated by the (SSMD) method converges a.s. to the unique optimal solution β^* of problem (StochOpt).

(b) For any $t \geq 1$, the vector $(\eta_0^*, \eta_1^*, \ldots, \eta_{t-1}^*)$ minimizes the upper bound of the error $\mathbb{E}[D_\omega(\beta_t, \beta^*)]$ given in Lemma 2 for all $(\eta_0, \eta_1, \ldots, \eta_{t-1}) \in \left(0, \frac{\mathfrak{L}_\omega}{2\mu_F}\right]^t$.

(c) The (SSMD) method attains the convergence rate $\mathcal{O}(1/t)$, i.e,

$$\mathbb{E}\left[\|\beta_t - \beta^*\|^2\right] \leq \left(\frac{C\mathfrak{L}_\omega}{\mu_\omega\mu_F}\right)^2 \frac{1}{t}, \quad \text{for all } t \geq 1.$$

(d) Let ϵ and ρ be arbitrary positive scalars and $T \triangleq \left(\frac{3C^2\mathfrak{L}_\omega^2}{2\mu_\omega\mu_F^2}\right)\frac{1}{\epsilon\rho}$ we have for all $t \geq T$

$$\text{Prob}\left(D_\omega(\beta_j, \beta^*) \leq \epsilon \text{ for all } j \geq t\right) \geq 1 - \rho.$$

Proof. (a) Note that the uniqueness of β^* is implied by Lemma 1. To show a.s. convergence, we apply Lemma 4. From the result of Lemma 2, we have for all $t \geq 0$

$$\mathbb{E}[D_\omega(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t] \leq \left(1 - \frac{2\mu_F}{\mathfrak{L}_\omega}\eta_t^*\right) D_\omega(\beta_t, \beta^*) + \frac{C^2\eta_t^{*2}}{2\mu_\omega}. \tag{2.18}$$

Let us define the following terms:

$$v_t \triangleq D_\omega(\beta_t, \beta^*), \quad \alpha_t \triangleq \frac{2\mu_F}{\mathfrak{L}_\omega}\eta_t^*, \quad \lambda_t \triangleq \frac{C^2\eta_t^{*2}}{2\mu_\omega}. \tag{2.19}$$

Note that since $\{\eta_t^*\}$ is non-increasing and that $\alpha_0 = \left(\frac{2\mu_F}{\mathfrak{L}_\omega}\right)\eta_0^* = \frac{8M^2\mu_\omega\mu_F^2}{C^2\mathfrak{L}_\omega}$, and the assumption $C^2\mathfrak{L}_\omega \geq 8M^2\mu_\omega\mu_F^2$, we can conclude that $0 \leq \alpha_t \leq 1$ for all $t \geq 0$. Moreover, from Lemma 3(c), we have that $\sum_{t=0}^{\infty}\alpha_t = \infty$ and $\sum_{t=0}^{\infty}\lambda_t < \infty$. Also, note that the definition of α_t and λ_t and that the self-tuned stepsize η_t^* has a limit of zero (see proof of Lemma 3, part (c)) imply that $\lim_{t\to\infty}\frac{\lambda_t}{\alpha_t} = 0$. Therefore, all conditions of Lemma 4 are satisfied indicating that $D_\omega(\beta_t, \beta^*) \to 0$ a.s.. Now, using the strong convexity of ω in (2.3), we have $\frac{\mu_\omega}{2}\|\beta_t - \beta^*\|^2 \leq D_\omega(\beta_t, \beta^*)$. Therefore, we conclude that β_t converges to β^* a.s..

(b) For any $t \geq 1$, let us define the function $er_t(\eta_0, \ldots, \eta_{t-1})$ given by the recursion (2.13) where $\theta \triangleq \frac{2\mu_F}{\mathfrak{L}_\omega}$, and $\delta \triangleq \frac{C^2}{2\mu_\omega}$. Also, let $er_0 \triangleq 2M^2\mathfrak{L}_\omega$. First note that for all $t \geq 0$, we have $\mathbb{E}[D_\omega(\beta_t, \beta^*)] \leq er_t(\eta_0, \ldots, \eta_{t-1})$ for any arbitrary $(\eta_0, \eta_1, \ldots, \eta_{t-1}) \in \left(0, \frac{\mathfrak{L}_\omega}{2\mu_F}\right]^t$ used in the (SSMD) method. To show this, taking expectations from both of the relation in Lemma 2, and from the definition of θ and δ we obtain

$$\mathbb{E}[D_\omega(\beta_{t+1}, \beta^*)] \leq (1 - \theta\eta_t)\,\mathbb{E}[D_\omega(\beta_t, \beta^*)] + \delta\eta_t^2,$$

for all $t \geq 0$. It is enough to show that $D_\omega(\beta_0, \beta^*) \leq er_0$. Note that from relation (2.4), and the triangle inequality we have

$$D_\omega(\beta_0, \beta^*) \leq \frac{\mathfrak{L}_\omega}{2}\|\beta_0 - \beta^*\|^2 \leq \frac{\mathfrak{L}_\omega}{2}\left(\|\beta_0\|^2 + \|\beta^*\|^2 + 2\|\beta_0\|\|\beta^*\|\right) \leq 2\mathfrak{L}_\omega M^2.$$

This implies that $D_\omega(\beta_0, \beta^*) \leq er_0$. Using induction and the relation in lemma 2, $\mathbb{E}[D_\omega(\beta_t, \beta^*)] \leq er_t(\eta_0, \ldots, \eta_{t-1})$ holds for all t implying that er_t is a well-defined upper bound. To complete the proof of this part, it suffices to show that the conditions of Lemma 3 hold. First we need to show that $er_0 \leq \frac{2\delta}{\theta^2}$. From the values of er_0, θ, δ, we have

$$\frac{er_0\theta^2}{2\delta} = \frac{8\mathfrak{L}_\omega M^2\mu_F^2\mu_\omega}{C^2\mathfrak{L}_\omega^2} \leq 1,$$

where the last relation follows by the assumption $C^2\mathfrak{L}_\omega \geq 8M^2\mu_\omega\mu_F^2$. Therefore, $er_0 \leq \frac{2\delta}{\theta^2}$ implying that the conditions of Lemma 3 hold. Hence, from part (a) in Lemma 3, we conclude that $(\eta_0^*, \eta_1^*, \ldots, \eta_{t-1}^*)$ minimizes the upper bound $er_t(\eta_0, \eta_1, \ldots, \eta_{t-1})$ for all $(\eta_0, \eta_1, \ldots, \eta_{t-1}) \in \left(0, \frac{\mathfrak{L}_\omega}{2\mu_F}\right]^t$.

(c) Following the proof of part (b), from Lemma 3(b), we obtain for all $t \geq 1$

$$\mathbb{E}[D_\omega(\beta_t, \beta^*)] \leq er_t(\eta_0, \eta_1, \ldots, \eta_{t-1}) \leq \frac{4\delta}{\theta^2}\frac{1}{t} = \left(\frac{C^2\mathfrak{L}_\omega^2}{2\mu_\omega\mu_F^2}\right)\frac{1}{t}. \tag{2.20}$$

Invoking the relation (2.3), we obtain the desired inequality.

(d) To show this result, we use the probabilistic bound given in Lemma 4. To this end, first we estimate the term $\sum_{i=t}^{\infty}\lambda_i$ where λ_i is given by (2.19). Using Lemma 3(b), we can write

$$\begin{aligned}\sum_{i=t}^{\infty}\lambda_i = \sum_{i=t}^{\infty}\frac{C^2}{2\mu_\omega}\eta_i^{*2} \leq \sum_{i=t}^{\infty}\frac{C^2}{2\mu_\omega}\left(\frac{\mathfrak{L}_\omega}{\mu_F\, i}\right)^2 &= \left(\frac{C^2\mathfrak{L}_\omega^2}{2\mu_\omega\mu_F^2}\right)\left(\frac{1}{t^2} + \sum_{i=t+1}^{\infty}\frac{1}{i^2}\right)\\ &\leq \left(\frac{C^2\mathfrak{L}_\omega^2}{2\mu_\omega\mu_F^2}\right)\left(\frac{1}{t} + \int_t^{\infty}\frac{1}{x^2}dx\right) = \left(\frac{C^2\mathfrak{L}_\omega^2}{\mu_\omega\mu_F^2}\right)\frac{1}{t}.\end{aligned}$$

From the preceding inequality, relation (2.20), and Lemma 4, we obtain the desired relation.

□

Comparison 1. Proposition 1 states that the self-tuned stepsizes not only guarantee the convergence of the (SSMD) method, but also the constant factor provided in part (c) is the minimum constant factor for any arbitrary stepsize rule within a given range. Let us for example compare this constant factor with that of the stochastic subgradient method under harmonic stepsize rules in [Nemirovski et al., 2009]. In that chapter (see relations (2.9) and (2.10)), under the harmonic update rule for stepsizes given by $\eta_t = \gamma/t$ for some constant $\gamma > 1/(2\mu_F)$, it is shown that

$$\mathbb{E}\left[\|\beta_t - \beta^*\|_2^2\right] \leq \max\left\{\frac{\gamma^2C^2}{2\mu_F\gamma - 1}, \|\beta_0 - \beta^*\|_2^2\right\}\frac{1}{t}. \tag{2.21}$$

Here we show that for any arbitrary $\gamma > \frac{1}{2\mu_F}$, the term $\frac{\gamma^2 C^2}{2\mu_F\gamma - 1}$ is larger than the constant factor of the self-tuned stepsizes that is $\left(\frac{C\mathfrak{L}_\omega}{\mu_\omega \mu_F}\right)^2$. Note that in the case of stochastic subgradient method, we set $\omega(\beta) := \frac{\|\beta\|_2^2}{2}$. This implies that $\mu_\omega = \mathfrak{L}_\omega = 1$ in the Euclidean norm space. We can write,

$$\frac{\text{Harmonic constant factor}}{\text{Self-tuned constant factor}} = \frac{\gamma^2 C^2 \mu_F^2}{(2\mu_F\gamma - 1)C^2} = \frac{\gamma^2 \mu_F^2}{2\mu_F\gamma - 1}.$$

Note that $\gamma^2\mu_F^2 - 2\mu_F\gamma + 1 = (\gamma\mu_F - 1)^2 > 0$ for all $\gamma > \frac{1}{2\mu_F}$. Therefore, the preceding relation implies that the harmonic constant factor in [Nemirovski et al., 2009] is larger than the self-tuned constant factor for any arbitrary $\gamma > \frac{1}{2\mu_F}$.

2.2.2 Self-tuned Stochastic Gradient Mirror Descent Methods

In this section, we consider the case where the objective function in problem (StochOpt) is differentiable and has Lipschitz gradients. Our goal here is to utilize this property and develop a self-tuned scheme that is characterized with the problem parameters and algorithm settings. To solve problem (StochOpt), we consider the stochastic gradient mirror descent method as follows

$$\beta_{t+1} := \underset{\beta\in\mathscr{B}}{\operatorname{argmin}}\{\eta_t\langle\nabla f(\beta_t,\xi_t), \beta - \beta_t\rangle + D_\omega(\beta_t,\beta)\}, \tag{SGMD}$$

for all $t \geq 0$, where $\nabla f(\beta_t, \xi_t)$ denotes the gradient of the stochastic function $f(\cdot, \xi_t)$ at β_t. Throughout this section, we let $F(x)$ have Lipschitz gradients with parameter $\mathfrak{L}_F > 0$. We also define the stochastic errors z_t as the difference between the sample gradient $\nabla f(\beta_t, \xi_t)$ and $\nabla F(\beta_t)$, i.e.,

$$z_t \triangleq \nabla f(\beta_t, \xi_t) - \nabla F(\beta_t). \tag{2.22}$$

We make the following assumption on the first and second moment of the stochastic errors.

Assumption 2 (First and second moment of stochastic gradients)**.** The errors z_t are such that a.s. we have $\mathbb{E}[z_t \mid \mathcal{F}_t] = 0$ for all $t \geq 0$. Moreover, there exists some $\nu > 0$ such that

$$\mathbb{E}\left[\|z_t\|_*^2 | \mathcal{F}_t\right] \leq \nu^2, \quad \text{for all } t \geq 0. \tag{2.23}$$

It is worth mentioning that the preceding assumption does not require boundedness of the gradients and can be seen weaker than the Assumption 1. Indeed, as it will be shown later in this section, utilizing the Lipschitzian property the convergence properties of the (SGMD) method can be established under this weaker assumption. Next, we have the following lemma that provides a recursive bound on the error of the algorithm. This result will play a key role in deriving the self-tuned stepsize rules in the sequel.

Lemma 5. [A recursive error bound for the (SGMD) scheme] Let Assumption 2 hold and let β_t be generated by the (SGMD) method. We have a.s. for all $t \geq 0$

$$\mathbb{E}[D_\omega(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t] \leq \left(1 - \frac{2\eta_t\mu_F}{\mathfrak{L}_\omega} + \frac{2\eta_t^2\mathfrak{L}_F^2}{\mu_\omega^2}\right) D_\omega(\beta_t, \beta^*) + \frac{\nu^2\eta_t^2}{\mu_\omega}, \tag{2.24}$$

where β^* is the unique optimal solution to problem (StochOpt).

Proof. By the first order optimality conditions for problem (StochOpt), we have

$$\langle \nabla F(\beta^*), \beta_{t+1} - \beta^* \rangle \geq 0, \quad \text{for all } t \geq 0.$$

Consider relation (2.7) and let $\beta := \beta^*$. Adding the resulting relation with $\eta_t\langle \nabla F(\beta^*), \beta_{t+1} - \beta^* \rangle \geq 0$, we obtain

$$\langle \nabla\omega(\beta_{t+1}) - \nabla\omega(\beta_t), \beta^* - \beta_{t+1} \rangle \geq \eta_t\langle \nabla f(\beta_t, \xi_t) - \nabla F(\beta^*), \beta_{t+1} - \beta^* \rangle. \tag{2.25}$$

From relation (2.2), we get

$$\langle \nabla\omega(\beta_{t+1}) - \nabla\omega(\beta_t), \beta^* - \beta_{t+1}\rangle = D_\omega(\beta_t, \beta^*) - D_\omega(\beta_{t+1}, \beta^*) - D_\omega(\beta_t, \beta_{t+1}).$$

Therefore, from relation (2.25) and relation (2.3) we have,

$$D_\omega(\beta_t, \beta^*) - D_\omega(\beta_{t+1}, \beta^*) - \frac{\mu_\omega}{2}\|\beta_t - \beta_{t+1}\|^2 \geq \eta_t\langle \nabla f(\beta_t, \xi_t) - \nabla F(\beta^*), \beta_{t+1} - \beta^*\rangle. \quad (2.26)$$

Next, we find a lower bound for the term on the right-hand side. By adding and subtracting $\eta_t\langle \nabla f(\beta_t, \xi_t) - \nabla F(\beta^*), \beta_t\rangle$, we get

$$\begin{aligned} \eta_t\langle \nabla f(\beta_t, \xi_t) - \nabla F(\beta^*), \beta_{t+1} - \beta^*\rangle &= \eta_t\langle \nabla f(\beta_t, \xi_t) - \nabla F(\beta^*), \beta_{t+1} - \beta_t\rangle \\ &\quad + \eta_t\langle \nabla f(\beta_t, \xi_t) - \nabla F(\beta^*), \beta_t - \beta^*\rangle \\ &\geq -\frac{\eta_t^2}{2\mu_\omega}\|\nabla f(\beta_t, \xi_t) - \nabla F(\beta^*)\|_*^2 - \frac{\mu_\omega}{2}\|\beta_{t+1} - \beta_t\|^2 \\ &\quad + \eta_t\langle \nabla f(\beta_t, \xi_t) - \nabla F(\beta^*), \beta_t - \beta^*\rangle, \end{aligned} \quad (2.27)$$

where the last inequality follows from Fenchel's inequality, i.e., $|\langle x, y\rangle| \leq \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|_*^2$. Combining (2.26) and (2.27) yields

$$\begin{aligned} D_\omega(\beta_{t+1}, \beta^*) \leq D_\omega(\beta_t, \beta^*) &- \eta_t\langle \nabla f(\beta_t, \xi_t) - \nabla F(\beta^*), \beta_t - \beta^*\rangle \\ &+ \frac{\eta_t^2}{2\mu_\omega}\|\nabla f(\beta_t, \xi_t) - \nabla F(\beta^*)\|_*^2. \end{aligned}$$

Using relation (2.22), and invoking the triangle inequality and relation $(a + b)^2 \leq 2a^2 + 2b^2$

for any $a, b \in \mathbb{R}$, we obtain

$$D_\omega(\beta_{t+1}, \beta^*) \leq D_\omega(\beta_t, \beta^*) - \eta_t \langle \nabla F(\beta_t) - \nabla F(\beta^*) + z_t, \beta_t - \beta^* \rangle$$
$$+ \frac{\eta_t^2}{\mu_\omega} \|\nabla F(\beta_t) - \nabla F(\beta^*)\|_*^2 + \frac{\eta_t^2}{\mu_\omega} \|z_t\|_*^2.$$

By taking the expectations on $\mathcal{F}_t$ from both sides of the preceding relation, and using Assumption 2, we have

$$\mathbb{E}[D_\omega(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t] \leq D_\omega(\beta_t, \beta^*) - \eta_t \langle \nabla F(\beta_t) - \nabla F(\beta^*), \beta_t - \beta^* \rangle$$
$$+ \frac{\eta_t^2}{\mu_\omega} \|\nabla F(\beta_t) - \nabla F(\beta^*)\|_*^2 + \frac{\nu^2 \eta_t^2}{\mu_\omega}.$$

Under Lipschitzian property of ∇F with parameter $\mathfrak{L}_F$, and strong convexity of F with parameter μ_F, we get

$$\mathbb{E}[D_\omega(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t] \leq D_\omega(\beta_t, \beta^*) - \eta_t \mu_F \|\beta_t - \beta^*\|^2 + \frac{\eta_t^2 \mathfrak{L}_F^2}{\mu_\omega} \|\beta_t - \beta^*\|^2 + \frac{\nu^2 \eta_t^2}{\mu_\omega}.$$

Recalling relations (2.3) and (2.4), we obtain the desired inequality. □

Inequality (2.24) provides a closed-form function for an upper bound of the error of the (SGMD) scheme. Comparing this relation with the result of Lemma 2, we observe that the inequalities differ from two aspects: (i) the *contraction term* multiplied by the term $D_\omega(\beta_t, \beta^*)$ in the nonsmooth case is smaller than that in the smooth case; (ii) the upper bound in the smooth case is independent of the bound on the gradient, i.e., constant C. Instead the relation is characterized by the bound on the stochastic errors, that is denoted by ν. Next, we present Self-tuned stepsizes for the (SGMD) method and show their properties.

Preposition 2. [Self-tuned stepsizes for (SGMD) scheme] Let $\{\beta_t\}$ be generated by the (SGMD) method. Let the function F be strongly convex with modulus μ_F and the set $\mathscr{B}$ be convex, closed, and bounded such that $\|\beta\| \leq M$ for all $\beta \in \mathscr{B}$ and some $M > 0$. Let Assumption 2 hold for some $\nu > 0$, and the stepsize η_t be given by

$$\eta_0^* := \frac{2\mu_F \mu_\omega^2 M^2}{\nu^2 \mu_\omega + 4\mathfrak{L}_\omega \mathfrak{L}_F^2 M^2}, \qquad \eta_t^* := \eta_{t-1}^* \left(1 - \frac{\mu_F}{\mathfrak{L}_\omega}\eta_{t-1}^*\right), \quad \text{for all } t \geq 1.$$

Then, the following hold:

(a) The sequence $\{\beta_t\}$ generated by the (SGMD) method converges a.s. to the unique optimal solution β^* of problem (StochOpt).

(b) For any $t \geq 1$, the vector $(\eta_0^*, \eta_1^*, \ldots, \eta_{t-1}^*)$ minimizes the upper bound of the error $\mathbb{E}[D_\omega(\beta_t, \beta^*)]$ given in Lemma 5 for all $(\eta_0, \eta_1, \ldots, \eta_{t-1}) \in \left(0, \frac{\mathfrak{L}_\omega}{2\mu_F}\right]^t$.

(c) The (SGMD) method attains the convergence rate $\mathcal{O}(1/t)$, i.e, for all $t \geq 1$

$$\mathbb{E}\left[\|\beta_t - \beta^*\|^2\right] \leq 2\left(\frac{\mathfrak{L}_\omega}{\mu_\omega \mu_F}\right)^2 \left(\nu^2 + \frac{4\mathfrak{L}_\omega \mathfrak{L}_F^2 M^2}{\mu_\omega}\right)\frac{1}{t}.$$

(d) Let ϵ and ρ be arbitrary positive scalars and $T \triangleq 2\left(\nu^2 + \frac{4\mathfrak{L}_\omega \mathfrak{L}_F^2 M^2}{\mu_\omega}\right)\left(\frac{3\mathfrak{L}_\omega^2}{2\mu_\omega \mu_F^2}\right)\frac{1}{\epsilon\rho}$ we have for all $t \geq T$

$$\text{Prob}\left(D_\omega(\beta_j, \beta^*) \leq \epsilon \text{ for all } j \geq t\right) \geq 1 - \rho.$$

Proof. Consider the inequality given in Lemma 5. Taking expectations from both sides and rearranging the terms, we can write

$$\mathbb{E}[D_\omega(\beta_{t+1}, \beta^*)] \leq \left(1 - \frac{2\eta_t \mu_F}{\mathfrak{L}_\omega}\right)\mathbb{E}[D_\omega(\beta_t, \beta^*)] + \frac{2\eta_t^2 \mathfrak{L}_F^2}{\mu_\omega^2}\mathbb{E}[D_\omega(\beta_t, \beta^*)] + \frac{\nu^2 \eta_t^2}{\mu_\omega}.$$

From relation (2.4), the triangle inequality, and the definition of constant M, we have

$$D_\omega(\beta_t, \beta^*) \leq \frac{\mathfrak{L}_\omega}{2}\|\beta_t - \beta^*\|^2 \leq \frac{\mathfrak{L}_\omega}{2}\left(\|\beta_t\|^2 + \|\beta^*\|^2 + 2\|\beta_t\|\|\beta^*\|\right) \leq 2\mathfrak{L}_\omega M^2.$$

From the preceding inequalities, we obtain the following relation

$$\mathbb{E}[D_\omega(\beta_{t+1},\beta^*)] \leq \left(1-\frac{2\eta_t\mu_F}{\mathfrak{L}_\omega}\right)\mathbb{E}[D_\omega(\beta_t,\beta^*)] + \left(\frac{8\mathfrak{L}_\omega\mathfrak{L}_F^2M^2}{\mu_\omega}+2\nu^2\right)\frac{\eta_t^2}{2\mu_\omega}.$$

Let us define $\bar{C}$ such that $\bar{C}^2 \triangleq \frac{8\mathfrak{L}_\omega\mathfrak{L}_F^2M^2}{\mu_\omega}+2\nu^2$. Note that the preceding inequality is similar to the relation (2.18) where C is replaced by the term $\bar{C}$. Therefore, the desired results here follow by only substituting C by $\bar{C}$ in Proposition 1. It is only remained to show that: (i) $\eta_0^* = (4\mu_F\mu_\omega M^2)/\bar{C}^2$, and (ii) the conditions of Proposition 1 also hold for $\bar{C}$. The relation (i) holds directly from definition of η_0^* given by Proposition 2 and the definition of $\bar{C}$. To show (ii), we need to verify that $\bar{C}^2\mathfrak{L}_\omega \geq 8M^2\mu_\omega\mu_F^2$. Since $\nu^2>0$, from definition of $\bar{C}$ we have

$$\frac{\bar{C}^2\mathfrak{L}_\omega}{8M^2\mu_\omega\mu_F^2} = \left(\frac{8\mathfrak{L}_\omega\mathfrak{L}_F^2M^2}{\mu_\omega}+2\nu^2\right)\frac{\mathfrak{L}_\omega}{8M^2\mu_\omega\mu_F^2} \geq \left(\frac{8\mathfrak{L}_\omega\mathfrak{L}_F^2M^2}{\mu_\omega}\right)\frac{\mathfrak{L}_\omega}{8M^2\mu_\omega\mu_F^2} = \left(\frac{\mathfrak{L}_\omega\mathfrak{L}_F}{\mu_\omega\mu_F}\right)^2 \geq 1,$$

where the last relation follows since $\mu_F \leq \mathfrak{L}_F$ and $\mu_\omega \leq \mathfrak{L}_\omega$. Therefore, the conditions of Proposition 1 hold for $\bar{C}$ and the desired results follow. □

2.2.3 Unifying Self-tuned Stepsizes

Recall that Proposition 1 provides self-tuned stepsize rules for the case where problem (StochOpt) is nonsmooth, while Proposition 2 provides stepsize rules when the problem is smooth. These update rules are characterized in terms of problem parameters such as $M, C, \nu, \mu_F, \mathfrak{L}_F$ and algorithm settings such as $\mu_\omega, \mathfrak{L}_\omega$. A challenge associated with implementing these schemes pertains to the applications where some of the problem parameters are not known in advance, or are challenging to estimate. In such cases, an important question is how we may employ such self-tuned stepsize rules? To address this question, in this section, our goal is to develop a unifying class of self-tuned stepsize rules that can be employed for solving problem (StochOpt) in both smooth and nonsmooth cases when some of the problem parameters are unavailable. Let us compare the stepsize rules in Proposition 1 and 2. We observe that although the initial stepsize η_0^* is different, both schemes share the same tuning rule given by $\eta_{t+1}^* := \eta_t^*\left(1 - \frac{\mu_F}{\mathfrak{L}_\omega}\eta_t^*\right)$. We also observe that the only problem parameter that is needed to be known for the tuning update rule is μ_F. This parameter is known in advance in many applications such as SVM. It is worth emphasizing that $\mathfrak{L}_\omega$ is not a problem parameter. It is the Lipschitzian parameter associated with the prox mapping and depends on the choice of the distance generating function $\omega(\beta)$. This function is user-specified. For example, for stochastic subgradient/gradient methods we set $\omega(\beta) := \frac{1}{2}\|\beta\|_2^2$, and therefore $\mathfrak{L}_\omega = 1$. In practice, when problem parameters such as M, C, ν, or $\mathfrak{L}_F$ are unavailable or difficult to estimate, the initial stepsize η_0^* cannot be evaluated. In such cases one may choose η_0^* arbitrarily and still use the update rule $\eta_{t+1}^* := \eta_t^*\left(1 - \frac{\mu_F}{\mathfrak{L}_\omega}\eta_t^*\right)$. We show that even under this relaxation, some of the main properties of the self-tuned stepsizes are preserved. This is presented by the following result.

Theorem 1. [Unifying self-tuned stepsize rules] Consider problem (StochOpt). Let the function F be strongly convex with modulus μ_F and the set $\mathscr{B}$ be convex, closed, and bounded. Suppose either of the following cases holds:

case (1): F is non-differentiable and Assumption 1 holds for some unknown $C > 0$.

case (2): F is continuously differentiable over $\mathscr{B}$ for all ξ, but ∇F is not Lipschitz over $\mathscr{B}$ and Assumption 1 holds.

case (3): F is differentiable over $\mathscr{B}$, it has Lipschitz gradients with an unknown parameter $\mathfrak{L}_F$, and Assumption 2 holds.

In case (1), let $\{\beta_t\}$ be generated by algorithm (SSMD). In cases (2) and (3) let $\{\beta_t\}$ be generated by algorithm (SGMD). In all these cases, let the stepsize η_t be given by

$$\eta_t := \eta_{t-1}\left(1 - \frac{\mu_F}{\mathfrak{L}_\omega}\eta_{t-1}\right), \quad \text{for all } t \geq 1,$$

where $0 < \eta_0 \leq \frac{\mathfrak{L}_\omega}{2\mu_F}$ is an arbitrary constant. Then: (i) $\{\beta_t\}$ converges to β^* a.s., and (ii) there exists a threshold $\bar{\eta} \leq \frac{\mathfrak{L}_\omega}{2\mu_F}$ such that for any $\eta_0 \leq \bar{\eta}$, an upper bound of the error $\mathbb{E}[D_\omega(\beta_t, \beta^*)]$ is minimized for all $(\eta_0, \eta_1, \ldots, \eta_{t-1}) \in \left(0, \frac{\mathfrak{L}_\omega}{2\mu_F}\right]^t$.

Proof. First, we show (i) and (ii) hold in case (1). Let C_{min} denote the minimum of all constants $C > 0$ that satisfy Assumption 1 (note that such a constant always exits). Let $\bar{C} \triangleq \max\left\{C_{min}, \sqrt{\frac{8M^2\mu_\omega\mu_F^2}{\mathfrak{L}_\omega}}\right\}$ and define $\bar{\eta} \triangleq \frac{4\mu_F\mu_\omega M^2}{\bar{C}^2}$. Note that $\bar{\eta} \leq \frac{\mathfrak{L}_\omega}{2\mu_F}$ from definition of $\bar{C}$. Let $0 < \eta_0 \leq \bar{\eta}$ be an arbitrary scalar and define $C_0 \triangleq \bar{C}\sqrt{\frac{\bar{\eta}}{\eta_0}}$. Note that since $C_0 \geq \bar{C} \geq C_{min}$, C_0 satisfies Assumption 1. Also, $C_0^2\mathfrak{L}_\omega \geq 8M^2\mu_\omega\mu_F^2$. Therefore, for $\eta_0 = \frac{4\mu_F\mu_\omega M^2}{{C_0}^2}$, we found a C_0 such that all conditions of Proposition 1 are met. Then we can apply Proposition 1 which implies that (i) and (ii) hold. Next, consider case (2). Note that since f is continuously differentiable, the set $\partial f(\beta, \xi)$ is a singleton, i.e., $\{\nabla f(\beta, \xi)\}$. From compactness of $\mathscr{B}$ and continuity of $\nabla f(\cdot, \xi)$, we conclude that Assumption 1 holds for some $C > 0$. Next, in a similar fashion to the proof of case (1), we can conclude that (i) and (ii) hold in case (2). The proof for case (3) can be done by invoking Proposition 2 similar to the proof for case (1). □

Remark 1. The unifying stepsize rule minimizes the mean squared error even when problem parameters are unknown. This suggests that self-tuned stepsizes are robust with respect to the choice of the initial stepsize. This indeed suggests that self-tuned stepsizes are robust with respect to the choice of the initial stepsize. We will demonstrate this property of the self-tuned stepsizes in our numerical experiments in Section 2.4. This can be seen as an important advantage in contrast with the classical harmonic stepsizes of the form $\frac{a}{(t+b)^c}$ that have been seen very sensitive to the choice of three parameters a, b and c (cf. Spall [2005]).

2.3 Self-tuned Randomized Block Coordinate SMD Methods

In many big data applications such as text classification, the dimensionality of the solution space, i.e., n, is huge. Consequently, each iteration of the mirror descent methods becomes computationally inefficient. To address this challenge, our goal is to develop randomized block coordinate variants of the self-tuned stochastic mirror descent method. We consider problem (StochOpt), where the set $\mathscr{B} \in \mathbb{R}^n$ has the block structure given by $\mathscr{B} \triangleq \prod_{i=1}^{l} \mathscr{B}_i$, where $\mathscr{B}_i \in \mathbb{R}^{n_i}$ and $n \triangleq \sum_{i=1}^{l} n_i$. We start with the case where the objective function is non-differentiable. Later, in Section 2.3.2, we discuss the case of differentiable objective functions with Lipschitz gradients. Let the distance generating function $\omega_i : \mathbb{R}^{n_i} \to \mathbb{R}$ be a continuously differentiable function. The Bregman divergence $D_{\omega_i} : \mathbb{R}^{n_i} \times \mathbb{R}^{n_i} \to \mathbb{R}$ associated with ω_i is given for $\beta_1, \beta_2 \in \mathscr{B}_i$ as

$$D_{\omega_i}(\beta_1, \beta_2) = \omega_i(\beta_2) - \omega_i(\beta_1) - \langle \nabla \omega_i(\beta_1), \beta_2 - \beta_1 \rangle.$$

Let $\nabla_{\beta_2} D_{\omega_i}(\cdot, \cdot)$ denote the partial derivative of $D_{\omega_i}(\beta_1, \beta_2)$ with respect to β_2. Then,

$$\nabla_{\beta_2} D_{\omega_i}(\beta_1, \beta_2) = \nabla \omega_i(\beta_2) - \nabla \omega_i(\beta_1), \quad \text{for all} \quad \beta_1, \beta_2 \in \mathscr{B}_i. \tag{2.28}$$

The Bregman divergence has the following property for all $\beta_1, \beta_2, \beta_3 \in \mathscr{B}_i$

$$D_{\omega_i}(\beta_1, \beta_2) - D_{\omega_i}(\beta_3, \beta_2) = D_{\omega_i}(\beta_1, \beta_3) + \langle \nabla \omega_i(\beta_3) - \nabla \omega_i(\beta_1), \beta_2 - \beta_3 \rangle. \tag{2.29}$$

We assume the distance generating function ω_i has Lipschitz gradients with parameter $\mathfrak{L}_{\omega_i}$ and is strongly convex with parameter μ_{ω_i}, i.e., for all $\beta_1, \beta_2, \beta_3 \in \mathscr{B}_i$

$$\frac{\mu_{\omega_i}}{2} \|\beta_2 - \beta_1\|^2 \leq D_{\omega_i}(\beta_1, \beta_2) \leq \frac{\mathfrak{L}_{\omega_i}}{2} \|\beta_2 - \beta_1\|^2. \tag{2.30}$$

Remark 2. Lipschitzian property of ω_i is a standard assumption in the literature of SMD methods; the convergence rate analysis provided in [Nedić and Lee, 2014; Dang and Lan, 2015b] relies on this property. Also note that for the stochastic gradient descent (SGD) method, we have $\mu_\omega = \mathfrak{L}_\omega = 1$.

The prox mapping $\mathcal{P}_i : \mathscr{B}_i \times \mathbb{R}^{n_i} \to \mathscr{B}_i$ is defined by

$$\mathcal{P}_i(\beta_1, \beta_2) = \underset{z \in \mathscr{B}_i}{\operatorname{argmin}} \{\langle \beta_2, z \rangle + D_i(\beta_1, z)\}, \tag{2.31}$$

for all $\beta_1 \in \mathscr{B}_i$ and $\beta_2 \in \mathbb{R}^{n_i}$. In the analysis, we use the following error function $\mathcal{L} : \mathscr{B} \times \mathscr{B} \to \mathbb{R}$ defined as

$$\mathcal{L}(\beta, z) \triangleq \sum\nolimits_{i=1}^{l} p_i^{-1} D_i(\beta^i, z^i), \quad \text{for all } \beta, z \in \mathscr{B}. \tag{2.32}$$

2.3.1 Self-tuned Randomized Block Subgradient SMD Method

Consider problem (StochOpt) where F is a non-differentiable convex function of β. Let $\mathbf{g}_t \in \partial F(\beta_t)$ denote a subgradient of function F at point $\beta_t \in \mathscr{B}$. Similarly, for any $\xi \in \Omega$, we let $\tilde{g}_t \in \partial f(\beta_t, \xi)$ denote a subgradient of function $f(\cdot, \xi)$ at point β_t. Throughout, we assume that F is strongly convex with parameter $\mu_F > 0$ over the set $\mathscr{B}$ with respect to the underlying norm $\|\cdot\|$.

Next we present the outline of the randomized block coordinate SMD method. Let P_b be a discrete probability distribution with probabilities $p_i > 0$ for $i = 1, \ldots, l$, where $\sum_{i=1}^{l} p_i = 1$. Given an initial vector $\beta_0 \in \mathscr{B}$, at iteration $t \geq 1$, random variable i_t is generated from the probability distribution P_b independently from random variable ξ. Then, only the i_tth block of β_t, i.e. $\beta_t^{i_t}$, is updated as follows:

$$\beta_{t+1}^i = \begin{cases} \mathcal{P}_{i_t}\big(\beta_t^{i_t}, \eta_t \tilde{g}_{i_t}(\beta_t)\big) & \text{if } i = i_t, \\ \beta_t^i & \text{if } i \neq i_t, \end{cases} \tag{RB-SSMD}$$

where $\tilde{g}_{i_t}(\beta_t)$ is the i_tth block of the subgradient of $f(\beta_t, \xi_t)$ and η_t is the stepsize. Throughout, let $\mathcal{F}_t = \{i_0, \xi_0, \ldots, i_{t-1}, \xi_{t-1}\}$. Next, we state the main assumptions.

Assumption 3. Let the stochastic subgradient $\tilde{g}(\beta) \in \partial f(\beta, \xi)$ be such that a.s. for all $\beta \in \mathscr{B}$, we have $\mathbb{E}[\tilde{g}(\beta)|\beta] = \mathbf{g}(\beta) \in \partial F(\beta)$. Moreover, for all $i = 1, \ldots, l$ and $\beta \in \mathscr{B}$, there exists a scalar $C_i > 0$ such that $\mathbb{E}\big[\|\tilde{g}_i(\beta)\|_{*_i}^2 |\beta\big] \leq C_i^2$.

Next, we develop a recursive inequality in terms of the error of the (RB-SSMD) scheme. Such a recursive inequality will be employed to develop a self-tuned stepsize rule.

Lemma 6. Let Assumption 3 hold and β_t be generated by the (RB-SSMD) scheme. Then for all $t \geq 0$,

$$\mathbb{E}[\mathcal{L}(\beta_{t+1}, \beta^*)|\mathcal{F}_t] \leq \left(1 - \eta_t 2\mu_F p_\wedge \mathfrak{L}_{max}^{-1}\right) \mathcal{L}(\beta_t, \beta^*) + \eta_t^2 \sum\nolimits_{i=1}^l C_i^2 (2\mu_{\omega_i})^{-1}. \tag{2.33}$$

Proof. At iteration t, we have $\beta_{t+1}^{i_t} = \mathcal{P}_{i_t}\big(\beta_t^{i_t}, \eta_t \tilde{g}_{i_t}(\beta_t)\big)$. Consider the definition of $\mathcal{P}_{i_t}$ given by (2.31). Writing the optimality condition, we have

$$\langle \eta_t \tilde{g}_{i_t} + \nabla D_{i_t}(\beta_t^{i_t}, \beta_{t+1}^{i_t}), \beta^{i_t} - \beta_{t+1}^{i_t} \rangle \geq 0, \quad \text{for all } \beta \in \mathscr{B}.$$

Using relations (2.28) and (2.29), and from the preceding relation,

$$D_{i_t}(\beta_t^{i_t}, \beta^{i_t}) - D_{i_t}(\beta_{t+1}^{i_t}, \beta^{i_t}) - D_{i_t}(\beta_t^{i_t}, \beta_{t+1}^{i_t}) \geq \eta_t \langle \tilde{g}_{i_t}, \beta_{t+1}^{i_t} - \beta^{i_t} \rangle, \quad \text{for all } \beta \in\mathscr{B}.$$

From the strong convexity of ω_{i_t} and relation (2.30), we have

$$D_{i_t}(\beta_t^{i_t}, \beta^{i_t}) - D_{i_t}(\beta_{t+1}^{i_t}, \beta^{i_t}) - 0.5\mu_{\omega_{i_t}} \|\beta_t^{i_t} - \beta_{t+1}^{i_t}\|_{i_t}^2 \geq \eta_t \langle \tilde{g}_{i_t}, \beta_{t+1}^{i_t} - \beta^{i_t} \rangle. \tag{2.34}$$

By adding and subtracting $\eta_t \langle \tilde{g}_{i_t}, \beta_t^{i_t} \rangle$ in the right-hand side, and using Fenchel's inequality, we have

$$\begin{aligned} \eta_t \langle \tilde{g}_{i_t}, \beta_{t+1}^{i_t} - \beta_t^{i_t} \rangle + \eta_t \langle \tilde{g}_{i_t}, \beta_t^{i_t} - \beta^{i_t} \rangle \geq -0.5\eta_t^2 \mu_{\omega_{i_t}}^{-1} \|\tilde{g}_{i_t}\|_{*i_t}^2 - 0.5\mu_{\omega_{i_t}} \|\beta_{t+1}^{i_t} - \beta_t^{i_t}\|_{i_t}^2 \\ + \eta_t \langle \tilde{g}_{i_t}, \beta_t^{i_t} - \beta^{i_t} \rangle. \end{aligned} \tag{2.35}$$

Combining (2.34) and (2.35) yields for all $\beta \in \mathscr{B}$

$$D_{i_t}(\beta_{t+1}^{i_t}, \beta^{i_t}) \leq D_{i_t}(\beta_t^{i_t}, \beta^{i_t}) + \eta_t \langle \tilde{g}_{i_t}, \beta^{i_t} - \beta_t^{i_t} \rangle + 0.5\eta_t^2 \mu_{\omega_{i_t}}^{-1} \|\tilde{g}_{i_t}\|_{*i_t}^2.$$

From the preceding relation, relation (2.32), and that $\beta_{t+1}^i = \beta_t^i$ for all $i \neq i_t$, we have

$$\begin{aligned} \mathcal{L}(\beta_{t+1}, \beta) &\leq \sum\nolimits_{i \neq i_t} p_i^{-1} D_i(\beta_t^i, \beta^i) + p_{i_t}^{-1}\big(D_{i_t}(\beta_t^{i_t}, \beta^{i_t}) + \eta_t \langle \tilde{g}_{i_t}, \beta^{i_t} - \beta_t^{i_t} \rangle + 0.5\eta_t^2 \mu_{\omega_{i_t}}^{-1} \|\tilde{g}_{i_t}\|_{*i_t}^2\big) \\ &= \mathcal{L}(\beta_t, \beta) + p_{i_t}^{-1} \left(\eta_t \langle \tilde{g}_{i_t}, \beta^{i_t} - \beta_t^{i_t} \rangle + 0.5\eta_t^2 \mu_{\omega_{i_t}}^{-1} \|\tilde{g}_{i_t}\|_{*i_t}^2\right). \end{aligned}$$

Taking conditional expectations from both sides of the preceding relation on $\mathcal{F}_t \cup \{i_t\}$, we get

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\beta_{t+1}, \beta) \mid \mathcal{F}_t \cup \{i_t\}] &\leq \mathcal{L}(\beta_t, \beta) + 0.5\eta_t^2 \mu_{\omega_{i_t}}^{-1} p_{i_t}^{-1} \mathbb{E}\big[\|\tilde{g}_{i_t}\|_{*i_t}^2 \mid \mathcal{F}_t \cup \{i_t\}\big] \\ &\quad + \frac{\eta_t}{p_{i_t}} \langle \mathbb{E}[\tilde{g}_{i_t} \mid \mathcal{F}_t \cup \{i_t\}], \beta^{i_t} - \beta_t^{i_t} \rangle \\ &\leq \mathcal{L}(\beta_t, \beta) + p_{i_t}^{-1} \eta_t \left\langle \mathbf{g}_{i_t}, \beta^{i_t} - \beta_t^{i_t} \right\rangle + p_{i_t}^{-1} \eta_t^2 \frac{C_{i_t}^2}{2\mu_{\omega_{i_t}}}, \end{aligned}$$

where we used Assumption 3. Taking expectations from previous inequality with respect to

i_t and setting $\beta := \beta^*$,

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t] &\leq \mathcal{L}(\beta_t, \beta^*) + \sum_{i=1}^{l} \frac{p_i}{p_i} \left(\eta_t \left\langle \mathbf{g}_i, \beta^{*i} - \beta_t^i \right\rangle + \eta_t^2 \frac{C_i^2}{2\mu_{\omega_i}} \right) \\ &= \mathcal{L}(\beta_t, \beta^*) + \eta_t \left\langle \mathbf{g}_t, \beta^* - \beta_t \right\rangle + \eta_t^2 \sum_{i=1}^{l} \frac{C_i^2}{2\mu_{\omega_i}},\end{aligned}$$

where we use the definition of $\langle \cdot, \cdot \rangle$ given in the notation. From strong convexity of function F, we have $\langle \mathbf{g}_t - \mathbf{g}^*, \beta_t - \beta^* \rangle \geq \mu_F \|\beta_t - \beta^*\|^2$. By optimality of β^*, we have $\langle \mathbf{g}^*, \beta_t - \beta^* \rangle \geq 0$. From the two preceding relations and the definition of norm,

$$\begin{aligned}\langle \mathbf{g}_t, \beta_t - \beta^* \rangle &\geq \mu_F \sum_{i=1}^{l} \|\beta_t^i - \beta^{*i}\|_i^2 \geq 2\mu_F \sum_{i=1}^{l} \frac{D_i(\beta_t^i, \beta^{*i})}{\mathfrak{L}_{\omega_i}} \\ &\geq 2\mu_F p_\wedge \mathfrak{L}_{max}^{-1} \sum_{i=1}^{l} p_i^{-1} D_i(\beta_t^i, \beta^{*i}) = 2\mu_F p_\wedge \mathfrak{L}_{max}^{-1} \mathcal{L}(\beta_t, \beta^*),\end{aligned}$$

where in the second inequality we used relation (2.30), and in the last relation we used the definition of function $\mathcal{L}$. From the preceding two relations, we obtain the desired inequality. □

Next, we present self-tuned stepsizes and their properties for the (RB-SSMD) method.

Preposition 3. Let $\{\beta_t\}$ be generated by the (RB-SSMD) method. Let the sets $\mathscr{B}_i$ be convex and closed such that $\|\beta^i\| \leq M_i$ for all $\beta^i \in \mathscr{B}_i$ and some $M_i > 0$, for all i. Let Assumption 3 hold for some C_i large enough such that $C_i^2 \mathfrak{L}_{\omega_i} \geq 8M_i^2 \mu_{\omega_i} \mu_F^2$ for all i. Let the stepsize η_t be given by

$$\begin{aligned}\eta_0^* &:= \frac{4\mu_F p_\wedge \sum_{i=1}^{l} p_i^{-1} \mathfrak{L}_{\omega_i} M_i^2}{\mathfrak{L}_{max} \sum_{i=1}^{l} \mu_{\omega_i}^{-1} C_i^2}, \\ \eta_t^* &:= \eta_{t-1}^* \left(1 - p_\wedge \mu_F \mathfrak{L}_{max}^{-1} \eta_{t-1}^*\right), \quad \text{for all } t \geq 1.\end{aligned}$$

Then, the following hold:

(a) The sequence $\{\beta_t\}$ converges a.s. to the unique optimal solution β^* of problem (StochOpt).

(b) For any $t \geq 1$, the vector $(\eta_0^*, \ldots, \eta_{t-1}^*)$ minimizes the upper bound of the error $\mathbb{E}[\mathcal{L}(\beta_t, \beta^*)]$ given in Lemma 6 for all $(\eta_0, \ldots, \eta_{t-1}) \in \left(0, \frac{\mathfrak{L}_{max}}{2p_\wedge \mu_F}\right]^t$.

(c) The (RB-SSMD) method attains the convergence rate $\mathcal{O}(1/t)$, i.e, for all $t \geq 1$ $\mathbb{E}[\|\beta_t - \beta^*\|^2] \leq \frac{p_\vee}{\mu_{min}} \sum_{i=1}^l \frac{C_i^2}{\mu_{\omega_i}} \left(\frac{\mathfrak{L}_{max}}{p_\wedge \mu_F}\right)^2 \frac{1}{t}$.

(d) Let ϵ and ρ be arbitrary positive scalars and $T \triangleq 1.5 \left(\frac{\mathfrak{L}_{max}}{p_\wedge \mu_F}\right)^2 \sum_{j=1}^l \frac{C_j^2}{\mu_{\omega_j}} \frac{1}{\epsilon\rho}$ we have for all $t \geq T$

$$\text{Prob}\left(\mathcal{L}(\beta_j, \beta^*) \leq \epsilon \text{ for all } j \geq t\right) \geq 1 - \rho.$$

Proof. (a) To show a.s. convergence, we apply Lemma 4. Consider the inequality (2.33) given by Lemma 6. Let us define

$$v_t \triangleq \mathcal{L}(\beta_t, \beta^*), \quad \alpha_t \triangleq \frac{2\mu_F p_\wedge}{\mathfrak{L}_{max}} \eta_t^*, \quad \lambda_t \triangleq \sum_{i=1}^l \frac{C_i^2}{2\mu_{\omega_i}} \eta_t^{*2}. \tag{2.36}$$

From definition of η_0^* and $C_i^2 \mathfrak{L}_{\omega_i} \geq 8M_i^2 \mu_{\omega_i} \mu_F^2$, we have

$$\alpha_0 = \frac{8\mu_F^2 p_\wedge^2 \sum_{i=1}^l \frac{\mathfrak{L}_{\omega_i} M_i^2}{p_i}}{\mathfrak{L}_{max}^2 \sum_{i=1}^l \frac{C_i^2}{\mu_{\omega_i}}} \leq \frac{p_\wedge^2 \sum_{i=1}^l \frac{\mathfrak{L}_{\omega_i}^2 C_i^2}{p_i \mu_{\omega_i}}}{\mathfrak{L}_{max}^2 \sum_{i=1}^l \frac{C_i^2}{\mu_{\omega_i}}} \leq p_\wedge < 1. \tag{2.37}$$

Therefore, since $\{\eta_t^*\}$ is non-increasing, we have $0 \leq \alpha_t \leq 1$ for all $t \geq 0$. Moreover, from Lemma 3(c), we have that $\sum_{t=0}^\infty \alpha_t = \infty$ and $\sum_{t=0}^\infty \lambda_t < \infty$. Also, the definition of α_t and λ_t and that the self-tuned stepsize η_t^* has a limit of zero imply that $\frac{\lambda_t}{\alpha_t} \to 0$. Therefore, all conditions of Lemma 4 are met and so $\mathcal{L}(\beta_t, \beta^*) \to 0$ a.s.. The definition of $\mathcal{L}$ and that $p_i > 0$ for all i imply that $D_i(\beta_t^i, \beta^{*i}) \to 0$ for all i. Using the strong convexity of ω_i (cf. (2.30)), we have $\frac{\mu_{\omega_i}}{2} \|\beta_t^i - \beta^{*i}\|^2 \leq D_i(\beta_t^i, \beta^{*i})$ for all i. We conclude that $\beta_t \to \beta^*$ a.s..

(b) For any $t \geq 1$, let us define the function $e_t(\eta_0, \ldots, \eta_{t-1})$ given by the recursion (2.13) where $\theta \triangleq \frac{2p_\wedge \mu_F}{\mathfrak{L}_{max}}$, and $\delta \triangleq \sum_{i=1}^{l} \frac{C_i^2}{2\mu_{\omega_i}}$. Also, let $e_0 \triangleq 2\sum_{i=1}^{l} p_i^{-1} \mathfrak{L}_{\omega_i} M_i^2$. Next, we show that $\mathcal{L}(\beta_0, \beta^*) \leq e_0$. Using the Lipschitizan property of $\nabla\omega_i$, and the triangle inequality, we have

$$\mathcal{L}(\beta_0, \beta^*) = \sum_{i=1}^{l} \frac{D_i(\beta_0^i, \beta^{*i})}{p_i} \leq \sum_{i=1}^{l} \frac{\mathfrak{L}_{\omega_i}}{2p_i}\|\beta_0^i - \beta^{*i}\|_i^2 \leq \sum_{i=1}^{l} \frac{\mathfrak{L}_{\omega_i}}{2p_i}(2\|\beta_0^i\|_i^2 + 2\|\beta^{*i}\|_i^2)$$
$$\leq \sum_{i=1}^{l} \frac{2\mathfrak{L}_{\omega_i} M_i^2}{p_i} = e_0.$$

From $\mathcal{L}(\beta_0, \beta^*) \leq e_0$, relations (2.13), (2.33) and using induction, it can be seen that $\mathbb{E}[\mathcal{L}(\beta_t, \beta^*)] \leq e_t(\eta_0, \ldots, \eta_{t-1})$ for all $t \geq 0$ and any arbitrary $(\eta_0, \ldots, \eta_{t-1}) \in \left(0, \frac{\mathfrak{L}_{max}}{2p_\wedge \mu_F}\right]^t$. Therefore, e_t is a well-defined upper bound for the algorithm. To complete the proof, it suffices to show that the conditions of Lemma 3 hold. First we show that $e_0 \leq \frac{2\delta}{\theta^2}$. From the values of e_0, η_0^*, θ, and δ, we have $\eta_0^* = \frac{\theta}{2\delta} e_0$. From the definition of α_0 in (2.36) and (2.37), we have $\alpha_0 = \theta\eta_0^* < 1$. By two preceding relations we obtain $e_0 \leq \frac{2\delta}{\theta^2}$. Hence, conditions of Lemma 3 hold. From Lemma 3(a), we conclude the desired result.

(c) Following the proof of part (b), from Lemma 3(b) and definitions of δ and θ in part (b), we obtain for all $t \geq 1$

$$\mathbb{E}[\mathcal{L}(\beta_t, \beta^*)] \leq e_t \leq \left(\frac{\mathfrak{L}_{max}}{p_\wedge \mu_F}\right)^2 \sum_{i=1}^{l} \frac{C_i^2}{2\mu_{\omega_i}} \frac{1}{t}. \tag{2.38}$$

Note that from strong convexity of ω_i we have

$$\mathcal{L}(\beta_t, \beta^*) = \sum\nolimits_{i=1}^{l} p_i^{-1} D_i(\beta_t^i, \beta^{*i}) \geq \sum\nolimits_{i=1}^{l} p_i^{-1} 0.5\mu_{\omega_i}\|\beta_t^i - \beta^{*i}\|_i^2 \geq \mu_{min}(2p_\vee)^{-1}\|\beta_t - \beta^*\|^2.$$

Combining the two preceding relations completes the proof.

(d) We use the probabilistic bound given in Lemma 4. First we estimate the term $\sum_{i=t}^{\infty} \lambda_i$

where λ_i is given by (2.36). Note that Lemma 3(b) implies $\eta_i^* \leq \frac{2}{\theta i}$. Therefore, we can write

$$\sum_{i=t}^{\infty} \lambda_i = \sum_{i=t}^{\infty} \sum_{j=1}^{l} \frac{C_j^2}{2\mu_{\omega_j}} \eta_i^{*2} \leq \sum_{j=1}^{l} \frac{C_j^2}{2\mu_{\omega_j}} \sum_{i=t}^{\infty} \left(\frac{\mathfrak{L}_{max}}{p_\wedge \mu_F \; i} \right)^2 \leq \left(\frac{\mathfrak{L}_{max}}{p_\wedge \mu_F} \right)^2 \sum_{j=1}^{l} \frac{C_j^2}{2\mu_{\omega_j}} \left(\frac{1}{t} + \int_t^{\infty} \frac{1}{x^2} dx \right)$$
$$= \left(\mathfrak{L}_{max} (p_\wedge \mu_F)^{-1} \right)^2 \sum_{j=1}^{l} C_j^2 {\mu_{\omega_j}}^{-1} \left(1/t \right). \tag{2.39}$$

By (2.39), (2.38), and Lemma 4, we obtain the desired relation. □

Under a uniform distribution, i.e., $p_i = \frac{1}{l}$ for $i = 1, \ldots, l$, Proposition 3 indicates that $\mathbb{E}[\|\beta_t - \beta^*\|^2] \to 0$ with the order of $\mathcal{O}\left(\frac{l}{t}\right)$.This is similar to the error bound derived in [Dang and Lan, 2015b] for stochastic block mirror descent (SBMD) method (cf. Corollary 2.5 in [Dang and Lan, 2015b]). Next, we compare the constant factor of the error bound derived in [Dang and Lan, 2015b] with that of (RB-SSMD) method.

Comparison 2. Let Assumption 3 hold for some unknown $C_i > 0$ for all i. Let β_t be generated by algorithm (RB-SSMD) where $\mathfrak{L}_{\omega_i} = \mathfrak{L}_\omega$ and $\mu_{\omega_i} = \mu_\omega$ for all $1 \leq i \leq l$ and $\bar{\beta}_t$ be generated by SBMD method in [Dang and Lan, 2015b]. Then, By Lemma 1, we have $\mathbb{E}\left[\|\bar{\beta}_t - \beta^*\|^2\right] \leq \frac{2}{\mu_F \mu_\omega} \mathbb{E}\left[F(\bar{\beta}_t) - F(\beta^*)\right]$ and by Corollary 2.5 in [Dang and Lan, 2015b], we have $\mathbb{E}\left[F(\bar{\beta}_t) - F(\beta^*)\right] \leq \frac{2l\mathfrak{L}_\omega}{\mu_F} \sum_{i=1}^{l} C_i^2 \left(\frac{1}{t+1}\right)$. Combining the preceding inequalities, we obtain for all $t \geq 1$

$$\mathbb{E}\left[\|\bar{\beta}_t - \beta^*\|^2\right] \leq \frac{4l\mathfrak{L}_\omega}{\mu_F^2 \mu_\omega} \sum_{i=1}^{l} C_i^2 \left(\frac{1}{t+1} \right). \tag{2.40}$$

On the other hand, by Proposition 3, we have for all $t \geq 1$

$$\mathbb{E}\left[\|\beta_t - \beta^*\|^2\right] \leq \frac{l\mathfrak{L}_\omega^2}{\mu_\omega^2 \mu_F^2} \sum_{i=1}^{l} C_i^2 \left(\frac{1}{t+1} \right). \tag{2.41}$$

Comparing (2.40) and (2.41), we note that the constant factor of the error bound of

(RB-SSMD) method is smaller when $\frac{\mathfrak{L}_\omega}{\mu_\omega} < 4$. In particular, for SGD method where $\mathfrak{L}_\omega = \mu_\omega = 1$, it can be four times better than the constant factor of SBMD in [Dang and Lan, 2015b].

2.3.2 Self-tuned Randomized Block Gradient SMD Method

In this section, we assume the objective function in problem (StochOpt) is differentiable and has Lipschitz gradients. Our goal is to utilize this property and develop a self-tuned scheme that is characterized with the problem parameters and algorithm settings. To solve problem (StochOpt), we consider the randomized block gradient SMD method as follows

$$\beta_{t+1}^i = \begin{cases} \mathcal{P}_{i_t}\big(\beta_t^{i_t}, \eta_t g_{i_t}(\beta_t)\big) & \text{if } i = i_t, \\ \beta_t^i & \text{if } i \neq i_t, \end{cases} \tag{RB-GSMD}$$

for all $t \geq 0$, where $g_{i_t}(\beta_t)$ is the i_tth block of the gradient of the stochastic function $f(\cdot, \xi_t)$ at β_t. Throughout this section, we let F have Lipschitz gradients with parameter $\mathfrak{L}_F > 0$. We also define the stochastic errors z_t^i as follows

$$z_t^i \triangleq g_i(\beta_t) - \nabla F_i(\beta_t), \quad \text{for all} \quad t \geq 0, \quad \text{and for all} \quad i = 1, \ldots, l. \tag{2.42}$$

Next, we state the main assumptions on stochastic gradients.

Assumption 4. The errors z_t^i are such that a.s. we have $\mathbb{E}[z_t^i \mid \mathcal{F}_t] = 0$ for all $t \geq 0$. Moreover, there exists some $\nu_i > 0$ for all i such that $\mathbb{E}[\|z_t^i\|_{*i}^2 | \mathcal{F}_t] \leq \nu_i^2$, for all $t \geq 0$.

Next, we have the lemma that provides a recursive bound on the error of the algorithm.

Lemma 7. Let Assumption 4 hold and β_t be generated by the (RB-GSMD) method. We have a.s. for all $t \geq 0$

$$\mathbb{E}[\mathcal{L}(\beta_{t+1}, \beta^*)|\mathcal{F}_t] \leq (1 - \eta_t 2\mu_F p_\wedge \mathfrak{L}_{max}^{-1} + \eta_t^2 2{\mathfrak{L}_F}^2 p_\vee \mu_{min}^{-2})\mathcal{L}(\beta_t, \beta^*) + \eta_t^2 \sum\nolimits_{i=1}^{l} \nu_i^2 \mu_{\omega_i}^{-1}. \quad (2.43)$$

Proof. Consider the update rule (RB-GSMD). Writing the first-order optimality condition, we have for all $\beta \in \mathscr{B}$

$$\langle \eta_t g_{i_t} + \nabla D_{i_t}(\beta_t^{i_t}, \beta_{t+1}^{i_t}), \beta^{i_t} - \beta_{t+1}^{i_t} \rangle \geq 0, \quad (2.44)$$

Using equation (2.28), from (2.44) we obtain for all $\beta \in \mathscr{B}$

$$\langle \nabla \omega_{i_t}(\beta_{t+1}^{i_t}) - \nabla \omega_{i_t}(\beta_t^{i_t}), \beta^{i_t} - \beta_{t+1}^{i_t} \rangle \geq \eta_t \langle g_{i_t}, \beta_{t+1}^{i_t} - \beta^{i_t} \rangle. \quad (2.45)$$

Let $\beta := \beta^*$ in relation (2.45). Adding and subtracting the term $\eta_t \langle \nabla F_{i_t}(\beta^*), \beta_{t+1}^{i_t} - \beta^{*i_t} \rangle$, we get

$$\begin{aligned} \langle \nabla \omega_{i_t}(\beta_{t+1}^{i_t}) - \nabla \omega_{i_t}(\beta_t^{i_t}), \beta^{*i_t} - \beta_{t+1}^{i_t} \rangle &\geq \eta_t \langle g_{i_t} - \nabla F_{i_t}(\beta^*), \beta_{t+1}^{i_t} - \beta^{*i_t} \rangle \\ &\quad + \eta_t \langle \nabla F_{i_t}(\beta^*), \beta_{t+1}^{i_t} - \beta^{*i_t} \rangle. \end{aligned} \quad (2.46)$$

From relation (2.29), we get

$$\langle \nabla \omega_{i_t}(\beta_{t+1}^{i_t}) - \nabla \omega_{i_t}(\beta_t^{i_t}), \beta^{*i_t} - \beta_{t+1}^{i_t} \rangle = D_{i_t}(\beta_t^{i_t}, \beta^{*i_t}) - D_{i_t}(\beta_{t+1}^{i_t}, \beta^{*i_t}) - D_{i_t}(\beta_t^{i_t}, \beta_{t+1}^{i_t}).$$

Therefore, from the preceding relation, (2.46), and relation (2.30),

$$\begin{aligned} &D_{i_t}(\beta_t^{i_t}, \beta^{*i_t}) - D_{i_t}(\beta_{t+1}^{i_t}, \beta^{*i_t}) - \frac{\mu_{\omega_{i_t}}}{2} \|\beta_t^{i_t} - \beta_{t+1}^{i_t}\|_{i_t}^2 - \eta_t \langle \nabla F_{i_t}(\beta^*), \beta_{t+1}^{i_t} - \beta^{*i_t} \rangle \geq \\ &\eta_t \langle g_{i_t} - \nabla F_{i_t}(\beta^*), \beta_{t+1}^{i_t} - \beta^{*i_t} \rangle \end{aligned} \quad (2.47)$$

Next, we find a lower bound for the right-hand side term. By adding and subtracting $\eta_t\langle g_{i_t} - \nabla F_{i_t}(\beta^*), \beta_t^{i_t}\rangle$, we get

$$\eta_t\langle g_{i_t} - \nabla F_{i_t}(\beta^*), \beta_{t+1}^{i_t} - \beta_t^{i_t}\rangle + \eta_t\langle g_{i_t} - \nabla F_{i_t}(\beta^*), \beta_t^{i_t} - \beta^{*i_t}\rangle \geq \frac{-\eta_t^2}{2\mu_{\omega_{i_t}}}\|g_{i_t} - \nabla F_{i_t}(\beta^*)\|_{*i_t}^2$$
$$- \frac{\mu_{\omega_{i_t}}}{2}\|\beta_{t+1}^{i_t} - \beta_t^{i_t}\|_{i_t}^2 + \eta_t\langle g_{i_t} - \nabla F_{i_t}(\beta^*), \beta_t^{i_t} - \beta^{*i_t}\rangle, \tag{2.48}$$

where the last inequality follows from Fenchel's inequality, i.e., $|\langle x, y\rangle| \leq \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|_*^2$. Combining (2.47) and (2.48) yields

$$D_{i_t}(\beta_{t+1}^{i_t}, \beta^{*i_t}) \leq D_{i_t}(\beta_t^{i_t}, \beta^{*i_t}) - \eta_t(\langle g_{i_t} - \nabla F_{i_t}(\beta^*), \beta_t^{i_t} - \beta^{*i_t}\rangle + \langle \nabla F_{i_t}(\beta^*), \beta_{t+1}^{i_t} - \beta^{*i_t}\rangle)$$
$$+ \frac{\eta_t^2\|g_{i_t} - \nabla F_{i_t}(\beta^*)\|_{*i_t}^2}{2\mu_{\omega_{i_t}}}.$$

Using relation (2.42), and invoking the triangle inequality and relation $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, we obtain

$$D_{i_t}(\beta_{t+1}^{i_t}, \beta^{*i_t}) \leq D_{i_t}(\beta_t^{i_t}, \beta^{*i_t}) - \eta_t\langle \nabla F_{i_t}(\beta^*), \beta_{t+1}^{i_t} - \beta^{*i_t}\rangle$$
$$- \eta_t\langle \nabla F_{i_t}(\beta_t) - \nabla F_{i_t}(\beta^*) + z_t^{i_t}, \beta_t^{i_t} - \beta^{*i_t}\rangle + \eta_t^2\mu_{\omega_{i_t}}^{-1}\|\nabla F_{i_t}(\beta_t) - \nabla F_{i_t}(\beta^*)\|_{*i_t}^2$$
$$+ \eta_t^2\mu_{\omega_{i_t}}^{-1}\|z_t^{i_t}\|_{*i_t}^2.$$

From the preceding relation, the definition of the function $\mathcal{L}$, and that $\beta_{t+1}^i = \beta_t^i$ for all $i \neq i_t$, we have

$$\mathcal{L}(\beta_{t+1}, \beta^*) = \sum_{i \neq i_t} p_i^{-1} D_i(\beta_{t+1}^i, \beta^{*i}) + p_{i_t}^{-1} D_{i_t}(\beta_{t+1}^{i_t}, \beta^{*i_t})$$
$$\leq \mathcal{L}(\beta_t, \beta^*) + p_{i_t}^{-1}\bigg(- \eta_t\langle \nabla F_{i_t}(\beta_t) - \nabla F_{i_t}(\beta^*) + z_t^{i_t}, \beta_t^{i_t} - \beta^{*i_t}\rangle$$
$$- \eta_t\langle \nabla F_{i_t}(\beta^*), \beta_{t+1}^{i_t} - \beta^{*i_t}\rangle + \eta_t^2\mu_{\omega_{i_t}}^{-1}\|\nabla F_{i_t}(\beta_t) - \nabla F_{i_t}(\beta^*)\|_{*i_t}^2 + \eta_t^2\mu_{\omega_{i_t}}^{-1}\|z_t^{i_t}\|_{*i_t}^2\bigg).$$

Taking conditional expectations from both sides of the preceding relation on $\mathcal{F}_t \cup \{i_t\}$, we get

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t \cup \{i_t\}] \leq \mathcal{L}(\beta_t, \beta^*) &+ p_{i_t}^{-1}\eta_t(-\langle \nabla F_{i_t}(\beta^*), \beta_{t+1}^{i_t} - \beta^{*i_t}\rangle \\ &+ \left\langle \mathbb{E}\left[z_t^{i_t} \mid \mathcal{F}_t \cup \{i_t\}\right], \beta^{*i_t} - \beta_t^{i_t}\right\rangle) \\ &+ p_{i_t}^{-1}\eta_t\langle \nabla F_{i_t}(\beta_t) - \nabla F_{i_t}(\beta^*), \beta^{*i_t} - \beta_t^{i_t}\rangle \\ &+ p_{i_t}^{-1}\eta_t^2\mu_{\omega_{i_t}}^{-1}\left(\|\nabla F_{i_t}(\beta_t) - \nabla F_{i_t}(\beta^*)\|_{*i_t}^2 + \mathbb{E}\left[\|z_t^{i_t}\|_{*i_t}^2 \mid \mathcal{F}_t \cup \{i_t\}\right]\right).\end{aligned}$$

Assumption 4 implies that $\mathbb{E}\left[z_t^{i_t} \mid \mathcal{F}_t\right] = 0$. Using that and the bound provided in Assumption 4, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t \cup \{i_t\}] \leq \mathcal{L}(\beta_t, \beta^*) &- p_{i_t}^{-1}\eta_t(\langle \nabla F_{i_t}(\beta^*), \beta_{t+1}^{i_t} - \beta^{*i_t}\rangle \\ &+ \langle \nabla F_{i_t}(\beta_t) - \nabla F_{i_t}(\beta^*), \beta^{*i_t} - \beta_t^{i_t}\rangle) \\ &+ p_{i_t}^{-1}\eta_t^2\mu_{\omega_{i_t}}^{-1}\left(\|\nabla F_{i_t}(\beta_t) - \nabla F_{i_t}(\beta^*)\|_{*i_t}^2 + \nu_{i_t}^2\right).\end{aligned}$$

Next, taking expectations with respect to i_t, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t] \leq \mathcal{L}(\beta_t, \beta^*) &+ \eta_t\left(\langle \nabla F(\beta_t) - \nabla F(\beta^*), \beta^* - \beta_t\rangle - \langle \nabla F(\beta^*), \beta_{t+1} - \beta^*\rangle\right) \\ &+ \eta_t^2\mu_{min}^{-1}\|\nabla F(\beta_t) - \nabla F(\beta^*)\|_*^2 + \eta_t^2\sum\nolimits_{i=1}^{l}\nu_i^2\mu_{\omega_i}^{-1},\end{aligned}$$

where we use the definition of $\langle\cdot,\cdot\rangle$ given in the notation. Using the optimality condition for problem (StochOpt) and under the Lipschitzian property of ∇F and strong convexity of F,

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\beta_{t+1}, \beta^*) \mid \mathcal{F}_t] \leq \mathcal{L}(\beta_t, \beta^*) - \eta_t\mu_F\|\beta_t - \beta^*\|^2 &+ \eta_t^2\mathfrak{L}_F^2{\mu_{min}}^{-1}\|\beta_t - \beta^*\|^2 \\ &+ \eta_t^2\sum\nolimits_{i=1}^{l}\nu_i^2{\mu_{\omega_i}}^{-1}.\end{aligned} \tag{2.49}$$

From the definition of norm $\|\cdot\|$, we can write

$$\begin{aligned}\|\beta_t - \beta^*\|^2 = \sum\nolimits_{i=1}^{l} \|\beta_t^i - \beta^{*i}\|_i^2 \geq 2\sum\nolimits_{i=1}^{l} D_i(\beta_t^i, \beta^{*i})\mathfrak{L}_{\omega_i}^{-1} \geq 2p_\wedge \mathfrak{L}_{max}^{-1} \sum\nolimits_{i=1}^{l} p_i^{-1} D_i(\beta_t^i, \beta^{*i}) \\ = 2p_\wedge \mathfrak{L}_{max}^{-1} \mathcal{L}(\beta_t, \beta^*),\end{aligned}$$

where in the first inequality we used relation (2.30), and in the last relation we used the definition of function $\mathcal{L}$. Similarly,

$$\|\beta_t - \beta^*\|^2 \leq 2\sum_{i=1}^{l} \frac{D_i(\beta_t^i, \beta^{*i})}{\mu_{\omega_i}} \leq \frac{2p_\vee}{\mu_{min}} \mathcal{L}(\beta_t, \beta^*),$$

From the last three relations, we obtain the desired inequality. □

Next, we present self-tuned stepsizes for the (RB-GSMD) method and show their properties.

Preposition 4. Let $\{\beta_t\}$ be generated by the (RB-GSMD) method. Let the set $\mathscr{B}_i$ be convex and closed such that $\|\beta^i\| \leq M_i$ for all $\beta_i \in \mathscr{B}_i$ and some $M_i > 0$. Let Assumption 4 hold for some $\nu_i > 0$, and the stepsize η_t be given by

$$\eta_0^* := \frac{4\mu_F p_\wedge \sum_{i=1}^{l} p_i^{-1} \mathfrak{L}_{\omega_i} M_i^2}{\mathfrak{L}_{max}\left(\frac{8\mathfrak{L}_F{}^2 p_\vee}{\mu_{min}{}^2} \sum_{i=1}^{l} p_i^{-1} \mathfrak{L}_{\omega_i} M_i^2 + \sum_{i=1}^{l} \frac{2\nu_i^2}{\mu_{\omega_i}}\right)},$$

$$\eta_t^* := \eta_{t-1}^* \left(1 - p_\wedge \mu_F \mathfrak{L}_{max}^{-1} \eta_{t-1}^*\right), \quad \text{for all } t \geq 1.$$

Then, the following hold:

(a) The sequence $\{\beta_t\}$ generated by the (RB-GSMD) method converges a.s. to the unique optimal solution β^* of problem (StochOpt).

(b) For any $t \geq 1$, the vector $(\eta_0^*, \eta_1^*, \ldots, \eta_{t-1}^*)$ minimizes the upper bound of the error $\mathbb{E}[D_\omega(\beta_t, \beta^*)]$ given in Lemma 5 for all $(\eta_0, \eta_1, \ldots, \eta_{t-1}) \in \left(0, \frac{\mathfrak{L}_{max}}{2\mu_F p_\wedge}\right]^t$.

(c) The (RB-GSMD) method attains the convergence rate $\mathcal{O}(1/t)$, i.e, for all $t \geq 1$

$$\mathbb{E}\left[\|\beta_t - \beta^*\|^2\right] \leq 2\left(\frac{\mathfrak{L}_{max}}{p_\wedge \mu_F}\right)^2 \left(4\frac{{\mathfrak{L}_F}^2 p_\vee}{{\mu_{min}}^2}\sum_{i=1}^{l} {p_i}^{-1}\mathfrak{L}_{\omega_i}M_i^2 + \sum_{i=1}^{l}\frac{\nu_i^2}{\mu_{\omega_i}}\right)\frac{1}{t}.$$

(d) Let ϵ and ρ be arbitrary positive scalars and $T \triangleq \frac{1.5}{\epsilon\rho}\left(\frac{\mathfrak{L}_{max}}{p_\wedge \mu_F}\right)^2 \left(\frac{8{\mathfrak{L}_F}^2 p_\vee}{{\mu_{min}}^2}\sum_{j=1}^{l} {p_j}^{-1}\mathfrak{L}_{\omega_j}M_j^2 + \sum_{j=1}^{l}\frac{2\nu_j^2}{\mu_{\omega_j}}\right)$ we have for all $t \geq T$

$$\text{Prob}\left(\mathcal{L}(\beta_j, \beta^*) \leq \epsilon \text{ for all } j \geq t\right) \geq 1 - \rho.$$

Proof. Consider relation (2.43). Taking expectations from both sides and rearranging the terms, we can write

$$\mathbb{E}[\mathcal{L}(\beta_{t+1}, \beta^*)] \leq \left(1 - \eta_t 2\mu_F p_\wedge \mathfrak{L}_{max}^{-1}\right)\mathbb{E}[\mathcal{L}(\beta_t, \beta^*)] + 2\eta_t^2{\mathfrak{L}_F}^2 p_\vee \mu_{min}^{-2}\mathbb{E}[\mathcal{L}(\beta_t, \beta^*)]$$
$$+ \eta_t^2 \sum\nolimits_{i=1}^{l} \nu_i^2 \mu_{\omega_i}^{-1}.$$

From relation (2.30), and the triangle inequality, we have

$$\mathcal{L}(\beta_t, \beta^*) \leq \sum_{i=1}^{l} {p_i}^{-1}\frac{\mathfrak{L}_{\omega_i}}{2}\|\beta_t^i - \beta^{*i}\|_i^2 \leq 2\sum_{i=1}^{l} {p_i}^{-1}\mathfrak{L}_{\omega_i}M_i^2.$$

From the preceding inequalities, we obtain

$$\mathbb{E}[\mathcal{L}(\beta_{t+1}, \beta^*)] \leq \left(1 - \eta_t 2\mu_F p_\wedge \mathfrak{L}_{max}^{-1}\right)\mathbb{E}[\mathcal{L}(\beta_t, \beta^*)] + \left(\frac{8{\mathfrak{L}_F}^2 p_\vee}{{\mu_{min}}^2}\sum_{i=1}^{l} {p_i}^{-1}\mathfrak{L}_{\omega_i}M_i^2 + \sum_{i=1}^{l}\frac{2\nu_i^2}{\mu_{\omega_i}}\right)\frac{1}{2}\eta_t^2.$$

Let us define $C^2 \triangleq \sum_{i=1}^{l} {\mu_{\omega_i}}^{-1}C_i^2$ and $\bar{C}^2$ such that $\bar{C}^2 \triangleq \frac{8{\mathfrak{L}_F}^2 p_\vee}{{\mu_{min}}^2}\sum_{i=1}^{l} {p_i}^{-1}\mathfrak{L}_{\omega_i}M_i^2 + \sum_{i=1}^{l}\frac{2\nu_i^2}{\mu_{\omega_i}}$.

Note that the preceding inequality is similar to the relation (2.33), where C^2 is replaced by the term $\bar{C}^2$. Therefore, the desired results here follow by only substituting C^2 by $\bar{C}^2$ in Proposition 3. It only remains to show that: (i) $\eta_0^* = \frac{4\mu_F p_\wedge \sum_{i=1}^{l} p_i^{-1}\mathfrak{L}_{\omega_i} M_i^2}{\mathfrak{L}_{max}\bar{C}^2}$, and (ii) the conditions of Proposition 3 also hold for α_0. The relation (i) holds directly from definition of η_0^* given by Proposition 2 and the definition of $\bar{C}$. To show (ii), we need to verify that $\alpha_0 < 1$. From definition of α_0 given by (2.36), we have

$$\begin{aligned}\alpha_0 &= 2\mu_F p_\wedge \mathfrak{L}_{max}^{-1}\eta_0^* = 2\mu_F p_\wedge \mathfrak{L}_{max}^{-1} \times \frac{4\mu_F p_\wedge \sum_{i=1}^{l} p_i^{-1}\mathfrak{L}_{\omega_i} M_i^2}{\mathfrak{L}_{max}\left(\frac{8{\mathfrak{L}_F}^2 p_\vee}{{\mu_{min}}^2}\sum_{i=1}^{l} p_i^{-1}\mathfrak{L}_{\omega_i} M_i^2 + \sum_{i=1}^{l}\frac{2\nu_i^2}{\mu_{\omega_i}}\right)} \\ &= \frac{\mu_F^2 p_\wedge^2}{\mathfrak{L}_{max}^2} \times \frac{\sum_{i=1}^{l} p_i^{-1}\mathfrak{L}_{\omega_i} M_i^2}{\left(\frac{{\mathfrak{L}_F}^2 p_\vee}{{\mu_{min}}^2}\sum_{i=1}^{l} p_i^{-1}\mathfrak{L}_{\omega_i} M_i^2\right)} = \frac{\mu_F^2}{\mathfrak{L}_F^2}\frac{p_\wedge^2}{p_\vee}\frac{\mu_{min}^2}{\mathfrak{L}_{max}^2} < 1,\end{aligned}$$

where the last relation follows since $\mu_F \leq \mathfrak{L}_F$ and $\mu_{\omega_i} \leq \mathfrak{L}_{\omega_i}$. Therefore, the conditions of Proposition 3 hold for α_0 and the desired results follow. □

2.4 Experimental Results

In this section, we analyze the performance of the self-tuned SMD schemes for solving the following soft-margin linear support vector machine problem:

$$\min \quad F(\beta) \triangleq \frac{1}{m}\sum_{i=1}^{m} L(\langle \beta, \mathbf{x}_i\rangle, y_i) + \frac{\lambda}{2}\|\beta\|_2^2 \,, \tag{2.50}$$

where $L(\langle \beta, \mathbf{x}_i\rangle, y_i) \triangleq max\{0, 1 - y_i\langle \beta, \mathbf{x}_i\rangle\}$ is the hinge-loss function. SVM is known as an effective classification framework and has been applied in real-world applications such as text categorization, image classification, etc. [Cristianini and Shawe-Taylor, 2000]. We use three binary classification data sets namely RCV1, Magic and Skin. The Reuters Corpus Volume I (RCV1) data set [Lewis et al., 2004] is a collection of news-wire stories produced by Reuters journalists from 1996-1997. The articles are categorized into four different classes including

Corporate/Industrial, Economics, Government/Social, and Markets. Here, the samples are documents and the features represent the existence or nonexistence of a given token in an article. We use a subset of the original data set with 199,328 samples and 138,921 features. The goal is to predict whether an article belongs to Markets class or not. The other data sets, Magic and Skin, are from UCI Machine Learning Repository. Magic data set provides some features to distinguish high-energy gamma particles from hadron particles using a gamma telescope and it includes 19,020 samples and 10 features. Skin segmentation data set classifies each pixel of scan photographs as skin or non-skin texture and is used in face and human detection applications. The goal is identifying the skin-like regions. It consists of 3 features, and 245,057 samples out of which 50,859 are the skin samples and 194,198 are non-skin samples. Note that (2.50) is a nonsmooth problem and $F(\beta)$ is a strongly convex function with parameter $\mu_F = \lambda$. In this section, we compare the unifying self-tuned stepsize rule given by Theorem 1 with harmonic stepsizes of the form $\eta_t = \frac{a}{(t+b)}$ where a and b are scalars [Spall, 2005]. Our goal is to compare the sensitivity of the harmonic stepsize rule with different choices of parameters a and b, with that of the unifying self-tuned stepsize rule with different initial stepsizes. We set $\omega = \frac{1}{2}\|\beta\|_2^2$ where $\mu_\omega = \mathfrak{L}_\omega = 1$. For any fixed value of λ, we use three different choices of η_0 for each data set, all within the interval $\left(0, \frac{\mathfrak{L}_\omega}{2\mu_F}\right]$ as we assumed in Theorem 1. These values are denoted by $\eta_0[1], \eta_0[2]$, and $\eta_0[3]$. Initial stepsizes for the RCV1 data set are selected according to Table 2.1.

Table 2.1: Initial stepsize values for RCV1 data set

λ	$\eta_0[1]$	$\eta_0[2] = \frac{\mathfrak{L}_\omega}{10\mu_F}$	$\eta_0[3] = \frac{\mathfrak{L}_\omega}{4\mu_F}$
0.001	0.9	100	250
0.01	0.9	10	25
0.1	0.9	1	2.5
1	0.01	0.1	0.25

For each experiment, the algorithm is run for $T = 10,000$ iterations. Spall [2005] [Ch. 4, pg. 113] considers using b that is about 5 to 10 percent of the total number of iterations.

Accordingly, we choose $b = 0.1 \times T$ and also $b = 0.2 \times T$ which is observed to be a better selection in some of the preliminary experiments. We select $a = \eta_0 b$ in order to start from the same initial stepsize as the self-tuned stepsize. In addition, we compare our proposed scheme with the harmonic stepsize η_0/t.

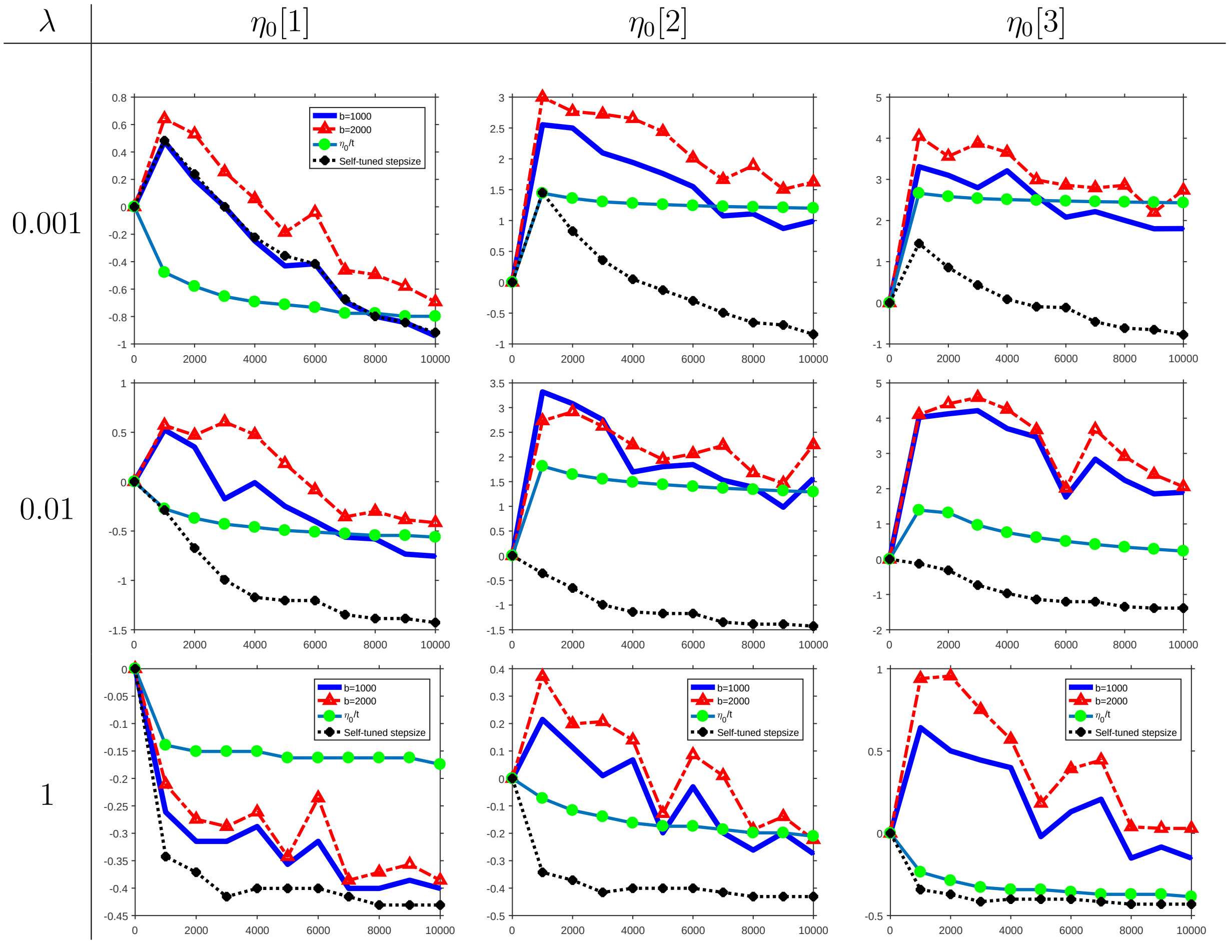

Figure 2.1: RCV1 data set

Figures 2.1-2.3 demonstrate the performance of these stepsize schemes in terms of logarithm of the averaged objective function F. In these plots, the blue and red curves correspond to the harmonic stepsize with parameter $b = 1000$ and $b = 2000$ respectively, and the green curves denote the stepsize η_0/t. The black curves represents the self-tuned stepsize rule.

We observe in Figures 2.1-2.3 that the self-tuned stepsize scheme outperforms the harmonic stepsize in most of the experiments. Importantly, the self-tune stepsize is significantly more robust with respect to (i) the choice of λ; (ii) the data set; and (iii) the initial value of the stepsize. It can be seen that the harmonic stepsize's performance varies for different data sets. While in some cases by increasing the tuning parameters a and b its performance improves, in other instances its performance deteriorates.

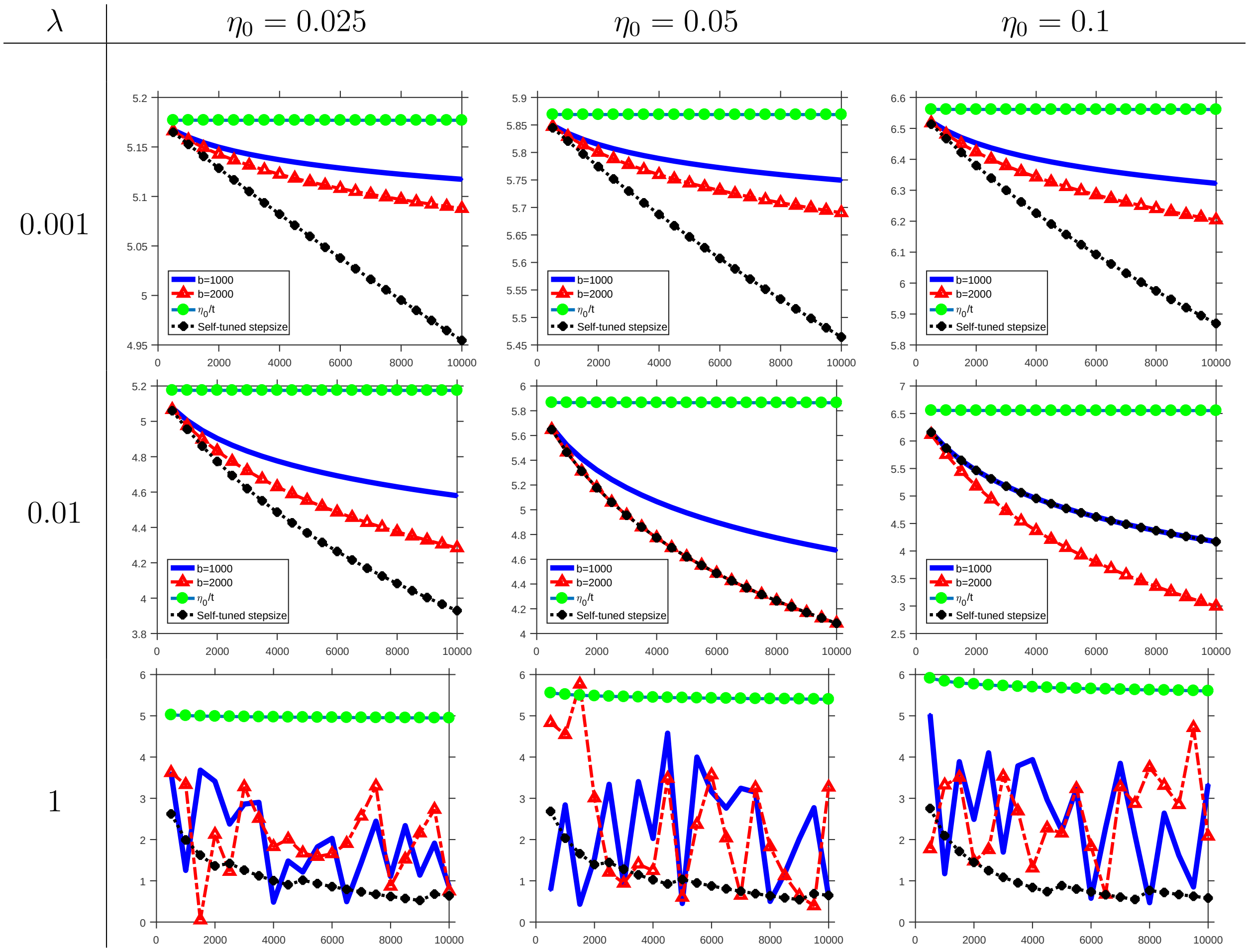

Figure 2.2: Magic data set

2.5 Concluding Remarks

We consider stochastic mirror descent (SMD) methods for solving canonical stochastic optimization problems with strongly convex objective functions. Much of the past research on SMD methods has focused on convergence and rate analysis in terms of order of the error

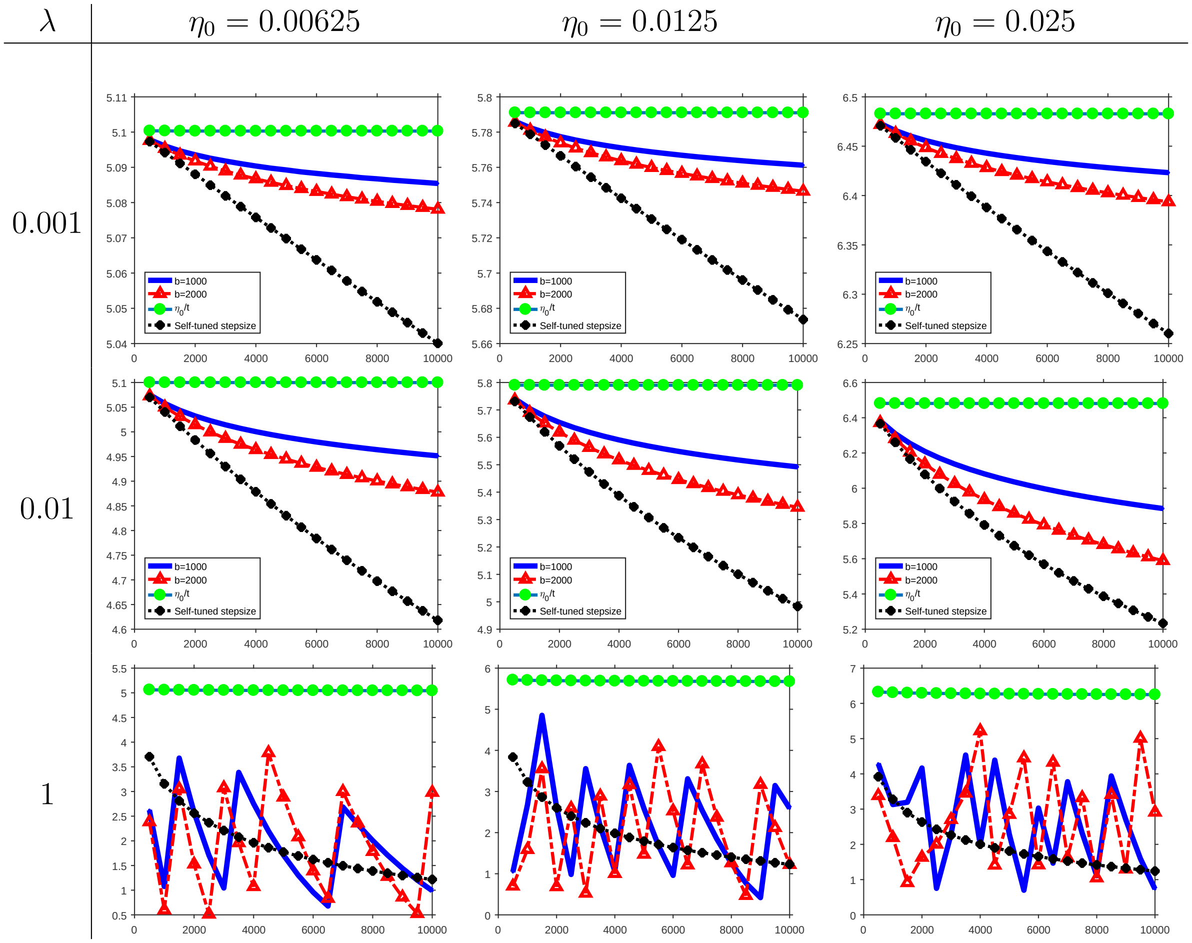

Figure 2.3: Skin data set

bounds. However, the stepsize choice plays a key role in the performance of the this class of algorithms. We consider nonsmooth, smooth, and high-dimensional stochastic optimization problems. We develop self-tuned stepsize rules for stochastic subgradient, gradient, and randomized block coordinate mirror descent methods accordingly. For each scheme, we prove almost sure convergence to the optimal solution of the problem and show that under the self-tuned stepsize rules, the error bound of the SMD scheme is minimized. In the case that some problem parameters are unknown, we develop a unifying self-tuned update rule for which an error bound of the scheme is minimized for any arbitrary and small enough initial stepsize. Moreover, we compare constant factor of our schemes with that of standard SMD methods and show that it can be improved up to four times under non-averaging schemes versus using the averaging scheme in [Dang and Lan, 2015b]. By applying our stepsize scheme to solve the linear SVM problem for three different data sets, we show that our scheme is superior over the well-known harmonic stepsizes and more robust w.r.t. the initial stepsize, choice of data set and problem parameters.

CHAPTER III

MIRROR DESCENT METHODS FOR MULTI-AGENT SEMIDEFINITE OPTIMIZATION PROBLEMS

This chapter addresses multi-agent problems over semidefinite matrix spaces which include cooperative multi-agent problems and non-cooperative Nash games. The goal is developing efficient first-order methods for addressing multi-user optimization problems on semidefinite matrix spaces. We develop mirror descent methods where we choose the distance generating function to be defined as the quantum entropy. These methods are single-loop first-order methods in the sense that they only require a gradient-type of update at each iteration. In the first part of the chapter, we propose a mirror descent incremental subgradient method for minimizing a convex function that consists of sum of component functions. This type of minimization over semidefinite matrix spaces arises in cooperative multi-agent problems such as sparse estimation of a covariance matrix. We show that the iterate generated by the algorithm converges asymptotically to the optimal solution and derive a non-asymptotic convergence rate. Motivated by non-cooperative Nash games in stochastic regimes, in the second part of the chapter, we consider Cartesian stochastic variational inequality (CSVI) problems where the variables are positive semidefinite matrices. In the literature of variational inequality (VI), much attention has been given to addressing VIs on vector spaces. There are a few methods addressing VIs on matrix spaces. Some of these methods have a two-loop framework and require solving a semidefinite optimization problem at each iteration. Others depend on assumptions that either does not hold in applications, or it is hard to verify. Motivated by this gap, we develop a stochastic mirror descent method that require different assumptions, i.e., monotonicity which holds in many applications. The originality of this work lies in the convergence analysis. Employing an auxiliary sequence of stochastic matrices and averaging techniques, we show that the iterate generated by the algorithm converges to a weak solution of the CSVI. Then, we derive a rate of convergence in terms of the expected value of a suitably defined gap function.

3.1 Problem Formulation and Background

First, we consider cooperative multi-agent problems. Decentralized optimization problems have a wide range of applications arising in data mining and machine learning [Nedić et al., 2017], wireless sensor networks [Durham et al., 2012], control [Ram et al., 2009b] and other areas in science and engineering [Xiao and Boyd, 2006] where decentralized processing of information is crucial for security purposes or for real-time decision making. In this chapter, we consider the following multi-agent finite-sum optimization problem which involves a network of multiple agents who cooperatively optimize a global objective,

$$\underset{X\in\mathcal{B}}{\text{minimize}} \ \sum_{i=1}^{m} f_i(X) \tag{3.1}$$

where $\mathcal{B} = \{X \in \mathbb{S}_n : X \succeq 0 \text{ and } \mathrm{tr}(X) = 1\}$, and $f_i : \mathcal{B} \to \mathbb{R}$ is a convex function. In decentralized optimization, the agents (players) need to communicate with their adjacent agents to spread the distributed information to every location in the network.

In the past two decades, there has been much interest in development of models and distributed algorithms for multi-agent optimization problems [Nedić and Ozdaglar, 2009; Lobel and Ozdaglar, 2011; Shi et al., 2015]. In particular, incremental gradient/subgradient methods and their accelerated aggregated variants [Nedić and Bertsekas, 2001; Ram et al., 2009a; Gürbuzbalaban et al., 2017] have been studied where a local gradient/subgradient is taken at each step of an iteration and is followed by communicating with adjacent agents.

Although each step is inexpensive, these methods usually require a large number of iterations to converge. Each iteration in decentralized optimization requires visiting all agents one by one which may cause a significant delay before a transfer of data begins. In this line of research, distributed proximal gradient methods [Bertsekas, 2011, 2015], and alternating direction method of multipliers (ADMM) [Chang et al., 2015; Makhdoumi and Ozdaglar, 2017] were developed and studied extensively as well. These methods have also been extended to applications where the network has a time-varying topology and/or there is a need to asynchronous implementations [Nedić, 2011; Nedić and Olshevsky, 2015]. More recently, Boţ and Böhm [2019] proposed an incremental mirror descent method with a stochastic sweeping of the component functions. While incremental gradient/subgradient methods and their accelerated aggregated variants are extensively studied in vector spaces, their performance and convergence analysis in matrix spaces have not been studied yet.

The sparse covariance inverse estimation is a specific application of finite-sum problem which sets a certain number of coefficients in the inverse covariance to zero to improve the stability of covariance matrix estimation. Lu [2010] developed two first-order methods including the adaptive spectral projected gradient and the adaptive Nesterov's smooth methods to solve the large scale covariance estimation problem. Hsieh et al. [2013] proposed a block coordinate descent (BCD) method with a superlinear convergence rate. In conic programming which is closely related to finite-sum problem, many first-order methods are combined with duality or penalty strategies [Lan et al., 2011; Necoara et al., 2019] to tackle complicated constraints. The aforementioned methods are projection based and do not scale with the problem size. A summary of these methods is given in Table 3.1.

Second, we consider non-cooperative multi-agent systems. VI problems which are very closely tied to the game theory were first introduced in the 1960s. They have a wide range of applications arising in engineering, finance, physics and economics (cf. [Facchinei and Pang, 2003]). Theory of VI can be used for formulating various equilibrium problems and analyzing them from the viewpoint of existence and uniqueness of solutions and stability. Particularly, in mathematical programming, VIs address problems such as optimization problems, complementarity problems and systems of nonlinear equations, to name a few [Scutari et al., 2010]. Given a set $\mathcal{X}$ and a mapping $F : \mathcal{X} \rightarrow \mathbb{R}^{n\times n}$, a VI problem denoted by $\text{VI}(\mathcal{X}, F)$ seeks a matrix $X^* \in \mathcal{X}$ such that $\text{tr}\big((X - X^*)^T F(X^*)\big) \geq 0, \quad \text{for all } X \in \mathcal{X}$. In this chapter, we consider Cartesian stochastic variational inequality problems where the set $\mathcal{X}$ is a Cartesian product of some component sets $\mathcal{X}_i$, i.e.,

$$\mathcal{X} = \{X | X \in \mathbb{S}_n : X = \text{diag}(X_1, \ldots, X_N),\ X_i \in \mathcal{X}_i\},$$
$$\text{where } \mathcal{X}_i = \{X_i | X_i \in \mathbb{S}^+_{n_i}, \text{tr}(X_i) = 1\} \quad \text{for all} \quad i = 1, \ldots, N. \tag{3.2}$$

Hence, we seek a matrix $X^* = \text{diag}(X^*_1, \ldots, X^*_N)$ which solves the following inequality for all $i = 1, \ldots, N$:

$$\text{tr}\big((X_i - X^*_i)^T F_i(X^*)\big) \geq 0, \quad \text{for all } X_i \in \mathcal{X}_i. \tag{3.3}$$

In particular, we study $\text{VI}(\mathcal{X}, F)$ where $F_i(X) = \mathbb{E}[\Phi_i(X, \xi_i(w))]$, i.e., the mapping F_i is the expected value of a stochastic mapping $\Phi_i : \mathcal{X} \times \mathbb{R}^{d_i} \rightarrow \mathbb{S}_n$ where the vector $\xi_i : \Omega \rightarrow \mathbb{R}^{d_i}$ is a random vector associated with a probability space represented by $(\Omega, \mathcal{F}, \mathbb{P})$. Here, Ω denotes the sample space, $\mathcal{F}$ denotes a σ-algebra on Ω, and $\mathbb{P}$ is the associated probability measure. Therefore, $X^* \in \mathcal{X}$ solves $\text{VI}(\mathcal{X}, F)$ if

$$\text{tr}\big((X_i - X^*_i)^T \mathbb{E}[\Phi_i(X^*, \xi(w))]\big) \geq 0,\ \text{for all } X_i \in \mathcal{X}_i. \tag{3.4}$$

Throughout, we assume that $\mathbb{E}[\Phi(X^*, \xi_i(w))]$ is well-defined (i.e., the expectation is finite). There are several challenges in solving CSVIs on semidefinite matrix spaces including presence of uncertainty, the semidefinite solution space and the Cartesian product structure. In what follows, we review some of the methods which address these challenges.

Stochastic Approximation (SA) schemes [Robbins and Monro, 1951] and their prox generalization [Nemirovski et al., 2009; Majlesinasab et al., 2019] shown to be very successful in solving optimization and variational inequality problems Jiang and Xu [2008] with uncertainties. While the convergence analysis of this class of solution methods relies on the monotonicity of the gradient mapping, the extragradient methods [Korpelevich, 1977; Dang and Lan, 2015a; Juditsky et al., 2011] depend on weaker assumptions, i.e., pseudo-monotone mappings to address VIs. Applying SA schemes to solve semidefinite optimization problems result in a two-loop framework and require projection onto a semidefinite cone at each iteration which increases the computational complexity.

Solving optimization problems with positive semidefinite variables is more challenging than solving problems in vector spaces because of the structure of problem constraints. Matrix exponential learning (MEL) which has strong ties to mirror descent methods is an optimization algorithm applied to positive semidefinite nonlinear problems. The distance generating function applied in MEL is the quantum entropy. Mertikopoulos et al. [2012] proposed an MEL based approach to solve the power allocation problem in MIMO multiple access channels. The convergence of MEL and its robustness w.r.t. uncertainties are investigated by Mertikopoulos and Moustakas [2016]. Although in the aforementioned studies, the problem can be formulated as an optimization problem, some practical cases such as multi-user MIMO maximization problem discussed in Section 1.2 cannot be treated as an optimization problem. Hence, Mertikopoulos et al. [2017] proposed an MEL based algorithm to solve N-player games under uncertain feedback and proved that it converges to a stable Nash equilibrium assuming that the mapping is strongly stable. However, in most applications including the game (1.7) this assumption is not met.

While the literature has focused on addressing finite-sum problem on vector spaces, there are applications defined over the set of semidefinite matrices (cf. Section 1.2). Also, in the

Table 3.1: Comparison of first-order schemes

Reference	Problem	Assumptions	Space	Scheme	Rate
Jiang and Xu [2008]	SVI	SM,S	Vector	SA	–
Juditsky et al. [2011]	SVI	MM,S/NS	Vector	Extragradient SMP	$\mathcal{O}\left(1/t\right)$
Mertikopoulos et al. [2012]	SOpt	C,S	Matrix	Exponential Learning	$e^{-\alpha t}(\alpha > 0)$
Koshal et al. [2013]	SVI	MM,S	Vector	Regularized Iterative SA	–
Yousefian et al. [2017]	SVI	MM,NS	Vector	Regularized Smooth SA	$\mathcal{O}\left(1/\sqrt{t}\right)$
Mertikopoulos et al. [2017]	SVI	SPM,S	Matrix	Exponential Learning	$\mathcal{O}\left(1/\lambda t\right)$
Yousefian et al. [2018]	CSVI	PM,S	Vector	Averaging B-SMP	$\mathcal{O}\left(1/t\right)$
Our work	CSVI	MM, NS	Matrix	A-M-SMD	$\mathcal{O}\left(1/\sqrt{t}\right)$
Lan et al. [2011]	Opt	C,S/NS	Matrix	Primal-dual Nesterov's methods	$\mathcal{O}\left(1/t\right)$
Hsieh et al. [2013]	Opt	NS,C	Matrix	BCD	superlinear
Bertsekas [2015]	finite-sum	C,S	Vector	Incremental Aggregated Proximal	Linear
Gürbuzbalaban et al. [2017]	finite-sum	C,S	Vector	Incremental Aggregated Gradient	Linear
Boţ and Böhm [2019]	finite-sum	C,NS	Vector	Incremental SMD	$\mathcal{O}\left(1/\sqrt{t}\right)$
Our work	finite-sum	MM, NS	Matrix	M-MDIS	$\mathcal{O}\left(1/\sqrt{t}\right)$

SM: *strongly monotone mapping*, MM: *merely monotone mapping*, PM: *psedue-monotone mapping*, C: *convex*,
SPM: *strongly psedue-monotone mapping*, S: *smooth function* NS: *nonsmooth function*,
Opt: *optimzation problem*, λ: *strong stability parameter*

VI regime, the focus has been more on addressing SVIs on vector spaces. In particular, CSVIs on matrix spaces which have applications in wireless networks and image retrieval (cf. Section 1.2) have not been studied yet. In this chapter, we consider finite-sum problem and CSVIs on matrix spaces where the mapping is merely monotone. We develop a matrix mirror descent incremental subgradient (M-MDIS) method to solve finite-sum problem (3.1) where we choose the distance generating function to be defined as the quantum entropy following Tsuda et al. [2005]. M-MDIS is a first-order method in the sense that only requires a gradient-type of update at each iteration. This is a single-loop algorithm meaning that it provides a closed-form solution for the projected point and hence it does not need to solve a projection problem at each iteration. We prove that M-MDIS method converges to the optimal solution of (3.1) asymptotically and derive a non-asymptotic convergence rate of $\mathcal{O}(1/\sqrt{t})$. Moreover, we develop an averaging matrix stochastic mirror descent (A-M-SMD) method to solve CSVI (3.4). A-M-SMD is also a first-order single-loop algorithm. To improve its robustness w.r.t. uncertainties, we apply the averaging technique in which $\overline{X}_t$ is defined as a weighted average $\overline{X}_t := \frac{\Gamma_t \overline{X}_{t-1} + \eta_t X_t}{\Gamma_t}$ where $\Gamma_t := \Gamma_{t-1} + \eta_t$ and η_t is the stepsize at iteration t. In this work, we have improved the MEL method of Mertikopoulos et al. [2017] based on the need to mitigate the assumption that mapping is strongly stable since it either does not hold in applications, or it is hard to verify. The originality of our work lies in the convergence analysis under monotonicity assumption. We establish convergence to a weak solution of the CSVI by introducing an auxiliary sequence. Then, we derive a convergence rate of $\mathcal{O}(1/\sqrt{t})$ in terms of the expected value of a suitably defined gap function. In Table 3.1, the distinctions between the existing methods and our work is summarized. We also applied the A-M-SMD method on the throughput maximization problem in wireless multi-user MIMO networks. Our results show that A-M-SMD scheme has a robust performance w.r.t. uncertainty and problem parameters and outperforms both non-averaging M-SMD and MEL methods.

3.2 Preliminaries

Suppose $\omega : \mathrm{dom}(\omega) \to \mathbb{R}$ is a strictly convex and differentiable function, where $\mathrm{dom}(\omega) \subseteq \mathbb{R}^{n\times n}$, and let $X, Y \in \mathrm{dom}(\omega)$. Then, Bregman divergence between X and Y is defined as $D(X,Y) := \omega(X) - \omega(Y) - \mathrm{tr}\big((X-Y)\nabla\omega(Y)^T\big)$. In what follows, our choice of ω is the quantum entropy [Vedral, 2002],

$$\omega(X) = \begin{cases} \mathrm{tr}(X\log X - X) & \text{if} \quad X \in \mathcal{B}, \\ +\infty & \text{otherwise}. \end{cases} \tag{3.5}$$

The Bregman divergence corresponding to the quantum entropy is called von Neumann divergence and is given by [Tsuda et al., 2005]

$$D(X,Y) = \mathrm{tr}(X\log X - X\log Y)\,. \tag{3.6}$$

In our analysis, we use the following property of ω.

Lemma 8. [Yu, 2013] The quantum entropy $\omega : \mathscr{X} \to \mathbb{R}$ is strongly convex with modulus 1 under the trace norm.

Since $\mathcal{B} \subset \mathscr{X}$, the quantum entropy $\omega : \mathcal{B} \to \mathbb{R}$ is also strongly convex with modulus 1 under the trace norm. Next, we derive the conjugate of the quantum entropy and its gradient.

Lemma 9 (Conjugate of von Neumann entropy)**.** Let $Y \in \mathbb{S}_n$ and $\omega(X)$ be defined as (3.5). Then, we have

$$\omega^*(Y) = \log(\mathrm{tr}(\exp(Y + I_n))) \quad (3.7a) \qquad\qquad \nabla\omega^*(Y) = \frac{\exp(Y + I_n)}{\mathrm{tr}(\exp(Y + I_n))}. \quad (3.7b)$$

Proof. Note that ω is a lower semi-continuous convex function on the linear space of all symmetric matrices. The conjugate of function ω is defined as

$$\begin{aligned} \omega^*(Y) &= \sup\{\mathrm{tr}(DY) - \omega(D) : \ D \in \mathcal{B}\} = \sup\{\mathrm{tr}(DY) - \mathrm{tr}(D \log D - D) : D \in \mathcal{B}\} \\ &= -\inf\{-\underbrace{\mathrm{tr}(D(Y + I_n)) + \mathrm{tr}(D \log D) : D \in \mathcal{B}}_{\text{Term 1}}\}. \end{aligned} \tag{3.8}$$

The minimizer of the above problem is $D = \dfrac{\exp(Y + I_n)}{\mathrm{tr}(\exp(Y + I_n))}$ which is called the Gibbs state (see Hiai and Petz [2014], Example 3.29). By plugging it into Term 1, we have (3.7a). The relation (3.7b) follows by standard matrix analysis and the fact that $\nabla_Y \mathrm{tr}(\exp(Y)) = \exp(Y)$ [Athans and Schweppe, 1965]. We observe that $\nabla\omega^*(Y)$ is a positive semidefinite matrix with trace equal to one, implying that $\nabla\omega^*(Y) \in \mathcal{B}$. □

Next, we show that the optimality conditions of a matrix constrained optimization problem can be formulated as a VI.

Lemma 10. Let $\mathcal{B} \subseteq \mathbb{R}^{n\times n}$ be a nonempty closed convex set, and let $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ be a differentiable convex function. Consider the optimization problem

$$\underset{\widetilde{X} \in \mathcal{B}}{\text{minimize}} \quad f(\widetilde{X}). \tag{3.9}$$

A matrix $\widetilde{X}^*$ is optimal to problem (3.9) iff $\widetilde{X}^* \in \mathcal{B}$ and $\operatorname{tr}\Big((Z - \widetilde{X}^*)^T \nabla f(\widetilde{X}^*)\Big) \geq 0$, for all $Z \in \mathcal{B}$.

Proof. ($\Rightarrow$) Assume $\widetilde{X}^*$ is optimal to problem (3.9). Assume by contradiction, there exists some $\hat{Z} \in \mathcal{B}$ such that $\operatorname{tr}\Big((\hat{Z} - \widetilde{X}^*)^T \nabla_{\widetilde{X}} f(\widetilde{X}^*)\Big) < 0$. Since f is continuously differentiable, by the first-order Taylor expansion, for all sufficiently small $0 < \alpha < 1$, we have

$$f(\widetilde{X}^* + \alpha(\hat{Z} - \widetilde{X}^*)) = f(X^*) + \operatorname{tr}\Big((\hat{Z} - \widetilde{X}^*)^T \nabla_{\widetilde{X}} f(\widetilde{X}^*)\Big) + o(\alpha) < f(X^*),$$

following the hypothesis $\operatorname{tr}\Big((\hat{Z} - \widetilde{X}^*)^T \nabla_{\widetilde{X}} f(\widetilde{X}^*)\Big) < 0$. Since $\mathcal{B}$ is convex and $X^*,\ \hat{Z} \in \mathcal{B}$, we have $\widetilde{X}^* + \alpha(\hat{Z} - \widetilde{X}^*) \in \mathcal{B}$ with smaller objective function value than the optimal matrix $\widetilde{X}^*$. This is a contradiction. Therefore, we must have $\operatorname{tr}\Big((Z - \widetilde{X}^*)^T \nabla_{\widetilde{X}} f(\widetilde{X}^*)\Big) \geq 0$ for all $Z \in \mathcal{B}$.

($\Leftarrow$) Now suppose that $\widetilde{X}^* \in \mathcal{B}$ and $\operatorname{tr}\Big((Z - \widetilde{X}^*)^T \nabla_{\widetilde{X}} f(\widetilde{X}^*)\Big) \geq 0$ for all $Z \in \mathcal{B}$. Since f is convex and by Lemma 12, we have

$$f(\widetilde{X}^*) + \operatorname{tr}\Big((Z - \widetilde{X}^*)^T \nabla_{\widetilde{X}} f(\widetilde{X}^*)\Big) \leq f(Z), \quad \text{for all} \quad Z \in \mathcal{B},$$

which implies for all $Z \in \mathcal{B}$,

$$f(Z) - f(\widetilde{X}^*) \geq \operatorname{tr}\Big((Z - \widetilde{X}^*)^T \nabla_{\widetilde{X}} f(\widetilde{X}^*)\Big) \geq 0,$$

where the last inequality follows by the hypothesis. Since $\widetilde{X}^* \in \mathcal{B}$, it follows that $\widetilde{X}^*$ is optimal. □

The next Lemma shows a set of sufficient conditions under which a Nash equilibrium can be obtained by solving a VI.

Lemma 11. [Nash equilibrium] Let $\mathcal{X}_i \in \mathbb{S}_{n_i}$ be a nonempty closed convex set and $f_i(X_i, X_{-i})$ be a differentiable convex function in X_i for all $i = 1, \cdots, N$, where $X_i \in \mathcal{X}_i$ and $X_{-i} \in \prod_{j \neq i} \mathcal{X}_j$. Then, $X^* \triangleq \text{diag}(X_1^*, \cdots, X_N^*)$ is a Nash equilibrium (NE) to game (G2) if and only if X^* solves $\text{VI}(\mathcal{X}, F)$, where

$$F(X) :\triangleq \text{diag}(\nabla_{X_1} f_1(X), \cdots, \nabla_{X_N} f_N(X)), \tag{3.10}$$

$$\mathcal{X} :\triangleq \{X | X = \text{diag}(X_1, \cdots, X_N),\ X_i \in \mathcal{X}_i,\ \text{for all } i\}. \tag{3.11}$$

Proof. First, suppose X^* is an NE to game (G2). We want to prove that X^* solves $\text{VI}(\mathcal{X}, F)$, i.e,

$\text{tr}\big((Z - X^*)^T F(X^*)\big) \geq 0$, for all $Z \in \mathcal{X}$. By optimality conditions of optimization problem $\underset{X_i \in \mathcal{X}_i}{min}\ f_i(X_i, X_{-i})$ and from Lemma 10, we know X^* is an NE if and only if $\text{tr}\big((Z_i - X_i^*)^T \nabla_{X_i} f_i(X^*)\big) \geq 0$ for all $Z_i \in \mathcal{X}_i$ and all $i = 1, \ldots, N$. Then, we obtain for all $i = 1, \cdots, N$

$$\text{tr}\big((Z_i - X_i^*)^T \nabla_{X_i} f_i(X^*)\big) = \sum_u \sum_v [Z_i - X_i^*]_{uv} [\nabla_{X_i} f_i(X^*)]_{uv} \geq 0. \tag{3.12}$$

Invoking the definition of mapping F given by (3.10) and from (3.12), we have

$\text{tr}\big((Z - X^*)^T F(X^*)\big) = \sum_{i,u,v} [Z_i - X_i^*]_{uv} [\nabla_{X_i} f_i(X^*)]_{uv} \geq 0$. From the definition of $\text{VI}(\mathcal{X}, F)$ and relation (3.3), we conclude that $X^* \in \text{SOL}(\mathcal{X}, F)$. Conversely, suppose $X^* \in \text{SOL}(\mathcal{X}, F)$. Then, $\text{tr}\big((Z - X^*)^T F(X^*)\big) \geq 0, \text{for all}\, Z \in \mathcal{X}$. Consider a fixed $i \in \{1, \ldots, N\}$ and a matrix $\bar{Z} \in \mathcal{X}$ given by (3.11) such that the only difference between X^* and $\bar{Z}$ is in i-th block, i.e.

$$\bar{Z} = \text{diag}\left([X_1^*], \ldots, \left[X_{i-1}^*\right], [Z_i], \left[X_{i+1}^*\right], \ldots, [X_N^*]\right),$$

where Z_i is an arbitrary matrix in $\mathcal{X}_i$. Then, we have

$$\bar{Z} - X^* = \mathrm{diag}\left(\mathbf{0}_{n_1 \times n_1}, \ldots, [Z_i - X_i^*], \ldots, \mathbf{0}_{n_N \times n_N}\right). \tag{3.13}$$

Therefore, substituting $\bar{Z} - X^*$ by term (3.13), we obtain

$$\mathrm{tr}\big((\bar{Z} - X^*)^T F(X^*)\big) = \sum_u \sum_v [(Z_i - X_i^*)]_{uv} [\nabla_{X_i} f_i(X^*)]_{uv} = \mathrm{tr}\big((Z_i - X_i^*)^T \nabla_{X_i} f_i(X^*)\big) \geq 0.$$

Since i was chosen arbitrarily, $\mathrm{tr}\big((Z_i - X_i^*)^T \nabla_{X_i} f_i(X^*)\big) \geq 0$ for any $i = 1, ..., N$. Hence, by applying Lemma 10 we conclude that X^* is a Nash equilibrium to game (G2). □

We make use of the following lemma in our analysis. Note that $\mathbb{R}^{n\times n}$ is a vector space with dimension n^2 [Axler, 1997].

Lemma 12. Let $[X]_{uv}$ denotes the elements of matrix X. If we rewrite matrices X, Z and $\nabla_X f(X)$ as vectors $x = ([X]_{11}, \ldots, [X]_{nn})^T$, $z = ([z]_{11}, \ldots, [z]_{nn})^T$, and $\nabla f(x) = ([\nabla_X f(X)]_{11}, \ldots, [\nabla_X f(X)]_{nn})^T$ respectively, it is trivial that

$$(z - x)^T \nabla f(x) = \sum_u \sum_v [(Z - X)]_{uv} [\nabla_X f(X)]_{uv} = \mathrm{tr}\big((Z - X)^T \nabla_X f(X)\big),$$

where the last inequality follows by relation $\mathrm{tr}\big(A^T B\big) = \sum_u \sum_v [A]_{uv} [B]_{uv}$.

3.3 Cooperative Multi-agent Problems

Consider the multi-agent optimization problem (3.1) on semidefinite matrix spaces. In this section, we present the mirror descent incremental subgradient method for solving (3.1). Algorithm 2 presents the outline of the M-MDIS method. The method maintains two matrices for each agent i: primal U_i and dual Y_i. The connection between the two matrices is via a function $U_i = \nabla\omega^*(Y_i)$ which projects Y_i onto the set $\mathcal{B}$ defined by (3.2). At each iteration t

and for any agent i, first, the subgradient of f_i is calculated at $U_{i-1,t}$, denoted by $\tilde{\nabla} f_i(U_{i-1,t})$. Next, we update the dual matrix by moving along the subgradient. Here η_t is a non-increasing step-size sequence. Then, $Y_{i,t}$ will be projected onto the set $\mathcal{B}$ using the closed-form solution (3.15). It should be noted that the update rule (3.15) is obtained by applying Lemma 9. Finally, the primal and dual matrices of agent m, i.e. $U_{m,t}$ and $Y_{m,t}$ are the input to the next iteration.

Algorithm 2 Matrix Mirror Descent Incremental Subgradient (M-MDIS)

```
1: **initialization**: pick feasible $X_0$ and $Y_{m,-1}$ arbitrarily.
2: **General step**: for any $t = 0, 1, 2, \cdots$ do the following:
  (a) $U_{0,t} = X_t$ and $Y_{0,t} = Y_{m,t-1}$
  (b) For i=1,...,$m$ do the following:
        $$Y_{i,t} = Y_{i-1,t} - \eta_t \tilde{\nabla} f_i(U_{i-1,t}) \quad (3.14)$$
        $$U_{i,t} = \frac{\exp(Y_{i,t} + \mathbf{I}_n)}{\mathrm{tr}(\exp(Y_{i,t} + \mathbf{I}_n))} \quad (3.15)$$
  (c) $X_{t+1} = U_{m,t}$.
```

Next, we state the main assumption and discuss its rationality.

Assumption 5. (Bounded subgradient) There exists a constant L_{f_i} for which $\|\tilde{\nabla} f_i(X)\|_2 \leq L_{f_i}$ for all $\tilde{\nabla} f_i(X) \in \partial f_i(X)$, and $X \in \mathcal{B}$.

Corollary 1 (Boundedness of subgradients)**.** For a proper convex function f_i and a nonempty and compact set $\mathcal{B} \subseteq \mathrm{int}(\mathrm{dom(f)})$, the union $\underset{X \in \mathcal{B}}{\cup} \partial f_i(X)$ is nonempty and bounded (Beck [2017], Theorem 3.16). Therefore, we conclude that Assumption 5 holds.

Note that since f_i is a convex function and $\mathcal{B}$ is a compact set, the above assumption holds. We use the following relations in convergence analysis,

$$Y_{i,t} \triangleq \tilde{\nabla} \omega(U_{i,t}) \in \partial \omega(U_{i,t}) \Leftrightarrow U_{i,t} \in \partial \omega^\star(Y_{i,t}). \tag{3.16}$$

It should be noted that the above relation holds because ω is a closed and convex function. Since $(A-B)^2 \in S_+^n$, we have $0 \leq \mathrm{tr}((A-B)^2) = \mathrm{tr}(A^2) - 2\mathrm{tr}(AB) + \mathrm{tr}(B^2)$. Therefore,

$$2\mathrm{tr}\big(A^TB\big) \leq \mathrm{tr}\big(A^2\big) + \mathrm{tr}\big(B^2\big) \leq (\mathrm{tr}(A))^2 + n\|B^2\|_2 = (\mathrm{tr}(A))^2 + n\|B\|_2^2, \tag{3.17}$$

where the last inequality follows by positive semidefinteness of matrix A and the relation $\mathrm{tr}(B) \leq n\|B\|_2$. Next, we prove the convergence of M-MDIS algorithm.

Theorem 2 (asymptotic convergence)**.** Consider Problem (3.1). Let Assumption 5 hold. Let $\{X_t\}$ be generated by the M-MDIS method with positive stepsize $\{\eta_t\}$. If $\lim_{T\to\infty} \frac{\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t} = 0$, then $f_T^{\min}$ converges to f^* as $T \to \infty$, where $f_T^{\min} = \min\limits_{t=0,\cdots,T} f(X_t)$.

Proof. Let $Y \in \cap_{i=1}^m \mathrm{dom} f_i$ be fixed. For every $i = 1, \cdots, m$ and every $t \geq 0$ we have

$$\begin{aligned}
D(Y, U_{i,t}) &= \omega(Y) - \omega(U_{i,t}) - \mathrm{tr}\Big(\tilde{\nabla}^T \omega(U_{i,t})(Y - U_{i,t})\Big) \\
&= \omega(Y) - \omega(U_{i,t}) - \mathrm{tr}\big((Y_{i,t})^T (Y - U_{i,t})\big) \\
&= \omega(Y) - \omega(U_{i,t}) - \mathrm{tr}\Big((Y_{i-1,t} - \eta_t \tilde{\nabla} f_i(U_{i-1,t}))^T (Y - U_{i,t})\Big) \\
&= \omega(Y) - \omega(U_{i,t}) - \mathrm{tr}\big((Y_{i-1,t})^T (Y - U_{i,t})\big) + \eta_t \mathrm{tr}\Big(\tilde{\nabla}^T f_i(U_{i-1,t})(Y - U_{i,t})\Big) \\
&= \omega(Y) - \omega(U_{i,t}) - \mathrm{tr}\Big(\tilde{\nabla}^T \omega(U_{i-1,t})(Y - U_{i,t})\Big) + \eta_t \mathrm{tr}\Big(\tilde{\nabla}^T f_i(U_{i-1,t})(Y - U_{i,t})\Big),
\end{aligned}$$

where we used relation (3.16) in the second and last equality and we applied the update rule of the Algorithm 2 in the third equality.

By adding and subtracting the term $\omega(U_{i-1,t}) + \tilde{\nabla}^T \omega(U_{i-1,t}) U_{i-1,t}$, we get

$$\begin{aligned}
D(Y, U_{i,t}) = \omega(Y) &- \omega(U_{i-1,t}) - \mathrm{tr}\Big(\tilde{\nabla}^T \omega(U_{i-1,t})(Y - U_{i-1,t})\Big) + \omega(U_{i-1,t}) - \omega(U_{i,t}) \\
&- \mathrm{tr}\Big(\tilde{\nabla}^T \omega(U_{i-1,t})(U_{i-1,t} - U_{i,t})\Big) + \mathrm{tr}\Big(\eta_t \tilde{\nabla}^T f_i(U_{i-1,t})(Y - U_{i,t})\Big) \\
&= D(Y, U_{i-1,t}) - D(U_{i,t}, U_{i-1,t}) + \eta_t \mathrm{tr}\Big(\tilde{\nabla}^T f_i(U_{i-1,t})(Y - U_{i,t})\Big).
\end{aligned}$$

By adding and subtracting the term $\eta_t \mathrm{tr}\Big(\tilde{\nabla}^T f_i(U_{i-1,t}) U_{i-1,t}\Big)$, we have

$$\begin{aligned} D(Y, U_{i,t}) &= D(Y, U_{i-1,t}) - D(U_{i,t}, U_{i-1,t}) + \eta_t \mathrm{tr}\Big(\tilde{\nabla}^T f_i(U_{i-1,t})(Y - U_{i-1,t})\Big) \\ &- \eta_t \mathrm{tr}\Big(\tilde{\nabla}^T f_i(U_{i-1,t})(U_{i,t} - U_{i-1,t})\Big) \leq D(Y, U_{i-1,t}) - D(U_{i,t}, U_{i-1,t}) \\ &+ \eta_t \left(f_i(Y) - f_i(U_{i-1,t})\right) + \eta_t \mathrm{tr}\Big(\tilde{\nabla}^T f_i(U_{i-1,t})(U_{i-1,t} - U_{i,t})\Big), \end{aligned} \tag{3.18}$$

where we used the definition of subgradient in the last relation. Using relation (3.17),

$$\eta_t \mathrm{tr}\Big(\tilde{\nabla}^T f_i(U_{i-1,t})(U_{i-1,t} - U_{i,t})\Big) \leq n\eta_t^2 \|\tilde{\nabla}^T f_i(U_{i-1,t})\|_2^2 + \frac{1}{4}(\mathrm{tr}(U_{i-1,t} - U_{i,t}))^2. \tag{3.19}$$

Plugging (3.19) into (3.18), we get

$$\begin{aligned} D(Y, U_{i,t}) &\leq D(Y, U_{i-1,t}) - D(U_{i,t}, U_{i-1,t}) + \eta_t(f_i(Y) - f_i(U_{i-1,t})) \\ &+ n\eta_t^2 \|\tilde{\nabla}^T f_i(U_{i-1,t})\|_2^2 + \frac{1}{4}(\mathrm{tr}(U_{i-1,t} - U_{i,t}))^2. \end{aligned}$$

Using that ω is 1-strongly convex, Lemma 8 and definition of Bregman divergence, we get

$$\begin{aligned} D(Y, U_{i,t}) &\leq D(Y, U_{i-1,t}) - D(U_{i,t}, U_{i-1,t}) + \eta_t \left(f_i(Y) - f_i(U_{i-1,t})\right) + n\eta_t^2 \|\tilde{\nabla}^T f_i(U_{i-1,t})\|_2^2 \\ &+ \frac{1}{2} D(U_{i,t}, U_{i-1,t}) = D(Y, U_{i-1,t}) + \eta_t \left(f_i(Y) - f_i(U_{i-1,t})\right) + n\eta_t^2 \|\tilde{\nabla}^T f_i(U_{i-1,t})\|_2^2 \\ &- \frac{1}{2} D(U_{i,t}, U_{i-1,t}). \end{aligned}$$

By Assumption 5, we have for any $i = 1, \cdots, m$ and $t \geq 0$

$$D(Y, U_{i,t}) \leq D(Y, U_{i-1,t}) + \eta_t \left(f_i(Y) - f_i(U_{i-1,t})\right) + n\eta_t^2 {L_{f_i}}^2 - \frac{1}{2} D(U_{i,t}, U_{i-1,t}).$$

Summing the above inequality for $i = 1, \cdots, m$, we get

$$D(Y, U_{m,t}) \leq D(Y, U_{0,t}) + \eta_t \sum_{i=1}^{m} (f_i(Y) - f_i(U_{i-1,t})) + n\eta_t^2 \sum_{i=1}^{m} {L_{f_i}}^2 - \sum_{i=1}^{m} \frac{1}{2} D(U_{i,t}, U_{i-1,t}).$$

Note that $U_{0,t} = X_t$. By adding and subtracting the term $\eta_t f(X_t)$, we have

$$\begin{aligned} D(Y, U_{m,t}) \leq D(Y, X_t) &+ \eta_t \sum_{i=1}^{m} (f_i(Y) - f_i(X_t)) + \eta_t \sum_{i=1}^{m} (f_i(X_t) - f_i(U_{i-1,t})) \\ &+ n\eta_t^2 \sum_{i=1}^{m} {L_{f_i}}^2 - \sum_{i=1}^{m} \frac{1}{2} D(U_{i,t}, U_{i-1,t}). \end{aligned} \tag{3.20}$$

By Assumption 5, we have f_i is continuous over $\mathcal{B}$ with parameter $L_{f_i} > 0$, i.e., $|f_i(A) - f_i(B)| \leq L_{f_i} \|A - B\|_2$. Therefore, we have

$$\begin{aligned} &\sum_{i=1}^{m} (f_i(X_t) - f_i(U_{i-1,t})) = \sum_{i=2}^{m} \sum_{j=1}^{i-1} (f_i(U_{j-1,t}) - f_i(U_{j,t})) \leq \sum_{i=2}^{m} \sum_{j=1}^{i-1} L_{f_i} \|U_{j-1,t} - U_{j,t}\|_2 \\ &\leq \left(\sum_{l=1}^{m} L_{f_l} \right) \sum_{i=1}^{m} \|U_{i-1,t} - U_{i,t}\|_2 = \left(\sum_{l=1}^{m} L_{f_l} \right) \sum_{i=1}^{m} \|\nabla \omega^*(Y_{i-1,t}) - \nabla \omega^*(Y_{i,t})\|_2 \\ &\leq \left(\sum_{l=1}^{m} L_{f_l} \right) \sum_{i=1}^{m} \|Y_{i-1,t} - Y_{i,t}\|_2, \end{aligned}$$

where the last inequality follows by Lipschitz continuity of $\nabla \omega^*$. Applying the update rule of the Algorithm 2, we have

$$\sum_{i=1}^{m} (f_i(X_t) - f_i(U_{i-1,t})) \leq \left(\sum_{l=1}^{m} L_{f_l} \right) \sum_{i=}^{m} \|\eta_t \tilde{\nabla} f_i(U_{i-1,t})\|_2 \leq \eta_t \left(\sum_{l=1}^{m} L_{f_l} \right) \left(\sum_{i=1}^{m} L_{f_i} \right), \tag{3.21}$$

where the last inequality follows by Assumption 5. Plugging (3.21) into (3.20), for any $t \geq 0$

$$\begin{aligned} D(Y, U_{m,t}) \leq D(Y, X_t) + \eta_t \sum_{i=1}^{m} (f_i(Y) - f_i(X_t)) + \eta_t^2 \left(\sum_{i=1}^{m} L_{f_i} \right)^2 \\ + n\eta_t^2 \sum_{i=1}^{m} {L_{f_i}}^2 - \sum_{i=1}^{m} \frac{1}{2} D(U_{i,t}, U_{i-1,t}). \end{aligned}$$

Since $\sum_{i=1}^{m} {L_{f_i}}^2 \leq (\sum_{i=1}^{m} L_{f_i})^2$, also $U_{m,t} = X_{t+1}$, and $Y_{m,t} = Y_{0,t+1}$, we get for any $t \geq 0$ that

$$D(Y, X_{t+1}) \leq D(Y, X_t) + \eta_t \sum_{i=1}^{m} (f_i(Y) - f_i(X_t)) + \eta_t^2 (n+1) \left(\sum_{i=1}^{m} L_{f_i} \right)^2,$$

where we used the fact that $D(U_{i,t}, U_{i-1,t}) \geq 0$. Let $Y := X^*$, summing up the inequality from $t = 0$ to $T - 1$, where $T \geq 1$ and rearranging, we get

$$D(X^*, X_T) + \sum_{t=0}^{T-1} \eta_t \left(\sum_{i=1}^{m} f_i(X_t) - \sum_{i=1}^{m} f_i(X^*) \right) \leq D(X^*, X_0) + \sum_{t=0}^{T-1} \eta_t^2 (n+1) \left(\sum_{i=1}^{m} L_{f_i} \right)^2.$$

By definition of $f_{T-1}^{\min}$, we have

$$\sum_{t=0}^{T-1} \eta_t \left(f_{T-1}^{\min} - f^* \right) \leq \sum_{t=0}^{T-1} \eta_t \left(\sum_{i=1}^{m} f_i(X_t) - \sum_{i=1}^{m} f_i(X^*) \right)$$

Since $D(X^*, X_T) \geq 0$, we get

$$f_{T-1}^{\min} - f^* \leq \frac{D(X^*, X_0) + (n+1) \left(\sum_{i=1}^{m} L_{f_i} \right)^2 \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t}. \tag{3.22}$$

By assumption, $\lim_{T \to \infty} \frac{\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t} = 0$ which implies $\sum_{t=0}^{T-1} \eta_t \to +\infty$. Therefore, $f_{T-1}^{\min} - f^* \to 0$, i.e., $f_{T-1}^{\min}$ converges to f^* as $T \to \infty$. □

Next, we present the convergence rate of the M-MDIS scheme.

Corollary 2. (Rate of convergence) Consider problem (3.1). Suppose Assumption 5 holds and let the sequence $\{X_t\}$ be generated by Algorithm 2. Given a fixed $T \geq 1$, let η_t be a sequence given by

$$\eta_t = \frac{1}{\sum_{i=1}^m L_{f_i}} \sqrt{\frac{D(X^*, X_0)}{n+1}} \frac{1}{\sqrt{T}}. \tag{3.23}$$

Then, we have

$$f_{T-1}^{\min} - f^* \leq 2 \left(\sum_{i=1}^m L_{f_i} \right) \sqrt{\frac{D(X^*, X_0)(n+1)}{T}} = \mathcal{O}(\frac{1}{\sqrt{T}}). \tag{3.24}$$

Proof. Assume that the number of iterations T is fixed and the stepsize is constant, i.e, $\eta_t = \eta$ for all $t \geq 0$, then it follows by (3.22) that

$$f_{T-1}^{\min} - f^* \leq \frac{D(X^*, X_0) + (n+1) \left(\sum_{i=1}^m L_{f_i} \right)^2 \sum_{t=0}^{T-1} \eta^2}{\sum_{t=0}^{T-1} \eta}. \tag{3.25}$$

Then, by minimizing the right-hand side of the above inequality over $\eta > 0$, we obtain the constant stepsize (3.23) for all $t \geq 0$. By plugging (3.23) into (3.25), we obtain the rate of the convergence of (3.24) for $T \geq 1$. □

3.4 Stochastic Non-cooperative Nash Games

In this section, we present the A-M-SMD scheme for solving (3.4). Algorithm 3 presents the outline of the A-M-SMD method. At each iteration t and for any user i, first, using an oracle, a realization of the stochastic mapping F is generated at X_t, denoted by $\Phi_i(X_t, \xi_t)$. Next, a matrix $Y_{i,t}$ is updated using (3.27). Here η_t is a non-increasing step-size sequence. Then, $Y_{i,t}$ will be projected onto the set $\mathcal{X}_i$ defined by (3.2) using the closed-form solution (3.28). It should be noted that the update rule (3.28) is obtained by applying Lemma 9. Then the averaged sequence $\overline{X}_{i,t+1}$ is generated using relations (3.29). Next, we state the main assumptions. Let us define the stochastic error at iteration t as

$$Z_{i,t} :\triangleq \Phi_i(X_t, \xi_t) - F_i(X_t) \quad \text{for all} \quad t \geq 0, \quad \text{and for all} \quad i = 1, \ldots, N. \tag{3.26}$$

Let $\mathcal{F}_t$ denote the history of the algorithm up to time t, i.e., $\mathcal{F}_t = \{X_0, \xi_0, \ldots, \xi_{t-1}\}$ for $t \geq 1$ and $\mathcal{F}_0 = \{X_0\}$.

Assumption 6. Let the following hold:

(a) The mapping $F(X) = \mathbb{E}[\Phi(X_t, \xi_t)]$ is monotone and continuous over the set $\mathcal{X}$.

(b) The stochastic mapping $\Phi_i(X_t, \xi_t)$ has a finite mean squared error, i.e, there exist scalars $C_i > 0$ such that $\mathbb{E}[\|\Phi_i(X_t, \xi_t)\|_2^2 | \mathcal{F}_t] \leq C_i^2$ for all $i = 1, \ldots, N$.

(c) The stochastic noise $Z_{i,t}$ has a zero mean, i.e., $\mathbb{E}[Z_{i,t} | \mathcal{F}_t] = \mathbf{0}$ for all $t \geq 0$ and for all $i = 1, \ldots, N$.

Algorithm 3 Averaging Matrix Stochastic Mirror Descent (A-M-SMD)

```
initialization: Set $Y_{i,0} := I_{n_i}/n_i$, a stepsize $\eta_0 > 0$, $\Gamma_0 = \eta_0$, let $X_{i,0} \in \mathcal{X}_i$ be a random initial matrix, and $\overline{X}_{i,0} = X_{i,0}$.
for $t = 0, 1, ..., T-1$ do
    for $i = 1, ..., N$ do
        Generate $\xi_t$ as realizations of the random variable $\xi$ and evaluate the mapping $\Phi_i(X_t, \xi_t)$. Let
            $Y_{i,t+1} := Y_{i,t} - \eta_t \Phi_i(X_t, \xi_t),$                                  (3.27)
            $X_{i,t+1} := \frac{\exp(Y_{i,t+1} + I_{n_i})}{\mathrm{tr}(\exp(Y_{i,t+1} + I_{n_i}))}.$        (3.28)
        Update $\Gamma_t$ and $\overline{X}_{i,t}$ using the following recursions:
            $\Gamma_{t+1} := \Gamma_t + \eta_{t+1},\ \overline{X}_{i,t+1} := \frac{\Gamma_t \overline{X}_{i,t} + \eta_{t+1} X_{i,t+1}}{\Gamma_{t+1}}.$   (3.29)
return $\overline{X}_T$.
```

3.4.1 Convergence and Rate Analysis

In this section, our interest lies in analyzing the convergence and deriving a rate statement for the sequence generated by the A-M-SMD method. Note that a solution of VI($\mathcal{X}, F$) is also referred to a strong solution. The convergence analysis is carried out by a gap function G defined subsequently. The definition of G is closely tied with a weak solution which is a counterpart of a strong solution. Next, we define a weak solution.

Definition 5. (Weak solution) The matrix $X_w^* \in \mathcal{X}$ is called a weak solution to VI($\mathcal{X}, F$) if it satisfies $\mathrm{tr}\big((X - X_w^*)^T F(X)\big) \geq 0$, for all $X \in \mathcal{X}$.

We let $\mathcal{X}_w^\star$ and $\mathcal{X}^*$ denote the set of weak solutions and strong solutions to VI($\mathcal{X}, F$), respectively.

Remark 3. Under Assumption 6(a), when the mapping F is monotone, any strong solution of problem (3.4) is a weak solution, i.e., $\mathcal{X}^* \subseteq \mathcal{X}_w^\star$. From continuity of F in Assumption 6(a), the converse is also true meaning that a weak solution is a strong solution. Moreover, for a monotone mapping F on a convex compact set e.g., $\mathcal{X}$, a weak solution always exists [Juditsky et al., 2011].

Unlike optimization problems where the objective function provides a metric for distinguishing solutions, there is no immediate analog in VI problems. However, different variants of gap function have been used in the analysis of variational inequalities (cf. Chapter 10 in [Facchinei and Pang, 2003]). Here we use the following gap function associated with a VI problem to derive a convergence rate.

Definition 6. (G function) Define the following function $G : \mathcal{X} \to \mathbb{R}$ as

$$G(X) = \sup_{Z \in \mathcal{X}} \mathrm{tr}\big((X - Z)^T F(Z)\big), \quad \text{for all } X \in \mathcal{X}.$$

The next lemma provides some properties of the G function.

Lemma 13. The function $G(X)$ given by Definition 6 is a well-defined gap function, i.e, (i) $G(X) \geq 0$ for all $X \in \mathcal{X}$; (ii) X_w^* is a weak solution to problem (3.4) iff $G(X_w^*) = 0$.

Proof. (i) For an arbitrary $X \in \mathcal{X}$, we have

$$G(X) = \sup_{Z \in \mathcal{X}} \operatorname{tr}\big((X - Z)^T F(Z)\big) \geq \operatorname{tr}\big((X - A)^T F(A)\big)\,,$$

for all $A \in \mathcal{X}$. For $A = X$, the above inequality suggests that $G(X) \geq \operatorname{tr}\big((X - X)^T F(X)\big) = 0$ implying that the function $G(X)$ is nonnegative for all $X \in \mathcal{X}$.

(ii) Assume X_w^* is a weak solution. By Definition 5, $\operatorname{tr}\big((X_w^* - X)^T F(X)\big) \leq 0$, for all $X \in \mathcal{X}$ which implies $G(X_w^*) = \sup\limits_{X \in \mathcal{X}} \operatorname{tr}\big((X_w^* - X)^T F(X)\big) \leq 0$. On the other hand, from Lemma 13(i), we get $G(X_w^*) \geq 0$. We conclude that $G(X_w^*) = 0$ for any weak solution X_w^*. Conversely, assume that there exists an X such that $G(X) = 0$. Therefore, $\sup\limits_{Z \in \mathcal{X}} \operatorname{tr}\big((X - Z)^T F(Z)\big) = 0$ which implies $\operatorname{tr}\big((Z - X)^T F(Z)\big) \geq 0$ for all $Z \in \mathcal{X}$. Therefore, X is a weak solution. □

Lemma 14. Assume the sequence η_t is non-increasing and the sequence $\overline{X}_{i,t}$ is given by the recursive rule (3.29) where $\Gamma_0 = \eta_0$ and $\overline{X}_{i,0} = X_{i,0}$. Then,

$$\overline{X}_{i,t} := \sum_{k=0}^{t} \left(\frac{\eta_k}{\sum_{k'=0}^{t} \eta_{k'}} \right) X_{i,k} \quad \text{for any } t \geq 0. \tag{3.30}$$

Proof. We use induction to prove (3.30). It is trivial that it holds for $t = 0$, since $\overline{X}_{i,0} = X_{i,0}$. Assume (3.30) holds for t. From (3.29), $\Gamma_t = \sum_{k'=0}^{t} \eta_{k'}$ which results in $\overline{X}_{i,t} = \frac{\sum_{k=0}^{t} \eta_k X_{i,k}}{\Gamma_t}$. From (3.29), we have

$$\overline{X}_{i,t+1} := \frac{\Gamma_t \overline{X}_{i,t} + \eta_{t+1} X_{i,t+1}}{\Gamma_{t+1}} = \frac{\sum_{k=0}^{t} \eta_k X_{i,k} + \eta_{t+1} X_{i,t+1}}{\Gamma_{t+1}} = \frac{\sum_{k=0}^{t+1} \eta_k X_{i,k}}{\sum_{k'=0}^{t+1} \eta_k'}.$$

□

Throughout, we use the notion of Fenchel coupling (Mertikopoulos and Sandholm [2016]):

$$H_i(Q_i, Y_i) \triangleq \omega_i(Q_i) + \omega_i^*(Y_i) - \operatorname{tr}\big({Q_i}^T Y_i\big) \,, \tag{3.31}$$

which provides a proximity measure between Q_i and $\nabla\omega_i^*(Y_i)$ and is equal to the associated Bregman divergence between Q and $\nabla\omega_i^*(Y_i)$.

Lemma 15. [Mertikopoulos et al., 2017] Let $\mathcal{X}_i$ be given by (3.2). For all matrices $X_i \in \mathcal{X}_i$ and for all $Y_i, Z_i \in \mathbb{S}_{n_i}$, the following holds

$$H_i(X_i, Y_i + Z_i) \leq H_i(X_i, Y_i) + \operatorname{tr}\big({Z_i}^T(\nabla\omega_i^*(Y_i) - X_i)\big) + \|Z_i\|_2^2. \tag{3.32}$$

Proof. Using the Fenchel coupling definition,

$$H(X, Y + Z) = \omega(X) + \omega^*(Y + Z) - \operatorname{tr}\big(X^T(Y + Z)\big) \,. \tag{3.33}$$

By strong convexity of ω w.r.t. trace norm (Lemma 8) and using duality between strong convexity and strong smoothness [Kakade et al., 2009], ω^* is 1-strongly smooth w.r.t. the spectral norm, i.e., $\omega^*(Y + Z) \leq \omega^*(Y) + \operatorname{tr}\big(Z^T\nabla\omega^*(Y)\big) + \|Z\|_2^2$. By plugging this inequality into (3.33) we have

$$\begin{aligned} H(X, Y + Z) &\leq \omega(X) + \omega^*(Y) + \operatorname{tr}\big(Z^T\nabla\omega^*(Y)\big) + \|Z\|_2^2 - \operatorname{tr}\big(X^T Y\big) - \operatorname{tr}\big(X^T Z\big) \\ &= H(X, Y) + \operatorname{tr}\big(Z^T(\nabla\omega^*(Y) - X)\big) + \|Z\|_2^2, \end{aligned}$$

where in the last relation, we used (3.31). □

Next, we develop an error bound for the G function given by Definition 6.

Lemma 16. Consider problem (3.4). Let $X_i \in \mathcal{X}_i$ and the sequence $\{\overline{X}_t\}$ be generated by

A-M-SMD algorithm. Suppose Assumption 6 holds. Then, for any $T \geq 1$,

$$\mathbb{E}[G(\overline{X}_T)] \leq \frac{2}{\sum_{t=0}^{T-1} \eta_t} \left(\sum_{i=1}^{N} \log(n_i + 1) + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^{N} C_i^2 \right). \tag{3.34}$$

Proof. From the definition of $Z_{i,t}$ in relation (3.26), the recursion in the A-M-SMD algorithm can be stated as

$$Y_{i,t+1} = Y_{i,t} - \eta_t (F_i(X_t) + Z_{i,t}). \tag{3.35}$$

Consider (3.32). From Algorithm 3 and (3.7b), we have $X_{i,t} = \nabla \omega_i^*(Y_{i,t})$. Let $Y_i := Y_{i,t}$ and $Z_i := -\eta_t(F_i(X_t) + Z_{i,t})$. From (3.35), we obtain

$$H_i(X_i, Y_{i,t+1}) \leq H_i(X_i, Y_{i,t}) - \eta_t \mathrm{tr}\big((X_{i,t} - X_i)^T (F_i(X_t) + Z_{i,t})\big) + \eta_t^2 \|F_i(X_t) + Z_{i,t}\|_2^2.$$

By adding and subtracting $\eta_t \mathrm{tr}\big((X_{i,t} - X_i)^T F_i(X)\big)$, we get

$$\begin{aligned} H_i(X_i, Y_{i,t+1}) \leq {} & H_i(X_i, Y_{i,t}) - \eta_t \mathrm{tr}\big((X_{i,t} - X_i)^T Z_{i,t}\big) - \eta_t \mathrm{tr}\big((X_{i,t} - X_i)^T (F_i(X_t) - F_i(X))\big) \\ & - \eta_t \mathrm{tr}\big((X_{i,t} - X_i)^T F_i(X)\big) + \eta_t^2 \|F_i(X_t) + Z_{i,t}\|_2^2. \end{aligned} \tag{3.36}$$

Let us define an auxiliary sequence $U_{i,t}$ such that $U_{i,t+1} :\triangleq U_{i,t} + \eta_t Z_{i,t}$, where $U_{i,0} = \mathbf{I}_{n_i}$ and define $V_{i,t} :\triangleq \nabla \omega_i^*(U_{i,t})$. From (3.36), invoking the definition of $Z_{i,t}$ and by adding and subtracting $V_{i,t}$, we obtain

$$\begin{aligned} \eta_t \mathrm{tr}\big((X_{i,t} - X_i)^T F_i(X)\big) \leq {} & H(X_i, Y_{i,t}) - H_i(X_i, Y_{i,t+1}) - \eta_t \mathrm{tr}\big((X_{i,t} - X_i)^T (F_i(X_t) - F_i(X))\big) \\ & + \eta_t \mathrm{tr}\big((V_{i,t} - X_{i,t})^T Z_{i,t}\big) + \eta_t \mathrm{tr}\big((X_i - V_{i,t})^T Z_{i,t}\big) + \eta_t^2 \|\Phi_{i,t}\|_2^2, \end{aligned} \tag{3.37}$$

where for simplicity of notation we use $\Phi_{i,t}$ to denote $\Phi_i(X_t, \xi_t)$. Then, we estimate the term

$\eta_t \mathrm{tr}\big((X_i - V_{i,t})^T Z_{i,t}\big)$. By Lemma 15 and setting $Y_i := U_{i,t}$ and $Z_i := \eta_t Z_{i,t}$, we get

$$\eta_t \mathrm{tr}\big((X_i - V_{i,t})^T Z_{i,t}\big) \le H_i(X_i, U_{i,t}) - H_i(X_i, U_{i,t+1}) + \eta_t^2 \|Z_{i,t}\|_2^2.$$

By plugging the above inequality into (3.37), we get

$$\begin{aligned}\eta_t \mathrm{tr}\big((X_{i,t} - X_i)^T F_i(X)\big) &\le H_i(X_i, Y_{i,t}) - H_i(X_i, Y_{i,t+1}) + H_i(X_i, U_{i,t}) - H_i(X_i, U_{i,t+1}) \\ &+ \eta_t^2 \|Z_{i,t}\|_2^2 + \eta_t \mathrm{tr}\big((V_{i,t} - X_{i,t})^T Z_{i,t}\big) + \eta_t^2 \|\Phi_{i,t}\|_2^2 - \eta_t \mathrm{tr}\big((X_{i,t} - X_i)^T (F_i(X_t) - F_i(X)\big)\,.\end{aligned}$$

Let us define $V_t := \mathrm{diag}\ (V_{1,t}, \ldots, V_{N,t})$. By summing the above inequality form $i = 1$ to N, we get

$$\begin{aligned}\eta_t \mathrm{tr}\big((X_t - X)^T F(X)\big) &\le \textstyle\sum_{i=1}^N H_i(X_i, Y_{i,t}) - \sum_{i=1}^N H_i(X_i, Y_{i,t+1}) + \sum_{i=1}^N H_i(X_i, U_{i,t}) \\ &- \textstyle\sum_{i=1}^N H_i(X_i, U_{i,t+1}) + \eta_t^2 \sum_{i=1}^N \|Z_{i,t}\|_2^2 + \eta_t \mathrm{tr}\big((V_t - X_t)^T Z_t\big) + \eta_t^2 \sum_{i=1}^N \|\Phi_{i,t}\|_2^2,\end{aligned}$$

where we used the monotonicity of mapping F, i.e. $\mathrm{tr}((X_t - X)(F(X_t) - F(X))) \ge 0$. By summing the above inequality form $t = 0$ to $T - 1$, we have

$$\begin{aligned}&\sum_{t=0}^{T-1} \eta_t \mathrm{tr}\big((X_t - X)^T F(X)\big) \le \sum_{i=1}^N H_i(X_i, Y_{i,0}) - \sum_{i=1}^N H_i(X_i, Y_{i,T}) + \sum_{i=1}^N H_i(X_i, U_{i,0}) \\ &- \sum_{i=1}^N H_i(X_i, U_{i,T}) + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^N \|Z_{i,t}\|_2^2 + \sum_{t=0}^{T-1} \eta_t \mathrm{tr}\big((V_t - X_t)^T Z_t\big) + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^N \|\Phi_{i,t}\|_2^2 \\ &\le \sum_{i=1}^N H_i(X_i, Y_{i,0}) + \sum_{i=1}^N H_i(X_i, U_{i,0}) + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^N \|Z_{i,t}\|_2^2 + \\ &\sum_{t=0}^{T-1} \eta_t \mathrm{tr}\big((V_t - X_t)^T Z_t\big) + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^N \|\Phi_{i,t}\|_2^2, \end{aligned} \tag{3.38}$$

where the last inequality holds by $H_i(X_i, Y_i) \ge 0$ implied by Fenchel's inequality. Recall that for $X_i \in \mathcal{X}_i$, $\mathrm{tr}(X_i) = 1$ and $-\log(n_i) \le \mathrm{tr}(X_i \log X_i) \le 0$ [Carlen, 2010]. By choosing

$Y_{i,0} = U_{i,0} = \mathbf{I}_{n_i}/n_i$ and from (3.5), (3.7a) and (3.31), we have

$$\begin{aligned} H_i(X_i, Y_{i,0}) = H_i(X_i, U_{i,0}) &= \operatorname{tr}(X_i \log X_i - X_i) + \log \operatorname{tr}\left(\exp(\mathbf{I}_{n_i} + \frac{\mathbf{I}_{n_i}}{n_i})\right) - \operatorname{tr}\left(\frac{X_i}{n_i}\right) \\ &\leq 0 - 1 + \log(n_i + 1) - \frac{1}{n_i} \leq \log(n_i + 1). \end{aligned}$$

Plugging the above inequality into (3.38) yields

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t \operatorname{tr}\big((X_t - X)^T F(X)\big) = \operatorname{tr}\left(\sum_{t=0}^{T-1} \eta_t (X_t - X)^T F(X)\right) \leq 2 \sum_{i=1}^{N} \log(n_i + 1) \qquad (3.39) \\ + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^{N} \|Z_{i,t}\|_2^2 + \sum_{t=0}^{T-1} \eta_t \operatorname{tr}\big((V_t - X_t)^T Z_t\big) + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^{N} \|\Phi_{i,t}\|_2^2. \end{aligned}$$

Let us define $\gamma_t :\triangleq \frac{\eta_t}{\sum_{k=0}^{T-1} \eta_k}$, then, we have $\overline{X}_T :\triangleq \sum_{t=0}^{T-1} \gamma_t X_t$ by Lemma 14. We divide both sides of (3.39) by $\sum_{t=0}^{T-1} \eta_t$. Then for all $X \in \mathcal{X}$,

$$\begin{aligned} \operatorname{tr}\left(\left(\sum_{t=0}^{T-1} \gamma_t X_t - X\right)^T F(X)\right) = \operatorname{tr}\Big(\big(\overline{X}_T - X\big)^T F(X)\Big) \leq \frac{1}{\sum_{t=0}^{T-1} \eta_t} \Bigg(2 \sum_{i=1}^{N} \log(n_i + 1) \\ + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^{N} \|Z_{i,t}\|_2^2 + \sum_{t=0}^{T-1} \eta_t \operatorname{tr}\big((V_t - X_t)^T Z_t\big) + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^{N} \|\Phi_{i,t}\|_2^2 \Bigg). \end{aligned}$$

Note that the set $\mathcal{X}$ is a convex set. Since $\gamma_t > 0$ and $\sum_{t=0}^{T-1} \gamma_t = 1$, $\overline{X}_T \in \mathcal{X}$. Now, we take the supremum over the set $\mathcal{X}$ with respect to X and use the definition of the G function given by Definition 6. Note that the right-hand side of the preceding inequality is independent of X.

$$\begin{aligned} G(\overline{X}_T) \leq \frac{1}{\sum_{t=0}^{T-1} \eta_t} \Bigg(2 \sum_{i=1}^{N} \log(n_i + 1) + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^{N} \|Z_{i,t}\|_2^2 + \sum_{t=0}^{T-1} \eta_t \operatorname{tr}\big((V_t - X_t)^T Z_t\big) \\ + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^{N} \|\Phi_{i,t}\|_2^2 \Bigg). \end{aligned}$$

By taking expectations on both sides, we get

$$\mathbb{E}[G(\overline{X}_T)] \leq \frac{1}{\sum_{t=0}^{T-1} \eta_t} \left(2 \sum_{i=1}^{N} \log(n_i + 1) + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^{N} \mathbb{E}[\|Z_{i,t}|\mathcal{F}_t\|_2^2] + \right.$$
$$\left. \sum_{t=0}^{T-1} \eta_t \mathbb{E}[\mathrm{tr}\big((V_t - X_t)^T Z_t|\mathcal{F}_t\big)] + \sum_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^{N} \mathbb{E}[\|\Phi_{i,t}|\mathcal{F}_t\|_2^2] \right).$$

By definition, both X_t and V_t are $\mathcal{F}_t$-measurable. Therefore, $V_t - X_t$ is $\mathcal{F}_t$-measurable. In addition, Z_t is $\mathcal{F}_{t+1}$-measurable. Thus, by Assumption 6(c), we have $\mathbb{E}[\mathrm{tr}\big((V_t - X_t)^T Z_t\big)\,|\mathcal{F}_t] = 0$. Applying Assumption 6(b), we have

$$\mathbb{E}[G(\overline{X}_T)] \leq \frac{2}{\sum_{t=0}^{T-1} \eta_t} \left(\sum_{i=1}^{N} \log(n_i + 1) + \sum\nolimits_{t=0}^{T-1} \eta_t^2 \sum_{i=1}^{N} C_i^2 \right).$$

□

Next, we present the convergence rate of the A-M-SMD scheme.

Theorem 3. Consider problem (3.4) and let the sequence $\{\overline{X}_t\}$ be generated by A-M-SMD algorithm. Suppose Assumption 6 holds. Given a fixed $T > 0$, let η_t be a sequence given by

$$\eta_t = \frac{1}{\sum_{i=1}^{N} C_i} \sqrt{\frac{\sum_{i=1}^{N} \log(n_i + 1)}{T}}, \quad \text{for all} \quad t \geq 0. \tag{3.40}$$

Then, we have,

$$\mathbb{E}[G(\overline{X}_T)] \leq 3 \sum_{i=1}^{N} C_i \sqrt{\frac{\sum_{i=1}^{N} \log(n_i + 1)}{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \tag{3.41}$$

Proof. Consider relation (3.34). Assume that the number of iterations T is fixed and $\eta_t = \eta$

for all $t \geq 0$, then, we get

$$\mathbb{E}[G(\overline{X}_T)] \leq \frac{2\left(\sum_{i=1}^N \log(n_i+1) + T\eta^2 \sum_{i=1}^N C_i^2\right)}{T\eta}.$$

Then, by minimizing the right-hand side of the above inequality over $\eta > 0$, we obtain the constant stepsize (3.40). By plugging (3.40) into (3.34), we obtain (3.41). □

3.5 Numerical Experiments

In this section, we examine the behavior of A-M-SMD method on throughput maximization problem in a multi-user MIMO wireless network as described in Section 1.2. First, we need to show that the Nash equilibrium of game (1.7) is a solution of VI($\mathcal{X}, F$). In order to apply Lemma 11, we need to prove that the throughput function $R_i(\mathbf{X}_i, \mathbf{X}_{-i})$ is a concave function. In the next Lemma, we show the sufficient conditions on two functions that guarantee the concavity of their composition. We use the following definitions in the proof.

Definition 7. (Matrix convex function) Let $\mathbb{C}^n$ be the complex vector space.

(a) An arbitrary matrix $\mathbf{A} \in \mathbb{H}_m$ is nonnegative if $\langle \mathbf{A}\mathbf{y}, \mathbf{y} \rangle \geq 0$ for all $\mathbf{y} \in \mathbb{C}^n$.

(b) For $\mathbf{A}, \mathbf{B} \in \mathbb{H}_m$ we write $\mathbf{A} \geq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is nonnegative.

(c) A function $f : \mathbb{H}_m \to \mathbb{H}_n$ is convex if $f(\lambda\mathbf{A} + (1-\lambda)\mathbf{B}) \leq \lambda f(\mathbf{A}) + (1-\lambda)f(\mathbf{B})$, for all $0 \leq \lambda \leq 1$.

(d) A function $f : \mathbb{H}_m \to \mathbb{H}_n$ is called matrix monotone increasing if $\mathbf{A} \geq \mathbf{B}$ implies $f(\mathbf{A}) \geq f(\mathbf{B})$. [Watkins, 1974]

(e) A function $f : \mathbb{H}_m \to \mathbb{R}$ is called matrix monotone increasing if $\mathbf{A} \geq \mathbf{B}$ implies $f(\mathbf{A}) \geq f(\mathbf{B})$. [Kwong, 1989]

Lemma 17. Suppose $h : \mathbb{H}_n \to \mathbb{R}$ and $g : \mathbb{H}_m \to \mathbb{H}_n$. Then, $f(\mathbf{X}) = h(g(\mathbf{X}))$ is concave if h is concave and matrix monotone increasing and g is concave.

Proof. Assume that $\mathbf{X}, \mathbf{Z} \in \mathbb{H}_m$, and $0 \leq \lambda \leq 1$. By convexity of $\mathbb{H}_m$, we have $\lambda\mathbf{X}+(1-\lambda)\mathbf{Z} \in \mathbb{H}_m$, and from concavity of g, we have

$$g(\lambda\mathbf{X} + (1-\lambda)\mathbf{Z}) \geq \lambda g(\mathbf{X}) + (1-\lambda)g(\mathbf{Z}). \tag{3.42}$$

Since h is matrix monotone increasing and by definition 7(e), we get

$$h\left(g(\lambda\mathbf{X} + (1-\lambda)\mathbf{Z})\right) \geq h\left(\lambda g(\mathbf{X}) + (1-\lambda)g(\mathbf{Z})\right) \geq \lambda h(g(\mathbf{X})) + (1-\lambda)h(g(\mathbf{Z})), \tag{3.43}$$

where the last inequality follows from concavity of h. Therefore,

$$h\left(g(\lambda\mathbf{X} + (1-\lambda)\mathbf{Z})\right) \geq \lambda h(g(\mathbf{X})) + (1-\lambda)h(g(\mathbf{Z})), \tag{3.44}$$

and we conclude that f is a concave function. □

Now, we apply Lemma 17 to show each player's objective function $R_i(\mathbf{X}_i, \mathbf{X}_{-i})$ is concave.

Lemma 18. The user's transmission throughput function $R_i(\mathbf{X}_i, \mathbf{X}_{-i})$ is concave in $\mathbf{X}_i$.

Proof. Let us define $\mathbf{W}(\mathbf{X}_i) = \mathbf{I}_{m_i} + \sum_{j\neq i} \mathbf{H}_{ji}\mathbf{X}_j\mathbf{H}_{ji}^{\dagger} + \mathbf{H}_{ii}\mathbf{X}_i\mathbf{H}_{ii}^{\dagger}$. The function $\mathbf{W}(\mathbf{X}_i)$ is a linear function in terms of $\mathbf{X}_i$. Note that every linear transformation T of the form $T : \mathbf{A} \rightarrow \sum_i \alpha_i \mathbf{H}_{ii}^{\dagger}\mathbf{A}^T\mathbf{H}_{ii}$ preserves Hermitian matrices [de Pillis, 1967], where α_i is a real scalar, and each $\mathbf{H}_{ii}$ is a certain matrix depending on T. Therefore, $\mathbf{W}(\mathbf{X}_i)$ is Hermitian. Therefore, by definition 7(c), $\mathbf{W}(\mathbf{X}_i)$ is both convex and concave in $\mathbf{X}_i$.

We also know that $\log\det(\mathbf{X}^{-1})$ is monotone decreasing [Vandenberghe et al., 1998], meaning that if $\mathbf{A} \geq \mathbf{B}$, then $\log\det(\mathbf{A}^{-1}) \leq \log\det(\mathbf{B}^{-1})$. Then, we have $\log\det(\mathbf{I}_{m_i}) = \log\det(\mathbf{A}\mathbf{A}^{-1}) = \log\det(\mathbf{A}) + \log\det(\mathbf{A}^{-1})$, which results in $\log(1) = 0 = \log\det(\mathbf{A}) + \log\det(\mathbf{A}^{-1})$. Therefore, $\log\det(\mathbf{A}) \geq \log\det(\mathbf{B})$ which means $\log\det(\mathbf{X})$ is monotone increasing.

We also know that $g(\mathbf{X}) = \log\det(\mathbf{X})$ is a concave function ([Boyd and Vandenberghe, 2004], page 74). From convexity of $\mathbf{W}(\mathbf{X}_i)$ and Lemma 17 , we conclude that $R_i(\mathbf{X}_i, \mathbf{X}_{-i}) = \log\det\Big(\mathbf{I}_{m_i} + \sum_j \mathbf{H}_{ji}\mathbf{X}_j\mathbf{H}_{ji}^{\dagger}\Big) - \log\det(\mathbf{W}_{-i})$ is a concave function in $\mathbf{X}_i$. □

The following Corollary shows that sufficient equilibrium conditions are satisfied, therefore a Nash equilibrium of game (1.7) is a solution of variational inequality problem (3.4).

Corollary 3. The Nash equilibrium of (1.7) is a solution of VI($\mathcal{X}, \mathbf{F}$) where $\mathcal{X} \triangleq \prod_i \mathcal{X}_i$ and $F(\mathbf{X}) \triangleq -\text{diag}\Big(\mathbf{H}_{11}^{\dagger}\mathbf{W}^{-1}\mathbf{H}_{11}, \cdots, \mathbf{H}_{NN}^{\dagger}\mathbf{W}^{-1}\mathbf{H}_{NN}\Big)$.

Proof. Please note that $\nabla_{\mathbf{X}_i} R_i(\mathbf{X}_i, \mathbf{X}_{-i}) = \nabla_{\mathbf{X}_i} \log\det\Big(\mathbf{I}_{m_i} + \sum_j \mathbf{H}_{ji}\mathbf{X}_j\mathbf{H}_{ji}^{\dagger}\Big)$ since the second term, $\log\det(\mathbf{W}_{-i})$, is independent of $\mathbf{X}_i$. Let us define $\mathbf{W} = \Big(\mathbf{I}_{m_i} + \sum_j \mathbf{H}_{ji}\mathbf{X}_j\mathbf{H}_{ji}^{\dagger}\Big)$. Then, we have $\nabla_{\mathbf{X}_i} R_i(\mathbf{X}_i, \mathbf{X}_{-i}) = \mathbf{H}_{ii}^{\dagger}\mathbf{W}^{-1}\mathbf{H}_{ii}$ (Mertikopoulos and Moustakas [2016]). By Lemma 18, each player's objective function $R_i(\mathbf{X}_i, \mathbf{X}_{-i})$ is concave in $\mathbf{X}_i$. We also know that $\mathcal{X}_i$ is a convex set. Therefore, using Lemma 11, we have sufficient conditions to state the game (1.7) as a variational inequality problem VI($\mathcal{X}, F$). □

Next two Lemmas show that the mapping $F(\mathbf{X})$ is monotone. Therefore, the sequence generated by A-M-SMD converges to the weak solution of variational inequality (3.4).

Lemma 19. Suppose $f : \mathbb{H}_m \to \mathbb{R}$ is a differentiable function and $\mathcal{X} \subseteq \mathbb{H}_m$. If f is a convex function, then ∇f is monotone, i.e., $\text{tr}\big(\big(\nabla_{\mathbf{X}}^T f(\mathbf{X}) - \nabla_{\mathbf{Z}}^T f(\mathbf{Z})\big)(\mathbf{X} - \mathbf{Z})\big) \geq 0$, for all $\mathbf{X}, \mathbf{Z} \in \mathcal{B}$.

Proof. By convexity of f and by Lemma 12, we have for arbitrary $\mathbf{X}, \mathbf{Z} \in \mathcal{X}$

$$f(\mathbf{Z}) + \text{tr}\big((\mathbf{X} - \mathbf{Z})^T \nabla_{\mathbf{Z}} f(\mathbf{Z})\big) \leq f(\mathbf{X}).$$

By choosing the points in reverse, we also have

$$f(\mathbf{X}) + \text{tr}\big((\mathbf{Z} - \mathbf{X})^T \nabla_{\mathbf{X}} f(\mathbf{X})\big) \leq f(\mathbf{Z}).$$

Summing the above inequalities, we get

$$f(\mathbf{Z}) + f(\mathbf{X}) + \operatorname{tr}\big((\mathbf{X}-\mathbf{Z})^T\nabla_{\mathbf{Z}} f(\mathbf{Z})\big) + \operatorname{tr}\big((\mathbf{Z}-\mathbf{X})^T\nabla_{\mathbf{X}} f(\mathbf{X})\big) \leq f(\mathbf{X}) + f(\mathbf{Z}),$$

and using the fact that $\operatorname{tr}(\mathbf{A}+\mathbf{B}) = \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B})$, we get the desired result. □

Lemma 20. Consider the function R_i given by (1.6) and its gradient $\nabla^T_{\mathbf{X}_i}\left(R_i(\mathbf{X}_i, \mathbf{X}_{-i})\right) = (\mathbf{H}^\dagger_{ii}\mathbf{W}^{-1}\mathbf{H}_{ii})^T$. The mapping $F(\mathbf{X}) \triangleq -\operatorname{diag}\left(\nabla_{\mathbf{X}_1} R_1(\mathbf{X}_1, \mathbf{X}_{-1}), \ldots, \nabla_{\mathbf{X}_N} R_N(\mathbf{X}_N, \mathbf{X}_{-N})\right) = -\operatorname{diag}\left(\mathbf{H}^\dagger_{11}\mathbf{W}^{-1}\mathbf{H}_{11}, \cdots, \mathbf{H}^\dagger_{NN}\mathbf{W}^{-1}\mathbf{H}_{NN}\right)$ is monotone.

Proof. The function $R_i(\mathbf{X}_i, \mathbf{X}_{-i})$ is concave in $\mathbf{X}_i$ by Lemma 18 and as a result $-R_i(\mathbf{X}_i, \mathbf{X}_{-i})$ is a convex function. Therefore, $\nabla^T_{\mathbf{X}_i}\left(-R_i(\mathbf{X}_i, \mathbf{X}_{-i})\right) = -(\mathbf{H}^\dagger_{ii}\mathbf{W}^{-1}\mathbf{H}_{ii})^T$ is monotone in $\mathbf{X}_i$ by Lemma 19. In other words,

$$\begin{aligned}&-\operatorname{tr}\big(\big(\nabla^T_{\mathbf{X}_i} R_i(\mathbf{X}_i, \mathbf{X}_{-i}) - \nabla^T_{\mathbf{Z}_i} R_i(\mathbf{Z}_i, \mathbf{Z}_{-i})\big)\left(\mathbf{X}_i - \mathbf{Z}_i\right)\big) = \\ &-\operatorname{tr}\Bigg(\left(\mathbf{H}^\dagger_{ii}\mathbf{W}^{-1}(\mathbf{X}_i)\mathbf{H}_{ii} - \mathbf{H}^\dagger_{ii}\mathbf{W}^{-1}(\mathbf{Z}_i)\mathbf{H}_{ii}\right)^T\left(\mathbf{X}_i - \mathbf{Z}_i\right)\Bigg) \geq 0, \quad \text{for all} \quad \mathbf{X}_i, \mathbf{Z}_i \in \mathcal{X}_i.\end{aligned}$$

Then, we have

$$\begin{aligned}&\operatorname{tr}((\mathbf{F}(\mathbf{X}) - \mathbf{F}(\mathbf{Z}))(\mathbf{X}-\mathbf{Z})) = \\ &\operatorname{tr}(-\operatorname{diag}(\nabla_{\mathbf{X}_1} R_1(\mathbf{X}_1, \mathbf{X}_{-1}) - \nabla_{\mathbf{Z}_1} R_1(\mathbf{Z}_1, \mathbf{Z}_{-1}), \ldots, \nabla_{\mathbf{X}_N} R_N(\mathbf{X}_N, \mathbf{X}_{-N}) - \nabla_{\mathbf{Z}_N} R_N(\mathbf{Z}_N, \mathbf{Z}_{-N})) \\ &\times \operatorname{diag}(\mathbf{X}_1 - \mathbf{Z}_1, \ldots, \mathbf{X}_N - \mathbf{Z}_N)) = \\ &\operatorname{tr}\Big(-\operatorname{diag}\Big(\mathbf{H}^\dagger_{11}\mathbf{W}^{-1}(\mathbf{X}_1)\mathbf{H}_{11} - \mathbf{H}^\dagger_{11}\mathbf{W}^{-1}(\mathbf{Z}_1)\mathbf{H}_{11}, \ldots, \mathbf{H}^\dagger_{NN}\mathbf{W}^{-1}(\mathbf{X}_N)\mathbf{H}_{NN} - \mathbf{H}^\dagger_{NN}\mathbf{W}^{-1}(\mathbf{Z}_N) \\ &\mathbf{H}_{NN}\Big) \times \operatorname{diag}\left(\mathbf{X}_1 - \mathbf{Z}_1, \ldots, \mathbf{X}_N - \mathbf{Z}_N\right)\Big) = \\ &-\sum_{i=1}^{N}\sum_{u=1}^{m_i}\sum_{v=1}^{m_i}[(\mathbf{H}^\dagger_{ii}\mathbf{W}^{-1}(\mathbf{X}_i)\mathbf{H}_{ii} - \mathbf{H}^\dagger_{ii}\mathbf{W}^{-1}(\mathbf{Z}_i)\mathbf{H}_{ii})^T]_{uv}[(\mathbf{X}_i - \mathbf{Z}_i)]_{uv} \geq 0.\end{aligned}$$

□

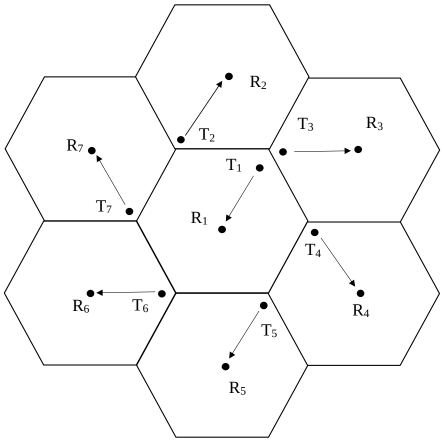

Figure 3.1: Multicell cellular system

Corollary 4. The sequence $\overline{\mathbf{X}}_t$ generated by A-M-SMD algorithm converges to the weak solution of VI$(\mathcal{X}, F)$.

3.5.1 Problem Parameters and Termination Criteria

We consider a MIMO multi-cell cellular network composed of seven hexagonal cells (each with a radius of 1 km) as Figure 3.1. We assume there is one MIMO link (user) in each cell which corresponds to the transmission from a transmitter (T) to a receiver (R). Following Scutari et al. [2009] we generate the channel matrices with a Rayleigh distribution, in other words, each element is generated as circularly symmetric Gaussian random variable with variance equal to the inverse of the square distance between the transmitters and receivers. In this regard, we normalize the distance between transmitters and receivers at first. The network can be considered as a 7-users game where each link (user) is a MIMO channel. Distance between different receivers and transmitters are shown in Table 3.2. It should be noted that the channel matrix between any pair of transmitter i and receiver j is a matrix with dimension of $m_j \times n_i$. In the experiments, we assume $m_j = m$ for all $j \in \{1, \dots, 7\}$ $n_i = n$ for all $i \in \{1, \dots, 7\}$. As mentioned before, p_{max} is the maximum average transmitted power in units of energy per transmission. In the experiments, the transmitters have a maximum power of 1 decibels of the measured power referenced to one milliwatt (dBm). We investigate the robustness of A-M-SMD algorithm under imperfect feedback. To simulate imperfections, the elements of $Z_{i,t}$ are generated as zero-mean circularly symmetric complex Gaussian random variables with variance equal to σ. In experiments, we apply the following gap function $Gap(\mathbf{X})$ which is equal to zero for a strong solution.

Table 3.2: Distance matrix (in terms of kilometer)

Transmitter Receiver	R1	R2	R3	R4	R5	R6	R7
T1	0.8944	1.0143	1.0568	1.1020	1.0143	1.0568	1.1020
T2	1.0143	0.8944	1.0568	2.1079	2.6940	2.6677	1.9964
T3	1.1020	1.9011	0.8944	1.0143	2.1079	2.7265	2.7203
T4	1.9964	2.6159	1.9493	0.8944	1.1020	2.1056	2.7620
T5	2.5635	2.6940	2.6677	1.9964	0.8944	1.0568	2.1079
T6	2.5270	2.1079	2.7265	2.7203	1.9011	0.8944	1.0143
T7	1.9011	1.1020	2.1056	2.7620	2.6159	1.9493	0.8944

Definition 8 (A gap function)**.** Define the following function $Gap : \mathscr{P}_+ \to \mathbb{R}$

$$Gap(\mathbf{X}) = \sup_{\mathbf{Z}\in\mathscr{P}_+} \operatorname{tr}\big((\mathbf{X}-\mathbf{Z})^T F(\mathbf{X})\big), \quad \text{for all } \mathbf{X} \in \mathscr{P}_+. \tag{3.45}$$

In the following lemma, we provide some properties of the Gap function.

Lemma 21 (Properties of the Gap function)**.** The function $Gap(\mathbf{X})$ given by Definition 8 is a well-defined gap function, in other words, (i) $Gap(\mathbf{X})$ is nonnegative for all $\mathbf{X} \in \mathscr{P}_+$; and (ii) $\mathbf{X}^*$ is a strong solution to problem (3.4) iff $Gap(\mathbf{X}^*) = 0$.

Proof. (i) For an arbitrary $\mathbf{X} \in \mathscr{P}_+$, we have

$$Gap(\mathbf{X}) = \sup_{\mathbf{Z}\in\mathscr{P}_+} \operatorname{tr}\big((\mathbf{X}-\mathbf{Z})^T F(\mathbf{X})\big) \geq \operatorname{tr}\big((\mathbf{X}-\mathbf{A})^T F(\mathbf{X})\big), \quad \text{for all } \mathbf{A} \in \mathscr{P}_+.$$

For $\mathbf{A} = \mathbf{X}$, the above inequality suggests that $Gap(\mathbf{X}) \geq \operatorname{tr}\big((\mathbf{X}-\mathbf{X})^T F(\mathbf{X})\big) = 0$ implying that the function $Gap(\mathbf{X})$ is nonnegative for all $\mathbf{X} \in \mathscr{P}_+$.

(ii) Assume $\mathbf{X}^*$ is a strong solution. By definition of VI$(\mathcal{X}, F)$ and relation (3.4), we have

$$\mathrm{tr}\big((\mathbf{X}^* - \mathbf{X})^T F(\mathbf{X}^*)\big) \leq 0, \quad \text{for all } \mathbf{X} \in \mathcal{X}$$

which implies

$$Gap(\mathbf{X}^*) = \sup_{\mathbf{X} \in \mathscr{P}_+} \mathrm{tr}\big((\mathbf{X}^* - \mathbf{X})^T F(\mathbf{X}^*)\big) \leq 0, \quad \text{for all } \mathbf{X} \in \mathcal{X}.$$

On the other hand, from Lemma 21(i), we get $Gap(\mathbf{X}^*) \geq 0$. We conclude that for any strong solution $\mathbf{X}^*$, we have $Gap(\mathbf{X}^*) = 0$. Conversely, assume that there exist an $\mathbf{X}$ such that $Gap(\mathbf{X}) = 0$. Therefore, $\sup_{\mathbf{Z} \in \mathscr{P}_+} \mathrm{tr}\big((\mathbf{X} - \mathbf{Z})^T F(\mathbf{X})\big) = 0$ which implies $\mathrm{tr}\big((\mathbf{X} - \mathbf{Z})^T F(\mathbf{X})\big) \leq 0$ for all $\mathbf{Z} \in \mathscr{P}_+$. Equivalently, we get $\mathrm{tr}\big((\mathbf{Z} - \mathbf{X})^T F(\mathbf{X})\big) \geq 0$ for all $\mathbf{Z} \in \mathscr{P}_+$ implying $\mathbf{X}$ is a strong solution. □

The algorithms are run for a fixed number of iterations T. We plot the gap function for different number of transmitter antennas (n) and receiver antennas (m). We also plot the gap function for different values of σ including $0.5, 1, 5$. We use MATLAB to run the algorithms and CVX software to solve the optimization problem (3.45). Computational experiments are performed using the same PC running on an Intel Core i5-520M 2.4 GHz processor with 4 GB RAM.

3.5.2 Averaging and Non-averaging Matrix Stochastic Mirror Descent Methods

First, we look into the first 100 iterations in one sample path to see the impact of averaging on the initial performance of matrix stochastic mirror descent (M-SMD) algorithm. Figure 3.2 compares the performance of averaging stochastic mirror descent (A-M-SMD) algorithm with M-SMD in the first 100 iterations. The pair of (n, m) denotes the number of transmitter and receiver antennas. The vertical axis displays the logarithm of gap function (3.45) while the horizontal axis displays the iteration number. In these plots, the blue (dash-dot) and black (solid) curves correspond to the M-SMD and A-M-SMD algorithms, respectively. We observe in Figure 3.2 that A-M-SMD algorithm outperforms the M-SMD in most of the experiments. Importantly, A-M-SMD is significantly more robust with respect to (i) the imperfections and uncertainty (σ); and (ii) problem size (the number of transmitter and receiver antennas). Then, we run both A-M-SMD algorithm and M-SMD for $T = 4000$ iterations and plotted their performance in Figure 3.3. In this figure, the vertical axis displays the logarithm of expected gap function (3.45) while the horizontal axis displays the iteration number. The expectation is taken over $\mathbf{Z}_t$, we repeat the algorithm for 10 sample paths and obtain the average of the gap function. For comparison purposes, we also plot the performance of M-SMD and A-M-SMD algorithms starting from a different initial point with better gap function value. This point is obtained by running the algorithm for 400 iteration and saving the best solution $\mathbf{X}$ to (3.45) and its corresponding $\mathbf{Y}$. In these plots, the blue (dash-dot) and magenta (solid diamond) curves correspond to the M-SMD with the initial solution $\mathbf{X}_0 = \mathbf{X}_0^1 = \mathbf{I}_n/n$ and $\mathbf{X}_0 = \mathbf{X}_0^2 = \mathbf{X}_{400}$ respectively, and the black (solid) and red (dash-dot triangle) curves display the A-M-SMD algorithm with the initial solution $\mathbf{X}_0 = \mathbf{X}_0^1 = \mathbf{I}_n/n$ and $\mathbf{X}_0 = \mathbf{X}_0^2 = \mathbf{X}_{400}$ respectively. As can be seen in Figure 3.3, A-M-SMD algorithm outperforms the M-SMD in all experiments. Particularly, A-M-SMD is significantly more robust with respect to (i) the imperfections (σ); and (ii) problem size. It is also observed that A-M-SMD algorithm converges to the strong solution with rate of convergence of $\mathcal{O}(1/T)$ while M-SMD does not converge for larger values of σ. Moreover, from Figure 3.3, it is evident that the A-M-SMD has better performance compared to M-SMD irrespective to the initial solution.

Stability of M-SMD and A-M-SMD: To compare the stability of two methods, we also plot the expected objective function value R_i against the iteration number in Figure 3.4. Here, we choose $n = m = 4$ and $\sigma = 10$. The algorithm is repeated for 10 sample paths and the average of objective function is obtained. Each plot represents the performance of both

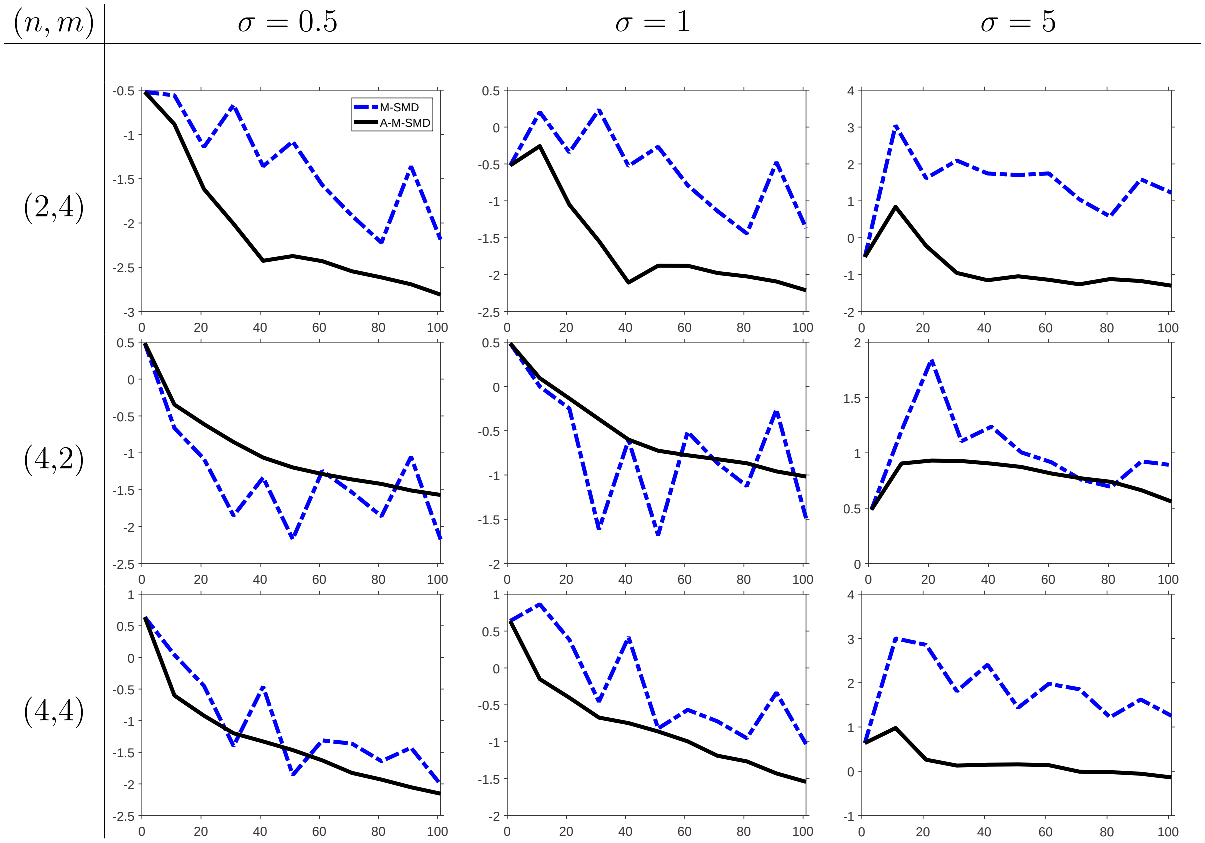

Figure 3.2: Comparison of M-SMD and A-M-SMD w.r.t. problem size (n, m) and uncertainty (σ) for 100 iterations

algorithms for one specific player $i \in \{1, \ldots, 7\}$. As an example, the first plot compares the stability of A-M-SMD (black solid curve) and M-SMD (blue dash-dot curve) for the first user. It can be seen that for all players, the A-M-SMD algorithm converges to a strong solution very fast while the M-SMD does not converge and oscillates significantly.

3.5.3 Matrix Exponential Learning

Mertikopoulos et al. [2017] proved the convergence of matrix exponential learning (MEL) algorithm under strong stability of mapping F assumption while, in practice, this assumption might not hold for the games and VIs. We proved the convergence of A-M-SMD without assuming strong stability. For comparison purposes, we need to regularize the mapping F

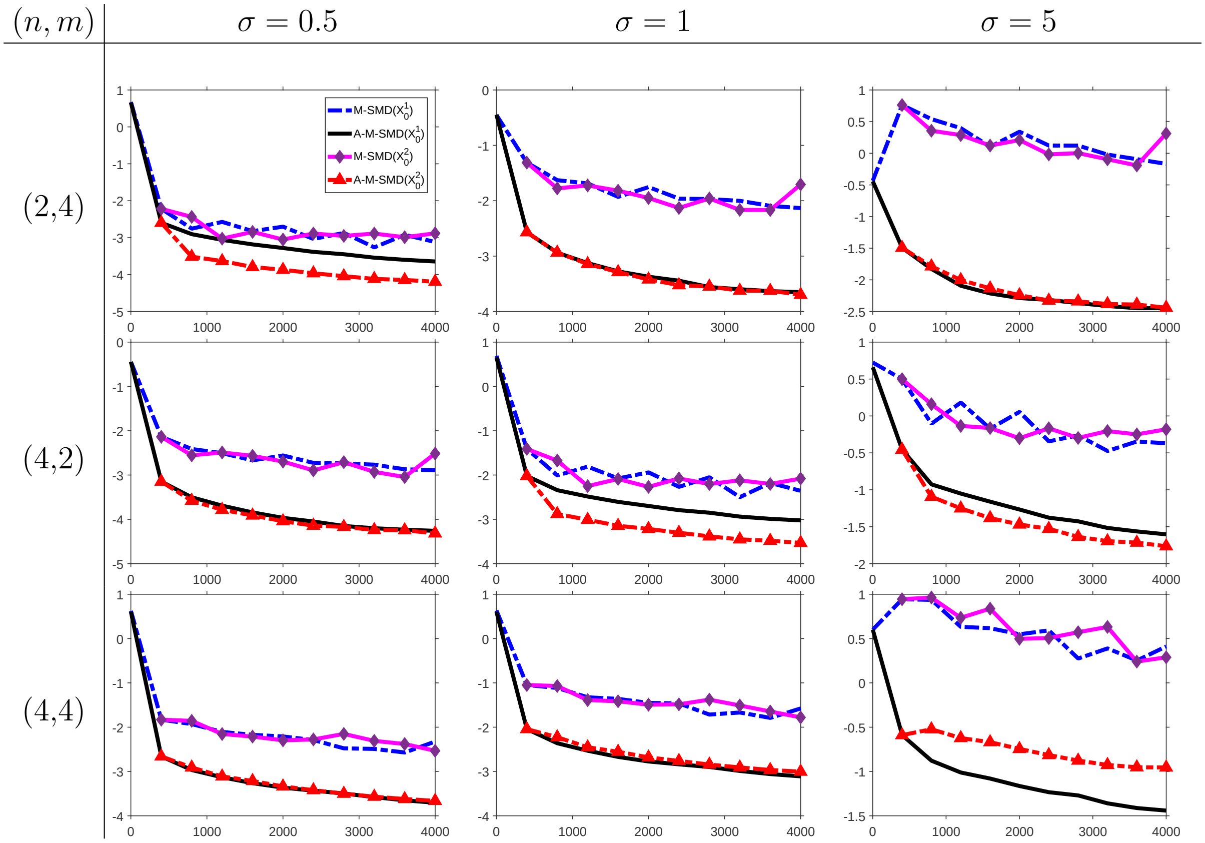

Figure 3.3: Comparison of M-SMD and A-M-SMD w.r.t. initial point (X_0), problem size (n,m), and uncertainty (σ) for 4000 iterations

by adding the gradient of a strongly convex function to it. Doing so, we obtain a strongly stable mapping (Facchinei and Pang [2003], Chapter 2). Let $\|\mathbf{A}\|_F$ denote the Frobenius norm of a matrix $\mathbf{A}$ which is defined as the square root of the sum of the absolute squares of its elements, i.e, $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(A^TA)} = \sqrt{\sum_u \sum_v |[A]_{uv}|^2}$ [Golub and Van Loan, 2012]. In the following Lemma, we show that the function $\frac{1}{2}\|\mathbf{A}\|_F^2$ is strongly convex.

Lemma 22. The function $h(\mathbf{A}) = \frac{1}{2}\|\mathbf{A}\|_F^2$ is strongly convex with parameter 1, i.e.,

$$\frac{1}{2}\|\mathbf{B}\|_F^2 \geq \frac{1}{2}\|\mathbf{A}\|_F^2 + \mathrm{tr}\big(\nabla_{\mathbf{A}}^T h(\mathbf{A})(\mathbf{B} - \mathbf{A})\big) + \frac{1}{2}\|\mathbf{A} - \mathbf{B}\|_F^2. \tag{3.46}$$

Proof. For an arbitrary matrix $\mathbf{A}$, we have $\nabla_{\mathbf{A}}\mathrm{tr}\big(\mathbf{A}^T\mathbf{A}\big) = \mathbf{A}$ [Athans and Schweppe, 1965].

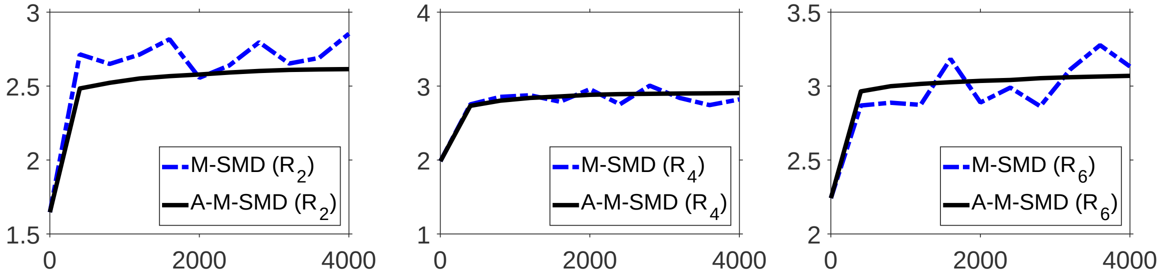

Figure 3.4: Comparison of stability of M-SMD and A-M-SMD in terms of users' objective function R_i for $i = 2, 4, 6$

That being said and using the definition of Frobenius norm, we have

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{A}\|_F^2 + \mathrm{tr}\big(\nabla_{\mathbf{A}}^T h(\mathbf{A})(\mathbf{B}-\mathbf{A})\big) + \frac{1}{2}\|\mathbf{A}-\mathbf{B}\|_F^2 = \\
&\frac{1}{2}\|\mathbf{A}\|_F^2 + \mathrm{tr}\big(\mathbf{A}^T(\mathbf{B}-\mathbf{A})\big) + \frac{1}{2}\mathrm{tr}\big((\mathbf{A}-\mathbf{B})^T(\mathbf{A}-\mathbf{B})\big) = \\
&\frac{1}{2}\|\mathbf{A}\|_F^2 + \mathrm{tr}\big(\mathbf{A}^T(\mathbf{B}-\mathbf{A})\big) + \frac{1}{2}\mathrm{tr}\big(\mathbf{A}^T\mathbf{A} - \mathbf{B}^T\mathbf{A} - \mathbf{A}^T\mathbf{B} + \mathbf{B}^T\mathbf{B}\big) = \\
&\frac{1}{2}\|\mathbf{A}\|_F^2 + \mathrm{tr}\big(\mathbf{A}^T\mathbf{B} - \mathbf{A}^T\mathbf{A}\big) + \frac{1}{2}\mathrm{tr}\big(\mathbf{A}^T\mathbf{A} - \mathbf{B}^T\mathbf{A} - \mathbf{A}^T\mathbf{B} + \mathbf{B}^T\mathbf{B}\big) = \\
&\frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\mathrm{tr}\big(\mathbf{A}^T\mathbf{B}\big) - \frac{1}{2}\mathrm{tr}\big(\mathbf{A}^T\mathbf{A}\big) - \frac{1}{2}\mathrm{tr}\big(\mathbf{B}^T\mathbf{A}\big) + \frac{1}{2}\mathrm{tr}\big(\mathbf{B}^T\mathbf{B}\big) = \\
&\frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\mathrm{tr}\big(\mathbf{A}^T\mathbf{B}\big) - \frac{1}{2}\|\mathbf{A}\|_F^2 - \frac{1}{2}\mathrm{tr}\big(\mathbf{A}^T\mathbf{B}\big) + \frac{1}{2}\|\mathbf{B}\|_F^2 = \frac{1}{2}\|\mathbf{B}\|_F^2.
\end{aligned}$$

Therefore, the inequality (3.46) holds in equality and we conclude that $h(\mathbf{A})$ is strongly convex with parameter 1. □

Note that $\nabla \frac{\lambda}{2}\|\mathbf{X}\|_F^2 = \lambda\mathbf{X}$. Therefore, to regularize the mapping F, we need to add the term $\lambda\mathbf{X}$ to it and consequently, the mapping $F' = F + \lambda\mathbf{X}$ is different from the original F. It should be noted for small values of λ, the algorithm converges very slowly. On the other hand, the solution which is obtained by using large values of λ is far from the solution to the original problem. Hence, we need to find a reasonable value of λ. For this reason, we tried three different values including $0.1, 0.5, 1$. The only difference between MEL and M-SMD algorithm is adding the term $\lambda\mathbf{X}$ to the mapping F.

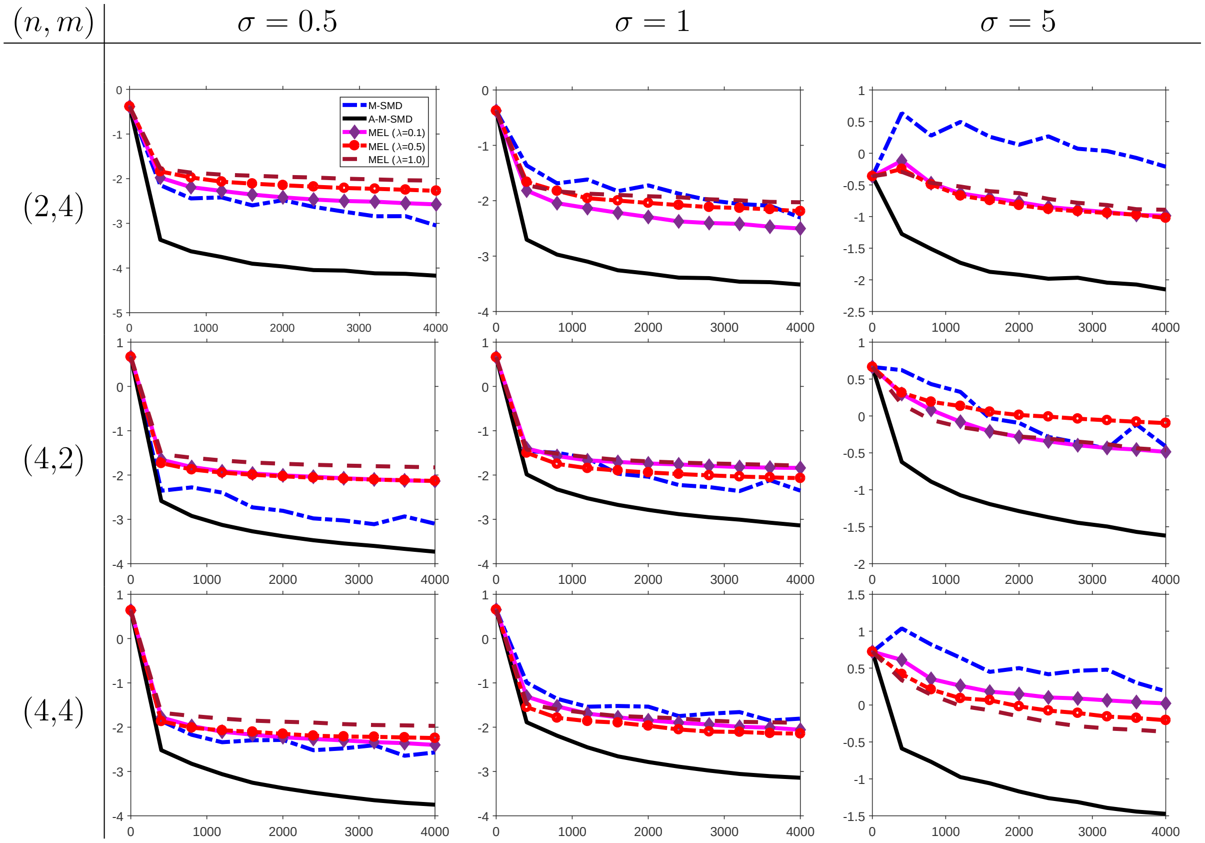

Figure 3.5: Comparison of M-SMD, A-M-SMD and MEL w.r.t. problem size (n, m), uncertainty (σ), and regularization parameter (λ) for 4000 iterations

For each experiment, the algorithm is run for $T = 4000$ iterations. We apply the well-known harmonic stepsize $\eta_t = \frac{1}{\sqrt{t}}$ for A-M-SMD and M-SMD, and harmonic stepsize $\eta_t = \frac{1}{t}$ for MEL. Figure 3.5 demonstrate the performance of A-M-SMD, M-SMD and MEL algorithms in terms of logarithm of expected value of gap function (3.45). The expectation is taken over $\mathbf{Z}_t$, we repeat the algorithm for 10 sample paths and obtain the average of gap function. In these plots, the blue (dash-dot) and black (solid) curves correspond to the M-SMD and A-M-SMD algorithms, respectively, the magenta (solid diamond), red (circle dashed) and brown (dashed) curves display MEL algorithm with $\lambda = 0.1, 0.5$ and 1. As can be seen in Figure 3.5, A-M-SMD algorithm outperforms the M-SMD and MEL algorithms in all experiments. It is evident that MEL algorithm converge slowly but faster than M-SMD.

Comparing three versions of MEL algorithm which apply large, moderate or small value of regularization parameter λ, it can be seen that MEL is not robust w.r.t this parameter since each one of MEL algorithms has a better performance than the other two in some cases.

3.6 Concluding Remarks

We consider multi-agent optimization problems on semidefinite matrix spaces. We develop mirror descent methods where we choose the distance generating function to be defined as the quantum entropy. These first-order single-loop methods include a mirror descent incremental subgradient method for minimizing a convex function that consists of sum of component functions and an averaging matrix stochastic mirror descent method for solving Cartesian stochastic variational inequality problems under monotonicity assumption of the mapping. We show that the iterate generated by M-MDIS algorithm converges asymptotically to the optimal solution and derive a non-asymptotic convergence rate. We also prove that A-M-SMD method converges to a weak solution of the CSVI with rate of $\mathcal{O}(1/\sqrt{T})$. Our numerical experiments performed on a wireless communication network display that the A-M-SMD method is significantly robust w.r.t. the problem size and uncertainty.

CHAPTER IV

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

This dissertation is motivated by applications in machine learning, statistical analysis, and signal processing where the problem can be formulated as a stochastic optimization, finite-sum, or an equilibrium problem and challenges such as uncertainty, high-dimensionality, and matrix structure of the decision variables may arise. We develop, analyze, and implement efficient computational methods to address the aforementioned challenges. In particular, we consider the stochastic mirror descent (SMD) methods that are among the popular avenues in solving stochastic optimization and variational inequality problems.

Much of the past research on SMD methods has focused on convergence and rate analysis in terms of order of the error bounds. However, the finite-time performance of these schemes is tied closely to the choice of the stepsize sequence. Motivated by this gap, in Chapter II, we consider nonsmooth, smooth, and high-dimensional stochastic optimization problems. We develop self-tuned stepsize rules for stochastic subgradient, gradient, and randomized block coordinate mirror descent methods accordingly which incorporate problem parameters, and are tuned as the algorithm goes on. For each scheme, we prove almost sure convergence to the optimal solution of the problem and show that under the self-tuned stepsize rules, the error bound of the stochastic mirror descent scheme is minimized. Moreover, in the case where problem parameters are unknown, we develop a unifying self-tuned update rule that can be applied in both smooth and nonsmooth regimes. We apply our unifying self-tuned stochastic mirror descent method on three classification datasets. The numerical experiments display that our scheme is significantly robust with respect to the uncertainty of data, problem parameters, and the initial stepsize.

In Chapter III, we focus on multi-user optimization problems over semidefinite matrix spaces. The first part of this chapter is motivated by statistical analysis applications such as distributed sparse estimation of covariance inverse matrix. This problem can be formulated as a finite-sum problem where the users (i.e., processors) can cooperatively optimize the likelihood estimation. We develop a mirror descent incremental subgradient (M-MDIS) method for solving the problem. We show that the iterate generated by M-MDIS algorithm converges asymptotically to the optimal solution and derive a non-asymptotic convergence rate. The second part of this chapter is motivated by wireless communication networks where there are transmitters and receivers that generate and detect the signals, respectively. An antenna enables a transmitter to send signals into the space, and enables a receiver to pick up signals from the space. In a multiple-input multiple-output (MIMO) wireless transmission system, multiple antennas are applied in transmitters and receivers in order to improve the performance. Each transmitter tries to maximize its information rate and competes with other transmitters. The transmit power of these transmitters are quantified by their covariance matrices which controls their variances as well. Therefore, the competition among the transmitters in the network can be characterized as a non-cooperative Nash game with positive semidefinite matrix variables. We develop a stochastic matrix mirror descent method equipped with convergence rate to compute the equilibrium of this type of games. The numerical experiments performed on a MIMO multi-cell cellular wireless network show that the proposed method is significantly robust with respect to the problem size and uncertainty.

In learning from data which has an important role in the areas of statistics, data mining, and engineering, the goal is to predict an output based on a number of features. In many real-world problems, the number of available features significantly exceeds the number of samples, but only a small number of features contribute to the response values. In order to cope with high dimensionality of data, one remedy which is proposed in the literature is making solutions sparse [Friedman et al., 2001]. The basic idea is to keep significant features with the strongest impact on the response values in the prediction model and remove the insignificant features. By doing so, we can make the data simpler and more concise, and consequently make the study and processing of the low dimensional samples more efficient. Sparsity helps to understand practical problems better by providing interpretable models. In other words, it reveals a clear relationship between the response variable and the features [Zhou et al., 2011]. That being said, one direction for future research can be developing self-tuned stochastic mirror descent made sparse algorithms.

Moreover, the convergence analysis of the SMD methods discussed in Chapter II requires the objective function to be strongly convex. However, this assumption is fairly restrictive and does not hold for applications such as minimizing the logistic regression loss function. Motivated by this gap, another direction for future research can be considering optimization problems with merely convex objectives and developing a regularized stochastic mirror descent made sparse algorithm, where the stepsize and the regularization parameter are updated iteratively.

Solving nonconvex optimization problems such as the problem of training deep neural networks has become increasingly important as the state-of-the-art in machine learning [Cui et al., 2020]. The global optimization of nonconvex objectives is an NP-hard problem in general [Jain and Kar, 2017]. As a result, a highly desirable goal in applications with nonconvex objectives is to find a local minimum of the objective function. The recent works of Agarwal et al. [2017] and Jin et al. [2019] propose two variants of gradient method for solving high-dimensional nonconvex optimization problems. The run-time of these methods depend quasi-linearly and linearly on the problem dimension. However, the convergence of these methods is only guaranteed to a saddle point. Developing computational algorithms which can solve smooth/nonsmooth nonconvex high-dimensional optimization problems and guarantee convergence to a local or global optimum can be another direction for future research.

Despite the recent advancements in first-order methods addressing problems over vector spaces such as SVRG [Johnson and Zhang, 2013] and SAGA [Defazio et al., 2014], there seem to be some shortcomings in the theory of the first-order methods for finite sum problems on semidefinite matrix spaces. One direction for future research can be developing a fast incremental mirror descent method with a linear convergence rate for strongly convex functions and rate of $\mathcal{O}(1/t)$ for convex functions. This method also can be applied for solving the sparse inverse covariance estimation problem where we need to estimate the inverse of the covariance matrix of a multivariate Gaussian distribution from a small set of samples.

Another direction for future research is developing a randomized block coordinate variant of averaging matrix stochastic mirror descent method discussed in Chapter III. This method can be applied to solve the multi-user maximization throughput problem for the case that there are a large number of the MIMO links in the wireless communication network.

Bibliography

Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., and Ma, T. (2017). Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199.

Athans, M. and Schweppe, F. C. (1965). Gradient matrices and matrix calculations. Technical report, MIT Lincoln Lab.

Axler, S. J. (1997). *Linear algebra done right*, volume 2. Springer.

Beck, A. (2017). *First-Order Methods in Optimization.* SIAM, Philadelphia, PA.

Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

Ben-Tal, A., Margalit, T., and Nemirovski, A. (2001). The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12(1):79–108.

Benveniste, A., Métivier, M., and Priouret, P. (1990). *Stochastic approximations and adaptive algorithms.* Springer-Verlag.

Bertsekas, D. P. (2011). Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163.

Bertsekas, D. P. (2015). Incremental aggregated proximal and augmented Lagrangian algorithms. Technical report, Laboratory for Information and Decision Systems Report LIDS-P-3176, MIT.

Bertsekas, D. P. and Tsitsiklis, J. N. (2000). Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642.

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820.

Bock, R., Chilingarian, A., Gaug, M., Hakl, F., Hengstebeck, T., Jiřina, M., Klaschka, J., Kotrč, E., Savickỳ, P., Towers, S., Vaicilius, A., and W., W. (2004). Methods for multidimensional event classification: a case study using images from a cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 516(2-3):511–528.

Boţ, R. I. and Böhm, A. (2019). An incremental mirror descent subgradient algorithm with random sweeping and proximal step. *Optimization*, 68(1):33–50.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Carlen, E. (2010). Trace inequalities and quantum entropy: an introductory course. *Entropy and the Quantum*, 529:73–140.

Chang, T.-H., Hong, M., and Wang, X. (2015). Multi-agent distributed optimization via inexact consensus ADMM. *IEEE Trans. Signal Processing*, 63(2):482–497.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Cui, Y., He, Z., and Pang, J.-S. (2020). Multicomposite nonconvex optimization for training deep neural networks. *SIAM Journal on Optimization*, 30(2):1693–1723.

Dang, C. D. and Lan, G. (2015a). On the convergence properties of non-Euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and applications*, 60(2):277–310.

Dang, C. D. and Lan, G. (2015b). Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization*, 25(2):856–881.

Darken, C. and Moody, J. (1992). Towards faster stochastic gradient search. In *Advances in neural information processing systems*, pages 1009–1016.

de Pillis, J. (1967). Linear transformations which preserve hermitian and positive semidefinite operators. *Pacific Journal of Mathematics*, 23(1):129–137.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654.

Durham, J. W., Franchi, A., and Bullo, F. (2012). Distributed pursuit-evasion without mapping or global localization via local frontiers. *Autonomous Robots*, 32(1):81–95.

Ermoliev, Y. (1983). Stochastic quasigradient methods and their application to system optimization. *Stochastics: An International Journal of Probability and Stochastic Processes*, 9(1-2):1–36.

Facchinei, F. and Pang, J.-S. (2003). *Finite-dimensional variational inequalities and complementarity problems. Vols. I,II.* Springer Series in Operations Research. Springer-Verlag, New York.

Fazel, M., Hindi, H., and Boyd, S. P. (2001). A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American Control Conference*, volume 6, pages 4734–4739. IEEE.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

George, A. P. and Powell, W. B. (2006). Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming. *Machine learning*, 65(1):167–198.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Gürbuzbalaban, M., Ozdaglar, A., and Parrilo, P. A. (2017). On the convergence rate of incremental aggregated gradient algorithms. *SIAM Journal on Optimization*, 27(2):1035–1048.

Hiai, F. and Petz, D. (2014). *Introduction to matrix analysis and applications.* Springer Science & Business Media.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., Ravikumar, P. K., and Poldrack, R. (2013). BIG & QUIC: Sparse inverse covariance estimation for a million variables. In *Advances in neural information processing systems*, pages 3165–3173.

Jain, P. and Kar, P. (2017). Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336.

Jiang, H. and Xu, H. (2008). Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Transactions on Automatic Control*, 53(6):1462–1475.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2019). Stochastic gradient descent escapes saddle points efficiently. *arXiv preprint arXiv:1902.04811.*

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323.

Juditsky, A., Nemirovski, A., and Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58.

Kakade, S., Shalev-Shwartz, S., and Tewari, A. (2009). On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, https://ttic.uchicago.edu/ shai/papers/KakadeShalevTewari09.pdf*.

Kesten, H. et al. (1958). Accelerated stochastic approximation. *The Annals of Mathematical Statistics*, 29(1):41–59.

Korpelevich, G. (1977). Extragradient method for finding saddle points and other problems. *Matekon*, 13(4):35–49.

Koshal, J., Nedić, A., and Shanbhag, U. V. (2013). Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Transactions on Automatic Control*, 58(3):594–609.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media.

Kwong, M. K. (1989). Some results on matrix monotone functions. *Linear Algebra and Its Applications*, 118:129–153.

Lan, G., Lu, Z., and Monteiro, R. D. (2011). Primal-dual first-order methods with $O(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

Lobel, I. and Ozdaglar, A. (2011). Distributed subgradient methods for convex optimization over random networks. *IEEE Transactions on Automatic Control*, 56(6):1291.

Lu, Z. (2010). Adaptive first-order methods for general sparse inverse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2000–2016.

Luo, Z. Q. and Tseng, P. (1992). On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35.

Luo, Z. Q. and Tseng, P. (1993). Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157—-178.

Majlesinasab, N., Yousefian, F., and Pourhabib, A. (2019). Self-tuned mirror descent schemes for smooth and nonsmooth high-dimensional stochastic optimization. *IEEE Transactions on Automatic Control*, 64(10):4377–4384.

Makhdoumi, A. and Ozdaglar, A. (2017). Convergence rate of distributed ADMM over networks. *IEEE Transactions on Automatic Control*, 62(10):5082–5095.

Mareček, J., Richtárik, P., and Takáč, M. (2015). Distributed block coordinate descent for minimizing partially separable functions. In *Numerical Analysis and Optimization*, pages 261–288. Springer.

Mertikopoulos, P., Belmega, E. V., and Moustakas, A. L. (2012). Matrix exponential learning: Distributed optimization in MIMO systems. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 3028–3032. IEEE.

Mertikopoulos, P., Belmega, E. V., Negrel, R., and Sanguinetti, L. (2017). Distributed stochastic optimization via matrix exponential learning. *IEEE Transactions on Signal Processing*, 65(9):2277–2290.

Mertikopoulos, P. and Moustakas, A. L. (2016). Learning in an uncertain world: MIMO covariance matrix optimization with imperfect feedback. *IEEE Transactions on Signal Processing*, 64(1):5–18.

Mertikopoulos, P. and Sandholm, W. H. (2016). Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324.

Necoara, I., Patrascu, A., and Glineur, F. (2019). Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. *Optimization Methods and Software*, 34(2):305–335.

Nedić, A. (2011). Asynchronous broadcast-based convex optimization over a network. *IEEE Transactions on Automatic Control*, 56(6):1337–1351.

Nedić, A. and Bertsekas, D. P. (2001). Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138.

Nedić, A. and Lee, S. (2014). On stochastic subgradient mirror-descent algorithm with weighted averaging. *SIAM Journal on Optimization*, 24(1):84–107.

Nedić, A. and Olshevsky, A. (2015). Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615.

Nedić, A., Olshevsky, A., and Uribe, C. A. (2017). Distributed learning for cooperative inference. *arXiv preprint arXiv:1704.02718.*

Nedić, A. and Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609.

Nesterov, Y. E. (2010). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341—362.

Nevelśon, M. B. and Hasḿinskii, R. Z. (1973). *Stochastic approximation and recursive estimation*, volume 47. American Mathematical Society.

Ortega, J. M. and Rheinboldt, W. C. (2000). *Iterative solution of nonlinear equations in several variables*. SIAM.

Pflug, G. C. (1988). Stepsize rules, stopping times and their implementation in stochastic quasigradient algorithms. *Numerical techniques for stochastic optimization*, pages 353–372.

Polyak, B. T. (1987). Introduction to optimization. *New York: Optimization Software, Inc.*

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Price, G. R. (1972). Extension of covariance selection mathematics. *Annals of human genetics*, 35(4):485–490.

Ram, S. S., Nedić, A., and Veeravalli, V. V. (2009a). Incremental stochastic subgradient algorithms for convex optimization. *SIAM Journal on Optimization*, 20(2):691–717.

Ram, S. S., Veeravalli, V. V., and Nedić, A. (2009b). Distributed non-autonomous power control through distributed convex optimization. In *IEEE INFOCOM 2009*, pages 3001–3005.

Richtárik, P. and Takáč, M. (2014). Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(2):1–38.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407.

Ruszczyński, A. and Syski, W. (1986). A method of aggregate stochastic subgradients with on-line stepsize rules for convex stochastic programming problems. In *Stochastic Programming 84 Part II*, pages 113–131. Springer.

Saridis, G. N. (1970). Learning applied to successive approximation algorithms. *IEEE Transactions on Systems Science and Cybernetics*, 6(2):97–103.

Scutari, G., Palomar, D. P., and Barbarossa, S. (2009). The MIMO iterative waterfilling algorithm. *IEEE Transactions on Signal Processing*, 57(5):1917–1935.

Scutari, G., Palomar, D. P., Facchinei, F., and Pang, J.-s. (2010). Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine*, 27(3):35–49.

Shi, W., Ling, Q., Wu, G., and Yin, W. (2015). Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966.

Spall, J. C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons.

Stengel, R. F. (2012). *Optimal control and estimation*. Courier Corporation.

Telatar, E. (1999). Capacity of multi-antenna Gaussian channels. *Transactions on Emerging Telecommunications Technologies*, 10(6):585–595.

Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming (Series B.)*, 117(1):387—423.

Tsuda, K., Rätsch, G., and Warmuth, M. K. (2005). Matrix exponentiated gradient updates for on-line learning and Bregman projection. *Journal of Machine Learning Research*, 6(Jun):995–1018.

Vandenberghe, L., Boyd, S., and Wu, S.-P. (1998). Determinant maximization with linear matrix inequality constraints. *SIAM journal on matrix analysis and applications*, 19(2):499–533.

Vedral, V. (2002). The role of relative entropy in quantum information theory. *Reviews of Modern Physics*, 74(1):197.

Wang, M. and Bertsekas, D. (2015). Incremental constraint projection methods for variational inequalities. *Mathematical Programming (Series A.)*, 150:321–363.

Watkins, W. (1974). Convex matrix functions. In *Proceedings of the American Mathematical Society*, volume 44, pages 31–34. JSTOR.

Xiao, L. and Boyd, S. (2006). Optimal scaling of a gradient method for distributed resource allocation. *Journal of optimization theory and applications*, 129(3):469–488.

Xu, Y. and Yin, W. (2013). A block coordinate descent method for multi-convex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789.

Yousefian, F., Nedić, A., and Shanbhag, U. V. (2012). On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67. An extended version of the paper available at: http://arxiv.org/abs/1105.4549.

Yousefian, F., Nedić, A., and Shanbhag, U. V. (2016). Self-tuned stochastic approximation schemes for non-Lipschitzian stochastic multi-user optimization and Nash games. *IEEE Transactions on Automatic Control*, 61(7):1753–1766.

Yousefian, F., Nedić, A., and Shanbhag, U. V. (2017). On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. *Mathematical Programming*, 165(1):391–431.

Yousefian, F., Nedić, A., and Shanbhag, U. V. (2018). On stochastic mirror-prox algorithms for stochastic Cartesian variational inequalities: randomized block coordinate and optimal averaging schemes. *Set-Valued and Variational Analysis*, 26(4):789–819.

Yu, Y.-L. (2013). The strong convexity of von Neumann's entropy. *Unpublished Manuscript, http://www.cs.cmu.edu/ yaoliang/mynotes/sc.pdf.*

Yudin, D. and Nemirovski, A. (1983). *Problem Complexity and Method Efficiency in Optimization.* Wiley-Interscience Series in discrete Mathematics. John Wiley and Sons.

Zhou, T., Tao, D., and Wu, X. (2011). Manifold elastic net: a unified framework for sparse dimension reduction. *Data Mining and Knowledge Discovery*, 22(3):340–371.

VITA

Nahidsadat Majlesinasab

Candidate for the Degree of

Doctor of Philosophy

Dissertation: SELF-TUNED, BLOCK-COORDINATE, AND INCREMENTAL MIRROR DESCENT METHODS WITH APPLICATIONS IN MACHINE LEARNING AND WIRELESS COMMUNICATIONS

Major Field: Industrial Engineering and Management

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Industrial Engineering and Management at Oklahoma State University, Stillwater, Oklahoma in July, 2020

Completed the requirements for the Master of Science in Industrial Engineering at Isfahan University of Technology, Isfahan, Iran in 2013

Completed the requirements for the Bachelor of Science in Industrial Engineering at Golpayegan College of Engineering, Golpayegan, Iran in 2010

Experience:

Employed by MODE Transportation Company in the position of Artificial Intelligence-Driven Lead Generation Intern in Dallas, Texas from January 2020 to July 2020

Employed by Oklahoma State University in the position of Instructor in Stillwater, Oklahoma during the Fall 2019 semester

Employed by Oklahoma State University in the position of Research/Teaching Assistant in Stillwater, Oklahoma from August 2015 to August 2019