RANDOM PROJECTION METHODS FOR STOCHASTIC CONVEX MINIMIZATION

BY

ÜMİT DENİZ TURSUN

DISSERTATION

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

Urbana, Illinois

Doctoral Committee:

Associate Professor Angelia Nedić, Chair
Associate Professor Carolyn L. Beck
Associate Professor Dusan M. Stipanović
Professor Petros G. Voulgaris

ABSTRACT

The first focus of this thesis is to solve a stochastic convex minimization problem over an arbitrary family of nonempty, closed and convex sets. The problem has random features. Gradient or subgradient of objective function carries stochastic errors. Number of constraint sets can be extensive or infinitely many. Constraint sets might not be known apriori yet revealed through random realizations or randomly chosen from a collection of constraint sets throughout the horizon as in online learning concept.

The traditional projection algorithms for solving minimization problems require projecting over complete collection of constraints at once or over a subset of them based on a predefined selection rule. But in practical applications either all of the constraints might not be known apriori or even if they are known projecting on the intersection set might be computationally prohibitive. We propose a two step gradient/subgradient iterative method with random projections. As the first step, a random gradient /subgradient projection is performed before observing the random constraint set realization. After taking random gradient /subgradient projection step we reach an intermittent point, which we obtained without considering the feasibility violation. Once the set realization is revealed or chosen within collection of constraint sets, the feasibility violation of intermittent point is corrected. We showed that projecting onto a random subcollection of them using our algorithm with diminishing stepsize is sufficient to converge to the solution set almost surely. Also the convergence of the algorithm for constant and nondiminishing nonsummable stepsizes are proved within an error bound. As the first set of experiments we tested the performance of the algorithm over a dynamic control system. We study three versions of the problem with correlated unknown-but-bounded additive noise, uncorrelated unknown-but-bounded additive noise and uncorrelated bounded output and peak input additive noise under fully known system description cases. It is essentially a robust least squares estimation problem where we recover state parameters from corrupted input and output data. We reformulated the linear least squares estimation problem as a stochastic convex minimization problem and then used the two step random projection algorithm to solve it. Although the problem has infinite number of constraints due to each realization of error term within bounded set, the algorithm goes through a finite subset of them

and converges to the solution set. We also prove the existence of solution and provide equivalent minimization formulations or upper bound for these three types of robust least squares problems. We used standard subgradient algorithm to gauge the performance of our method. The implementation results are comparable to the ones found in literature.

Our next focus is to solve a stochastic convex feasibility problem. We explored an algorithmic approach to solve both consistent and inconsistent convex feasibility problems for closed convex uncertain sets. We concentrated our attention on uncertain nature of sets and finding a feasible point using a random subcollection of them. The sets we consider might carry uncertainty due to inaccurate or imprecise spatial, spectral, stochastic information and confidence levels. For this objective we consider a stochastic optimization problem of minimizing an expected weighted proximity function over a collection of closed, convex sets. We show that the proposed algorithm converges to a point in the solution set when solution set is nonempty. In case of inconsistent feasibility problem i.e. the intersection of closed convex constraint sets being empty the algorithm minimizes the proximity function. The projection onto a subcollection of sets approach can be viewed as somewhere between random implementation of alternating projection method and parallel projection method. But our method is not deterministic. It uses random projections onto sets that carry additive bounded noise. Each realization within the bounded disturbances has equal chance of occurence. The conventional approach of set theoretic estimation problems provide solution that confirm with constraint sets known a priori or observed. But it fails to take into account that sets built on a priori or observed data may carry disturbances or have erroneously predicted statistical information, which may result in inconsistent sets. The attributes of original signal such as amplitude bound, region of support, band-limitedness that are used to built sets in estimation problems may not be accurate. Additionally sets that are built using moments, spectral properties, distribution and bounds information are based on predicted stochastic estimations. The overly conservative confidence bounds or statistical assumptions may cause inconsistencies. Also noise pertubations in measurements or random variations in the impulse response of a system can cause inconsistencies. Our algorithm projects onto a subcollection of sets some of which carry a random realization of noise on it. The implementation results show that the algorithm converges to the solution asymptotically even if the algorithm projects onto a random subcollection of sets at each iteration.

All in all this thesis work presents iterative methods to solve stochastic convex minimization problems and stochastic convex set intersection problems. The almost sure convergence of algorithms are proven. And the performance of them are shown on numerical experiments.

Dedicated to:

Kerem Sarıca

My little nephew who died due to undiagnosed Hypoplastic Left Heart Syndrome.

He gave me a new sense of responsibility in life.

01/13/2013 - 01/27/2013

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor Professor Angelia Nedić. Her unequivocal expertise, patience and serene sensibility enlightened my path. I was truly lucky to have her as my mentor. I am forever grateful to Professor Imad L. Al-Qadi. He gave me his full support through hard times. I would like to thank my committee members Prof. Carolyn Beck, Prof. Dusan Stipanović, Prof. Petros G. Voulgaris and Prof. Hayri Önal for their time and attention.

Ada, my baby, your smile always cheered me up and kept me going. I am going to be always in debt to my aunt Güler Çelik and my friend Amy Gilbert for taking care of my baby while I was working long hours. I would not have been able to complete this journey without unconditional love and support of my family. My mother Makbule Tursun, my father Hüseyin Tursun, my brother Murat Tursun, my sisters Derya and Demet Tursun were there for me whenever I needed help. Last but not the least my husband, my companion, my constant in life, Hasan was standing by me every step of the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

This thesis study is focused on one main technique, "Random Projection Algorithm under Noise". The proposed algorithms are built upon random projection method to solve convex stochastic smooth/nonsmooth minimization problems and stochastic convex feasibility problem.

In this chapter we are going to present the motivation that led us to study this method as well as a summary of our contributions to the field.

1.1 Motivation

The essence of algorithms that are proposed in this work is iterative random projection method to solve stochastic convex optimization/feasibility problems. Random projection is the technique of projecting a set of points to a randomly chosen low-dimensional space. It has been extensively investigated in theory of learning after Johnson and Lindenstrauss (1984) who proved that random projection approximately preserves key properties of solution sets.

The fundamental idea in this work is analogous to robust concept learning, asserting that a relatively small number of set revelations are sufficient to converge to the optimum/feasible solution set. Robust concept learning aims for finding intersections of examples by reducing the dimensionality of examples, while preserving concepts using random projection, which is essentially a feasibility problem as in Arriaga and Vempala (2006). Convex feasibility problem, which has been surveyed in detail by Bauschke and Borwein (1996), is a special case of the convex stochastic smooth/nonsmooth minimization problem that we focus on. The determination of a common point of convex sets by the method of successive projection first proposed by Bregman (1965) and Gubin et al. (1967). The projection algorithms for convex feasibility problems have broad applicability in many areas of mathematics and physical sciences such as computerized tomography, signal and image processing as in studies by Aharoni and Censor (1989), and Combettes (1996). We ex-

plored an algorithmic approach to solve both consistent and inconsistent convex feasibility problems for closed convex uncertain sets.

One of the employed algorithmic approaches for solving convex feasibility problems is subgradient projection algorithm. It can be either "cyclic" or "weighted" controlled. The classical cyclic subgradient projection method has somewhat similar essence to this work in chapter 3. But classical cyclic subgradient projection method requires sparse set functions depending on very few variables as in Censor and Lent (1982). But our algorithm with random feasibility updates can handle convex feasibility problems as well as convex optimality problems without any provision on the constraint set. When the number of sets is too large to handle, projecting onto intersection of sets is computationally formidable. Whereas our algorithm allows us to project onto a randomly chosen constraint set at each step and is proven that it is converging to the solution set within finite steps.

The robust optimization approach which is extensively covered in Ben-Tal et al. (2009) introduces the robust counterparts concept of uncertain problems that requires semi-infinite programming techniques and thus can be intractable even when all instances of the uncertain problem are easy to solve. Yet projection on one combination of uncertainity set that defines a certain constraint set within all possibilities allows us to formulate and solve the problem with less computational effort. We tested our random projection algorithm under noise on a system identification problem with bounded noise and known system description. The same least squares problem where input matrix and output vector carry unknown but bounded noise was studied by Ghaoui and Lebret (1997). They minimize the worst-case residual error using semidefinite programming input matrix with additive perturbations has lower-triangular toeplitz structure. Based on the implementation results our algorithm has application potential in control problems with uncertain system description and/or system identification with noise which were studied by quite a few researhers and so far where only solution boundaries are achieved as in Bertsekas and Rhodes (1971) and in robust optimization context El Ghaoui and Calafiore (2002).

1.2 Contribution

The contribution of this work is using random gradient/subgradient projection method for stochastic optimization/feasibility problems that utilize random projections even if when gradient/subgradient noise is present. The closest works are Polyak (2001) that uses random gradient projections for the special class

of convex feasibility problems and Bertsekas and Tsitsiklis (2000) that investigates the unconstrained minimization problem with error. The distinction of our method lies in the fact that it handles uncertain objective with constrained case with either large number of constraints or not completely known collection of constraint sets. Another advantage of our method is that gradient/subgradient or gradient/subgradient-analog methods that are based on some type of deterministic or stochastic descent argument assumes that f is bounded below. Yet in our case to establish the almost sure convergence of the proposed algorithms with random projections, we only require Lipschitz continous gradient for differentiable case and uniformly bounded subgradient norm over the universal set that we project.

We also focused on solving the same stochastic problem for nonsmooth objective case. The well-studied subgradient method is very much like the ordinary gradient method with a few differences. It is applicable to nondifferentiable f with step lengths usually fixed ahead of time that does not necessarily decrease f monotonically. Problem scaling and conditioning affect the performance closely. The idea of applying gradient methods with constant step-length for unsmooth functions was first suggested by Shor (1964) in his PhD dissertation for finite dimensional unconstrained problems. Later the convergence of the same case for diminishing stepsize was proved by Ermol'ev (1966) and convergence by geometric progression rate for not summable diminishing step length was proved by Polyak (1967). But the algorithm we propose uses the uncertain subgradient projection direction to minimize the given objective initially, while second level of projection after the constraint set is revealed decreases the feasibility violation. Therefore one of the main contributions of this work is that projection on sets are not limited to projection onto certain super half-spaces, which contain the convex sets.

Notation: A vector is a column vector. We use x^T to denote the transpose of a vector x, and $\|x\|$ to denote the standard Euclidean norm. Minimum distance of a vector $\bar{x}$ to a closed convex set X is $\mathrm{dist}(\bar{x}, X)$. The projection of a vector $\bar{x}$ on a closed convex set X is represented as $\Pi[\bar{x}] = \mathrm{argmin}_{x \in X} \|x - \bar{x}\|^2$. Probability distribution of a random variable Z and expectation of a random variable Z are indicated by $\mathsf{Pr}[Z]$ and $\mathsf{E}[Z]$ respectively.

Chapter 2

SMOOTH STOCHASTIC CONVEX MINIMIZATION: RANDOM PROJECTION ALGORITHM UNDER NOISE

2.1 Introduction

The focus of this chapter is a smooth stochastic convex minimization problem over an arbitrary (possibly infinite) collection of nonempty, closed and convex sets $\{X_i, i \in \mathscr{M}\}$ in $\mathbb{R}^n$. Our objective is to solve the problem by using a two step random projection algorithm.

Stochastic optimization problems have random variables in objective functions. And/or they have random constraints. Firstly the proposed algorithm takes a gradient projection step reaching an intermittent point. The calculated gradient is uncertain carrying a stochastic error term. Just before the second step of the algorithm one of the constraint set is revealed or chosen randomly. Then the feasibility violation of intermittent point is remedied using a subgradient projection onto the revealed/chosen set. The proposed algorithm is in essence generating a random path through a subcollection of constraint sets. So our algorithm is suitable to solve stochastic convex optimization problems with random objective and constraints.

The error/noise term accompanying the gradient can originate from computing, measurement, Monte Carlo sampling error, etc. It is typical for problems for which the gradient is obtained using measurements of extramal control and experimental design systems. It is also common for mean risk type functions for problems of adaptation, learning, pattern recognition. Measurement errors are typically random. But the information about the bound of errors is usually available. In problems of adaptation, online learning, pattern recognition, it is required to minimize mean risk type functions that usually contain an error term. The distribution of error is not specified but rather a sample of it is given. Then the exact computation of gradient is, in principle, impossible. So approximate gradient, which is calculated based on a sample is used instead. The stochastic error term we have is random, independent and centered with bounded deterministic variance.

The feasibility set of the problem is specified as the intersection of possibly infinitely many constraint

sets that are not known in advance. At each algorithm iteration one of them is revealed or chosen. Neither projection on more than one constraint set at a time is needed nor knowledge of complete collection of constraint sets is required. The proposed algorithm is particularly advantageous when the projection on each individual constraint set is easy, on the contrary to the projection on complete collection of constraint sets at once being computationally prohibitive. Especially robust optimization problems that calls for Semi-Definite Programming techniques, with infinite number of set possibilities can be solved with finite number of iterations using this algorithm setup. The algorithm is applicable to online learning problems where the gradient is not defined exactly and constraint sets are revealed through the horizon. In additon to learning problems the algorithm is a viable option when projecting on intersection of a vast number or infinite number of constraint sets is required.

We present the convergence results of algorithm for diminishing square summable, constant and nondiminishing nonsummable stepsizes. It is proved that the algorithm is converging to a random point in optimal set for diminishing square summable stepsize. We have discovered that for constant stepsize error bounds are proportional to gradient and subgradient norm bounds, the set regularity constant, stepsize as well as the variance of noise. In addition we have provided per-iteration and asymptotic error bounds on the expected performance of the algorithm along the averages of the iterates for nondiminishing nonsummable stepsize. Although we have established almost sure convergence properties of the algorithm for constraint sets defined by the convex inequalities, we have also established that when projection on each set is easy in the sense that we have a closed form expression for the projection operation i.e. the constraint sets are defined by convex equalities, the convergence results are also applicable.

In the next section we present the problem formulation and algorithm description as well as the assumptions used throughout this chapter. Section 2.3 is devoted to investigate the convergence properties of the method for differentiable objective function, f with Lipschitz gradients. Implementation details and results for this algorithm are demonstrated for a linear control system with bounded input and output noises in Chapter 4.

2.2 Problem Formulation and Algorithm Description

We consider the following convex constrained minimization problem,

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in X, \quad X \triangleq X_0 \cap (\cap_{i \in \mathscr{M}} X_i), \\ &\text{with } X_i = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0\} \quad \forall i \in \mathscr{M}. \end{aligned} \tag{2.1}$$

where f is convex and each set $X_i \subseteq \mathbb{R}^n$ is nonempty, closed and convex.

The first step of the algorithm that we propose takes a gradient step and reaches an intermittent point υ_k. Then we observe a realization of random variable ω_k, which is a random sample of ω at time k drawn from an arbitrary set of $\omega \in \mathscr{M}$. And we calculate next iterate value x_k using a subgradient step of $g^+_{\omega_k}(x)$ at $x = \upsilon_k$. The subgradient step minimizes randomly chosen "feasibility violation" function $g^+_i(x)$, where $g^+_i(x) = \max\{g_i(x), 0\}$. Computing both the intermittent point υ_k and the new iterate x_k involve a projection operation onto the set X_0.

The iterate process is given by

$$\begin{aligned} \upsilon_k &= \Pi_{X_0}\left[x_{k-1} - \alpha_k\left(\nabla f(x_{k-1}) + \varepsilon_k\right)\right] \\ x_k &= \Pi_{X_0}\left[\upsilon_k - \beta \frac{g^+_{\omega_k}(\upsilon_k)}{\|d_k\|^2} d_k\right] \quad \text{for all } k \geq 1, \end{aligned} \tag{2.2}$$

where $d_k \in \partial g^+_{\omega_k}(\upsilon_k)$, $\alpha_k > 0$ is a deterministic stepsize, and β is also a deterministic parameter with $0 < \beta < 2$. The initial point $x_0 \in X_0$ is selected randomly with an arbitrary distribution. The absolute random noise, ε_k can be interpreted as the stochastic error associated with the evaluation of the gradient $\nabla f(x)$ at $x = x_{k-1}$. Stochastic error affecting objective function is common for wide array of applications such as in robust predictive filters for dynamic systems as in Bertsekas and Rhodes (1971).

We let f^* and X^* denote the optimal value and optimal set of problem (2.1) respectively,

$$f^* = \inf_{x \in X} f(x), \qquad X^* = \{x \in X \mid f(x) = f^*\}. \tag{2.3}$$

The assumptions that are used throughout the chapter for random projection algorithm under noise for smooth objective function (2.2) are introduced below.

Assumption 1. *The functions f and every g_i are defined and convex over some open set that contains the set X_0. The subgradients s_{g_i} are uniformly bounded over the set X_0,*

$$\|s_{g_i}(x)\| \leq C_g \quad \textit{for all } x \in X_0 \textit{ and } \forall i \in \mathscr{M},$$

where C_g is a positive scalar.

The following assumption relates the distance from a point x to the set X with expected feasibility violation for inequality constraint $g^+_{\omega_k}(x)$.

Assumption 2. *The global error bound on the distance between an arbitrary point in set X_0 and its projection on a nonempty convex set determined by convex inequality ω_k, is measured in terms of a residual $g^+_{\omega_k}(x) := max\{0, g_{\omega_k}(x)\}$ as follows*

$$dist^2(x,X) \leq c\mathsf{E}\left[\left(g^+_{\omega_k}(x)\right)^2\right] \quad \textit{for all } x \in X_0,$$

where $c \geq 0$ is some scalar.

This Assumption 2 is related to metric regularity, specifically a finite collection of sets $X_i = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0\}$ is linearly metrically regular with respect to their representations if there exists a constant $\gamma > 0$ such that $\operatorname{dist}(x,X) \leq \gamma \max_{i \in \mathscr{M}} g_i^+(x)$ for all x in $\mathbb{R}^n$ as it is stated in Facchinei and Pang (2003), Vol. I, Section 6.8. The Assumption 2 is satisfied when there is a scalar γ such that the following global error bound holds

$$\operatorname{dist}(x,X) \leq \gamma \max\left\{\operatorname{dist}(x,X_0), \max_{i \in \mathscr{M}} g_i^+(x)\right\} \quad \text{for all } x \in \mathbb{R}^n.$$

A global error bound on the distance between an arbitrary point in the n-dimensional real space $\mathbb{R}^n$ and its projection on a nonempty convex set determined by m convex inequalities is measured in terms of a residual as in Mangasarian (1996) where residual $g^+_{\omega_k}(x) := \max\{0, g_{\omega_k}(x)\}$ is an indication of violations of the inequalities.

We let $\mathscr{F}_k$ denote the history of the method run up to time k,

$$\mathscr{F}_k = \{x_0, (\varepsilon_t, 1 \leq t \leq k), (\omega_t, 1 \leq t \leq k)\} \quad \text{for } k \geq 1,$$

with $\mathscr{F}_0 = \{x_0\}$. Using this notation, we now specify our assumption on the stochastic errors ε_k.

Assumption 3. *The stochastic errors ε_k are random, independent, centered and have bounded variance*

$$\mathsf{E}\left[\varepsilon_k \mid \mathscr{F}_{k\text{-}1}\right] = 0, \qquad \mathsf{E}\left[\|\varepsilon_k\|^2 \mid \mathscr{F}_{k\text{-}1}\right] \leq \nu_k \quad \textit{for all } k \geq 1,$$

where the scalars ν_k are deterministic and the sequence $\{\varepsilon_k\}$ is independent of $\{\omega_k\}$ and x_0.

This concludes the assumptions further used in the text for the proposed algorithm for the optimization problem (2.1) with convex inequality constraint sets.

Next we will introduce a few well-known lemmas for Euclidean projection operation.

Lemma 1. *The non-expansive property of Euclidean projection operation on a closed convex set $Y \subseteq \mathbb{R}^n$ is given as*

$$\|\Pi_Y[x] - z\| \leq \|x - z\| \quad \textit{for all } z \in Y, \textit{ and } x \in \mathbb{R}^n. \tag{2.4}$$

The proof of this result is presented in the book by Facchinei and Pang (2003) (Vol. I, page 77). Another variation of the non-expansive property of Euclidean projection operation is presented in the book by Polyak (1987) (page 121).

Lemma 2. *For a closed convex set $Y \subseteq \mathbb{R}^n$*

$$\|\Pi_Y[x] - \Pi_Y[z]\| \leq \|x - z\| \quad \textit{for any } x, z \in \mathbb{R}^n. \tag{2.5}$$

The strictly non-expansive property of Euclidean projection operation is as follows.

Lemma 3. *For a nonempty closed convex set $Y \subseteq \mathbb{R}^n$*

$$\|\Pi_Y[x] - z\|^2 \leq \|x - z\|^2 - \|x - \Pi_Y[x]\|^2 \quad \textit{for all } z \in Y,\ x \in \mathbb{R}^n. \tag{2.6}$$

The proof of this result can be found in Facchinei and Pang (2003) (Vol. II, 12.1.13 Lemma, page 1120).

In order to investigate the random characteristics of sequences, the following supermartingale convergence result due to Robbins and Siegmund (1971) (see also Polyak (1987), Lemma 11, page 50) is used.

Theorem 1. *Let v_k, u_k, a_k, b_k be sequences of nonnegative random variables that may be dependent and let*

$$\mathsf{E}\left[v_{k+1} \mid \mathscr{F}_k\right] \leq (1+a_k)v_k - u_k + b_k \qquad a.s. \quad \text{for all } k \geq 0,$$
$$\sum_{k=0}^{\infty} a_k < \infty \qquad a.s., \qquad \sum_{k=0}^{\infty} b_k < \infty \qquad a.s.,$$

where $\mathscr{F}_k$ denotes the collection $v_0, ..., v_k$, $u_0, ..., u_k$, $a_0, ..., a_k$, $b_0, ..., b_k$. Then

$$\lim v_k \to v \qquad a.s., \qquad \sum_{k=0}^{\infty} u_k < \infty \qquad a.s.,$$

where $v \geq 0$ is some random variable.

The next theorem is Danskin's Theorem from Bertsekas et al. (2003) (Proposition 4.5.1, page 245), which relates the subdifferential set to convex hull of a "max" function.

Theorem 2. *If $\phi(x,z)$ is differentiable with respect to x for all the points of the maximizing set $z \in Z_0$, where*

$$Z_0(x) = \left\{ \bar{z} : \phi(x,\bar{z}) = \max_{z \in Z} \phi(x,z) \right\},$$

then the subdifferential of $f(x)$ is given by

$$\partial f(x) = \operatorname{conv}\left\{ \nabla_x \phi(x,z) : z \in Z_0(x) \right\}.$$

Consequently the subdifferential of feasibility violation function is

$$\partial g^+(x) = \left\{ \alpha \partial g(x) \mid \alpha \in [0,1] \right\}.$$

Hence the direction to decrease the feasibility violation at k^{th} intermittent point υ_k is chosen as $d_k \in \partial g^+(\upsilon_k)$ where convex hull containing

$$\partial g^+(\upsilon_k) = \begin{cases} \partial g_i(\upsilon_k) & \text{if } g(\upsilon_k) > 0, \\ \alpha \partial g_i(\upsilon_k) \mid \alpha \in [0,1] & \text{if } g(\upsilon_k) \leq 0. \end{cases}$$

2.2.1 Algorithm Modification when Projection on Individual Sets, X_i are available in closed-form

Constrained minimization problem that we aim to solve can also be modified to cover constraint functions having convex level sets, when the projection on each set X_i has a closed form expression for the projection operation. Thus the targeted minimization problem takes the form

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in X, \quad X \triangleq \cap_{i \in \mathscr{M}} X_i, \end{aligned}$$

where each set $X_i \subseteq \mathbb{R}^n$ is nonempty, closed and convex, while the function f is defined and convex over an open set that contains the set $\cup_{i \in \mathscr{M}} X_i$. Degree of infeasibility indicator defined previously for algorithm (2.2), $g_i^+(x) = \max\{d(x, X_i)\}$, takes nonnegative values. Based on definition for a closed convex set, $C \subset \mathbb{R}^n$, $d(x) = \min\{\|y - x\| \mid y \in C\} = \|\Pi_C(x) - x\|$, $\langle x - \Pi_C(x), y - \Pi_C(x) \rangle \leq 0 \quad \forall y \in C$ and subdifferential of distance function to the set C evaluated at $x = x^0$ is

$$\partial d\left(x^0\right) = \left\{ \frac{x^0 - \Pi_C\left(x^0\right)}{\|x^0 - \Pi_C(x^0)\|} \right\} \quad \text{if } x^0 \notin C,$$

so we have

$$d_k \in \partial g_{\omega_k}^+(\upsilon_k) = \left\{ \frac{\upsilon_k - \Pi_{X_{\omega_k}}[\upsilon_k]}{\left\| \upsilon_k - \Pi_{X_{\omega_k}}[\upsilon_k] \right\|} \right\} \quad \text{if } \upsilon_k \notin X_{\omega_k}.$$

In case of $\upsilon_k \notin X_{\omega_k}$ and $\beta = 1$ the algorithm takes the form

$$x_k = [\upsilon_k] - \beta \frac{\left\| \upsilon_k - \Pi_{X_{\omega_k}}[\upsilon_k] \right\| \dfrac{\left[\upsilon_k - \Pi_{X_{\omega_k}}[\upsilon_k] \right]}{\left\| \upsilon_k - \Pi_{X_{\omega_k}}[\upsilon_k] \right\|}}{\dfrac{\left[\upsilon_k - \Pi_{X_{\omega_k}}[\upsilon_k] \right]^T \left[\upsilon_k - \Pi_{X_{\omega_k}}[\upsilon_k] \right]}{\left\| \upsilon_k - \Pi_{X_{\omega_k}}[\upsilon_k] \right\|^2}} = \Pi_{X_{\omega_k}}[\upsilon_k].$$

Therefore the modified algorithm is

$$x_k = \quad \Pi_{X_{\omega_k}}\left[x_{k-1} - \alpha_k\left(\bigtriangledown f(x_{k-1}) + \varepsilon_k\right)\right] \quad \text{for all } k \geq 1,$$

where

$$X_i = \{x \in \mathbb{R}^n \mid g_i(x) = d(x, X_i) = \inf\{\|x - x_0\| \mid x_0 \in X_i\} \leq 0\} \text{ for any } i \in \mathscr{M}.$$

2.3 Convergence Results for Differentiable Objective Function

In this section, we show convergence behavior of method (2.2) for differentiable objective function f with various stepsizes. For upcoming convergence results, we assume f has Lipschitz continuous gradients with constant L over the set X_0, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \qquad for\, all\, x, y \in X_0. \tag{2.7}$$

Throughout this section we assume that Assumptions 1-3 hold.

2.3.1 Preliminary Results

Firstly we are going to establish preliminary results to be used in convergence analysis of method (2.2). The first result relates the distance between an iterate point and any point in set X.

Lemma 4. *Let X be a closed convex set and y be defined as follows*

$$y = \Pi_{X_0}\left[\upsilon - \beta \frac{g^+(\upsilon)}{\|d\|^2} d\right],$$

where $d \in \partial g^+(\upsilon)$ and $\upsilon = \Pi_{X_0}[x - \alpha(\nabla f(x) + \varepsilon)]$. Then, we have

$$\begin{aligned}\|y - \bar{x}\|^2 \leq \|x - \bar{x}\|^2 + 2\alpha(f(\bar{x}) - f(x)) + 2\alpha\varepsilon^T(\bar{x} - x) - \|x - \upsilon\|^2 + 2\alpha\nabla f(x)^T(x - \upsilon) \\ + 2\alpha\varepsilon^T(x - \upsilon) + \left(\beta^2 - 2\beta\right)\frac{g^+(\upsilon)^2}{\|d\|^2} \quad for\, all\, \bar{x} \in X.\end{aligned}$$

Proof. We use strictly non-expansive property of the projection operation (2.6), in order to establish the distance between $y = \Pi_{X_0}\left[\upsilon - \beta \frac{g^+(\upsilon)}{\|d\|^2} d\right]$ and a point $\bar{x} \in X$. Then we get

$$\left\|\Pi_{X_0}\left[\upsilon - \beta \frac{g^+(\upsilon)}{\|d\|^2} d\right] - \bar{x}\right\|^2 \leq \left\|\left[\upsilon - \beta \frac{g^+(\upsilon)}{\|d\|^2} d\right] - \bar{x}\right\|^2 - \left\|\left[\upsilon - \beta \frac{g^+(\upsilon)}{\|d\|^2} d\right] - \Pi_{X_0}\left[\upsilon - \beta \frac{g^+(\upsilon)}{\|d\|^2} d\right]\right\|^2 \tag{2.8}$$

for all $\bar{x} \in X$, where $\upsilon = \Pi_{X_0}[x - \alpha(\nabla f(x) + \varepsilon)]$.

By using subgradient property and the subdifferential of feasibility violation function $g^+(\upsilon)$ being able to be presented as a convex hull, the relation $(\bar{x} - \upsilon)^T d \leq g^+(\bar{x}) - g^+(\upsilon)$ holds. Therefore the first right hand side term of the inequality above is

$$\left\| (\upsilon - \bar{x}) - \beta \frac{g^+(\upsilon)}{\|d\|^2} d \right\|^2 \leq \|\upsilon - \bar{x}\|^2 + 2\beta \left(g^+(\bar{x}) - g^+(\upsilon) \right) \frac{g^+(\upsilon)}{\|d\|^2} + \beta^2 \frac{g^+(\upsilon)^2}{\|d\|^2}.$$

Then we use strict non-expansive property of the projection operation (2.6) to estimate the term $\|\upsilon - \bar{x}\|^2$ above

$$\|\upsilon - \bar{x}\|^2 = \|\Pi_{X_0}[\vartheta] - \bar{x}\|^2 \leq \|\vartheta - \bar{x}\|^2 - \|\vartheta - \Pi_{X_0}[\vartheta]\|^2 \quad \text{for all } \bar{x} \in X,$$

where $\vartheta = x - \alpha(\nabla f(x) + \varepsilon)$.

Thus the estimated bound for inequality (2.8) is

$$\begin{aligned} \|y - \bar{x}\|^2 \leq \|\vartheta - \bar{x}\|^2 - \|\vartheta - \Pi_{X_0}[\vartheta]\|^2 + 2\beta \left(g^+(\bar{x}) - g^+(\upsilon) \right) \frac{g^+(\upsilon)}{\|d\|^2} + \beta^2 \frac{g^+(\upsilon)^2}{\|d\|^2} \\ - \left\| \left[\upsilon - \beta \frac{g^+(\upsilon)}{\|d\|^2} d \right] - \Pi_{X_0} \left[\upsilon - \beta \frac{g^+(\upsilon)}{\|d\|^2} d \right] \right\|^2 \quad \text{for all } \bar{x} \in X. \end{aligned}$$

Although the last term is tightening the bound on the estimation, it is going to be dropped from this point forward. For a feasible point $\bar{x} \in X$ the feasibility violation $g^+(\bar{x})$ vanishes. Therefore the preceding inequality above yields

$$\|y - \bar{x}\|^2 \leq \|x - \alpha(\nabla f(x) + \varepsilon) - \bar{x}\|^2 - \|x - \alpha(\nabla f(x) + \varepsilon) - \upsilon\|^2 + \left(\beta^2 - 2\beta \right) \frac{g^+(\upsilon)^2}{\|d\|^2}, \tag{2.9}$$

where $\upsilon = \Pi_{X_0}[x - \alpha(\nabla f(x) + \varepsilon)]$.

The terms $\|x - \alpha(\nabla f(x) + \varepsilon) - \bar{x}\|^2$ and $\|x - \alpha(\nabla f(x) + \varepsilon) - \upsilon\|^2$ can be bounded using convexity of f as follows:

$$\|x - \alpha(\nabla f(x) + \varepsilon) - \bar{x}\|^2 \leq \|x - \bar{x}\|^2 + 2\alpha(f(\bar{x}) - f(x)) + 2\alpha\varepsilon^T(\bar{x} - x) + \alpha^2 \|\nabla f(x) + \varepsilon\|^2, \tag{2.10}$$

$$\|x - \alpha(\nabla f(x) + \varepsilon) - \upsilon\|^2 = \|x - \upsilon\|^2 + 2\alpha\nabla f(x)^T(\upsilon - x) + 2\alpha\varepsilon^T(\upsilon - x) + \alpha^2 \|\nabla f(x) + \varepsilon\|^2. \tag{2.11}$$

We plug-in the equalities (2.10) and (2.11) into (2.9) and we get

$$\|y-\bar{x}\|^2 \leq \|x-\bar{x}\|^2 + 2\alpha\left(f(\bar{x})-f(x)\right) + 2\alpha\varepsilon^T(\bar{x}-x) - \|x-\upsilon\|^2 + 2\alpha\nabla f(x)^T(x-\upsilon)$$
$$+ 2\alpha\varepsilon^T(x-\upsilon) + \left(\beta^2-2\beta\right)\frac{g^+(\upsilon)^2}{\|d\|^2} \quad \text{for all } \bar{x}\in X.$$

□

The term $\beta\dfrac{g^+(\upsilon)}{\|d\|^2}$ in Lemma 4 can be interpreted as Polyak's stepsize as it is proposed in Polyak (1987) (chapter 5, page 142) for pure feasibility problems with the minimal value of the function known that is $g^+(\upsilon^*)$ being zero. Polyak's stepsize in general form is given by

$$\gamma_k = \beta\frac{f(x_k)-f^*}{\|\partial f(x_k)\|^2}$$

where β is bounded away from zero and 2, which ensures convergence with the rate of geometric progression for classical gradient/subgradient method. Even if $f^*=0$ is not known an estimate of it can be used and updated at each iteration point.

Possible function with known optimal value of $f^*=0$ can be a minimization of the function

$$f(x)=\sum_{i=1}^{n}\left|(a^i,x)-b_i\right|$$

for the system of compatible linear equations $\quad (a^i,x)=b_i, \quad i=1,\ldots,n, \quad x\in\mathbb{R}^n.$

But it is important to note that ratio of progression gets close to linear rate if the problem is ill-posed.

Next lemma provides another auxiliary relation for further use. It is built for the iterate obtained after one step of the algorithm and two arbitrary points.

Lemma 5. *Let the function f be differentiable over the set X_0 with Lipschitz continous gradients with constant L. Then, we have almost surely*

$$\mathsf{E}\left[\|x_k-\bar{x}\|^2 \mid \mathscr{F}_{k-1}\right] \leq A_{\eta,k}\|x_{k-1}-\bar{x}\|^2 + B_{\eta,k}\|x_{k-1}-z_{k-1}\|^2 + 2\alpha_k\left(f(\bar{x})-f(z_{k-1})\right) - \|x_{k-1}-\upsilon_k\|^2$$
$$+ (8+8\eta)\alpha_k^2\|\nabla f(\bar{x})\|^2 + 4\alpha_k^2\nu_k + \frac{\left(\beta^2-2\beta\right)}{C_g^2}\mathsf{E}\left[g^+_{\omega_k}(\upsilon_k)^2 \mid \mathscr{F}_{k-1}\right]$$

for all $\bar{x}\in X$ and $k\geq 1$, where $A_{\eta,k}=1+8\alpha_k^2L^2(1+\eta)$, $B_{\eta,k}=\alpha_kL+\dfrac{1}{4\eta}$ and $\eta>0$ is arbitrary.

Proof. We start with the result that we had in Lemma 4 for differentiable functions,

$$\begin{aligned}\|y-\bar{x}\|^2 \leq \|x-\bar{x}\|^2 + 2\alpha\left(f(\bar{x})-f(x)\right) + 2\alpha\varepsilon^T(\bar{x}-x) - \|x-\upsilon\|^2 + 2\alpha\nabla f(x)^T(x-\upsilon) \\ + 2\alpha\varepsilon^T(x-\upsilon) + \left(\beta^2-2\beta\right)\frac{g^+(\upsilon)^2}{\|d\|^2} \quad \text{for all } \bar{x}\in X,\end{aligned} \tag{2.12}$$

where $\upsilon = \Pi_{X_0}[x-\alpha(\nabla f(x)+\varepsilon)]$.

In order to estimate $f(\bar{x})-f(x)$ we represent the term as shown below

$$f(\bar{x})-f(x) = f(\bar{x})-f(z)+f(z)-f(x) \quad \text{for all } z\in X_0.$$

If $\bar{x}$ is the optimal solution and $z=\Pi_{X_0}[x]$ is any feasible solution of the problem then

$$f(z) \leq f(x)+\nabla f(x)^T(z-x)+\frac{L}{2}\|x-z\|^2 = f(x)+\left(\nabla f(x)-\nabla f(\bar{x})+\nabla f(\bar{x})\right)^T(z-x)+\frac{L}{2}\|x-z\|^2.$$

Using Cauchy–Schwarz and triangle inequalities we get

$$f(z) \leq f(x)+\|\nabla f(x)-\nabla f(\bar{x})\|\,\|x-z\|+\|\nabla f(\bar{x})\|\,\|x-z\|+\frac{L}{2}\|x-z\|^2.$$

Then using Lipschitz condition $2\alpha\left(f(\bar{x})-f(x)\right)$ can be estimated as

$$2\alpha\left(f(\bar{x})-f(x)\right) \leq 2\alpha L\|x-\bar{x}\|\,\|x-z\|+2\alpha\|\nabla f(\bar{x})\|\,\|x-z\|+\alpha L\|x-z\|^2+2\alpha\left(f(\bar{x})-f(z)\right). \tag{2.13}$$

For the first and second terms of inequality (2.13) we use the relation $2|a||b| \leq \beta|a|^2+\frac{1}{\beta}|b|^2$, where $\beta = 8\eta$ and $\eta>0$ is arbitrary, then we get

$$2\alpha L\|x-\bar{x}\|\,\|x-z\| \leq 8\eta\alpha^2L^2\|x-\bar{x}\|^2+\frac{1}{8\eta}\|x-z\|^2, \tag{2.14}$$

$$2\alpha\|\nabla f(\bar{x})\|\|x-z\| \leq 8\eta\alpha^2\|\nabla f(\bar{x})\|^2+\frac{1}{8\eta}\|x-z\|^2. \tag{2.15}$$

The term $2\alpha\nabla f(x)^T(x-\upsilon)$ of (2.12) can also be estimated as

$$\begin{aligned} 2\alpha\nabla f(x)^T(x-\upsilon) &\leq 2\alpha\|\nabla f(x)\|\|x-\upsilon\| \leq 2\alpha\|\nabla f(x)\|\|\vartheta-x\| = 2\alpha^2\|\nabla f(x)\|\|\nabla f(x)+\varepsilon\| \\ &\leq 2\alpha^2\|\nabla f(x)\|^2 + 2\alpha^2\|\nabla f(x)\|\|\varepsilon\| \leq 3\alpha^2\|\nabla f(x)\|^2 + \alpha^2\|\varepsilon\|^2, \end{aligned} \tag{2.16}$$

where we use $2|a||b| \leq |a|^2+|b|^2$ and $\vartheta = x-\alpha(\nabla f(x)+\varepsilon)$.

The term $\|\nabla f(x)\|^2$ can be estimated using Lipschitz property and Minkowski inequality $\|f+g\|^p \leq 2^{p-1}(\|f\|^p+\|g\|^p)$ for $1 \leq p \leq \infty$ as follows

$$\|\nabla f(x)\|^2 = \|\nabla f(x)-\nabla f(\bar{x})+\nabla f(\bar{x})\|^2 \leq 2L\|x-\bar{x}\|^2 + 2\|\nabla f(\bar{x})\|^2. \tag{2.17}$$

The term $2\alpha\varepsilon^T(x-\upsilon)$ with $\vartheta = x-\alpha(\nabla f(x)+\varepsilon)$ and $\upsilon = \Pi_{X_0}[x-\alpha(\nabla f(x)+\varepsilon)]$ of (2.12) can also be estimated as

$$\begin{aligned} 2\alpha\varepsilon^T(x-\upsilon) &\leq 2\alpha\|\varepsilon\|\|\upsilon-x\| \leq 2\alpha\|\varepsilon\|\|\vartheta-x\| = 2\alpha^2\|\varepsilon\|\|\nabla f(x)+\varepsilon\| \\ &\leq 2\alpha^2\|\varepsilon\|\|\nabla f(x)\| + 2\alpha^2\|\varepsilon\|^2 \leq 3\alpha^2\|\varepsilon\|^2 + \alpha^2\|\nabla f(x)\|^2 \\ &\leq 3\alpha^2\|\varepsilon\|^2 + 2\alpha^2L\|x-\bar{x}\|^2 + 2\alpha^2\|\nabla f(\bar{x})\|^2, \end{aligned} \tag{2.18}$$

where we use relation $2|a||b| \leq |a|^2+|b|^2$.

We use above bounds while arranging the terms within (2.12) and we get

$$\begin{aligned} \|y-\bar{x}\|^2 \leq{}& A_\eta\|x-\bar{x}\|^2 + B_\eta\|x-z\|^2 + 2\alpha(f(\bar{x})-f(z)) + 2\alpha\varepsilon^T(\bar{x}-x) - \|x-\upsilon\|^2 \\ &+ (8+8\eta)\alpha^2\|\nabla f(\bar{x})\|^2 + 4\alpha^2\|\varepsilon\|^2 + (\beta^2-2\beta)\frac{g^+(\upsilon)^2}{\|d\|^2} \quad \text{for all } \bar{x}\in X, \end{aligned}$$

where $A_\eta = 1+8\eta\alpha^2L^2+8\alpha^2L$, $B_\eta = \alpha L + \dfrac{1}{4\eta}$ and $\upsilon = \Pi_{X_0}[x-\alpha(\nabla f(x)+\varepsilon)]$.

We use the definition of the iterate $\{x_k\}$ in (2.2) and the following identifications: $y=x_k$, $\upsilon=\upsilon_k$, $x=x_{k-1}$, $\varepsilon=\varepsilon_k$, $\alpha=\alpha_k$, $z=\Pi_{\omega_k}[x_{k-1}]$, $g^+(\upsilon)=g^+_{\omega_k}(\upsilon)$, and $d_k \in \partial g^+_{\omega_k}(\upsilon_k)$ and we get

$$\begin{aligned} \|x_k-\bar{x}\|^2 \leq{}& A_\eta\|x_{k-1}-\bar{x}\|^2 + B_\eta\|x_{k-1}-z_{k-1}\|^2 + 2\alpha_k(f(\bar{x})-f(z_{k-1})) + 2\alpha_k\varepsilon_k^T(\bar{x}-x_{k-1}) - \|x_{k-1}-\upsilon_k\|^2 \\ &+ (8+8\eta)\alpha_k^2\|\nabla f(\bar{x})\|^2 + 4\alpha_k^2\|\varepsilon_k\|^2 + (\beta^2-2\beta)\frac{g^+_{\omega_k}(\upsilon_k)^2}{C_g^2} \quad \text{for all } \bar{x}\in X \text{ and } k\geq 1. \end{aligned}$$

where we also use $\|d_k\|^2 \leq C_g^2$.

The constraint sample path until time $k-1$ was revealed, therefore taking expectation conditioned on $\mathscr{F}_{k-1}$ requires only to find expectation of terms that belong to time k, since x_{k-1} and z_{k-1} are fully determined. Hence we have almost surely

$$\begin{aligned}\mathsf{E}\left[\|x_k-\bar{x}\|^2 \mid \mathscr{F}_{k-1}\right] \leq & A_\eta \|x_{k-1}-\bar{x}\|^2 + B_\eta \|x_{k-1}-z_{k-1}\|^2 + 2\alpha_k \left(f(\bar{x})-f(z_{k-1})\right) \\ & + 2\alpha_k \mathsf{E}\left[\varepsilon_k \mid \mathscr{F}_{k-1}\right]^T (\bar{x}-x_{k-1}) - \|x_{k-1}-\upsilon_k\|^2 + (8+8\eta)\alpha_k^2 \|\nabla f(\bar{x})\|^2 \\ & + 4\alpha_k^2 \mathsf{E}\left[\|\varepsilon_k\|^2 \mid \mathscr{F}_{k-1}\right] + \frac{\left(\beta^2-2\beta\right)}{C_g^2} \mathsf{E}\left[g_{\omega_k}^+(\upsilon_k)^2 \mid \mathscr{F}_{k-1}\right].\end{aligned}$$

For the terms that are related to noise we take into account Assumption 3 and the above relation reduces into following inequality,

$$\begin{aligned}\mathsf{E}\left[\|x_k-\bar{x}\|^2 \mid \mathscr{F}_{k-1}\right] \leq & A_\eta \|x_{k-1}-\bar{x}\|^2 + B_\eta \|x_{k-1}-z_{k-1}\|^2 + 2\alpha_k \left(f(\bar{x})-f(z_{k-1})\right) - \|x_{k-1}-\upsilon_k\|^2 \\ & + (8+8\eta)\alpha_k^2 \|\nabla f(\bar{x})\|^2 + 4\alpha_k^2 \nu_k + \frac{\left(\beta^2-2\beta\right)}{C_g^2} \mathsf{E}\left[g_{\omega_k}^+(\upsilon_k)^2 \mid \mathscr{F}_{k-1}\right].\end{aligned}$$

for all $\bar{x} \in X$ and $k \geq 1$, where $A_{\eta,k} = 1+8\alpha_k^2 L^2(1+\eta)$, $B_{\eta,k} = \alpha_k L + \dfrac{1}{4\eta}$ and $\eta > 0$ is arbitrary. □

The following proposition is an auxiliary result to be used for the convergence analysis of algorithm with various types of stepsizes and differentiable objective function.

Proposition 3. *Let the function f have Lipschitz gradients over the set X_0. Assume that problem* (2.1) *has a nonempty optimal set X^*. Then, the iterates $\{x_k\}$ and any point x^* in the optimal set X^* generated by method* (2.2) *satisfy the following relation almost surely*

$$\begin{aligned}\mathsf{E}\left[\|x_k-x^*\|^2 \mid \mathscr{F}_{k-1}\right] \leq & \left(1+\alpha_k^2 L^2 A_{\tau,\eta}\right)\|x_{k-1}-x^*\|^2 + \frac{\left(\beta^2-2\beta\right)}{2cC_g^2}\|x_{k-1}-z_{k-1}\|^2 + 2\alpha_k\left(f(x^*)-f(z_{k-1})\right) \\ & - \|x_{k-1}-\upsilon_k\|^2 + A_{\tau,\eta}\alpha_k^2\|\nabla f(x^*)\|^2 + D_\tau \alpha_k^2 \nu_k \quad \textit{for all } k \geq \tilde{k},\end{aligned}$$

where $\alpha_{\tilde{k}} L \leq \dfrac{1}{4\eta}$, $z_{k-1} = \Pi_X[x_{k-1}]$, $A_{\eta,\tau} = 8+8\eta-8\left(\beta^2-2\beta\right)$, $D_\tau = 4-4\left(\beta^2-2\beta\right)$, $\tau = 4$ *and* $\eta = \dfrac{2cC_g^2}{(2\beta-\beta^2)}$.

Proof. The relation that is proposed in Lemma 5 constitutes the origin of this propositon,

$$\mathsf{E}\left[\|x_k-\bar{x}\|^2 \mid \mathscr{F}_{\mathrm{k-1}}\right] \leq A_{\eta,k}\|x_{k-1}-\bar{x}\|^2 + B_{\eta,k}\|x_{k-1}-z_{k-1}\|^2 + 2\alpha_k\left(f(\bar{x})-f(z_{k-1})\right) - \|x_{k-1}-\upsilon_k\|^2 + (8+8\eta)\,\alpha_k^2\|\nabla f(\bar{x})\|^2 + 4\alpha_k^2\nu_k + \frac{(\beta^2-2\beta)}{C_g^2}\mathsf{E}\left[g_{\omega_k}^+(\upsilon_k)^2 \mid \mathscr{F}_{\mathrm{k-1}}\right] \tag{2.19}$$

for all $\bar{x}\in X$ and $k\geq 1$, where $A_{\eta,k}=1+8\alpha_k^2L^2(1+\eta)$, $B_{\eta,k}=\alpha_kL+\dfrac{1}{4\eta}$ and $\eta>0$ is arbitrary.

We are going to start by relating the residual of intermittent point of k^{th} iterate to the residual of $(k-1)^{th}$ iterate as follows

$$\begin{aligned}\left(g_{\omega_k}^+(\upsilon_k)\right)^2 &= \left(\left(g_{\omega_k}^+(\upsilon_k)-g_{\omega_k}^+(x_{k-1})\right)+g_{\omega_k}^+(x_{k-1})\right)^2\\ &\geq 2\left(g_{\omega_k}^+(\upsilon_k)-g_{\omega_k}^+(x_{k-1})\right)g_{\omega_k}^+(x_{k-1})+\left(g_{\omega_k}^+(x_{k-1})\right)^2\\ &\geq -2\left|g_{\omega_k}^+(\upsilon_k)-g_{\omega_k}^+(x_{k-1})\right|g_{\omega_k}^+(x_{k-1})+\left(g_{\omega_k}^+(x_{k-1})\right)^2.\end{aligned}$$

Based on the subgradient boundedness assumption, and due to the fact that $x_{k-1},\upsilon_k\in X_0$ we have

$$\begin{aligned}2\left|g_{\omega_k}^+(\upsilon_k)-g_{\omega_k}^+(x_{k-1})\right|g_{\omega_k}^+(x_{k-1}) &\leq 2C_g\|\upsilon_k-x_{k-1}\|\,g_{\omega_k}^+(x_{k-1}) \leq 2C_g\alpha_k\|\nabla f(x_{k-1})+\varepsilon_k\|\,g_{\omega_k}^+(x_{k-1})\\ &\leq \tau C_g^2\alpha_k^2\|\nabla f(x_{k-1})+\varepsilon_k\|^2+\frac{1}{\tau}\left(g_{\omega_k}^+(x_{k-1})\right)^2,\end{aligned}$$

where we use $2|a||b|\leq\tau|a|^2+\dfrac{1}{\tau}|b|^2$ and $\tau\geq 1$ is arbitrary.

By using triangle inequality $\|a+b\|^2\leq(\|a\|+\|b\|)^2$ the relation above can be represented as follows:

$$\begin{aligned}2\left|g_{\omega_k}^+(\upsilon_k)-g_{\omega_k}^+(x_{k-1})\right|g_{\omega_k}^+(x_{k-1}) &\leq \tau C_g^2\alpha_k^2\|(\nabla f(x_{k-1})+\varepsilon_k)\|^2+\frac{1}{\tau}\left(g_{\omega_k}^+(x_{k-1})\right)^2\\ &\leq \tau C_g^2\alpha_k^2\|\nabla f(x_{k-1})\|^2+\tau C_g^2\alpha_k^2\|\varepsilon_k\|^2+\frac{1}{\tau}\left(g_{\omega_k}^+(x_{k-1})\right)^2.\end{aligned} \tag{2.20}$$

By using the relation $\|a\pm b\|^2\leq 2\|a\|^2+2\|b\|^2$ and Lipschitz gradient property we get

$$\|\nabla f(x_{k-1})\|^2\leq 2\|\nabla f(x_{k-1})-\nabla f(\bar{x})\|^2+2\|\nabla f(\bar{x})\|^2\,2L^2\|x_{k-1}-\bar{x}\|^2+2\|\nabla f(\bar{x})\|^2. \tag{2.21}$$

By combining the preceding inequalities (2.20) and (2.21), we have

$$\begin{aligned}\left(g^{+}_{\omega_k}(\upsilon_k)\right)^2 &\geq -2\left|g^{+}_{\omega_k}(\upsilon_k)-g^{+}_{\omega_k}(x_{k-1})\right| g^{+}_{\omega_k}(x_{k-1})+\left(g^{+}_{\omega_k}(x_{k-1})\right)^2.\\ &\geq -2\tau C_g^2\alpha_k^2\left(L^2\|x_{k-1}-\bar{x}\|^2+\|\nabla f(\bar{x})\|^2\right)-\tau C_g^2\alpha_k^2\|\varepsilon_k\|^2+\frac{\tau-1}{\tau}\left(g^{+}_{\omega_k}(x_{k-1})\right)^2.\end{aligned}$$

Due to β having values between $0<\beta<2$, the deterministic coefficient $\beta^2-2\beta$ has a negative value and we have almost surely

$$\begin{aligned}\frac{(\beta^2-2\beta)}{C_g^2}\mathsf{E}\left[g^{+}_{\omega_k}(\upsilon_k)^2\mid\mathscr{F}_{k-1}\right]&\leq -2\tau\alpha_k^2(\beta^2-2\beta)\left(L^2\|x_{k-1}-\bar{x}\|^2+\|\nabla f(\bar{x})\|^2\right)-\tau\alpha_k^2\nu_k(\beta^2-2\beta)\\ &\quad+\frac{\tau-1}{\tau}\frac{(\beta^2-2\beta)}{C_g^2}\mathsf{E}\left[\left(g^{+}_{\omega_k}(x_{k-1})\right)^2\mid\mathscr{F}_{k-1}\right].\end{aligned}$$

Therefore we can use the relation above within inequality (2.19) to obtain almost surely

$$\begin{aligned}\mathsf{E}\left[\|x_k-\bar{x}\|^2\mid\mathscr{F}_{k-1}\right]&\leq\left(1+\alpha_k^2L^2\left(8+8\eta-2\tau(\beta^2-2\beta)\right)\right)\|x_{k-1}-\bar{x}\|^2+\left(\alpha_kL+\frac{1}{4\eta}\right)\|x_{k-1}-z_{k-1}\|^2\\ &\quad+2\alpha_k(f(\bar{x})-f(z_{k-1}))-\|x_{k-1}-\upsilon_k\|^2+\left(8+8\eta-2\tau(\beta^2-2\beta)\right)\alpha_k^2\|\nabla f(\bar{x})\|^2\\ &\quad+\left(4-\tau(\beta^2-2\beta)\right)\alpha_k^2\nu_k+\frac{(\beta^2-2\beta)}{C_g^2}\frac{\tau-1}{\tau}\mathsf{E}\left[\left(g^{+}_{\omega_k}(x_{k-1})\right)^2\mid\mathscr{F}_{k-1}\right]\end{aligned}$$

for all $\bar{x}\in X$ and $k\geq 1$.

Since $z_{k-1}=\Pi_X[x_{k-1}]$, it follows that $\|x_{k-1}-z_{k-1}\|=\text{dist}\,(x_{k-1},X)$. Based on Assumption 2 $\text{dist}^{\,2}(x_{k-1},X)\leq c\mathsf{E}\left[\left(g^{+}_{\omega_k}(x_{k-1})\right)^2\right]$, so we obtain for all $\bar{x}\in X$ and $k\geq 1$,

$$\begin{aligned}\mathsf{E}\left[\|x_k-\bar{x}\|^2\mid\mathscr{F}_{k-1}\right]&\leq\left(1+\alpha_k^2L^2A_{\eta,\tau}\right)\|x_{k-1}-\bar{x}\|^2+B_{\eta,\tau}\text{dist}^2(x_{k-1},X)+2\alpha_k(f(\bar{x})-f(z_{k-1}))\\ &\quad-\|x_{k-1}-\upsilon_k\|^2+A_{\eta,\tau}\alpha_k^2\|\nabla f(\bar{x})\|^2+D_\tau\alpha_k^2\nu_k,\end{aligned}$$

where $A_{\eta,\tau}=8+8\eta-2\tau(\beta^2-2\beta)$, $B_{\eta,\tau}=\alpha_kL+\dfrac{1}{4\eta}+\dfrac{(\beta^2-2\beta)}{cC_g^2}\dfrac{\tau-1}{\tau}$, and $D_\tau=4-\tau(\beta^2-2\beta)$.

Since $\alpha_k\to 0$, by choosing k large enough so that $\alpha_kL\leq\dfrac{1}{4\eta}$, we have $\alpha_kL+\dfrac{1}{4\eta}\leq\dfrac{1}{2\eta}$. Therefore

$$\alpha_kL+\frac{1}{4\eta}+\frac{(\beta^2-2\beta)}{cC_g^2}\frac{\tau-1}{\tau}\leq\frac{1}{2\eta}+\frac{(\beta^2-2\beta)}{cC_g^2}\frac{\tau-1}{\tau}.$$

Since $\tau \geq 1$ and $\eta > 0$ are arbitrary, choosing $\tau = 4$ and $\eta = -2\left(\frac{(\beta^2 - 2\beta)}{cC_g^2}\right)^{-1}$ yields

$$\frac{1}{2\eta} + \frac{(\beta^2 - 2\beta)}{cC_g^2}\frac{\tau - 1}{\tau} = \frac{1}{2}\frac{(\beta^2 - 2\beta)}{cC_g^2}.$$

Therefore, we have almost surely for all $\bar{x} \in X$ and $k \geq \tilde{k}$, where $\tilde{k}$ is large enough,

$$\begin{aligned}\mathsf{E}\left[\|x_k - \bar{x}\|^2 \mid \mathscr{F}_{k-1}\right] \leq \left(1 + \alpha_k^2 L^2 A_{\tau,\eta}\right)\|x_{k-1} - \bar{x}\|^2 + \frac{(\beta^2 - 2\beta)}{2cC_g^2}\text{dist}^{\,2}(x_{k-1}, X) + 2\alpha_k\left(f(\bar{x}) - f(z_{k-1})\right) \\ - \|x_{k-1} - \upsilon_k\|^2 + A_{\tau,\eta}\alpha_k^2\|\nabla f(\bar{x})\|^2 + D_\tau \alpha_k^2 \nu_k,\end{aligned}$$

where $\tau = 4$, and $\eta = -2\left(\frac{(\beta^2 - 2\beta)}{cC_g^2}\right)^{-1}$.

Then once we let $\bar{x} = x^*$ with $x^* \in X^*$ and $\text{dist}^{\,2}(x_{k-1}, X) = \|x_{k-1} - z_{k-1}\|^2$ we obtain almost surely

$$\begin{aligned}\mathsf{E}\left[\|x_k - x^*\|^2 \mid \mathscr{F}_{k-1}\right] \leq \left(1 + \alpha_k^2 L^2 A_{\tau,\eta}\right)\|x_{k-1} - x^*\|^2 + \frac{(\beta^2 - 2\beta)}{2cC_g^2}\|x_{k-1} - z_{k-1}\|^2 + 2\alpha_k\left(f(x^*) - f(z_{k-1})\right) \\ - \|x_{k-1} - \upsilon_k\|^2 + A_{\tau,\eta}\alpha_k^2\|\nabla f(x^*)\|^2 + D_\tau\alpha_k^2\nu_k \quad \text{for all } k \geq \tilde{k}.\end{aligned}$$

□

2.3.2 Almost Sure Convergence Results

In this section, we would like to show that the proposed algorithm converges to the solution set almost surely for not summable but square summable stepsize. The convergence of the method (2.2) for a deterministic diminishing stepsize α_k is established in the next proposition. As it is indicated in the next proposition, the algorithm has almost sure convergence.

Proposition 4. *Let the function f have Lipschitz gradients over the set X_0. Let the stepsize be not summable but square summable such that $\sum_{k=1}^{\infty}\alpha_k = \infty$, $\sum_{k=1}^{\infty}\alpha_k^2 < \infty$ and $\sum_{k=1}^{\infty}\alpha_k^2\nu_k < \infty$. Assume that problem* (2.1) *has a nonempty optimal set X^*. Then the iterates $\{x_k\}$ generated by method* (2.2) *converge almost surely to some random point in the optimal set X^*.*

Proof. We start with the result of Proposition 3

$$\mathsf{E}\left[\|x_k - x^*\|^2 \mid \mathscr{F}_{k-1}\right] \leq \left(1+\alpha_k^2 L^2 A_1\right)\|x_{k-1}-x^*\|^2 + \frac{\left(\beta^2-2\beta\right)}{2cC_g^2}\|x_{k-1}-z_{k-1}\|^2 + 2\alpha_k\left(f(x^*)-f(z_{k-1})\right)$$
$$-\|x_{k-1}-\upsilon_k\|^2 + A_1\alpha_k^2\|\nabla f(x^*)\|^2 + D_1\alpha_k^2\nu_k \quad \text{for all } k \geq \tilde{k}, \tag{2.22}$$

where $A_1 = 8+8\eta-8\left(\beta^2-2\beta\right)$, $D_1 = 4-4\left(\beta^2-2\beta\right)$, $\tau = 4$ and $\eta = -2\dfrac{cC_g^2}{\left(\beta^2-2\beta\right)}$.

The problem (2.1) is convex, therefore the gradient mapping $\nabla f(x^*)$ is singleton over the optimal set that is closed and convex due to Assumption 1, (see Facchinei and Pang (2003), Volume I, Corollary 2.3.7). Additionally, since $z_{k-1} \in X$, we have $f(x^*)-f(z_{k-1}) \leq 0$. Under the assumption $\sum_{k=1}^{\infty}\alpha_k^2 < \infty$, the relation (2.22) satisfies the conditions of Theorem 13. As a result the sequence $\{\|x_k - x^*\|\}$ is convergent almost surely for every $x^* \in X^*$, and

$$\sum_{k=1}^{\infty} 2\alpha_k(f(z_{k-1})-f(x^*)) < \infty, \qquad \sum_{k=1}^{\infty}\|x_{k-1}-z_{k-1}\|^2 < \infty \qquad a.s.$$

The preceding relations and the condition $\sum_{k=1}^{\infty}\alpha_k = \infty$ imply that

$$\liminf_{k\to\infty}(f(z_{k-1})-f(x^*)) = 0 \qquad a.s.,$$
$$\lim_{k\to\infty}\|x_{k-1}-z_{k-1}\| = 0 \qquad a.s.$$

We have already concluded that $\{\|x_k - x^*\|\}$ is convergent *a.s.* for every $x^* \in X^*$. We can further conclude that $\{\|z_k - x^*\|\}$ is also convergent *a.s.* for every $x^* \in X^*$ as well. This inherently implies that the sequence $\{z_k\}$ is *a.s.* bounded and has accumulation points. Additionally taking into account the continuity of f the sequence $\{z_k\}$ has an accumulation point in the set X^* *a.s.* Besides $\{\|z_k - x^*\|\}$ converges *a.s.* for every $x^* \in X^*$. Therefore $\{z_k\}$ converges almost surely to a random point in set X^*. In view of relation (2.22) we reach the conclusion that $\{x_k\}$ converges *a.s.* to a random point in X^*. □

The feasibility step of the method (2.2) includes a stepsize parameter β that was chosen to be constant. We would like to note that one may use a time-varying parameter β_k only if $\sum_{k=1}^{\infty}\beta_k(2-\beta_k) \leq \infty$. In case of time-varying parameter β_k is introduced then $A_{1,k} = 8+\dfrac{16cC_g^2}{2\beta_k-\beta_k^2}+8\left(2\beta_k-\beta_k^2\right)$ and $D_{1,k} = 4+4(2\beta_k-\beta_k^2)$

terms are prone pushing the terms $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ to infinity. When we apply the Supermartingale Convergence Theorem (Theorem 13) to the result of Proposition 3 with time-varying parameter β_k , then we get

$$\mathsf{E}\left[\|x_k - x^*\|^2 \mid \mathscr{F}_{k-1}\right] \leq \left(1 + \alpha_k^2 L^2 A_{1,k}\right) \|x_{k-1} - x^*\|^2 + \frac{\left(\beta_k^2 - 2\beta_k\right)}{2cC_g^2} \|x_{k-1} - z_{k-1}\|^2 + 2\alpha_k \left(f(x^*) - f(z_{k-1})\right)$$
$$- \|x_{k-1} - \upsilon_k\|^2 + A_1 \alpha_k^2 \|\nabla f(x^*)\|^2 + D_{1,k} \alpha_k^2 \nu_k \quad \text{for all } k \geq \tilde{k}, \tag{2.23}$$

where $A_{1,k} = 8 + \dfrac{16cC_g^2}{2\beta_k - \beta_k^2} + 8\left(2\beta_k - \beta_k^2\right)$ and $D_{1,k} = 4 + 4(2\beta_k - \beta_k^2)$ $\tau = 4$ and $\eta = \dfrac{2cC_g^2}{\left(2\beta_k - \beta_k^2\right)}$. If we impose the requirement of β_k being such that $\sum_{k=1}^{\infty} \beta_k (2 - \beta_k) \leq \infty$ in addition to the stepsize be not summable but square summable such that $\sum_{k=1}^{\infty} \alpha_k = \infty$, $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$ and $\sum_{k=1}^{\infty} \alpha_k^2 \nu_k < \infty$ we can conclude that $\{\|x_k - x^*\|\}$ is convergent *a.s.* for every $x^* \in X^*$.

2.3.3 Convergence Analysis and Error Bound

The error bound of the sequence generated by the method (2.2) for a constant stepsize, $\bar{\alpha}$ is established in this section. We are going to start by constructing an auxiliary Lemma.

Lemma 6. *Let the stepsize be such that* $0 < \alpha_k \leq \bar{\alpha}$ *for some scalar* $\bar{\alpha} > 0$ *with* $\bar{\alpha}L \leq \dfrac{1}{4\eta}$. *Assume that problem* (2.1) *has a nonempty optimal set* X^*. *Let* $\{x_k\}$ *be the iterate sequence generated by the method* (2.2)*, and define the weighted averages*

$$\hat{x}_t = \frac{1}{S_t}\sum_{k=1}^{t}\alpha_k x_{k-1} \quad \textit{and} \quad \hat{z}_t = \frac{1}{S_t}\sum_{k=1}^{t}\alpha_k z_{k-1} \quad \textit{with} \quad S_t = \sum_{k=1}^{t}\alpha_k \quad \textit{for any } t \geq 1.$$

Then the bound is

$$\begin{aligned}\mathsf{E}\left[f\left(\hat{z}_t\right)\right] - f^* + \frac{\beta\left(2-\beta\right)}{4\bar{\alpha}cC_g^2}\mathsf{E}\left[\left\|\hat{x}_t - \hat{z}_t\right\|^2\right] \leq \frac{1}{2S_t}\mathsf{E}\left[\operatorname{dist}^2\left(x_0, X^*\right)\right] + L^2\sum_{k=1}^{t}\frac{A\alpha_k^2}{2S_t}\mathsf{E}\left[\left\|x_{k-1} - x^*\right\|^2\right] \\ - \frac{1}{2S_t}\sum_{k=1}^{t}\mathsf{E}\left[\left\|x_{k-1} - \upsilon_k\right\|^2\right] + \frac{\left(2 - 2\left(\beta^2 - 2\beta\right)\right)}{S_t}\sum_{k=1}^{t}\alpha_k^2\nu_k \\ + \frac{1}{2S_t}\left\|\nabla f\left(x^*\right)\right\|^2\sum_{k=1}^{t}A\alpha_k^2.\end{aligned}$$

where $A = 8\left(1 + \eta - \beta^2 + 2\beta\right)$, *and* $\eta = -2\left(\dfrac{\beta^2 - 2\beta}{cC_g^2}\right)^{-1}$.

Proof. When we take the total expectation of Proposition 3 with $x = x^* \in X^*$, we get

$$\begin{aligned}\mathsf{E}\left[\left\|x_k - x^*\right\|^2\right] \leq \left(1 + \alpha_k^2 L^2 A\right)\mathsf{E}\left[\left\|x_{k-1} - x^*\right\|^2\right] + \frac{\left(\beta^2 - 2\beta\right)}{2cC_g^2}\mathsf{E}\left[\left\|x_{k-1} - z_{k-1}\right\|^2\right] \\ + 2\alpha_k\left(f\left(x^*\right) - \mathsf{E}\left[f\left(z_{k-1}\right)\right]\right) - \mathsf{E}\left[\left\|x_{k-1} - \upsilon_k\right\|^2\right] + A\alpha_k^2\left\|\nabla f\left(x^*\right)\right\|^2 \\ + \left(4 - 4\left(\beta^2 - 2\beta\right)\right)\alpha_k^2\nu_k \text{ for all } k \geq 1,\end{aligned}$$

where $A = 8\left(1 + \eta + 2\beta - \beta^2\right)$, $\eta = -2\left(\dfrac{\beta^2 - 2\beta}{cC_g^2}\right)^{-1}$.

For $\alpha_k \leq \bar{\alpha}$, we transform above inequality into

$$\begin{aligned}2\alpha_k\left(\mathsf{E}\left[f\left(z_{k-1}\right)\right] - f^*\right) + \frac{\beta\left(2-\beta\right)}{4\bar{\alpha}cC_g^2}\mathsf{E}\left[\left\|x_{k-1} - z_{k-1}\right\|^2\right] \leq \left(1 + \alpha_k^2 L^2 A\right)\mathsf{E}\left[\left\|x_{k-1} - x^*\right\|^2\right] - \mathsf{E}\left[\left\|x_k - x^*\right\|^2\right] \\ - \mathsf{E}\left[\left\|x_{k-1} - \upsilon_k\right\|^2\right] + \left(4 - 4\left(\beta^2 - 2\beta\right)\right)\alpha_k^2\nu_k + A\alpha_k^2\left\|\nabla f\left(x^*\right)\right\|^2.\end{aligned}$$

As we sum the preceding inequality from $k=1$ to $k=t$ we obtain

$$2\sum_{k=1}^{t}\alpha_k\left(\mathsf{E}\left[f\left(z_{k-1}\right)\right]-f^*\right)+\frac{1}{2\bar{\alpha}}\frac{\beta\left(2-\beta\right)}{cC_g^2}\sum_{k=1}^{t}\alpha_k\mathsf{E}\left[\left\|x_{k-1}-z_{k-1}\right\|^2\right]\leq\sum_{k=1}^{t}\left(1+\alpha_k^2L^2A\right)\mathsf{E}\left[\left\|x_{k-1}-x^*\right\|^2\right]$$
$$-\sum_{k=1}^{t}\mathsf{E}\left[\left\|x_k-x^*\right\|^2\right]-\sum_{k=1}^{t}\mathsf{E}\left[\left\|x_{k-1}-\upsilon_k\right\|^2\right]+\left(4-4\left(\beta^2-2\beta\right)\right)\sum_{k=1}^{t}\alpha_k^2\nu_k+\left\|\nabla f\left(x^*\right)\right\|^2A\sum_{k=1}^{t}\alpha_k^2.$$

Some of the terms cancel out and we get

$$2\sum_{k=1}^{t}\alpha_k\left(\mathsf{E}\left[f\left(z_{k-1}\right)\right]-f^*\right)+\frac{1}{2\bar{\alpha}}\frac{\beta\left(2-\beta\right)}{cC_g^2}\sum_{k=1}^{t}\alpha_k\mathsf{E}\left[\left\|x_{k-1}-z_{k-1}\right\|^2\right]\leq\mathsf{E}\left[\left\|x_0-x^*\right\|^2\right]-\mathsf{E}\left[\left\|x_t-x^*\right\|^2\right]$$
$$+L^2A\sum_{k=1}^{t}\alpha_k^2\mathsf{E}\left[\left\|x_{k-1}-x^*\right\|^2\right]-\sum_{k=1}^{t}\mathsf{E}\left[\left\|x_{k-1}-\upsilon_k\right\|^2\right]+\left(4-4\left(\beta^2-2\beta\right)\right)\sum_{k=1}^{t}\alpha_k^2\nu_k+\left\|\nabla f\left(x^*\right)\right\|^2A\sum_{k=1}^{t}\alpha_k^2.$$

The term $\left\|x_{k-1}-x^*\right\|$ is $\operatorname{dist}\left(x_{k-1},X^*\right)$. For $t\to\infty$ the distance between x_t and x^* are small enough that we can assume $\left\|x_t-x^*\right\|\to 0$.

The term $\sum_{k=1}^{t}\mathsf{E}\left[\left\|x_{k-1}-x^*\right\|^2\right]$ is tightening the bound but it is going to be dropped from this point forward. For $S_t=\sum_{k=1}^{t}\alpha_k$ and dividing the preceding inequality by $2S_t$, we further reach

$$\sum_{k=1}^{t}\frac{\alpha_k}{S_t}\left(\mathsf{E}\left[f\left(z_{k-1}\right)\right]-f^*\right)+\frac{\beta\left(2-\beta\right)}{4\bar{\alpha}cC_g^2}\sum_{k=1}^{t}\frac{\alpha_k}{S_t}\mathsf{E}\left[\left\|x_{k-1}-z_{k-1}\right\|^2\right]\leq\frac{1}{2S_t}\mathsf{E}\left[\operatorname{dist}^2\left(x_0,X^*\right)\right]$$
$$+L^2\sum_{k=1}^{t}\frac{A\alpha_k^2}{2S_t}\mathsf{E}\left[\left\|x_{k-1}-x^*\right\|^2\right]-\frac{1}{2S_t}\sum_{k=1}^{t}\mathsf{E}\left[\left\|x_{k-1}-\upsilon_k\right\|^2\right]+\frac{\left(4-4\left(\beta^2-2\beta\right)\right)}{2S_t}\sum_{k=1}^{t}\alpha_k^2\nu_k$$
$$+\frac{1}{2S_t}\left\|\nabla f\left(x^*\right)\right\|^2\sum_{k=1}^{t}A\alpha_k^2.$$

As it can be observed that the terms $\dfrac{\alpha_k}{\sum_{k=1}^{t}\alpha_k}$, $k=1,\ldots,t$ are convex weights while f and squared norm are convex functions. If we use average values of $\hat{z}=\dfrac{1}{S_t}\sum_{k=1}^{t}\alpha_k z_{k-1}$ and $\hat{x}_t=\dfrac{1}{S_t}\sum_{k=1}^{t}\alpha_k x_{k-1}$ then for any $t\geq 1$,

$$\mathsf{E}\left[f\left(\hat{z}_t\right)\right]-f^*+\frac{\beta\left(2-\beta\right)}{4\bar{\alpha}cC_g^2}\mathsf{E}\left[\left\|\hat{x}_t-\hat{z}_t\right\|^2\right]\leq\frac{1}{2S_t}\mathsf{E}\left[\operatorname{dist}^2\left(x_0,X^*\right)\right]+AL^2\sum_{k=1}^{t}\frac{\alpha_k^2}{2S_t}\mathsf{E}\left[\left\|x_{k-1}-x^*\right\|^2\right]$$
$$-\frac{1}{2S_t}\sum_{k=1}^{t}\mathsf{E}\left[\left\|x_{k-1}-\upsilon_k\right\|^2\right]+\frac{\left(2-2\left(\beta^2-2\beta\right)\right)}{S_t}\sum_{k=1}^{t}\alpha_k^2\nu_k+\frac{1}{2S_t}\left\|\nabla f\left(x^*\right)\right\|^2\frac{1}{2S_t}\sum_{k=1}^{t}A\alpha_k^2.$$

□

2.3.3.1 Constant Stepsize

The next proposition is going to provide error bounds on the performance of the algorithm (2.2) for a constant stepsize using Lemma 6.

Proposition 5. *Assume that problem* (2.1) *has a nonempty optimal set* X^*. *Let* $\{x_k\}$ *be the iterate sequence generated by the method* (2.2). *Also let* $\hat{x}_t = \frac{1}{S_t}\sum_{k=1}^{t}\alpha_k x_{k-1}$ *with* $S_t = \sum_{k=1}^{t}\alpha_k$ *for* $t \geq 1$. *If the stepsize is constant, i.e.,* $\alpha_k = \bar{\alpha}$ *with* $\bar{\alpha}L \leq \frac{1}{4\eta}$, *and the stochastic errors* ε_k *have constant variance, i.e.,* $\nu_k = \bar{\nu}$ *for all k, then we have the following error bound for all* $t \geq 1$,where $\hat{z}_t = \frac{1}{S_t}\sum_{k=1}^{t}\alpha_k z_{k-1}$ *with* $z_{k-1} = \Pi_X\left[x_{k-1}\right]$.

$$0 \leq \mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right] \leq \frac{2cC_g^2}{t\beta(2-\beta)}\mathsf{E}\left[\mathrm{dist}^2(x_0, X^*)\right] + \frac{8t\bar{\alpha}^2\bar{\nu}cC_g^2}{\beta(2-\beta)}(1+\beta(2-\beta)) + \frac{16\bar{\alpha}^2cC_g^2\|\nabla f(x^*)\|^2}{\beta(2-\beta)} + \frac{32\bar{\alpha}^2c^2C_g^4\|\nabla f(x^*)\|^2}{(\beta(2-\beta))^2} + 16\bar{\alpha}^2cC_g^2\|\nabla f(x^*)\|^2$$

and

$$|\mathsf{E}[f(\hat{x}_t)] - f^*| \leq \|\nabla f(x^*)\|\sqrt{\mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right]} + \frac{1}{2t\bar{\alpha}}\mathsf{E}\left[\mathrm{dist}^2(x_0, X^*)\right] + 2\bar{\alpha}\bar{\nu}(1+\beta(2-\beta)) + 4\bar{\alpha}\|\nabla f(x^*)\|^2 D$$

where $D = 1+\beta(2-\beta) + \frac{2cC_g^2}{\beta(2-\beta)}$.

Proof. Let $\alpha_k = \bar{\alpha}$ and $\nu_k = \bar{\nu}$ in Lemma 6 for all $t \geq 1$, $S_t = t\bar{\alpha}$ and using Assumption 1, we get

$$\mathsf{E}[f(\hat{z}_t)] - f^* + \frac{\beta(2-\beta)}{4\bar{\alpha}cC_g^2}\mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right] \leq \frac{1}{2t\bar{\alpha}}\mathsf{E}\left[\mathrm{dist}^2(x_0, X^*)\right] + \frac{AL^2\bar{\alpha}}{2}\sum_{k=1}^{t}\mathsf{E}\left[\|x_{k-1} - x^*\|^2\right] - \frac{1}{2t\bar{\alpha}}\sum_{k=1}^{t}\mathsf{E}\left[\|x_{k-1} - \upsilon_k\|^2\right] + \left(2 - 2\left(\beta^2 - 2\beta\right)\right)\bar{\alpha}\bar{\nu} + \frac{A\bar{\alpha}}{2}\|\nabla f(x^*)\|^2,$$

where $A = 8\left(1+\eta-\beta^2+2\beta\right)$, $\eta = -2\left(\frac{\beta^2-2\beta}{cC_g^2}\right)^{-1}$, $\hat{z}_t = \sum_{k=1}^{t} z_{k-1}$ and $z_{k-1} = \Pi_X\left[x_{k-1}\right]$.

The terms $\frac{AL^2\bar{\alpha}}{2}\sum_{k=1}^{t}\mathsf{E}\left[\|x_{k-1} - x^*\|^2\right]$ and $-\frac{1}{2t\bar{\alpha}}\sum_{k=1}^{t}\mathsf{E}\left[\|x_{k-1} - \upsilon_k\|^2\right]$ cancel each other as number of iterations increases. Since set X is convex and $\hat{z}_t$ is convex combination of $z_{k-1} \in X$, it follows that $\hat{z}_t \in X$.

Therefore $\mathsf{E}[f(\hat{z}_t)] - f^* \geq 0$ which leads to

$$\begin{aligned} 0 \leq \mathsf{E}[f(\hat{z}_t)] - f^* \leq \frac{1}{2t\bar{\alpha}} \mathsf{E}\left[\mathrm{dist}^2(x_0, X^*)\right] + 2\bar{\alpha}\bar{\nu}(1+\beta(2-\beta)) \\ + 4\bar{\alpha}\|\nabla f(x^*)\|^2 \left(1 + \beta(2-\beta) + \frac{2cC_g^2}{\beta(2-\beta)}\right). \end{aligned} \tag{2.24}$$

Also $\frac{\beta(2-\beta)}{4\bar{\alpha}cC_g^2} \mathsf{E}\left[\left\|\hat{x}_t - \hat{\hat{z}}_t\right\|^2\right] \geq 0$ therefore

$$\begin{aligned} 0 \leq \mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right] \leq \frac{2cC_g^2}{t\beta(2-\beta)} \mathsf{E}\left[\mathrm{dist}^2(x_0, X^*)\right] + \frac{8\bar{\alpha}^2\bar{\nu}cC_g^2}{\beta(2-\beta)}(1+\beta(2-\beta)) + \frac{16\bar{\alpha}^2cC_g^2\|\nabla f(x^*)\|^2}{\beta(2-\beta)} \\ + \frac{32\bar{\alpha}^2c^2C_g^4\|\nabla f(x^*)\|^2}{(\beta(2-\beta))^2} + 16\bar{\alpha}^2cC_g^2\|\nabla f(x^*)\|^2. \end{aligned}$$

Now we are going to prove the second part of proposition $|\mathsf{E}[f(\hat{x}_t)] - f^*|$. We can express $|\mathsf{E}[f(\hat{x}_t)] - f^*|$ using Jensen's inequality as follows

$$|\mathsf{E}[f(\hat{x}_t)] - f^*| \leq |\mathsf{E}[f(\hat{x}_t)] - \mathsf{E}[f(\hat{z}_t)]| + \mathsf{E}[f(\hat{z}_t)] - f^* \leq |\mathsf{E}[f(\hat{x}_t) - f(\hat{z}_t)]| + \mathsf{E}[f(\hat{z}_t)] - f^*.$$

Since $\hat{x}_t, \hat{z}_t \in X_0$ for all $t \geq 1$, we have

$$|\mathsf{E}[f(\hat{x}_t) - f(\hat{z}_t)]| \leq \|\nabla f(x^*)\| \mathsf{E}[\|\hat{x}_t - \hat{z}_t\|] \quad \text{for all } t \geq 0.$$

When we combine the last two inequalities, they yield

$$|\mathsf{E}[f(\hat{x}_t)] - f^*| \leq \|\nabla f(x^*)\| \mathsf{E}[\|\hat{x}_t - \hat{z}_t\|] + \mathsf{E}[f(\hat{z}_t)] - f^* \quad \text{for all } t \geq 0. \tag{2.25}$$

Once more we use Jensen's Inequality and we get

$$\sqrt{(\mathsf{E}[\|\hat{x}_t - \hat{z}_t\|])^2} \leq \sqrt{\mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right]}.$$

Then we use the preceding relation within inequality (2.25) and we obtain

$$|\mathsf{E}[f(\hat{x}_t)] - f^*| \leq \|\nabla f(x^*)\| \sqrt{\mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right]} + \mathsf{E}[f(\hat{z}_t)] - f^* \quad \text{for all } t \geq 0. \tag{2.26}$$

When we combine (2.24) and (2.26) we get

$$\begin{aligned} |\mathsf{E}\left[f\left(\hat{x}_t\right)\right]-f^*| \leq \|\nabla f\left(x^*\right)\| \sqrt{\mathsf{E}\left[\|\hat{x}_t-\hat{z}_t\|^2\right]}+\frac{1}{2t\bar{\alpha}}\mathsf{E}\left[\operatorname{dist}^2\left(x_0,X^*\right)\right] \\ +2\bar{\alpha}\bar{\nu}\left(1+\beta\left(2-\beta\right)\right)+4\bar{\alpha}\left\|\nabla f\left(x^*\right)\right\|^2 D \end{aligned} \tag{2.27}$$

where $D=1+\beta\left(2-\beta\right)+\dfrac{2cC_g^2}{\beta\left(2-\beta\right)}$.

□

For a fixed stepsize it can be seen that the expected proximity of average iterate to the solution set depends on regularity constant of Assumption 2. Also the initial point affects the bound on the function values of expected average iterate and optimal set, as the number of iterations increases its contribution diminishes.The relation between expected distance between weighted average of iterates, $\hat{x}_t=\dfrac{1}{S_t}\sum_{k=1}^{t}\alpha_k x_{k-1}$ and set X is equal or less than projection of weighted averages of $\hat{z}_t=\dfrac{1}{S_t}\sum_{k=1}^{t}\alpha_k z_{k-1}$, i.e. $\mathsf{E}\left[\operatorname{dist}^2\left(\hat{x}_t,X\right)\right]\leq \mathsf{E}\left[\|\hat{x}_t-\hat{z}_t\|^2\right]$. Thus, Proposition 5 provides error bounds for expected distance between average of iterates and set X with a multiple of constant step size since as number of iterations, $t\to\infty$ than the term $\dfrac{2cC_g^2}{t\beta\left(2-\beta\right)}\mathsf{E}\left[\operatorname{dist}^2\left(x_0,X^*\right)\right]$ vanishes.

2.3.3.2 Nondiminishing Nonsummable Stepsize

Proposition 6. *Assume that problem* (2.1) *has a nonempty optimal set* X^*. *Let* $\{x_k\}$ *be the iterate sequence generated by the method* (2.2). *Also define the weighted averages* $\hat{x}_t = \frac{1}{S_t}\sum_{k=1}^{t}\alpha_k x_{k-1}$, $\hat{z}_t = \frac{1}{S_t}\sum_{k=1}^{t}\alpha_k z_{k-1}$ *with* $S_t = \sum_{k=1}^{t}\alpha_k$ *for* $t \geq 1$. *If the stepsize satisfies* $\lim_{k\to\infty}\alpha_k = \hat{\alpha} \geq 0$ *and* $\sum_{k=1}^{\infty}\alpha_k = \infty$, *then we have the following asymptotic error bounds;*

$$\limsup_{t\to\infty}\mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right] \leq \frac{8B\bar{\alpha}\bar{\nu}\hat{\alpha}cC_g^2}{\beta(2-\beta)} + \frac{16\bar{\alpha}\hat{\alpha}cD\|\nabla f(x^*)\|^2 C_g^2}{\beta(2-\beta)}$$

$$\limsup_{t\to\infty}|\mathsf{E}[f(\hat{x}_t)] - f^*| \leq \|\nabla f(x^*)\| C_g\sqrt{\frac{8B\bar{\alpha}\bar{\nu}\hat{\alpha}cC_g^2}{\beta(2-\beta)} + \frac{16\bar{\alpha}\hat{\alpha}cD\|\nabla f(x^*)\|^2 C_g^2}{\beta(2-\beta)}} + 2B\bar{\nu}\hat{\alpha} + 2D\|\nabla f(x^*)\|^2\hat{\alpha}$$

where $\bar{\alpha} = \max_k \alpha_k$ *and* $\bar{\nu} \geq \max_k \nu_k$.

Proof. $\mathsf{E}[f(\hat{z}_t)] - f^* \geq 0$ for all $t \geq 1$ using Lemma (6)

$$0 \leq \mathsf{E}[f(\hat{z}_t)] - f^* \leq \frac{1}{2S_t}\mathsf{E}\left[\mathrm{dist}^2(x_0, X^*)\right] + \frac{2B}{S_t}\sum_{k=1}^{t}\alpha_k^2\nu_k + \frac{4D\|\nabla f(x^*)\|^2}{S_t}\sum_{k=1}^{t}\alpha_k^2,$$

$$\frac{\beta(2-\beta)}{4\bar{\alpha}cC_g^2}\mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right] \leq \frac{1}{2S_t}\mathsf{E}\left[\mathrm{dist}^2(x_0, X^*)\right] + \frac{2B}{S_t}\sum_{k=1}^{t}\alpha_k^2\nu_k + \frac{4D\|\nabla f(x^*)\|^2}{S_t}\sum_{k=1}^{t}\alpha_k^2,$$

where $B = 1 + \beta(2-\beta)$, $D = 1 + \beta(2-\beta) + \frac{2cC_g^2}{\beta(2-\beta)}$, $\bar{\alpha} = \max_k \alpha_k$ and $\bar{\nu} \geq \max_k \nu_k$.

Since $\lim_{t\to\infty} S_t = \infty$, $\lim_{t\to\infty}\frac{\sum_{k=1}^{t}\alpha_k^2}{S_t} = \hat{\alpha}$, and $\limsup_{t\to\infty}\frac{\sum_{k=1}^{t}\nu_k\alpha_k^2}{\sum_{k=1}^{t}\alpha_k} \leq \bar{\nu}\hat{\alpha}$ by letting $t \to \infty$ and noting that $\bar{z}_t \in X$, we attain

$$0 \leq \limsup_{t\to\infty}\mathsf{E}[f(\hat{z}_t)] - f^* \leq 2B\bar{\nu}\hat{\alpha} + 4D\|\nabla f(x^*)\|^2\hat{\alpha}$$

$$\limsup_{t\to\infty}\mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right] \leq \frac{8B\bar{\alpha}\bar{\nu}\hat{\alpha}cC_g^2}{\beta(2-\beta)} + \frac{16\bar{\alpha}\hat{\alpha}cD\|\nabla f(x^*)\|^2 C_g^2}{\beta(2-\beta)}$$

As it was shown in Proposition 5 we can express $|\mathsf{E}[f(\hat{x}_t)] - f^*|$ as in (2.26)

$$|\mathsf{E}[f(\hat{x}_t)] - f^*| \leq \|\nabla f(x^*)\|\sqrt{\mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right]} + \mathsf{E}[f(\hat{z}_t)] - f^* \quad \text{for all } t \geq 1.$$

For $t \to \infty$ above inequality can be expressed as

$$\limsup_{t\to\infty} |\mathsf{E}[f(\hat{x}_t)] - f^*| \leq \|\nabla f(x^*)\| \limsup_{t\to\infty} \sqrt{\mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right]} + \limsup_{t\to\infty} \mathsf{E}[f(\hat{z}_t)] - f^*.$$

We combine two inequalities above, using (2.27) and we get

$$\begin{aligned}\limsup_{t\to\infty} |\mathsf{E}[f(\hat{x}_t)] - f^*| &\leq \|\nabla f(x^*)\| \limsup_{t\to\infty} \sqrt{\mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right]} + \limsup_{t\to\infty} \mathsf{E}[f(\hat{z}_t)] - f^* \\ &\leq \|\nabla f(x^*)\| C_g \sqrt{\frac{8B\bar{\alpha}\bar{\nu}\hat{\alpha}cC_g^2}{\beta(2-\beta)} + \frac{16\bar{\alpha}\hat{\alpha}cD\|\nabla f(x^*)\|^2 C_g^2}{\beta(2-\beta)}} + 2B\bar{\nu}\hat{\alpha} + 4D\|\nabla f(x^*)\|^2 \hat{\alpha}\end{aligned}$$

□

Expected distance between weighted average of iterates, $\hat{x}_t = \frac{1}{S_t}\sum_{k=1}^{t} \alpha_k x_{k-1}$ and set X is equal or less than projection of weighted averages of $\hat{z}_t = \frac{1}{S_t}\sum_{k=1}^{t} \alpha_k z_{k-1}$, i.e. $\mathsf{E}\left[\text{dist}^2(\hat{x}_t, X)\right] \leq \mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right]$. Thus, Proposition 5 provides error bounds for expected distance between average of iterates and set X with a multiple of constant step size since as number of iterations, $t \to \infty$ than the term $\frac{2cC_g^2}{t\beta(2-\beta)}\mathsf{E}\left[\text{dist}^2(x_0, X^*)\right]$ vanishes. That is an indication of irrelevance of initial point. Accordingly the asymptotic difference between the expected function value at averaged iterates, and the optimal value is also multiple of the stepsize. Another key point to notice is that for the case of constant stepsize and the stochastic errors ε_k with constant variance, as $t \to \infty$ contribution of error term for the bound asymptotic difference between the expected function value at averaged iterates, and the optimal value is insignificant.

When the set X_0 is bounded in Proposition 6 , by substituting D value and using $\mathsf{E}\left[\text{dist}^2(\hat{x}_t, X)\right] \leq \mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right]$ we obtain

$$\mathsf{E}\left[\text{dist}^2(\hat{x}_t, X)\right] \leq \frac{2cC_g^2D^2}{t\beta(2-\beta)} + \frac{8\bar{\alpha}^2\bar{\nu}cC_g^2}{\beta(2-\beta)} + 8c\bar{\alpha}^2\|\nabla f(x^*)\|^2 C_g^2\left(1 + \frac{1}{\beta(2-\beta)} + \frac{cC_g^2}{\beta^2(2-\beta)^2}\right)$$

for all $t \geq 1$, where $D = \max_{x,y\in X_0}\|x-y\|$. When $\beta = 1$ the error bound has its optimal value in terms of β that yields

$$\mathsf{E}\left[\text{dist}^2(\hat{x}_t, X)\right] \leq \frac{2cC_g^2D^2}{t} + 8\bar{\alpha}^2\bar{\nu}cC_g^2 + 8c\bar{\alpha}^2\|\nabla f(x^*)\|^2 C_g^2\left(2 + cC_g^2\right).$$

Therefore, for all $t \geq 1$,

$$\mathsf{E}\left[\operatorname{dist}\left(\hat{x}_t, X\right)\right] \leq \frac{\sqrt{2c}C_g D}{\sqrt{t}} + 2\bar{\alpha}C_g\left(\sqrt{2\bar{\nu}c} + \left\|\nabla f\left(x^*\right)\right\|\sqrt{2c\left(2+cC_g^2\right)}\right).$$

As number of iterations increases the first term diminishes and asymptotic error bound becomes a multiple of stepsize. When $\beta = 1$ asymptotic difference between the expected function value at averaged iterates, and the optimal value is also multiple of the stepsize value $\bar{\alpha}$.

$$\begin{aligned}\left|\mathsf{E}\left[f\left(\hat{x}_t\right)\right] - f^*\right| \leq{}& \frac{\sqrt{2c}\left\|\nabla f\left(x^*\right)\right\| C_g D}{\sqrt{t}} + \frac{D^2}{2t\bar{\alpha}} + \frac{2\bar{\alpha}\bar{\nu}}{t} \\ &+ 2\bar{\alpha}\left\|\nabla f\left(x^*\right)\right\| C_g\left(\sqrt{2\bar{\nu}c} + \left\|\nabla f\left(x^*\right)\right\|\sqrt{2c\left(2+cC_g^2\right)}\right) \\ &+ \bar{\alpha}\left\|\nabla f\left(x^*\right)\right\|^2\left(2 + \frac{2cC_g^2}{\beta\left(2-\beta\right)} + 2\beta\left(2-\beta\right)\right).\end{aligned}$$

Chapter 3

NONSMOOTH STOCHASTIC CONVEX MINIMIZATION: RANDOM PROJECTION ALGORITHM UNDER NOISE

3.1 Introduction

The focus of this chapter is a nonsmooth stochastic convex minimization problem over an arbitrary (possibly infinite) collection of nonempty, closed and convex sets $\{X_i, i \in \mathscr{M}\}$ in $\mathbb{R}^n$. Our objective is to solve the problem by using a two step random subgradient projection algorithm. The stochastic subgradient methods are used to handle the problems as defined above and a variety of robust design or decision problems with uncertain data as in Shor (1998), Polyak (1987). The algorithm we propose is an alternative solution technique for these type of problems.

Stochastic subgradient method is essentialy the ordinary subgradient method with noisy subgradients and constraints. The noise can be due to computation and/or measurement error and error that arises in Monte Carlo evaluation of a function that is defined as an expected value. But the ordinary stochastic subgradient method is prone to the problems of slow convergence. And it requires keeping track of best point among stochastic processes of sequences $\{x^k\}$, and the associated function value f^k_{best} since the ordinary stochastic subgradient method is not monotonically decreasing.

Our algorithm firstly takes a subgradient projection step reaching an intermittent point. The calculated subgradient is uncertain carrying a stochastic error term. Just before the second step of the algorithm one of the constraint set is revealed or chosen randomly. Then the feasibility violation of intermittent point is remedied using a subgradient projection onto the revealed/chosen set. The proposed algorithm is in essence generating a random path through a subcollection of constraint sets. So our algorithm is suitable to solve nonsmooth convex stochastic optimization problems with random objective and constraints.

This chapter is dedicated to show the details of the proposed stochastic random projection subgradient algorithm and its convergence properties.

3.2 Nonsmooth Problem Formulation and Algorithm Description

We would like to focus on the following convex constrained nonsmooth minimization problem for an arbitrary collection $\{X_i, i \in \mathscr{M}\}$ of nonempty, closed and convex sets in $\mathbb{R}^n$,

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in X, \quad X \triangleq X_0 \cap (\cap_{i\in\mathscr{M}} X_i), \\ &\text{with } X_i = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0\} \quad \forall i \in \mathscr{M}. \end{aligned} \tag{3.1}$$

The algorithm is essentially same as (2.2) presented in section 2.2 except we calculate next iterate value, x_k using subgradient projection of x_{k-1} firstly and then the feasibility violation of υ_k is corrected at the next step. Computing both the intermittent point υ_k and the new iterate x_k involve the projection on the set X_0. The iterate process is given by

$$\begin{aligned} \upsilon_k &= \Pi_{X_0}\left[x_{k-1} - \alpha_k\left(s_f(x_{k-1}) + \varepsilon_k\right)\right] \\ x_k &= \Pi_{X_0}\left[\upsilon_k - \beta \frac{g^+_{\omega_k}(\upsilon_k)}{\|d_k\|^2} d_k\right] \quad \text{for all } k \geq 1, \end{aligned} \tag{3.2}$$

where $s_f(x_{k-1}) \in \partial f(x_{k-1})$ and $d_k \in \partial g^+_{\omega_k}(\upsilon_k)$. The scalar $\alpha_k > 0$ is a deterministic stepsize and β is also a deterministic parameter with $0 < \beta < 2$. The initial point $x_0 \in X_0$ is selected randomly with an arbitrary distribution. The absolute random noise ε_k can be interpreted as the stochastic error associated with the evaluation of the subgradient $\partial f(x)$ at $x = x_{k-1}$.

A subgradient $s_f(x)$ of a convex function f at $x = \hat{x} \in \mathrm{dom} f$ satisfies the following relation as it is shown in Polyak (1987) (page 127)

$$f(\hat{x}) \geq f(x) + s_f(x)^T(\hat{x} - x) \quad \text{for all } x \in \mathrm{dom} f. \tag{3.3}$$

The assumptions for global error bound Assumption, 2, and stochastic errors, Assumption 3, that are used throughout this chapter are same as in section 2.2. But uniformly bounded subgradients assumption for random projection algorithm under noise for nonsmooth objective functions is different than the one in section 2.2 and it is presented below.

Assumption 4. *The functions f and every g_i are defined and convex over some open set that contains the set X_0. The subgradients $s_f(x)$ and $s_{g_i}(x)$ are uniformly bounded over the set X_0,*

$$\left\|s_f(x))\right\| \leq C_f, \qquad \|s_{g_i}(x)\| \leq C_g \quad \textit{for all } x \in X_0 \textit{ and } \forall i \in \mathscr{M}, \tag{3.4}$$

where C_f and C_g are positive scalars.

The function f is defined and convex over an open set that contains set X_0, therefore the subdifferential set $\partial f(x)$ is nonempty for all $x \in X_0$, hence the method is well defined. We assume the subgradient norms of f are uniformly bounded over X_0 for some positive scalar C_f, which is equivalent to f being Lipschitz continuous with constant C_f, as follows

$$\begin{aligned} &|f(x) - f(y)| \leq C_f \|x - y\| \quad \text{for all } x, y, \\ &\left\|s_f(x)\right\| \leq C_f \quad \text{for all } s_f(x) \in \partial f(x) \text{ and } x \in X_0. \end{aligned} \tag{3.5}$$

The Lipschitz continuity is essentially asserting that the function in question has a bounded slope which in consequentially leads to subdifferentials of Lipschitz continuous functions being bounded sets as it is proved in Vinter (2010) (proposition 4.7.1, page154) and which is presented below.

Proposition 7. *For a lower semicontinuous function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, and a point $x \in \mathbb{R}^n$, assume that f is Lipschitz continuous on a neighborhood of x with Lipschitz constant L. Then*

- *$\partial f(x)$ is nonempty and $\partial f(x) \subset LB$,*
- *$\partial^\infty f(x) = 0$*

This concludes the assumptions further used in this chapter for the proposed algorithm to solve the optimization problem (3.1).

Algorithm can be modified when projections on individual sets, X_i are available in closed form. We let $s_f(x)$ to denote a subgradient of f at x, and $\partial f(x)$ to denote the set of all subgradients of f at x. Then the algorithm takes the form below,

$$x_k = \Pi_{X_0}\left[\Pi_{X_{\omega_k}}\left[x_{k-1} - \alpha_k\left(s_f(x_{k-1}) + \varepsilon_k\right)\right]\right] \quad \text{for all } k \geq 1.$$

where

$$X_i = \{x \in \mathbb{R}^n \mid g_i(x) = d(x, X_i) = \inf\{\|x - x_0\| \mid x_0 \in X_i\} \leq 0\} \text{ for any } i \in \mathscr{M}.$$

3.3 Convergence Results for Nonsmooth Objective Function

In this section, we show convergence behavior of method (3.2) for nonsmooth objective function f with various stepsizes.

3.3.1 Preliminary Results

In this subsection, preliminary results to be used in convergence analysis of method (3.2) are presented.

The first result relates the distance between an iterate point and any point in set X.

Lemma 7. *Let X_0 be a closed convex set and y is obtained using algorithm* (3.2) *as follows*

$$y = \Pi_{X_0}\left[\Pi_{X_0}\left[x - \alpha\left(s_f(x) + \varepsilon\right)\right] - \beta\frac{g^+(\upsilon)}{\|d\|^2}d\right],$$

where $s_f(x) \in \partial f(x)$ and $d \in \partial g^+(\upsilon)$. We have

$$\begin{aligned}\|y - \bar{x}\|^2 \leq \|x - \bar{x}\|^2 + 2\alpha\left(f(\bar{x}) - f(x)\right) + 2\alpha\varepsilon^T(\bar{x} - x) - \|x - \upsilon\|^2 + 2\alpha s_f(x)^T(x - \upsilon)\\ + 2\alpha\varepsilon^T(x - \upsilon) + \left(\beta^2 - 2\beta\right)\frac{g^+(\upsilon)^2}{\|d\|^2} \quad \textit{for all } \bar{x} \in X,\end{aligned}$$

where $\vartheta = x - \alpha\left(s_f(x) + \varepsilon\right)$ and $\upsilon = \Pi_{X_0}\left[x - \alpha\left(s_f(x) + \varepsilon\right)\right]$.

Proof. The proof procedure for Lemma 7 follows a similar track of Proof for Lemma 4 except objective function being nonsmooth. Therefore the details of proof is not presented in detail. □

Next lemma is an intermediate step to be used for the convergence analysis. It shows a snapshot of algorithm at a point in time that we have an iterate point and expect to move to the next iterate, which is moving towards an arbitrary point $\bar{x}$ for all $\bar{x} \in X$.

Lemma 8. *Let y be obtained using algorithm* (3.2) *as follows*

$$y = \Pi_{X_0}\left[\Pi_{X_0}\left[x - \alpha\left(s_f(x) + \varepsilon\right)\right] - \beta\frac{g_\omega^+(\upsilon)}{\|d\|^2}d\right],$$

where $\vartheta = x - \alpha\left(s_f(x) + \varepsilon\right)$, $\upsilon = \Pi_{X_0}\left[x - \alpha\left(s_f(x) + \varepsilon\right)\right]$ *and* $s_f \in \partial f$. *Then,*

$$\begin{aligned}\|y-\bar{x}\|^2 \leq \|x-\bar{x}\|^2 + A_\eta\|x-z\|^2 + 2\alpha\varepsilon^T(\bar{x}-x) - \|x-\upsilon\|^2 + \alpha^2\|\varepsilon\|^2 + 2\alpha(f(\bar{x}) - f(z)) \\ + 2\alpha\varepsilon^T(\bar{x}-\upsilon) + \alpha^2 C_f^2(40\eta+3) + \left(\beta^2 - 2\beta\right)\frac{g^+(\upsilon)}{\|d\|^2} \quad \textit{for all } \bar{x} \in X,\end{aligned}$$

where $z = \Pi_X[x]$, $A_\eta = \dfrac{1}{4\eta} + \alpha C_f$ *and* $\eta > 0$ *is arbitrary.*

Proof. The proof follows a similar path as proof of Lemma 5. Therefore it is going to be explained briefly. We start with the result that we had in Lemma 7.

$$\begin{aligned}\|y-\bar{x}\|^2 \leq \|x-\bar{x}\|^2 + 2\alpha\left(f(\bar{x}) - f(x)\right) + 2\alpha\varepsilon^T(\bar{x}-x) - \|x-\upsilon\|^2 + 2\alpha s_f(x)^T(x-\upsilon) \\ + 2\alpha\varepsilon^T(x-\upsilon) + \left(\beta^2 - 2\beta\right)\frac{g^+(\upsilon)^2}{\|d\|^2} \quad \text{for all } \bar{x} \in X.\end{aligned} \tag{3.6}$$

First term to estimate is $\left(f(\bar{x}) - f(x)\right)$. If $\bar{x}$ is the optimal solution and $z = \Pi_X[x]$ is any feasible solution of the problem and using Cauchy-Schwarz and triangle inequalities we have

$$\begin{aligned}f(\bar{x}) - f(x) = f(\bar{x}) - f(z) + f(z) - f(x) &\leq f(\bar{x}) - f(z) + \left(s_f(x) - s_f(\bar{x}) + s_f(\bar{x})\right)^T(z-x) \\ &\leq f(\bar{x}) - f(z) + \left\|s_f(x) - s_f(\bar{x})\right\|\|z-x\| + \left\|s_f(\bar{x})\right\|\|z-x\|.\end{aligned}$$

Then the term $2\alpha\left(f(\bar{x}) - f(x)\right)$ can be denoted as

$$\begin{aligned}2\alpha\left(f(\bar{x}) - f(x)\right) &\leq 2\alpha\left(f(\bar{x}) - f(z)\right) + 2\alpha\left\|s_f(x) - s_f(\bar{x})\right\|\|x-z\| + 2\alpha\left\|s_f(\bar{x})\right\|\|x-z\| \\ &\leq 2\alpha\left(f(\bar{x}) - f(z)\right) + 40\eta\alpha^2 C_f^2 + \frac{1}{4\eta}\|x-z\|^2,\end{aligned}$$

where $2|a||b| \leq \beta|a|^2 + \dfrac{1}{\beta}|b|^2$, $\beta = 8\eta$ and $\eta > 0$ is arbitrary.

The term $2\alpha s_f(x)^T(x-\upsilon)$ of (7) can be estimated due to Cauchy-Schwarz inequality, nonexpansiveness property of projection operation and Minkowski inequality as follows

$$\begin{aligned}2\alpha s_f(x)^T(x-\upsilon) &\leq 2\alpha\left\|s_f(x)\right\|\|x-\upsilon\| \leq 2\alpha\left\|s_f(x)\right\|\|\vartheta - x\| = 2\alpha^2\left\|s_f(x)\right\|\left\|s_f(x) + \varepsilon\right\| \\ &\leq 3\alpha^2\left\|s_f(x)\right\|^2 + 2\alpha^2\|\varepsilon\|^2 \leq 3\alpha^2 C_f^2 + \alpha^2\|\varepsilon\|^2.\end{aligned}$$

Thus we get the final form as

$$\|y-\bar{x}\|^2 \leq \|x-\bar{x}\|^2 + A_\eta \|x-z\|^2 + 2\alpha\varepsilon^T(\bar{x}-x) - \|x-\upsilon\|^2 + \alpha^2\|\varepsilon\|^2 + 2\alpha(f(\bar{x})-f(z)) + 2\alpha\varepsilon^T(\bar{x}-\upsilon) + \alpha^2 C_f^2(40\eta+3) + \left(\beta^2-2\beta\right)\frac{g^+(\upsilon)}{\|d\|^2} \quad \text{for all } \bar{x}\in X,$$

where $A_\eta = \dfrac{1}{4\eta} + \alpha C_f$. □

Next Lemma provides a bound on the expected distance between a point in the solution set and any iterate point of the algorithm run along the path.

Lemma 9. *Let y obtained using algorithm* (3.2) *as follows*

$$y = \Pi_{X_0}\left[\Pi_{X_0}\left[x-\alpha\left(s_f(x)+\varepsilon\right)\right] - \beta\frac{g^+_\omega(\upsilon)}{\|d\|^2}d\right].$$

Then, for $z_{k-1} = \Pi_{X_0}[x_{k-1}]$

$$\mathsf{E}\left[\|x_k-\bar{x}\|^2 \mid \mathscr{F}_{k-1}\right] \leq \|x_{k-1}-\bar{x}\|^2 + A_\eta\|x_{k-1}-z_{k-1}\|^2 - \|x_{k-1}-\upsilon_k\|^2 + \alpha_k^2\upsilon_k + 2\alpha_k(f(\bar{x})-f(z_{k-1})) + \alpha_k^2 C_f^2(40\eta+3) + \frac{\left(\beta^2-2\beta\right)}{C_g^2}\mathsf{E}\left[g^+_{\omega_k}(\upsilon_k)^2 \mid \mathscr{F}_{k-1}\right] \quad \textit{for all } \bar{x}\in X, \textit{ and } k\geq 1,$$

where $\upsilon_k = \Pi_{X_0}\left[x_{k-1}-\alpha_k\left(s_f(x_{k-1})+\varepsilon_k\right)\right]$, $A_\eta = \dfrac{1}{4\eta} + \alpha C_f$ *and* $\eta > 0$ *is arbitrary.*

Proof. We start with the result that we had in Lemma 8

$$\|y-\bar{x}\|^2 \leq \|x-\bar{x}\|^2 + A_\eta\|x-z\|^2 + 2\alpha\varepsilon^T(\bar{x}-x) - \|x-\upsilon\|^2 + \alpha^2\|\varepsilon\|^2 + 2\alpha(f(\bar{x})-f(z)) + 2\alpha\varepsilon^T(\bar{x}-\upsilon) + \alpha^2 C_f^2(40\eta+3) + \left(\beta^2-2\beta\right)\frac{g^+(\upsilon)}{\|d\|^2} \quad \text{for all } \bar{x}\in X,$$

where $A_\eta = \dfrac{1}{4\eta} + \alpha C_f$.

By the uniform boundedness of subgradient $d \in \partial g^+(\upsilon)$ and we have $\|d\|^2 \leq C_g^2$ for all $x \in X_0$ and $i \in M$. It transforms above inequality into

$$\|y-\bar{x}\|^2 \leq \|x-\bar{x}\|^2 + A_\eta\|x-z\|^2 + 2\alpha\varepsilon^T(\bar{x}-x) - \|x-\upsilon\|^2 + \alpha^2\|\varepsilon\|^2 + 2\alpha(f(\bar{x})-f(z)) + 2\alpha\varepsilon^T(\bar{x}-\upsilon) + \alpha^2 C_f^2(40\eta+3) + \left(\beta^2-2\beta\right)\frac{g^+(\upsilon)^2}{C_g^2} \quad \text{for all } \bar{x}\in X \text{ and } k\geq 1.$$

Let $y = x_k$, $\upsilon = \upsilon_k$, $x = x_{k-1}$, $\varepsilon = \varepsilon_k$, $\alpha = \alpha_k$, $z_{k-1} = \Pi_{X_0}[x_{k-1}]$, $g^+(\upsilon) = g^+_{\omega_k}(\upsilon_k)$, and $d_k \in \partial g^+_{\omega_k}(\upsilon_k)$. We take into account Assumption 3 and then expected bound on how far current iterate is away from $\bar{x}$, the optimal solution based on path until time $k-1$ is as follows

$$\begin{aligned}\mathsf{E}\left[\|x_k - \bar{x}\|^2 \mid \mathscr{F}_{\mathrm{k}-1}\right] \leq \|x_{k-1} - \bar{x}\|^2 + A_\eta \|x_{k-1} - z_{k-1}\|^2 - \|x_{k-1} - \upsilon_k\|^2 + \alpha_k^2 \nu_k + 2\alpha_k (f(\bar{x}) - f(z_{k-1})) \\ + \alpha_k^2 C_f^2 (40\eta + 3) + \frac{(\beta^2 - 2\beta)}{C_g^2} \mathsf{E}\left[g^+_{\omega_k}(\upsilon_k)^2 \mid \mathscr{F}_{\mathrm{k}-1}\right] \quad \text{for all } \bar{x} \in X, \text{ and } k \geq 1.\end{aligned}$$

where $\mathsf{E}\left[\|\varepsilon_k\|^2 \mid \mathscr{F}_{k\text{-}1}\right] \leq \nu_k$ and the scalar ν_k is a deterministic constant. □

The result stated in Lemma 9 is going to be used as basis for the following proposition.

Proposition 8. *Let y be given by*

$$y = \Pi_{X_0}\left[\Pi_{X_0}\left[x - \alpha\left(s_f(x) + \varepsilon\right)\right] - \beta \frac{g^+(\upsilon)}{\|d\|^2} d\right],$$

where $s_f(x) \in \partial f(x)$ *and* $d \in \partial g^+(\upsilon)$. *Let Assumptions 2, 3, and 4 hold. Then, for the iterates of the subgradient method* (3.2) *we have for all* $\bar{x} \in X$, $\bar{x} = x^*$ *with* $x^* \in X^*$ *and* $dist^2(x_{k-1}, X) = \|x_{k-1} - z_{k-1}\|^2$,

$$\begin{aligned}\mathsf{E}\left[\|x_k - x^*\|^2 \mid \mathscr{F}_{\mathrm{k}-1}\right] \leq \|x_{k-1} - x^*\|^2 - \|x_{k-1} - \upsilon_k\|^2 + \left(1 + 8\beta - 4\beta^2\right)\alpha_k^2 \nu_k - 2\alpha_k \left(f(z_{k-1}) - f(x^*)\right) \\ + \alpha_k^2 C_f^2 \left(40\eta + 3 + 8\beta - 4\beta^2\right) + \frac{(\beta^2 - 2\beta)}{2cC_g^2} \|x_{k-1} - z_{k-1}\|^2 \quad \textit{for all } k \geq \tilde{k}.\end{aligned}$$

where $\tau = 4$ *and* $\eta = \dfrac{2cC_g^2}{\beta(2-\beta)}$.

Proof. We start with the result that we had in Lemma 9 for non-differentiable functions, f

$$\begin{aligned}\mathsf{E}\left[\|x_k - \bar{x}\|^2 \mid \mathscr{F}_{\mathrm{k}-1}\right] \leq \|x_{k-1} - \bar{x}\|^2 + A_\eta \|x_{k-1} - z_{k-1}\|^2 - \|x_{k-1} - \upsilon_k\|^2 + \alpha_k^2 \nu_k + 2\alpha_k (f(\bar{x}) - f(z_{k-1})) \\ + \alpha_k^2 C_f^2 (40\eta + 3) + \frac{(\beta^2 - 2\beta)}{C_g^2} \mathsf{E}\left[g^+_{\omega_k}(\upsilon_k)^2 \mid \mathscr{F}_{\mathrm{k}-1}\right] \quad \text{for all } \bar{x} \in X, \text{ and } k \geq 1.\end{aligned}$$

To relate residual of $(k-1)^{th}$ iterate and residual of intermittent point of k^{th} iterate, we use

$$\begin{aligned}
\left(g^+_{\omega_k}(\upsilon_k)\right)^2 &= \left(\left(g^+_{\omega_k}(\upsilon_k) - g^+_{\omega_k}(x_{k-1})\right) + g^+_{\omega_k}(x_{k-1})\right)^2 \\
&\geq 2\left(g^+_{\omega_k}(\upsilon_k) - g^+_{\omega_k}(x_{k-1})\right) g^+_{\omega_k}(x_{k-1}) + \left(g^+_{\omega_k}(x_{k-1})\right)^2 \\
&\geq -2\left|g^+_{\omega_k}(\upsilon_k) - g^+_{\omega_k}(x_{k-1})\right| g^+_{\omega_k}(x_{k-1}) + \left(g^+_{\omega_k}(x_{k-1})\right)^2 \\
&\geq -2C_g\alpha_k \left\|s_f(x_{k-1}) + \varepsilon_k\right\| g^+_{\omega_k}(x_{k-1}) + \left(g^+_{\omega_k}(x_{k-1})\right)^2 \\
&\geq -\tau C_g^2\alpha_k^2 \left\|s_f(x_{k-1}) + \varepsilon_k\right\|^2 - \frac{1}{\tau}\left(g^+_{\omega_k}(x_{k-1})\right)^2 + \left(g^+_{\omega_k}(x_{k-1})\right)^2 \\
&\geq -\tau C_g^2\alpha_k^2 C_f^2 - \tau C_g^2\alpha_k^2 \left\|\varepsilon_k\right\|^2 + \frac{\tau-1}{\tau}\left(g^+_{\omega_k}(x_{k-1})\right)^2.
\end{aligned}$$

For each path of constraint realizations the relation above is expected to be

$$\mathsf{E}\left[g^+_{\omega_k}(\upsilon_k)^2 \mid \mathscr{F}_{\mathrm{k-1}}\right] \geq -\tau C_g^2\alpha_k^2 C_f^2 - \tau C_g^2\alpha_k^2\nu_k + \frac{\tau-1}{\tau}\mathsf{E}\left[\left(g^+_{\omega_k}(x_{k-1})\right)^2 \mid \mathscr{F}_{\mathrm{k-1}}\right].$$

Deterministic coefficient $(\beta^2 - 2\beta)$ has a negative value therefore

$$\begin{aligned}
\frac{(\beta^2-2\beta)}{C_g^2}\mathsf{E}\left[g^+_{\omega_k}(\upsilon_k)^2 \mid \mathscr{F}_{\mathrm{k-1}}\right] \leq &-(\beta^2-2\beta)\tau\alpha_k^2 C_f^2 - (\beta^2-2\beta)\tau\alpha_k^2\nu_k \\
&+ \frac{\beta^2-2\beta}{C_g^2}\frac{\tau-1}{\tau}\mathsf{E}\left[\left(g^+_{\omega_k}(x_{k-1})\right)^2 \mid \mathscr{F}_{\mathrm{k-1}}\right].
\end{aligned}$$

Therefore we can use the relation above within Lemma 9

$$\begin{aligned}
\mathsf{E}\left[\|x_k - \bar{x}\|^2 \mid \mathscr{F}_{\mathrm{k-1}}\right] \leq &\|x_{k-1} - \bar{x}\|^2 + A_\eta\|x_{k-1} - z_{k-1}\|^2 - \|x_{k-1} - \upsilon_k\|^2 + \left(1 + (2\beta-\beta^2)\tau\right)\alpha_k^2\nu_k \\
&+ 2\alpha_k(f(\bar{x}) - f(z_{k-1})) + \alpha_k^2 C_f^2\left(40\eta + 3 + (2\beta-\beta^2)\tau\right) \\
&+ \frac{\beta^2-2\beta}{C_g^2}\frac{\tau-1}{\tau}\mathsf{E}\left[\left(g^+_{\omega_k}(x_{k-1})\right)^2 \mid \mathscr{F}_{\mathrm{k-1}}\right].
\end{aligned}$$

The distance between the current iterate and set X is $\|x_{k-1} - z_{k-1}\| = \mathrm{dist}(x_{k-1}, X)$, since $z_{k-1} = \Pi_X[x_{k-1}]$. And due to Assumption 2, we have $\mathrm{dist}^2(x_{k-1}, X) \leq c\mathsf{E}\left[\left(g^+_{\omega_k}(x_{k-1})\right)^2 \mid \mathscr{F}_{k-1}\right]$

$$\begin{aligned}
\mathsf{E}\left[\|x_k - \bar{x}\|^2 \mid \mathscr{F}_{\mathrm{k-1}}\right] \leq &\|x_{k-1} - \bar{x}\|^2 - \|x_{k-1} - \upsilon_k\|^2 + \left(1 + (2\beta-\beta^2)\tau\right)\alpha_k^2\nu_k + 2\alpha_k\left(f(\bar{x}) - f(z_{k-1})\right) \\
&+ \alpha_k^2 C_f^2\left(40\eta + 3 + (2\beta-\beta^2)\tau\right) + \left(\alpha_k C_f + \frac{1}{4\eta} + \frac{\beta^2-2\beta}{cC_g^2}\frac{\tau-1}{\tau}\right)\mathrm{dist}^2(x_{k-1}, X).
\end{aligned}$$

Since $\alpha_k \to 0$, by choosing k large enough so that $\alpha_k C_f \leq \frac{1}{4\eta}$, we have $\alpha_k C_f + \frac{1}{4\eta} \leq \frac{1}{2\eta}$. Choosing $\tau = 4$ and $\eta = -2\left(\frac{\beta^2 - 2\beta}{cC_g^2}\right)^{-1}$ we have

$$\alpha_k C_f + \frac{1}{4\eta} + \frac{\beta^2 - 2\beta}{cC_g^2}\frac{\tau - 1}{\tau} \leq \frac{1}{2\eta} + \frac{(\beta^2 - 2\beta)}{cC_g^2}\frac{\tau - 1}{\tau} = \frac{(\beta^2 - 2\beta)}{2cC_g^2}.$$

We let $\bar{x} = x^*$ with $x^* \in X^*$ and $\text{dist}^2(x_{k-1}, X) = \|x_{k-1} - z_{k-1}\|^2$,

$$\begin{aligned}\mathsf{E}\left[\|x_k - x^*\|^2 \mid \mathscr{F}_{\text{k}-1}\right] \leq \|x_{k-1} - x^*\|^2 - \|x_{k-1} - v_k\|^2 + \left(1 + 8\beta - 4\beta^2\right)\alpha_k^2 v_k - 2\alpha_k\left(f(z_{k-1}) - f(x^*)\right) \\ + \alpha_k^2 C_f^2\left(40\eta + 3 + 8\beta - 4\beta^2\right) + \frac{(\beta^2 - 2\beta)}{2cC_g^2}\|x_{k-1} - z_{k-1}\|^2 \quad \text{for all } k \geq \tilde{k}.\end{aligned}$$

□

3.3.2 Almost Sure Convergence Results for Nonsmooth Case

In this section, we would like to show that the proposed algorithm converges to the solution set almost surely for not summable but square summable stepsize. The convergence of the method (3.2) for a deterministic diminishing stepsize α_k is established in the next proposition. As it is indicated in the next proposition, the algorithm has almost sure convergence.

Proposition 9. *Let the function f is nonsmooth and convex. Let the stepsize be not summable but square summable such that $\sum_{k=1}^{\infty}\alpha_k = \infty$, $\sum_{k=1}^{\infty}\alpha_k^2 < \infty$ and $\sum_{k=1}^{\infty}\alpha_k^2 v_k < \infty$. Assume that problem* (3.1) *has a nonempty optimal set X^*. Then the iterates $\{x_k\}$ generated by method* (3.2) *converge almost surely to some random point in the optimal set X^**

Proof. We start with the result of Proposition 8

$$\begin{aligned}\mathsf{E}\left[\|x_k - x^*\|^2 \mid \mathscr{F}_{\text{k}-1}\right] \leq \|x_{k-1} - x^*\|^2 - \|x_{k-1} - v_k\|^2 + \left(1 + 8\beta - 4\beta^2\right)\alpha_k^2 v_k - 2\alpha_k\left(f(z_{k-1}) - f(x^*)\right) \\ + \alpha_k^2 C_f^2\left(40\eta + 3 + 8\beta - 4\beta^2\right) + \frac{(\beta^2 - 2\beta)}{2cC_g^2}\|x_{k-1} - z_{k-1}\|^2 \quad \text{for all } k \geq \tilde{k}.\end{aligned}$$

where $\tau = 4$ and $\eta = \dfrac{2cC_g^2}{\beta(2-\beta)}$.

Since $z_{k-1} \in X$, we have $f(z_{k-1}) - f(x^*) \geq 0$. Under the assumption $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, the inequality above satisfies the conditions of Theorem 1. As a result the sequence $\{\|x_k - x^*\|\}$ is convergent almost surely for every $x^* \in X^*$, and

$$\sum_{k=1}^{\infty} 2\alpha_k (f(z_{k-1}) - f(x^*)) < \infty, \qquad \sum_{k=1}^{\infty} \|x_{k-1} - z_{k-1}\|^2 < \infty \qquad a.s.$$

The preceding relations and the condition $\sum_{k=1}^{\infty} \alpha_k = \infty$ imply that

$$\liminf_{k\to\infty} (f(z_{k-1}) - f(x^*)) = 0 \qquad a.s., \tag{3.7}$$

$$\lim_{k\to\infty} \|x_{k-1} - z_{k-1}\| = 0 \qquad a.s. \tag{3.8}$$

We have already concluded that $\{\|x_k - x^*\|\}$ is convergent *a.s.* for every $x^* \in X^*$. When we take into account equation (3.8), we can further conclude that $\{\|z_k - x^*\|\}$ is also convergent *a.s.* for every $x^* \in X^*$ as well. This inherently implies that the sequence $\{z_k\}$ is *a.s.* bounded and has accumulation points. Additionally taking into account the continuity of f and relation (3.7), the sequence $\{z_k\}$ has an accumulation point in the set X^* *a.s.* Besides $\{\|z_k - x^*\|\}$ converges *a.s.* for every $x^* \in X^*$. Therefore $\{z_k\}$ converges almost surely to a random point in set X^*. All in all in view of relation (3.8) we reach the conclusion that $\{x_k\}$ converges *a.s.* to a random point in X^*. □

3.3.3 Convergence Analysis and Error Bound for Nonsmooth Case

Next auxiliary lemma is constructed in order to establish error bounds on the performance of the subgradient algorithm (3.2) for constant and nondiminishing stepsizes.

Lemma 10. *Let Assumptions 2, 3 and 4 hold. Let the stepsize be such that* $0 < \alpha_k \leq \bar{\alpha}$ *for some scalar* $\bar{\alpha} > 0$ *and all* $\bar{\alpha}L \leq \dfrac{1}{4\eta}$. *Assume that problem* (3.1) *has a nonempty optimal set* X^*. *Let* $\{x_k\}$ *be the iterate sequence generated by the method* (3.2)*, and define the weighted averages*

$$\hat{x}_t = \frac{1}{S_t} \sum_{k=1}^{t} \alpha_k x_{k-1} \text{ and } \hat{z}_t = \frac{1}{S_t} \sum_{k=1}^{t} \alpha_k z_{k-1} \text{ with } S_t = \sum_{k=1}^{t} \alpha_k,$$

Then, we have for all $t \geq \tilde{k}$,

$$\mathsf{E}\left[f\left(\hat{z}_t\right)\right] - f^* + \frac{2\beta-\beta^2}{4\bar{\alpha}cC_g^2}\mathsf{E}\left[\left\|\hat{x}_t - \hat{z}_t\right\|^2\right] \leq \frac{1}{2S_t}\mathsf{E}\left[dist^2\left(x_{\tilde{k}-1}, X^*\right)\right] - \frac{1}{2S_t}\sum_{k=1}^{t}\mathsf{E}\left[\left\|x_{k-1} - \upsilon_k\right\|^2\right] + \frac{A}{2S_t}\sum_{k=1}^{t}\alpha_k^2\nu_k + \frac{C_f^2 B}{2S_t}\sum_{k=1}^{t}\alpha_k^2,$$

where $A = \left(1+8\beta-4\beta^2\right)$, $B = \left(40\eta+3+8\beta-4\beta^2\right)$, *and* $\eta = \dfrac{2cC_g^2}{\beta\left(2-\beta\right)}$.

Proof. The proof procedure of Lemma 10 follows similar steps of Lemma 6. Therefore the details are not provided. □

3.3.3.1 Constant Stepsize

The error bound of the sequence generated by the method (3.2) for a constant stepsize, $\bar{\alpha}$ is established in this section. The next proposition is going to provide error bounds on the performance of the subgradient algorithm (3.2) by using Lemma 10.

Proposition 10. *Let Assumptions 2, 3 and 4 hold. Let* $\{x_k\}$ *be the iterate sequence generated by the method* (3.2) *and define the weighted averages*

$$\hat{x}_t = \frac{1}{S_t}\sum_{k=1}^{t}\alpha_k x_{k-1} \text{ and } \hat{z}_t = \frac{1}{S_t}\sum_{k=1}^{t}\alpha_k z_{k-1} \text{ with } S_t = \sum_{k=1}^{t}\alpha_k,$$

If the stepsize is constant, i.e., $\alpha_k = \bar{\alpha}$ *for all* $k \geq 1$, *and the stochastic errors* ε_k *has constant variance, i.e.,* $\nu_k = \bar{\nu}$ *for all k then we have the following error bound for all* $t \geq 1$,

$$\mathsf{E}\left[\left\|\hat{x}_t - \hat{z}_t\right\|^2\right] \leq \frac{2cC_g^2}{t\beta\left(2-\beta\right)}\mathsf{E}\left[dist^2\left(x_0, X^*\right)\right] + \frac{2A\bar{\alpha}^2\bar{\nu}cC_g^2}{\beta\left(2-\beta\right)} + \frac{2B\bar{\alpha}^2 cC_f^2 C_g^2}{\beta\left(2-\beta\right)},$$

$$\left|\mathsf{E}\left[f\left(\hat{x}_t\right)\right] - f^*\right| \leq C_f\sqrt{\mathsf{E}\left[\left\|\hat{x}_t - \hat{z}_t\right\|^2\right]} + \frac{1}{2t\bar{\alpha}}\mathsf{E}\left[dist^2\left(x_0, X^*\right)\right] + \frac{A\bar{\alpha}\bar{\nu}}{2} + \frac{B\bar{\alpha}C_f^2}{2},$$

where $A = \left(1+8\beta-4\beta^2\right)$, $B = \left(40\eta+3+8\beta-4\beta^2\right)$, *and* $\eta = \dfrac{cC_g^2}{\beta\left(2-\beta\right)}$

Proof. The proof procedure of Proposition 10 follows a similar track as Proposition 5. □

3.3.3.2 Nondiminishing Nonsummable Stepsize

Proposition 11. *Let Assumptions 2, 3 and 4 hold. Assume that problem* (3.1) *has a nonempty optimal set* X^*. *Let* $\{x_k\}$ *be the iterate sequence generated by the method* (3.2)*, and define the weighted averages*

$$\hat{x}_t = \frac{1}{S_t}\sum_{k=\tilde{k}}^{t} \alpha_k x_{k-1} \text{ and } \hat{z}_t = \frac{1}{S_t}\sum_{k=\tilde{k}}^{t} \alpha_k z_{k-1} \text{ with } S_t = \sum_{k=\tilde{k}}^{t} \alpha_k,$$

If the stepsize satisfies $lim_{k\to\infty}\alpha_k = \hat{\alpha} \geq 0$ *and* $\sum_{k=1}^{\infty} \alpha_k = \infty$, *then we have the following asymptotic error bounds;*

$$\limsup_{t\to\infty} \mathsf{E}\left[\|\hat{x}_t - \hat{z}_t\|^2\right] \leq \frac{2A\bar{\nu}\hat{\alpha}cC_g^2}{\beta(2-\beta)} + \frac{2B\hat{\alpha}cC_f^2C_g^2}{\beta(2-\beta)},$$

$$\limsup_{t\to\infty} |\mathsf{E}[f(\hat{x}_t)] - f^*| \leq C_f C_g \sqrt{\frac{2A\bar{\nu}\hat{\alpha}c}{\beta(2-\beta)} + \frac{2B\hat{\alpha}cC_f^2}{\beta(2-\beta)}} + \frac{A\bar{\nu}\hat{\alpha}}{2} + \frac{B\hat{\alpha}C_f^2}{2},$$

where $\bar{\alpha} = max_k\alpha_k$ *and* $\bar{\nu} \geq max_k\nu_k$, $A = \left(1+8\beta-4\beta^2\right)$, $B = \left(40\eta+3+8\beta-4\beta^2\right)$, *and* $\eta = \dfrac{cC_g^2}{\beta(2-\beta)}$.

Proof. The proof procedure of Proposition11 follows a similar track as Propostion 6. □

Chapter 4

STOCHASTIC RANDOM PROJECTION ALGORITHM: PARAMETER ESTIMATION UNDER BOUNDED DATA UNCERTAINITIES

4.1 Introduction

The algorithm (2.2) is tested on a dynamic control system problem. We study three versions of the problem with correlated unknown-but-bounded additive noise, uncorrelated unknown-but-bounded additive noise and uncorrelated bounded output and peak input additive noise under fully known system description cases. It is essentially a linear least squares estimation problem where we recover state parameters from corrupted input and output data. More specifically, assume $U \in \mathbb{R}^{m\times n}$ is a given full rank input matrix with $m \geq n$ and $y \in \mathbb{R}^m$ is a given output vector. The $\psi \in \mathbb{R}^{m\times n}$ input noise and $\varphi \in \mathbb{R}^m$ output noise terms are additive and belong to bounded sets. The input and output are linearly related via an unknown vector of parameters $h \in \mathbb{R}^n$. Due to noise affecting input and output, a residual, $\mathbf{U}(\psi)h - \mathbf{y}(\varphi)$ emerges. Our aim is to minimize the worst case residual. The problem we solve is

$$\phi(\mathbf{U}(\psi), \mathbf{y}(\varphi), \rho) \triangleq \min\max \|\mathbf{U}(\psi)h - \mathbf{y}(\varphi)\|, \tag{4.1}$$

where error terms belong to one of the following bounded sets;

- Correlated additive input-output

$$\|\psi, \varphi\| \leq \rho \text{ for } \rho \geq 0, \tag{4.2}$$

- Uncorrelated additive input-output

$$\|\psi\| \leq \rho \ \|\varphi\| \leq \rho \text{ for } \rho \geq 0, \tag{4.3}$$

- Uncorrelated additive output-peak input

$$\|\psi\|_\infty \leq \rho, \|\varphi\| \leq \rho \text{ for } \rho \geq 0. \tag{4.4}$$

We reformulated the linear least squares estimation problem as a stochastic convex minimization problem and then used a two step random projection algorithm to solve it. Although the problem has infinite number of constraints due to each realization of error term within bounded set, the algorithm goes through a finite subset of them and converges to the solution set.

There are alternative optimization criteria such as regularized least squares, ridge regression, total least squares and robust estimation that have been proposed to solve least squares estimation problem with errors in data matrix. Regularized least squares and ridge regression methods require to know apriori statistical properties of unobservable random error variables, which is not a viable requirement for practical applications. Total least-squares (TLS) method also known as orthogonal regression or errors-in-variables method allows for data errors besides observation errors on the contrary to standard least-squares (LS) method, Golub and Van Loan (1980). TLS approach produces reasonable solutions only when independent and equally sized errors exist in all data, Van Huffel and Vandewalle (1987). If this prerequisite does not hold for error set, it overemphasize the effect of noise leading to conservative results. On the other hand our projection algorithm randomly chooses a finite number of disturbances and does not require any statistical prior condition or information.

Robust estimation method treats uncertainty as deterministic and describes it in terms of bounded sets. The robust optimization approach which is extensively covered in Ben-Tal et al. (2009) introduces the robust counterparts concept of uncertain problems that requires semi-infinite programming techniques and thus can be intractable even when all instances of the uncertain problem are easy to solve.

The authors of El Ghaoui and Lebret (1997) have formulated and solved a similar estimation problem. The problem (4.12) is referred as structured robust least squares (SRLS) formulation by El Ghaoui and Lebret (1997). They assume that the noise sets are bounded by $\rho = 1$ and (SRLS) problem is defined as follows.

$$\phi_S(\mathbf{A}, \mathbf{b}, \rho) \triangleq \min_x \max_{\|\delta\|=1} \|\mathbf{A}(\delta)x - \mathbf{b}(\delta)\|, \tag{4.5}$$

where

$$\mathbf{A}(\delta) \triangleq A_0 + \sum_{i=1}^{p} \delta_i A_i \quad \mathbf{b}(\delta) \triangleq b_0 + \sum_{i=1}^{p} \delta_i b_i.$$

Initially they solve a one-dimensional convex differentiable function in order to calculate the squared worst-case residual for $\rho = 1$. Then they formulate and solve a semidefinite programming (SDP) problem instead of (4.5). The computational complexity of the approach they used is $O\left(nm^2 + m^3.5\right)$ But it is only applicable for correlated input-output noise case where $\rho = 1$. But in the case of uncorrelated output-peak input noise they present only upper and lower bounds for worst-case residual.

The distinction of our algorithm is that we were able to solve system estimation problem (4.1) for variations of bounded sets (4.2), (4.3), (4.4) with less computational effort. Projecting on only one combination of uncertainity within all possibilities of error set at each iteration provides us the computational convenience.

4.2 Problem Description

In this section we explain the min-max problems ; (4.1)-(4.2), (4.1)-(4.3), (4.1)-(4.4); that we solve using the algorithm (2.2). We also present equivalent minimization formulations or upper bound for these three types of robust least squares problems. Subgradient algorithm solutions are used to gauge the performance of our method. Equivalent formulations are needed for standard subgradient algorithm implementation.

Let input matrix $U \in \mathbb{R}^{m\times n}$, input noise $\psi \in \mathbb{R}^{m\times n}$ with $m \geq n$, output matrix $y \in \mathbb{R}^m$, output noise $\varphi \in \mathbb{R}^m$and $\rho \geq 0$. The input and output are linearly related via an unknown vector of parameters $h \in \mathbb{R}^n$. Due to noise affecting input and output, a residual, $\mathbf{U}(\psi)h - \mathbf{y}(\varphi)$ emerges. Our aim is to minimize the worst case residual.

Problem 1 Correlated Bounded Additive Noise

$$\phi\left(\mathbf{U}(\psi), \mathbf{y}(\varphi), \rho\right) \triangleq \min_{h} \max_{\|\psi,\varphi\|\leq\rho} \|\mathbf{U}(\psi)h - \mathbf{y}(\varphi)\|. \tag{4.6}$$

Problem can be represented as a minimization problem using the operator norm definition as follows

$$\|(\mathbf{U}+\psi)\mathbf{h} - (\mathbf{y}+\varphi)\| = \max_{\|x\|\leq 1} x^T\left(\mathbf{U}\mathbf{h} - \mathbf{y}\right) + x^T\left(\psi\mathbf{h} - \varphi\right).$$

Therefore

$$\max_{\|\psi,\varphi\|\leq\rho} \|(\mathbf{U}+\psi)\,\mathbf{h}-(\mathbf{y}+\varphi)\| = \|\mathbf{U}h-\mathbf{y}\|+\rho\left\|\begin{bmatrix} x \\ 1 \end{bmatrix}\right\|.$$

So the problem (4.6) in minimization form is

$$\min_{h} \|\mathbf{U}h-\mathbf{y}\|+\rho\left\|\begin{bmatrix} x \\ 1 \end{bmatrix}\right\|. \tag{4.7}$$

Problem 2 Uncorrelated Bounded Additive Noise

$$\phi\left(\mathbf{U}(\psi),\mathbf{y}(\varphi),\rho\right) \triangleq \min_{h} \max_{\|\psi\|\leq\rho,\|\varphi\|\leq\rho} \|\mathbf{U}(\psi)\,h-\mathbf{y}(\varphi)\|. \tag{4.8}$$

Equivalent minimization problem formulation for the minimization of worst case problem can be achieved as follows

$$\begin{aligned}\|(\mathbf{U}+\psi)\,\mathbf{h}-(\mathbf{y}+\varphi)\| &\leq \|\mathbf{U}h-\mathbf{y}\|+\|\psi\|\,\|h\|+\|\varphi\| \\ &\leq \|\mathbf{U}h-\mathbf{y}\|+\rho\,\|h\|+\rho.\end{aligned}$$

The upper bound is achieved when the disturbances are

$$\psi_m = \frac{(\mathbf{U}\mathbf{h}-\mathbf{y})}{\|\mathbf{U}\mathbf{h}-\mathbf{y}\|}\frac{\mathbf{h}}{\|\mathbf{h}\|}\rho, \quad \varphi_m = -\frac{(\mathbf{U}\mathbf{h}-\mathbf{y})}{\|\mathbf{U}\mathbf{h}-\mathbf{y}\|}\rho.$$

Therefore we can claim that the problem (4.8) in minimization form is as follows

$$\min_{h} \|\mathbf{U}h-\mathbf{y}\|+\rho\,\|h\|+\rho. \tag{4.9}$$

Problem 3 Uncorrelated Peak Input and Output Bounded Additive Noise

The third case is for uncorrelated bounded output and peak input. The peak disturbance over matrix U is bounded so the set for noise is modified as $\|\psi\|_\infty \leq \rho, \|\varphi\| \leq \rho$

$$\phi\left(\mathbf{U}(\psi),\mathbf{y}(\varphi),\rho\right) \triangleq \min_{h} \max_{\|\psi\|_\infty\leq\rho,\|\varphi\|\leq\rho} \|\mathbf{U}(\psi)\,h-\mathbf{y}(\varphi)\|. \tag{4.10}$$

The norm equivalence relation for matrix $\psi \in \mathbb{R}^{m\times n}$, with $m \geq n$ is

$$\frac{1}{\sqrt{n}} \|\psi\|_\infty \leq \|\psi\| \leq \sqrt{m}\, \|\psi\|_\infty .$$

Therefore we can get an upper bound for the problem (4.10) as follows

$$\begin{aligned} &\min_h \max_{\|\psi\|_\infty \leq \rho, \|\varphi\| \leq \rho} \|\mathbf{U}(\psi) h - \mathbf{y}(\varphi)\| \\ &\leq \min_h \|\mathbf{U}h - \mathbf{y}\| + \sqrt{m}\rho \, \|h\| + \rho . \end{aligned} \tag{4.11}$$

4.3 System Estimation Problem for Discrete-Time Systems with Bounded Noise and Known System Description

4.3.1 Problem Definition

The input/output characteristics of a dynamic system describes how an external input, affects the system output. For linear time-invariant, LTI systems a complete response characterization of the relaxed linear system to any input signal is defined by impulse response. The concept of the impulse response, is a basic time domain characterization of a linear time-invariant system. We seek to estimate the impulse response h, of an LTI system assuming the system is single input, U and single output, y through the convolution equation

$$Uh = y,$$

where

$$h = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ \vdots \\ h_n \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}, \quad U_{n\times n} = \begin{bmatrix} u_1 & 0 & 0 & \dots & 0 \\ u_2 & u_1 & 0 & \dots & 0 \\ u_3 & u_2 & u_1 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_n & u_{n-1} & u_{n-2} & \dots & u_1 \end{bmatrix}.$$

U is a lower-triangular Toeplitz matrix ($U_{i,j} = U_{i-1,j-1}$) whose first column is nominal input vector, u. Assuming U and y are known exactly leads to a linear equation in h, which can be computed with standard least squares methods. In practice, however, both y and U may be subject to measurement and/or process

noise that is deterministic. The actual value of y is $y+\varphi$ and that of U is $U+\psi$, where φ and ψ are unknown-but-bounded perturbations with $\varphi,\psi\in\mathbb{R}^n$. The perturbed matrices for input and output are

$$\mathbf{U}(\psi)=U+\sum_{i=1}^{n}\psi_i U_i,\quad \mathbf{y}(\varphi)=\mathbf{y}+\sum_{i=1}^{n}\varphi_i\mathbf{e}_i,$$

where e_i is the i^{th} column of the $n\times n$ identity matrix and U_i are lower-triangular Toeplitz matrices with first column equal to e_i. We are going to estimate the state of a linear dynamic system with noise-corrupted observations, when input disturbances and observation errors are unknown except for the fact that they belong to given bounded sets.

In signal processing and control theory, Bounded-Input Bounded-Output (BIBO) stability is a form of stability for linear signals and systems that have bounded output for every input to the system that is bounded. BIBO stability is equivalent to p-stability for finite-dimensional LTI state-space systems. A p-stable system is characterized by the requirement that every input of finite p-norm gives rise to an output of finite l-norm. This stability criterion is the basis to the assumption that input and output energies are bounded.

(4.6) and (4.10) versions of robust least squares estimation problem are investigated by El Ghaoui and Lebret (1997) using semidefinite programming (SDP). We use the same nominal values that they used. Ghaoui and Lebret (1997) addresses the first version of the problem (4.6) using a Structured Robust Least Squares, (SRLS) approach that minimizes worst-case residual $r(\mathbf{U}(\psi),\mathbf{y}(\varphi),\rho,\mathbf{h})$. For $\rho\geq 0$ and $h\in\mathbb{R}^n$ the square of structured worst-case residual is defined as

$$r^2(\mathbf{U}(\psi),\mathbf{y}(\varphi),\rho,\mathbf{h})\triangleq\max_{\|\psi,\varphi\|\leq\rho}\|\mathbf{U}(\psi)h-\mathbf{y}(\varphi)\|^2.$$

Minimizing the square of worst case residual is formulated as

$$\phi_S(\mathbf{U}(\psi),\mathbf{y}(\varphi),\rho)\triangleq\min_{h}\max_{\|\psi,\varphi\|\leq\rho}\|\mathbf{U}(\psi)h-\mathbf{y}(\varphi)\|^2. \tag{4.12}$$

Although this problem was solved by El Ghaoui and Lebret (1997) using general purpose SDP solvers, it was mentioned that more efficient special interior-point methods are called for and this direction of research is left as future work. One of the advantages of the proposed algorithm (2.2) is that it is not reliant upon a solver or special interior-point methods.

4.3.2 Existence of Solution and Strongly Convex Nature of Robust System Estimation Problem

The general form of the original least squares problem with bounded noise is

$$\begin{aligned} P_\zeta \qquad & \min_{h\in\mathbb{R}^n} f(h,\zeta) \\ & \text{subject to } \|\zeta\| \le \rho, \end{aligned} \tag{4.13}$$

where $\zeta = (\psi, \varphi) \in \mathbb{R}^n$ is the data matrix that belongs to set

$$\mathscr{U} = \left\{ \zeta = (\psi,\varphi) \mid \left\| \begin{bmatrix} \psi, & \varphi \end{bmatrix} \right\| \le \rho,\ \rho \ge 0 \right\}.$$

The vector h is feasible solution to $P = \left\{P_\zeta\right\}_{\zeta\in\mathscr{U}}$, if it satisfies all possible realizations of perturbation set, Zhou et al. (1995). The robust counterpart problem that minimizes the worst case residual as defined in Ben-Tal et al. (2009) is

$$P^* \qquad \min_{h\in\mathbb{R}^n} \{\sup f(h,\zeta) : \forall \zeta \in \mathscr{U}\}.$$

Before presenting our approach we would like to investigate the existence of solution to robust counterpart problem defined as above. Based on Weierstrass Theorem Bertsekas et al. (2003) if a closed proper function $f : \mathbb{R}^n \to (-\infty, \infty]$ coercive then the set of minima of f over $\mathbb{R}^n$ is nonempty and compact. The sufficiency conditions for existence of solution to robust counterpart problem, P^* is presented as follows. The open form square of $f(h,\zeta)$ i.e. least squares objective is

$$\tilde{f}(h,\zeta) = \begin{bmatrix} \mathbf{h} \end{bmatrix}^T \left[U(\zeta)^T U(\zeta)\right] \begin{bmatrix} \mathbf{h} \end{bmatrix} - \left[2y(\zeta)^T U(\zeta)\right] \begin{bmatrix} \mathbf{h} \end{bmatrix} + y(\zeta)^T y(\zeta) \quad \forall \zeta \in \mathscr{U},$$

where the Hessian is $\left[\nabla^2 \tilde{f}(h,\zeta)\right] = \left[U(\zeta)^T U(\zeta)\right] \succ 0$ for all $h \in \mathbb{R}^n$ for each $\zeta \in \mathscr{U}$ remembering additive nature of perturbations to lower-triangular Toeplitz matrix $U(\zeta)$. $\tilde{f}(h,\zeta)$ function is defined by a unique lower triangular matrix $U(\zeta)$, with strictly positive diagonal elements, that allows the Cholesky decomposition of $M(\zeta) = U(\zeta)^T U(\zeta)$. The matrix U and vector y are input and output of the system originating from nominal values that are positive for a working relaxed system with linearly added bounded perturbations. Impulse response of an LTI system is defined for zero stored energy case. A linear dynamic system is relaxed if it has no stored energy, so that its response to a zero input is a zero output. Implied in the above is the fact that a system will always exhibit the same response to any given input signal when

the stored energy is zero. Since the relaxed physical LTI system can not have negative or zero input when there exists a positive output vector, lower triangular toeplitz matrix U whose first column is nominal input vector u or perturbations added nominal input has always strictly positive diagonal elements that leads to conclusion of $M = U^T U$ and perturbed matrix $M(\zeta) = U(\zeta)^T U(\zeta)$ being symmetric positive definite matrices. Thus $\tilde{f}(h,\zeta) \to \infty$ if and only if we have a case of infinite impulse response (IIR) filters, which have internal feedback and may continue to respond indefinitely which is out of scope of the system defined here. Consequently $\tilde{f}(h,\zeta) < \infty$ for at least one $h \in \mathbb{R}^n$. Hence $\tilde{f}(h,\zeta)$ is a proper function.

The supremum function over the set $\mathscr{U}$, $\sup_{\zeta \in \mathscr{U}} f(h,\zeta)$ is convex since the intersection operation preseves convexity then every $\tilde{f}(h,\zeta)$ which are the outcome of perturbations set $\mathscr{U}$ is convex. Due to a convex function implying continuity over $\mathbb{R}^n$ and quadratic functions being closed, we can conclude that $\sup_{\zeta \in \mathscr{U}} \tilde{f}(h,\zeta)$ is convex and closed Boyd and Vandenberghe (2004), §4.4. $\tilde{f}$ is strongly convex if and only if there exist $\alpha > 0$ such that

$$(\nabla f(x) - \nabla f(y))^T (x-y) \geq \alpha \|x-y\|^2 \quad \forall x, y \in \mathbb{R}^n,$$

Bertsekas et al. (2003). Therefore

$$\left(2\left[U(\zeta)^T U(\zeta)\right]\begin{bmatrix} \mathbf{x} \end{bmatrix} - 2\left[U(\zeta)^T U(\zeta)\right]\begin{bmatrix} \mathbf{y} \end{bmatrix}\right)^T \left(\begin{bmatrix} \mathbf{x} \end{bmatrix} - \begin{bmatrix} \mathbf{y} \end{bmatrix}\right) \geq \alpha \|x-y\|^2$$

leads to the conclusion that if the minimum eigenvalue of $\left[U(\zeta)^T U(\zeta)\right]$ matrix is greater than or equal to any $\frac{\alpha}{2}$ value that is positive then $\tilde{f}(h)$ is a strongly convex function. Ostrowski-Elsner theorem Stewart and Sun (1990), Theorem 1.1 states that λ be an eigenvalue of any matrix A of algebraic multiplicity m then for any norm $\|.\|$ and all sufficiently small $\varepsilon > 0$ there is a $\delta > 0$ such that $\|E\| < \delta$, the disk $\mathscr{D}(\lambda,\varepsilon) = \{\zeta \in \mathbb{C} : |\zeta - \lambda| \leq \varepsilon\}$ contains exactly m eigenvalues of perturbed matrix $\tilde{A}$. The intutiton of Elsner's theorem is that if any m disks are isolated from the others, and then their union contains exactly m eigenvalues of $\tilde{A}$ i.e. eigenvalues of A and $\tilde{A}$ can be grouped into nearby pairs. The Elsner's theorem is for general case. Yet the distance between eigenvalues greatly varies based on the structure of matrices. That is why there are individual results for different classes of matrices one of them is for normal and diagonalizable matrices in Stewart and Sun (1990), §3.1. A normal matrix is any matrix satisfying $A^T A = AA^T$. Any normal matrix can be diagonalized by a unitary transformation, therefore the normal matrices are also included in the category of diagonalizable matrices. Therefore symmetric positive definite matrices $M = \left[U^T U\right]$

and $M(\zeta) = \left[U(\zeta)^T U(\zeta)\right]$ in our system identification problem comply with Hoffman-Wielandt theorem in Stewart and Sun (1990), Theorem 3.1. $M$and $M(\zeta)$ are normal matrices. Then the 2-norm matching distance between eigenvalues of unperturbed matrix M and perturbed matrix $M(\zeta)$ is bounded as

$$md_2(M, M(\zeta)) \leq \|M(\zeta) - M\|_F .$$

The 2-norm matching distance is defined as $md_2(M, M(\zeta)) = \min_{\pi} \sum_i \sqrt{\left|\lambda_{\pi(i)}(\zeta) - \lambda_i\right|^2}$ where π ranges over all permutations of the integers $1, 2, \ldots, n$. Since the perturbation set is compact, we can conclude that there exist $\frac{\alpha}{2} > 0$ smaller than the minimum eigenvalue of any perturbed matrix $M(\zeta)$. At any point on the boundary of a strongly convex set one can associate an enclosing ball with fixed radius R such that x is on the boundary of the ball as well Vial (1982). Therefore $\left\{\sup \tilde{f}(h, \zeta) : \forall \zeta \in \mathscr{U}\right\}$ is strongly convex and as a consequence coercive due to the result stated in Aubin (1998), $\tilde{f}$ is strongly convex then it is coercive. Hence we can conclude that P^* is convex in lieu of the fact that for a convex function $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ over nonempty convex set $C \subseteq \mathbb{R}^n \times \mathbb{R}^p$ the $g(x) = \inf_{z \in C} f(x, z)$ is convex and the set of minima of f over $\mathbb{R}^n$ is nonempty and compact. As a result we proved that M is strongly convex and according to results above $M(\zeta)$ has eigenvalues including the minimum one are within neighborhood of eigenvalues of unperturbed matrix. The disks defined by Elsner theorem contain one eigenvalue each. Since M is real the eigenvalues in the disks must be real and are contained in the intersection of the disks with the real line. Additonally $M(\zeta)$ being symmetric positive definite ensures that in the interval containing minimum eigenvalue is going to be on positive real line whatever the perturbation level is. This ensures strong convexity. Hence, there exist $x_0 = [h_0, \zeta_0] \in \mathbb{R}^n \times \mathscr{U}$ such that

$$f(x_0) = \min_{h \in \mathbb{R}^n} \{\sup f(h, \zeta) : \forall \zeta \in \mathscr{U}\} .$$

4.3.3 Convergence of Algorithm for Strongly Convex Objective Function

We already showed that the problem (4.13) is strongly convex. As a special case we would like to present the following proposition to prove that the algorithm (2.2) converges almost surely to the solution set for strongly convex problems.

Proposition 12. *Let the function f have uniformly bounded gradients over the set X_0 with a scalar C_f and is assumed to be strongly convex with constant l. Let the stepsize be not summable but square summable such that $\sum_{k=1}^{\infty}\alpha_k=\infty$, $\sum_{k=1}^{\infty}\alpha_k^2<\infty$ and $\sum_{k=1}^{\infty}\alpha_k^2\nu_k<\infty$. Assume that problem* (2.1) *has a nonempty optimal set X^*. Then the iterates $\{x_k\}$ generated by method* (2.2) *converge almost surely to some random point in the optimal set X^*.*

Proof. We start with the result of Proposition 3

$$\begin{aligned}&\mathsf{E}\left[\|x_k-x^*\|^2\mid F_{\mathrm{k}-1}\right]\leq\left(1+\alpha_k^2L^2A_{\tau,\eta}\right)\|x_{k-1}-x^*\|^2+\frac{\left(\beta^2-2\beta\right)}{2cC_g^2}\|x_{k-1}-z_{k-1}\|^2\\&+2\alpha_k\left(f\left(x^*\right)-f\left(z_{k-1}\right)\right)-\|x_{k-1}-\upsilon_k\|^2+B_{\tau,\eta}\alpha_k^2\|\nabla f\left(x^*\right)\|^2+\left(4-4\left(\beta^2-2\beta\right)\right)\alpha_k^2\nu_k\end{aligned}\tag{4.14}$$

for all $k\geq\tilde{k}$, where $A_{\tau,\eta}=B_{\tau,\eta}=\left(8+8\eta-2\tau\left(\beta^2-2\beta\right)\right)$, $\tau=4$ and $\eta=-2\left(\dfrac{\left(\beta^2-2\beta\right)}{cC_g^2}\right)^{-1}$.

As the first step we can drop the terms $\dfrac{\left(\beta^2-2\beta\right)}{2cC_g^2}\|x_{k-1}-z_{k-1}\|^2$ and $-\|x_{k-1}-\upsilon_k\|^2$ although both of which are tightening the bound on the algorithm convergence.

For a point in optimal set X^* the relation $f\left(x\right)-f\left(x^*\right)\geq\dfrac{l}{2}\|x-x^*\|^2$ holds for a differentiable strongly convex function with constant l, Polyak (1987)(Page 11). Therefore

$$2\alpha_k\left(f\left(z_{k-1}\right)-f\left(x^*\right)\right)\geq\alpha_k l\|z_{k-1}-x^*\|^2.$$

The relation (4.14) satisfies the conditions of Theorem 1 for the following nonnegative random variables

$$\begin{aligned}&a_k=\alpha_k^2L^2A_1,\quad v_k=\|x_{k-1}-x^*\|^2,\quad u_k=\alpha_k l\|z_{k-1}-x^*\|^2,\\&b_k=B_1\alpha_k^2C_f^2+4\alpha_k^2\nu_k\left(1+2\beta-\beta^2\right).\end{aligned}$$

Due to diminishing nature of the stepsize α_k we have

$$\left(B_1 C_f^2 + 4\nu_k \left(1 + 2\beta - \beta^2\right)\right) \sum_{k=0}^{\infty} \alpha_k^2 \leq \infty \quad \text{a.s.}$$

and

$$L^2 A_1 \sum_{k=0}^{\infty} \alpha_k^2 \leq \infty \quad \text{a.s.},$$

where $A_1 = B_1 = 8\left(1 + \dfrac{2cC_g^2}{2\beta - \beta^2} + 2\beta - \beta^2\right)$ and $0 < \beta < 2$.

As a result the sequence $\{\|x_k - x^*\|\}$ is convergent almost surely for every $x^* \in X^*$ to $\upsilon \geq 0$, which is some nonnegative random variable.

$$\lim \{\|x_k - x^*\|\} \to \upsilon$$

We can further conclude that $\{\|z_k - x^*\|\}$ is also convergent *a.s.* for every $x^* \in X^*$ as well.

$$l \sum_{k=1}^{\infty} \alpha_k \|z_{k-1} - x^*\|^2 < \infty \qquad a.s.$$

This implies that the sequence $\{z_k\}$ is *a.s.* bounded and has accumulation points. Besides $\{\|z_k - x^*\|\}$ converges *a.s.* for every $x^* \in X^*$. Therefore $\{z_k\}$ converges almost surely to a random point in set X^*. We reach the conclusion that $\{x_k\}$ converges *a.s.* to a random point in the solution set, X^*. □

4.3.4 Problem Reformulation and Implementation

We formulated the problem (4.6) in the context of convex optimization with a linear objective and quadratic constraints as follows:

$$\begin{aligned} &\min_{\mathbf{h}\in\mathbb{R}^n,\vartheta\in\mathbb{R}_+} \vartheta \\ &\text{subject to} \quad \|U(\psi)\mathbf{h}-y(\varphi)\|^2 \le \vartheta \quad \text{such that} \quad \sqrt{\sum_{i=1}^{n}\psi_i^2+\sum_{j=1}^{n}\varphi_j^2}\le\rho \quad \text{for every } (\psi,\varphi), \end{aligned} \tag{4.15}$$

where the variable of the model is defined as $x^T = \begin{bmatrix} h_1 & h_2 & \dots & h_n & \vartheta \end{bmatrix}$ with $\mathbf{h}\in\mathbb{R}^n$ and $\vartheta\in\mathbb{R}_+$. The constraint functions $g_{\psi,\varphi}(\mathbf{h},\vartheta) = \|U(\psi)\mathbf{h}-y(\varphi)\|^2 - \vartheta$ are convex quadratic functions that are defined over $\mathbb{R}^n\times\mathbb{R}$ as

$$\begin{aligned} g_{\psi,\varphi}(x) = g_{\psi,\varphi}(\mathbf{h},\vartheta) = &\begin{bmatrix} \mathbf{h} \\ \vartheta \end{bmatrix}^T \begin{bmatrix} \mathbf{U}(\psi)^{\mathbf{T}}\mathbf{U}(\psi) & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \vartheta \end{bmatrix} \\ &- \begin{bmatrix} 2y(\varphi)^T U(\psi) & 1 \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \vartheta \end{bmatrix} + y(\varphi)^T y(\varphi). \end{aligned}$$

The algorithm projects onto $\mathbb{R}^n\times\mathbb{R}_+$ as below

$$\begin{aligned} \upsilon_k &= \Pi_{\mathbb{R}^n\times\mathbb{R}_+}\left[x_{k-1}-\alpha_k\left(\bigtriangledown f(x_{k-1})+\varepsilon_k\right)\right] \\ x_k &= \Pi_{\mathbb{R}^n\times\mathbb{R}_+}\left[\upsilon_k-\beta\frac{g^+_{\psi_k,\varphi_k}(\upsilon_k)}{\|d_k\|^2}d_k\right] \quad \text{for all } k\ge 1, \end{aligned} \tag{4.16}$$

where deterministic parameter β being $0<\beta<2$.

For algorithm (2.2) we have already showed that the initial point, x_0 does not affect the convergence properties. But for implementation purposes $x_0\in\mathbb{R}^{n+1}$ is selected randomly with an arbitrary distribution.

For this application we do not take into account noise accompanying the gradient of f since the objective function has a simple linear form. Hereby the gradient vector is $\bigtriangledown f(x) = \begin{bmatrix} 0 & \dots & 0 & 1 \end{bmatrix}^T$ for all x.

The first step of the iteration, k is taking a gradient step using the objective function and reaching an intermittent point $\upsilon_k^T = \begin{bmatrix} \upsilon_{\mathbf{k},\mathbf{h}} & \upsilon^+_{k,\vartheta} \end{bmatrix}$ where $\upsilon^+_{k,\vartheta} = \max\{\upsilon_{k,\vartheta},0\}$. Then the feasibility violation function

is calculated using intermittent point as follows; $g^{+}_{\psi_k,\varphi_k}(\upsilon_k) = \max\left\{g_{\psi_k,\varphi_k}(\upsilon_k), 0\right\}$. Therefore the feasibility violation function has the form of $f(x) = \max_{z\in Z}\phi(x,z)$ that arises in a variety of contexts in optimization applications with Z being a compact subset of $\mathbb{R}^m$and $\phi : \mathbb{R}^n \times Z \mapsto \mathbb{R}$ being continuous and such that $\phi(x,z) : \mathbb{R}^n \mapsto \mathbb{R}$ is convex for each $z \in Z$. Danskin's Theorem (Proposition 4.5.1, Bertsekas et al. (2003)) provides information about the derivatives of a function in preceding stated form. According to Danskin's Theorem if $\phi(x,z)$ is differentiable with respect to x for all the points of the maximizing set $z \in Z_0(x)$, where

$$Z_0(x) = \left\{\bar{z} : \phi(x,\bar{z}) = \max_{z\in Z}\phi(x,z)\right\},$$

then the subdifferential of $f(x)$ is given by

$$\partial f(x) = \operatorname{conv}\left\{\nabla_x \phi(x,z) : z \in Z_0(x)\right\}.$$

And consequently the subdifferential of feasibility violation function is

$$\partial g^{+}_{\psi,\varphi}(x) = \left\{\alpha \partial g_{\psi,\varphi}(x) \mid \alpha \in [0,1], i = 1,\ldots,n+1\right\}.$$

Hence the direction to decrease the feasibility violation at k^{th} intermittent point $\upsilon_k^T = \begin{bmatrix} \upsilon_{\mathbf{k},\mathbf{h}} & \upsilon^{+}_{k,\vartheta} \end{bmatrix}$ is chosen as $d_k \in \partial g^{+}_{\psi_k,\varphi_k}(\upsilon_k)$ where convex hull containing

$$\partial g^{+}_{\psi_k,\varphi_k}(\upsilon_k) = \begin{cases} \begin{bmatrix} 2\mathbf{U}(\psi_\mathbf{k})^{\mathbf{T}}\mathbf{U}(\psi_\mathbf{k}) & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \begin{bmatrix} \upsilon_{\mathbf{k},\mathbf{h}} \\ \upsilon^{+}_{k,\vartheta} \end{bmatrix} - \begin{bmatrix} \mathbf{2U}(\psi_\mathbf{k})\mathbf{y}(\varphi_\mathbf{k}) \\ 1 \end{bmatrix} \\ \text{if } g_{\psi_k,\varphi_k}(\upsilon_k) > 0, \\ \alpha\left[\begin{bmatrix} 2\mathbf{U}(\psi_\mathbf{k})^{\mathbf{T}}\mathbf{U}(\psi_\mathbf{k}) & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \begin{bmatrix} \upsilon_{\mathbf{k},\mathbf{h}} \\ \upsilon^{+}_{k,\vartheta} \end{bmatrix} - \begin{bmatrix} \mathbf{2U}(\psi_\mathbf{k})\mathbf{y}(\varphi_\mathbf{k}) \\ 1 \end{bmatrix}\right] \\ \alpha \in [0,1], \text{ if } g_{\psi_k,\varphi_k}(\upsilon_k) \leq 0. \end{cases}$$

We consider the following nominal values for y and u that are also used in §7.5 Robust Identification

example of Ghaoui and Lebret (1997):

$$u = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}^T, \qquad y = \begin{bmatrix} 4 & 5 & 6 \end{bmatrix}^T$$

For all three version of the problem each realization of perturbation leads to a specific quadratic constraint $g_{\psi,\varphi}(h,\vartheta)$. Therefore the elements of bounded disturbance sets defined by $\{\left\| \begin{bmatrix} \psi, \varphi \end{bmatrix} \right\| \leq \rho\}$, $\{\|\psi\| \leq \rho, \|\varphi\| \leq \rho\}$, $\{\|\psi\|_\infty \leq \rho, \|\varphi\| \leq \rho\}$, where $\rho \geq 0$ generate an infinite collection of constraints. We generated the random noise accompanying the input and output within the bounded sets as defined above.

Computational algorithms that rely on repeated random sampling to obtain numerical results should have enough samples to inspect the performance of the method. Therefore for the sake of having enough path to represent the performance of the algorithm we ran Monte Carlo simulation of algorithm for 95% confidence interval. We created in total of 101 paths for each ρ value.

For comparison purposes we solved the problems by implementing ordinary subgradient method onto the equivalent formulations (4.7), (4.9), and the upper bound formulation (4.11). Additionally the algorithm results are compared with respect to ordinary subgradient algorithm and related comparisons and statistics are presented in Table 4.1 and Table 4.2 [1] . Subgradient algorithm solutions are used as benchmarks. Subgradient algorithm took the average solution vector of MonteCarlo runs and used it as initial point. Although subgradient algorithm generally stabilized after about 800 iterations, we ran a constant iteration cycle of 5000. And we kept track of the function values as well as the solution points. And we used the best function value to compare it with random projection algorithm's converged function value. The error is calculated as function value difference between algorithm solution and subgradient algorithm. [2]. Stepsize of the algorithm is defined as $\alpha_k = a/k^b$, k being the number of iteration of algorithm which is increased as the bound on the perturbations is increased by $100 \times \rho^2$. Although the input norm intervals were chosen densely due to representation purposes we only present the integer values of ρ.

The performance of proposed random projection algorithm (RPA) (2.2) for problems (4.6), (4.8) and (4.10) using stepsize $\alpha_k = 1/k^{0.75}$ with respect to subgradient algorithm (SA) are presented in figures 4.1, 4.2, 4.3 for $\rho \leq 5$.

The Random Projection Algorithm performance for stepsize $\alpha_k = 1/k^{0.95}$ is also presented for $\rho \leq 10$

[1] $Residual = \|Uh - y\| + \rho \left\| \begin{bmatrix} h \\ 1 \end{bmatrix} \right\|$

[2] Error Residual $= F_{\text{random projection algorithm}} - F_{\text{subgradient algorithm}}$

below. As it can be seen the algorithm function values are closely follow the Subgradient Algorithm function values. And for 95% confidence interval the converged function values of RPA do not vary much although for each Monte Carlo run a different initial point and a random constraint path were chosen.

4.3.5 Figures and Tables

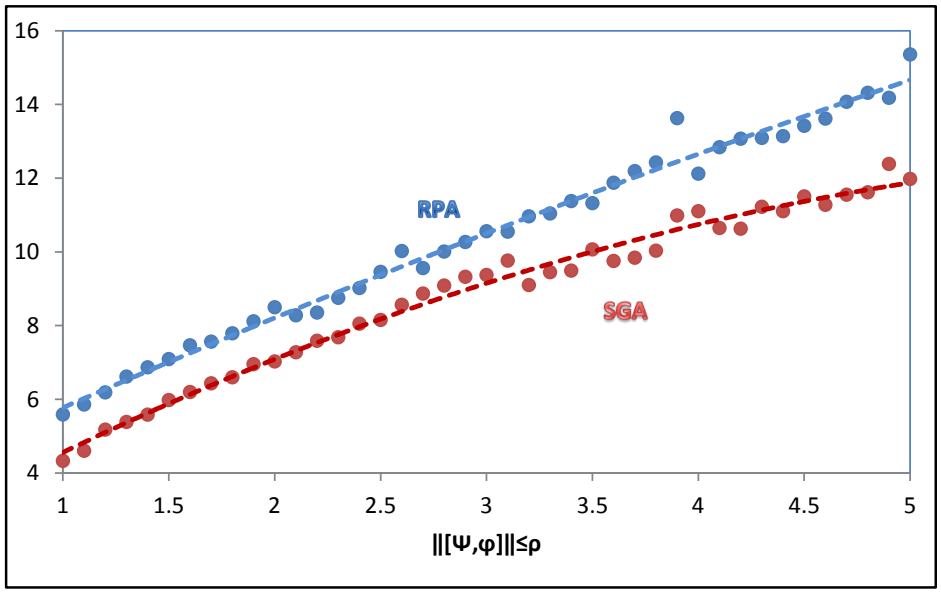

Figure 4.1: RPA vs. SA
$\alpha_k = 1/k^{0.75}$
$\|\psi, \varphi\| \leq \rho$

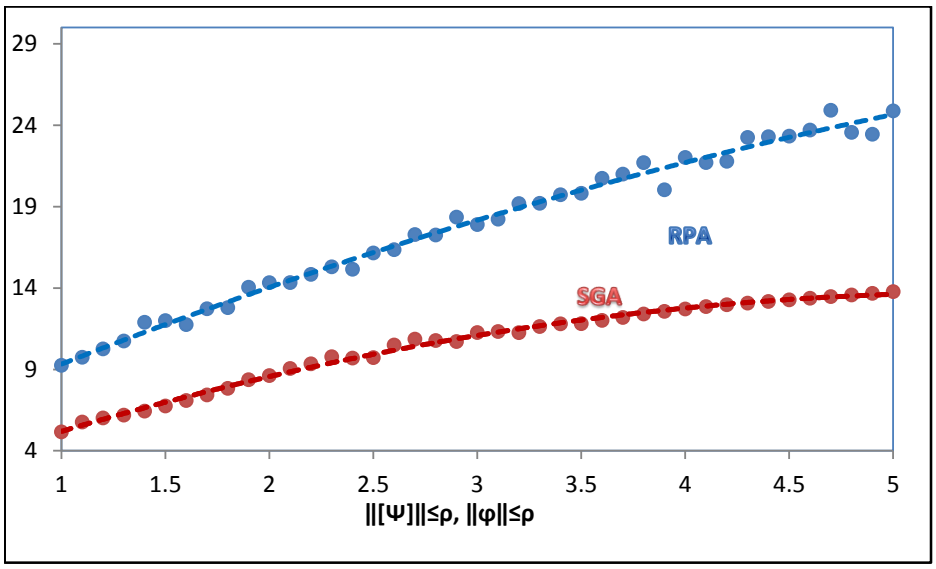

Figure 4.2: RPA vs. SA
$\alpha_k = 1/k^{0.75}$
$\|\psi\| \leq \rho, \|\varphi\| \leq \rho$

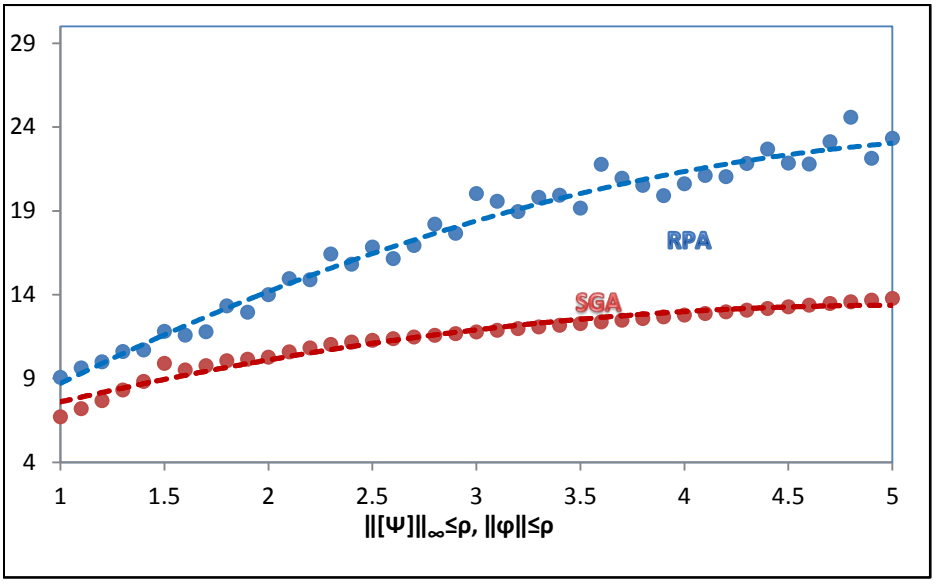

Figure 4.3: RPA vs. SA
$\alpha_k = 1/k^{0.75}$
$\|\psi\|_\infty \leq \rho, \|\varphi\| \leq \rho$

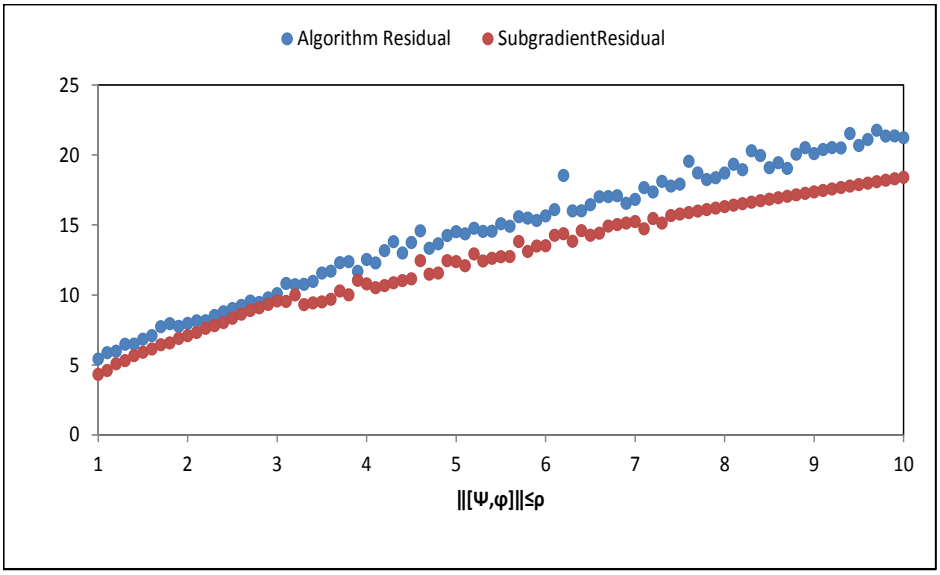

Figure 4.4: RPA vs. SA
$\alpha_k = 1/k^{0.95}$
$\|\psi, \varphi\| \leq \rho$

Table 4.1: Function Value Comparison

$\alpha_k = a/k^b$	ρ	1	2	3	4	5	6	7	8	9	10
$\alpha_k = 1/k^{0.75}$	f_{alg}[a]	5.588	8.498	10.561	12.122	15.385	15.415	16.671	19.089	20.925	22.398
	f_{sg}[b]	4.33	7.028	9.375	11.104	11.979	13.814	15.246	16.312	17.364	18.406
$\alpha_k = 1/k^{0.95}$	f_{alg}	5.423	7.969	10.097	12.536	14.523	15.651	16.829	18.713	20.102	21.236
	f_{sg}	4.329	7.092	9.577	10.792	12.384	13.52	15.245	16.313	17.364	18.405
$\alpha_k = 10/k^{0.75}$	f_{alg}	6.449	10.587	12.208	13.537	16.58	18.418	21.887	21.415	22.793	26.347
	f_{sg}	4.329	7.14	9.577	11.128	12.309	13.242	15.246	16.313	16.998	18.120
$\alpha_k = 10/k^{0.95}$	f_{alg}	5.809	8.596	11.137	13.319	15.504	16.081	17.366	19.114	21.048	22.167
	f_{sg}	4.329	6.99	9.578	10.358	12.069	14.156	15.239	16.313	17.364	18.406
$\alpha_k = 100/k^{0.75}$	f_{alg}	6.164	10.714	15.355	17.403	23.241	25.295	28.05	26.689	30.846	39.559
	f_{sg}	4.329	7.0129	9.319	10.353	13.028	13.526	15.246	16.105	17.364	18.405
$\alpha_k = 100/k^{0.95}$	f_{alg}	6.077	10.921	13.953	17.203	20.215	20.19	22.242	22.965	25.254	26.331
	f_{sg}	4.329	7.084	9.318	10.343	12.356	14.155	15.245	15.824	17.364	18.406

[a] Average function value based on 101 simulation paths.

[b] Ordinary subgradient algorithm function value for problem, 5000 iterations

Table 4.2: Algorithm Deviation Statistics from Subgradient Algorithm

$\alpha_k = a/k^b$	ρ	1	2	3	4	5	6	7	8	9	10
$\alpha_k = 1/k^{0.75}$	Err_{ave}[a]	1.258	1.471	1.186	1.018	3.379	1.6	1.426	2.77	3.56	3.98
	$StdDev$[b]	1.783	2.43	2.72	2.106	5.514	2.449	1.936	4.626	6.065	7.363
$\alpha_k = 1/k^{0.95}$	Err_{ave}	1.094	0.876	0.519	1.744	2.138	2.1309	1.584	2.399	2.737	2.83
	$StdDev$	2.135	1.684	2.055	2.833	2.707	3.847	2.181	3.687	3.962	3.717
$\alpha_k = 10/k^{0.75}$	Err_{ave}	2.119	3.446	2.631	2.409	4.269	5.176	6.641	5.103	5.795	8.227
	$StdDev$	2.186	6.46	4.726	3.769	6.675	7.511	11.128	7.388	8.469	28.155
$\alpha_k = 10/k^{0.95}$	Err_{ave}	1.48	1.605	1.559	2.962	3.434	1.926	2.126	2.801	3.684	3.761
	$StdDev$	2.071	2.077	4.007	7.585	5.952	3.354	3.769	3.508	5.644	4.8
$\alpha_k = 100/k^{0.75}$	Err_{ave}	1.834	3.588	6.036	7.049	10.213	11.769	12.804	12.584	13.482	21.253
	$StdDev$	2.378	4.308	9.558	7.174	13.94	15.449	24.743	15.856	16.21	24.033
$\alpha_k = 100/k^{0.95}$	Err_{ave}	1.748	3.837	4.634	6.86	7.859	6.034	6.996	7.141	7.889	7.926
	$StdDev$	1.782	4.817	7.267	9.942	17.269	9.312	9.205	9.08	12.508	9.836

[a]Error Average is defined as average function value difference between random projection algorithm and subgradient algorithm

[b]Standard Deviation of Error

Chapter 5

STOCHASTIC CONVEX SET INTERSECTION PROBLEM: RANDOM FEASIBILITY PROJECTION ALGORITHM

5.1 Introduction

Convex feasibility problem has broad applicability in diverse areas of mathematics and physical sciences. The mathematical formulation of the convex feasibility problem in N-dimensional euclidean space is as follows.

$$\text{Find any } x \in X \text{ such that,}$$

$$X = \cap_{i=1}^{\mathscr{M}} X_i \text{ for all } X_i \in \mathbb{R}^n.$$

This problem originally emerged with finding a starting feasible point in simplex algorithm for linear programming problems but we can trace back the roots of the problem even earlier than that. Cimmino's famous algorithm, Cimmino (1938), is considered to be the ancestor of feasibility problem. He considers a system of linear algebraic equations $Ax = b$ where A is a nonsingular real $n \times n$ matrix and $b \in \mathbb{R}^n$. If $a_i^T = [a_{i1}, a_{i2} \ldots, a_{in}]$ denotes the ith row of A, the solution point x^* is the unique intersection point of the n hyperplanes defined as

$$\langle a_i, x \rangle = b_i, \quad i = 1, 2, \ldots, n.$$

The algorithm takes a step by reflecting the iterate with respect to the hyperplanes

$$x_i^{(k+1)} = x^{(k)} + 2\frac{b_i - \langle a_i, x^{(k)} \rangle}{\|a_i\|^2} a_i.$$

The next iterate is constructed by convex weights $m_1, \ldots, m_n$ multiplied by each reflection point, $x_i^{(k+1)}$. Cimmino showed that the iterates $\{x^{(k)}\}$ converges to a solution even in the case of a singular but consistent system provided that rank$(A) \geq 2$. In the singular case, Cimmino obtains a bound on the relative error in

the Euclidean norm, showing the linear rate of convergence of his method. One important aspect of Cimmino's algorithm is that even if the linear system is inconsistent the sequence $\{x^{(k)}\}$ converge to a weighted least-squares solution. Polish mathematician Kaczmarz (1993) published a similar approach a year before Cimmino. In his method, the current approximation $x^{(k)}$ is orthogonally projected onto the hyperplanes sequentially. The projection onto the nth hyperplane is taken as the new iteration $x^{(k+1)}$. The sequence generated by this method converges to the solution of $Ax = b$ as $k \to \infty$. The methods of Cimmino and Kaczmarz are closely related. Then the progress by Cimmino and Kaczmarz is further improved to find a point which satisfies a finite number of halfspaces defined by linear inequalities. The "feasible point" is used as initial iteration point for Simplex algorithm. Solving the linear feasibility problem in this context was given by Agmon (1954) and Motzkin and Schoenberg (1954). The projection algorithms, which were referred as relaxation algorithms by Agmon (1954) and Motzkin and Schoenberg (1954) were used to find a starting feasible point for system of linear inequalities in order to initialize the simplex algorithm. Generalizations for feasibility problem onto convex sets in real n-dimensional spaces were first given in Eremin (1969) and Jakubowich (1966). The classical projection method for the case of finite intersecting closed convex sets in a real Hilbert space was first introduced in Bregman (1965) and Bregman (1967). He showed that, given an arbitrary starting point x_0, the sequence generated by the projection algorithm converges weakly to a point in nonempty feasible set. The scheme that uses varying weighted averages of relaxed projections onto approximating halfspaces by Flåm and Zowe (1990) is worth mentioning too. A complete and exhaustive survey work on algorithms for solving convex feasibility problem is given by Bauschke and Borwein (1996).

The most frequently used technique to solve convex feasibility problem is algorithmic projection to generate a sequence converging to the any point of solution set. Projection operation onto the sets is the key element of the algorithms defined. The sets can be simple so that the projection onto set X_i can be calculated explicitly. In case projecting onto set X_i is not possible then an approximating superset of X_i usually in the form of lower level set of the original set is used. Some of the application fields are as follows.

- Approximation theory where usually sets are closed subspaces with applications in linear prediction theory, partial differential equations (Dirichlet problem), complex analysis (Bergman kernels, conformal mappings), Deutsch (1992).

- Discrete image reconstruction and signal restoration models where sets are halfspaces or hyperplanes with applications in medical imaging, radiation therapy treatment planning, electron microscopy, Cen-

sor (1988), Trussell and Civanlar (1984), Censor and Herman (1987) .

- Continuous image reconstruction models where the approximating superset is usually an infinite-dimensional Hilbert space with applications in signal processing, computerized tomography, Hermann (1980), Herman (1995), Youla and Webb (1982), Censor and Herman (1987) Oskoui-Fard and Stark (1988), Eggermont et al. (1981).

- Subgradient algorithms where some of the sets are approximated through a superset with applications in solution of convex inequalities, minimization of convex nonsmooth functions, Censor and Lent (1982), Shor et al. (1985), Dem'Yanov et al. (1985).

We explored an algorithmic approach to solve both consistent and inconsistent convex feasibility problems for closed convex uncertain sets. We got inspiration from Nedić (2010). But our focus is on uncertain nature of sets and finding a feasible point using a random subcollection of closed, convex uncertain sets. For this objective we consider a stochastic optimization problem of minimizing an expected proximity function over a collection of closed, convex sets. We would like to show that the proposed algorithm converge to a point in the solution set when solution set is nonempty. In case of inconsistent feasibility problem i.e. the intersection of closed convex constraint sets being empty the algorithm minimizes a weighted proximity function. The projection onto a subcollection of sets approach can be viewed as somewhere between random implementation of alternating projection method and parallel projection method. But our method is not deterministic, the algorithm that we propose utilizes random projections. In general sense, our work in this chapter is related to convex feasibility problems resulting from random sampling Alamo et al. (2009), Calafiore (2010).

For numerical example we solved a signal deconvolution problem, which is an estimation problem. It is mainly implemented for recovery of original images, which has been passed through a gaussian blur and further corrupted by additive noise. Image/signal recovery problem is to recover input vector h from the observation of data signals, x. It is fundamentally an inverse problem in which an original signal is inferred from the observed signal. It has wide spectrum of applications such as ultrasonic imaging, nondestructive testing, digital radiology. There is extensive amount of literature on image recovery applications Andrews and Hunt (1977), Capricelli and Combettes (2007), Combettes (1997), Combettes (1996).

The earlier versions of image recovery algorithms share the common objective of producing a solution

consistent with a collection of affine or affine inequality constraints. But many useful constraints encountered in practice are nonaffine. Therefore the extension from affine sets to convex set formulations broadened the applicability of feasibility algorithms. The serial projections onto convex sets (PCOS) became popular and had been introduced in image reconstruction and image recovery by Lent and Tuy (1981) and Youla and Webb (1982). The convex feasibility problem is a central problem in applied mathematics as finding a common point of closed and convex sets or solving a system of convex inequalities as in Bauschke and Borwein (1996), Censor (1984) and Combettes (1993). The fundamental concepts of Fejeŕ-monotocity, admissibility, nonexpansive mapping, and bounded regularity were introduced to the field of convex feasibility by the pioneer works of Gubin et al. (1967) and Browder (1967). Gubin et al. (1967) have established the first convergence rate result assuming that the intersection set has nonempty interior. In image recovery problems spatial and spectral information can be incorporated in the form of convex sets but usually carries uncertainty that are characterized by estimated statistical values. Therefore in most cases even the sets are uncertain. Our method is a novel approach where randomly chosen batches of an arbitrary collection $\{X_i, i \in \mathscr{M}\}$ of nonempty, closed, convex uncertain sets in $\mathbb{R}^n$ are to be activated at each iteration as opposed to being activated in periodic order as in Browders admissable control. Sets that we define are not estimated using statistical models.

The sets we consider might carry uncertainty due to inaccurate or imprecise spatial and spectral information. For example the sets that were used in image restoration work by Youla and Webb (1982) mostly depend on the attributes of the original signal. The sets used in Youla and Webb (1982) are amplitude bounds, region of support, band-limitedness, energy that are not known exactly. The attributes of the noise associated with sets in Combettes and Trussell (1991) are also based on predicted stochastic information and confidence levels. If the stochastic information carries errors then the sets assumed would not represent the real sets. So you might get inconsistent sets although in reality the intersection set is nonempty. Or due to conservative estimation of confidence intervals the sets may not intersect. But in our formulation we have only a bound on the uncertainty parameter. Each realization within the bounded disturbances has equal chance of occurence.

Notation: A vector is a column vector. We use x^T to denote the transpose of a vector x, and $\|x\|$ to denote the standard Euclidean norm. Minimum distance of a vector $\bar{x}$ to a closed convex set X is $\text{dist}(\bar{x}, X)$. The projection of a vector $\bar{x}$ on a closed convex set X is represented as $\Pi[\bar{x}] = \text{argmin}_{x \in X} \|x - \bar{x}\|^2$. Probability

distribution of a random variable Z and expectation of a random variable Z are indicated by $\mathsf{Pr}[Z]$ and $\mathsf{E}[Z]$ respectively. We often abbreviate *almost surely* by *a.s.*.

5.2 Problem Formulation and Algorithm Description

We consider the following convex constrained minimization problem for a collection $\{X_i, i \in \mathscr{M}\}$ of nonempty, closed, convex and possibly uncertain sets in $\mathbb{R}^n$. We would like to use an iterative algorithm to find a common point for the sets when such a point exists. Also, we would like to determine if a solution does not exist based on the behavior of the algorithm.

$$\begin{aligned} &\text{determine a point such that } x \in X, \quad X \triangleq (\cap_{i\in\mathscr{M}} X_i), \\ &\text{with } X_i = \{x \in \mathbb{R}^n, \omega \in \mathbb{R}^n, \|\omega_i\| \leq \rho_i \mid g_i(x,\omega_i) \leq 0\} \quad \forall i \in \mathscr{M}, \end{aligned}$$

where ω is a bounded random variable. This problem is related to the following stochastic optimization problem of expected average residual of x with respect to sets X_i. It is called the expected weighted proximity function, $\mathscr{R} : \mathbb{R}^n \to \mathbb{R}_+$

$$\begin{aligned} &\text{minimize } \mathscr{R}(x) = \frac{1}{2}\sum_{i\in\mathscr{M}} w_i \mathsf{E}\left[\|x - \Pi_{X_i}[x]\|^2\right] \\ &\text{subject to } x \in \mathbb{R}^n, \end{aligned} \tag{5.1}$$

where w_i is the weight of a particular set X_i, within the collection of nonempty, closed and convex sets, $\mathscr{M} = \{1, \ldots, m, \ldots M\}$. Weights, w_i are chosen to be strictly convex such that

$$\sum_i w_i = 1 \quad \textit{and} \quad w_i > 0 \quad \forall i \in \mathscr{M}.$$

The number of sets are finite but sets might carry uncertainty due to inaccurate or imprecise spatial and spectral information, stochastic perdiction etc.

The smaller the value $\mathscr{R}(x)$ is, the closer the point x is satisfying all the properties in weighted least-squares sense. In other words the function $\mathscr{R}(x)$ measures the degree of unfeasibility.

We assume that the projection on each set X_i is available in a closed form. Such sets are hyperplanes,

balls and the sets given by linear inequalities.

Let X^* and $\mathscr{R}^*$ denote the optimal set and the optimal value of problem (5.1). So X^* is the set of minimizers of $\mathscr{R}$. X^* is the set of weighted least squares feasible points in consistent case. The solution set is defined as $X^* = \cap_{i=1}^{M} X_i$ for the consistent case where $\mathscr{R}(x^*) = 0$ for all $x^* \in X^*$. And for the inconsistent case X^* is the set of minimizers where $\mathscr{R}^* \neq 0$. We do not necessarily require $\cap_{i=1}^{M} X_i, \neq \emptyset$. When incompatible constraints are present then $X^* = \cap_{i=1}^{M} X_i = \emptyset$ or it coincides with the set of stationary points of $\mathscr{R}(x)$. Then the weighted average of the squares of distances to constraint sets does not take the value of zero. The proximity function is defined as weighted least squares feasibility problem, which indicates the degree of unfeasibility of a signal, x. When consistency is not certain then the goal is to find the minimum of expected proximity function (5.1) where

$$X^* = \{x \mid \mathscr{R}(x^*) \leq \mathscr{R}(x)\} \quad \forall x \in \mathbb{R}^n.$$

We consider a stochastic random projection algorithm. At time k, we have an iterate x_{k-1} and a random subcollection of m sets are either observed or chosen. The uncertain sets within the batch of m sets may carry additive noise realizations of random variables, ω_{i_k}. The iterate process is given by

$$x_k = x_{k-1} - \alpha_k \left(x_{k-1} - \sum_{i=1}^{m} w_i \Pi_{X_{\omega_{i_k}}} [x_{k-1}] \right) \quad \text{for all } k \geq 1, \tag{5.2}$$

where $\alpha_k > 0$ is a deterministic stepsize. Unless information about solution x^* being closer to a certain set is available, we suggest using equal weights for each set. The initial point $x_0 \in \mathbb{R}^n$ is selected randomly. The random variables ω_{i_k} are random samples of ω_i that are drawn from bounded noise sets defined by $\|\omega_i\| \leq \rho_i$ at iteration k.

We would like to point out some features of the expected weighted proximity function (5.1). Firstly, we define a determinate function. If we knew exactly which realizations of random variables, ω_i are going to be revealed, the function $R(x, \omega_i)$ is defined as follows

$$R(x, \omega_i) = \frac{1}{2} \sum_{i=1}^{m} w_i \left\| x - \Pi_{X_{\omega_i}} [x] \right\|^2 \quad \text{for all } x \in \mathbb{R}^n, \|\omega_i\| \leq \rho_i, \tag{5.3}$$

where $\sum_i^m w_i = 1 \quad \textit{and} \quad w_i > 0$. $R(x, \omega_i)$ is convex and differentiable in x for every realization of ω_i.

The proximity function for every batch of closed, convex sets, we define one $R(x,\omega_i)$ where $\|\omega_i\| \leq \rho_i$. Therefore we can claim the following relation between functions (5.1) and (5.3)

$$\mathscr{R}(x) = \mathsf{E}[R(x,\omega_i)]. \tag{5.4}$$

The gradient of $R(x,\omega_i)$ is

$$\nabla_x R(x,\omega_i) = \sum_{i=1}^{m} w_i \left(x - \Pi_{X_{\omega_i}}[x]\right) \text{ for } x \in \mathbb{R}^n. \tag{5.5}$$

The gradient of $\mathscr{R}(x)$ is

$$\nabla\mathscr{R}(x) = \mathsf{E}[\nabla_x R(x,\omega_i)] = \sum_{i=1}^{m} w_i (x - \mathsf{E}[\Pi_{X_i}[x]]) \quad \text{for } x \in \mathbb{R}^n. \tag{5.6}$$

Now we would like to introduce the assumptions that are used throughout the chapter for random feasibility projection algorithm.

We assume the following for the sets X_i.

Assumption 5. *The sets defined as*

$$X_i = \{x \in \mathbb{R}^n, \omega \in \mathbb{R}^n, \|\omega_i\| \leq \rho_i \mid g_i(x,\omega_i) \leq 0\}$$

are closed and convex for every realization of bounded noise ω_i and for each $i \in \mathscr{M}$.

For the random sequence of sets chosen and noise realizations we assume the following.

Assumption 6. *At each iteration a random subcollection of m sets is either chosen or revealed out of M sets. For the sets that carry uncertain additive noise, disturbance realizations of ω_{i_t} are going to be revealed at each step. Both the chosen sets m_t and revealed noise component ω_{i_t} are independent of past as well as the initial point x_0. We let $\mathscr{F}_k$ denote the history of the method run up to time k,*

$$\mathscr{F}_k = \{x_0, (m_t, 1 \leq t \leq k), (\omega_{i_t}, 1 \leq t \leq k)\} \qquad \textit{for } k \geq 1,$$

with $\mathscr{F}_0 = \{x_0\}$.

This concludes the assumptions further used in the text for the proposed algorithm for the optimization problem (2.1) with convex inequality constraint sets.

Next we will introduce a few well-known lemmas for Euclidean projection operation.

Lemma 11. *The non-expansive property of Euclidean projection operation on a closed convex set* $Y \subseteq \mathbb{R}^n$ *can be shown as*

$$\|\Pi_Y[x] - z\| \leq \|x - z\| \quad \textit{for all } z \in Y, \textit{ and } x \in \mathbb{R}^n \tag{5.7}$$

and

$$(\Pi_Y[x] - \Pi_Y[z])^T (x - z) \geq \|\Pi_Y[x] - \Pi_Y[z]\|^2 \quad \textit{for any } x, z \in \mathbb{R}^n. \tag{5.8}$$

The proofs of these results are presented in the book by Facchinei and Pang (2003) (Vol. I, page 77). Another variation of the non-expansive property of Euclidean projection operation is presented in the book by Polyak (1987) (page 121).

Lemma 12. *For a closed convex set* $Y \subseteq \mathbb{R}^n$

$$\|\Pi_Y[x] - \Pi_Y[z]\| \leq \|x - z\| \quad \textit{for any } x, z \in \mathbb{R}^n. \tag{5.9}$$

The strictly non-expansive property of Euclidean projection operation is as follows.

Lemma 13. *For a nonempty closed convex set* $Y \subseteq \mathbb{R}^n$

$$\|\Pi_Y[x] - z\|^2 \leq \|x - z\|^2 - \|x - \Pi_Y[x]\|^2 \quad \textit{for all } z \in Y, \; x \in \mathbb{R}^n. \tag{5.10}$$

The proof of this result can be found in Facchinei and Pang (2003) (Vol. II, 12.1.13 Lemma, page 1120).

Since the projection operation is a nonexpansive mapping, the set of all fixed points, which is defined by

$$\text{Fix}T = \{x \in \mathbb{R}^n : x = \Pi[x]\}, \tag{5.11}$$

is always closed and convex. The proof of this result can be found in Goebel and Kirk (1990) (Lemma 3.4).

In order to investigate the random characteristics of sequences, the following supermartingale convergence result due to Robbins and Siegmund (1971) (see also Polyak (1987), Lemma 11, page 50) is used.

Theorem 13. *Let v_k, u_k, a_k, b_k be sequences of nonnegative random variables that may be dependent and let*

$$\mathsf{E}\left[v_{k+1} \mid \mathscr{F}_k\right] \leq (1+a_k)v_k - u_k + b_k \qquad a.s. \quad \textit{for all } k \geq 0,$$
$$\sum_{k=0}^{\infty} a_k < \infty \qquad a.s.\,, \qquad \sum_{k=0}^{\infty} b_k < \infty \qquad a.s.,$$

where $\mathscr{F}_k$ denotes the collection $v_0,...,v_k$, $u_0,...,u_k$, $a_0,...,a_k$, $b_0,...,b_k$. Then

$$\lim_{k\to\infty} v_k \to v \qquad a.s.\,, \qquad \sum_{k=0}^{\infty} u_k < \infty \qquad a.s.,$$

where $v \geq 0$ is some random variable.

Both $\mathscr{R}(x)$ and $R(x)$ have Lipschitz gradients with constant $L = 1$ such that

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

$$\|\nabla \mathscr{R}(x) - \nabla \mathscr{R}(y)\| \leq \|x-y\| \quad \text{for all } x, y \in \mathbb{R}^n. \tag{5.12}$$

Based on relations presented in Polyak (1987) (section 1.1.3, page 7) we can claim that

$$\begin{aligned}
&\mathscr{R}(x) \leq \mathscr{R}(y) + \nabla \mathscr{R}(y)^T (x-y) + \frac{1}{2}\|x-y\|^2,\\
&\mathscr{R}(x) \geq \mathscr{R}(y) + \nabla \mathscr{R}(y)^T (x-y) - \frac{1}{2}\|\nabla \mathscr{R}(x) - \nabla \mathscr{R}(y)\|^2,\\
&(y-x)^T (\nabla \mathscr{R}(y) - \nabla \mathscr{R}(x)) \leq \|y-x\|^2,\\
&(y-x)^T (\nabla \mathscr{R}(y) - \nabla \mathscr{R}(x)) \geq \|\nabla \mathscr{R}(y) - \nabla \mathscr{R}(x)\|^2 \quad \text{for all } x, y \in \mathbb{R}^n.
\end{aligned} \tag{5.13}$$

We concluded presenting the relations that are going to be used to show the convergence properties of algorithm (5.2).

5.3 Convergence Results for Random Convex Feasibility Algorithm

In this section we would like to show that the iterate sequences $\{x_k\}$ obtained by proposed algorithm (5.2) is converging to the solution of problem (5.1) almost surely for a diminishing stepsize $\alpha_k > 0$ such that

$$\sum_{k=1}^{\infty} \alpha_k = \infty \text{ and } \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

And we are going to investigate existence of solution set.

5.3.1 Preliminary Results

We start with proving the converging behavior of $\mathscr{R}(x_k)$ and boundedness of $\|\nabla \mathscr{R}(x_{k-1})\|$.

Proposition 14. *Let Assumptions 5 -6 hold. The stepsize is not summable but square summable such that* $\sum_{k=1}^{\infty} \alpha_k = \infty$ *and* $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$. *Then, for the iterates* $\{x_k\}$ *generated by method* (5.2)*, we have almost surely*

$$\lim_{k\to\infty} \mathscr{R}(x_k) = r \text{ for a random scalar } r \geq \mathscr{R}^* \quad a.s.$$

$$\liminf_{k\to 0} \|\nabla \mathscr{R}(x_{k-1})\| = 0 \quad a.s.$$

Proof. We are going to start with relating two consecutive iterates using relation (5.13) as follows

$$\mathscr{R}(x_k) \leq \mathscr{R}(x_{k-1}) + \nabla \mathscr{R}(x_{k-1})^T (x_k - x_{k-1}) + \frac{1}{2} \|x_k - x_{k-1}\|^2 \quad \text{for all } k. \tag{5.14}$$

And when we use the algorithm (5.2) the difference between two iterations using a random batch of m sets each time is

$$x_k - x_{k-1} = -\alpha_k \left(x_{k-1} - \sum_{i=1}^{m} w_i \Pi_{X_{\omega_{i_k}}} \right). \tag{5.15}$$

The gradient of $R(x, \omega_i)$ for randomly chosen sets of m is

$$\nabla_x R(x, \omega_i) = x - \sum_{i=1}^{m} w_i \Pi_{X_{\omega_i}}[x] \text{ for } x \in \mathbb{R}^n$$

and for $\mathscr{R}(x)$ is

$$\nabla \mathscr{R}(x) = \mathsf{E}[\nabla_x R(x, \omega_i)] = x - \sum_{i=1}^{m} w_i \mathsf{E}\left[\Pi_{X_{\omega_i}}[x]\right] \quad \text{for } x \in \mathbb{R}^n,$$

where

$$\sum_{i=1}^{m} w_i = 1.$$

Therefore the iterate relation (5.15) for $(x_k - x_{k-1})$ can be written as

$$x_k - x_{k-1} = -\alpha_k \left(\nabla_x R\left(x, \omega_{i_k}\right)\right) \quad \text{for all } k \geq 1. \tag{5.16}$$

For randomly chosen m sets we can write relation (5.14) as follows

$$\mathscr{R}\left(x_k\right) \leq \mathscr{R}\left(x_{k-1}\right) - \alpha_k \nabla \mathscr{R}\left(x_{k-1}\right)^T \nabla_x R\left(x_{k-1}, \omega_{i_k}\right) + \frac{{\alpha_k}^2}{2} \left\|\nabla_x R\left(x_{k-1}, \omega_{i_k}\right)\right\|^2$$

We would like to remind that m sets chosen within M of them and the realizations of disturbances ω_i of uncertain sets at iteration k are independent of the past path followed. We take expectation based on algorithm path up until iteration k and we get

$$\begin{aligned} \mathsf{E}\left[\mathscr{R}\left(x_k\right) \mid \mathscr{F}_{k-1}\right] \leq \mathscr{R}\left(x_{k-1}\right) - \alpha_k \nabla \mathscr{R}\left(x_{k-1}\right)^T \mathsf{E}\left[\nabla_x R\left(x_{k-1}, \omega_{i_k}\right) \mid \mathscr{F}_{k-1}\right] \\ + \frac{\alpha_k^2}{2} \mathsf{E}\left[\left\|\nabla_x R\left(x_{k-1}, \omega_{i_k}\right)\right\|^2 \mid \mathscr{F}_{k-1}\right]. \end{aligned} \tag{5.17}$$

From the definitions of $\mathscr{R}\left(x\right)$ and $R\left(x, \omega_i\right)$ we have the relations below

$$\begin{aligned} \mathsf{E}\left[\nabla_x R\left(x_{k-1}\omega_k\right) \mid \mathscr{F}_{k-1}\right] &= \nabla \mathscr{R}\left(x_{k-1}\right), \\ \mathsf{E}\left[\left\|\nabla_x R\left(x_{k-1}\omega_k\right)\right\|^2 \mid \mathscr{F}_{k-1}\right] &= 2\mathscr{R}\left(x_{k-1}\right) \quad \text{for any } x \in \mathbb{R}^n. \end{aligned}$$

Using relations above in inequality (5.17), we get

$$\mathsf{E}\left[\mathscr{R}\left(x_k\right) \mid \mathscr{F}_{k-1}\right] \leq \left(1 + {\alpha_k}^2\right) \mathscr{R}\left(x_{k-1}\right) - \alpha_k \left\|\nabla \mathscr{R}\left(x_{k-1}\right)\right\|^2.$$

The supermartingale convergence result due to Robbins and Siegmund (1971) is applicable to the inequality above since $\mathscr{R}\left(x_k\right) \geq 0$ for all x. Recall that the stepsize is square summable, $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$. Then we can claim that $\mathscr{R}\left(x_k\right) \to r$ for random variable $r \geq 0$ almost surely. Also $\sum_{k=1}^{\infty} \{\alpha_k \left\|\nabla \mathscr{R}\left(x_{k-1}\right)\right\|^2\} < \infty$. Since the stepsize we choose is not summable, $\sum_{k=1}^{\infty} \alpha_k = \infty$, we have $\liminf_{k\to 0} \left\|\nabla \mathscr{R}\left(x_{k-1}\right)\right\| = 0$ almost surely. □

5.3.2 Almost Sure Convergence Result

In this section, we would like to show that the proposed algorithm converges to the solution set almost surely for not summable but square summable stepsize. As it is indicated in the next proposition, the algorithm has almost sure convergence.

Proposition 15. *In addition to Assumptions above, lets assume that the optimal set X^* is nonempty. Then, the seqeunce $\{x_k\}$ generated by method* (5.2)*, converges almost surely to a random point in the set X^* .*

Proof. Let z be any point in X^*. We are going to start by writing the distance between any iterate point x_k for$k \geq 1$ and a random point z in the solution set using the definition of algorithm (5.2).

$$\begin{aligned}\|x_k - z\|^2 &= \left\| x_{k-1} - z - \alpha_k \left(x_{k-1} - \sum_{i=1}^{m} w_i \Pi_{X_{\omega_{i_k}}} [x_{k-1}] \right) \right\|^2 \\ &= \|x_{k-1} - z\|^2 + \alpha_k^2 \left\| x_{k-1} - \sum_{i=1}^{m} w_i \Pi_{X_{\omega_{i_k}}} [x_{k-1}] \right\|^2 - 2\alpha_k \left(x_{k-1} - \sum_{i=1}^{m} w_i \Pi_{X_{\omega_{i_k}}} [x_{k-1}] \right)^T (x_{k-1} - z).\end{aligned} \tag{5.18}$$

Since $\nabla_x R(x_{k-1}, \omega_{i_k}) = x_{k-1} - \sum_{i=1}^{m} w_i \Pi_{X_{\omega_{i_k}}} [x_{k-1}]$ where $\sum_{i=1}^{m} w_i = 1$, equality (5.18) can be written as

$$\|x_k - z\|^2 = \|x_{k-1} - z\|^2 + \alpha_k^2 \|\nabla_x R(x_{k-1}, \omega_{i_k})\|^2 - 2\alpha_k (\nabla_x R(x_{k-1}, \omega_{i_k}))^T (x_{k-1} - z). \tag{5.19}$$

We already showed that $\mathsf{E}[\nabla_x R(x_{k-1}\omega_k) \mid \mathscr{F}_{k-1}] = \nabla \mathscr{R}(x_{k-1})$ and $\mathsf{E}\left[\|\nabla_x R(x_{k-1}\omega_k)\|^2 \mid \mathscr{F}_{k-1}\right] = 2\mathscr{R}(x_{k-1})$.

When we take the expectation of inequality (5.19) conditioned on the past path until $\mathscr{F}_{k-1}$, we get

$$\mathsf{E}\left[\|x_k - z\|^2 \mid \mathscr{F}_{k-1}\right] = \|x_{k-1} - z\|^2 + 2\alpha_k^2 \mathscr{R}(x_{k-1}) - 2\alpha_k \nabla \mathscr{R}(x_{k-1})^T (x_{k-1} - z). \tag{5.20}$$

Since $\mathscr{R}(x_k)$ is convex, we have

$$\nabla \mathscr{R}(x_{k-1})^T (x_{k-1} - z) \geq \mathscr{R}(x_{k-1}) - \mathscr{R}(z).$$

We can use above inequality in (5.20) and we get

$$\mathsf{E}\left[\|x_k - z\|^2 \mid \mathscr{F}_{k-1}\right] \leq \|x_{k-1} - z\|^2 + 2\alpha_k^2 \mathscr{R}(x_{k-1}) - 2\alpha_k (\mathscr{R}(x_{k-1}) - \mathscr{R}(z)). \tag{5.21}$$

In Proposition 14 we showed that $\mathscr{R}(x_k) \to r$ for random variable $r \geq 0$ almost surely. Therefore $\mathscr{R}(x_k)$ is bounded a.s. And the stepsize we choose is square summable so $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$.
Therefore $\sum_{k=1}^{\infty} \alpha_k^2 \mathscr{R}(x_{k-1}) < \infty$ almost surely. Since z be any point in X^*, $\mathscr{R}(x_{k-1}) - \mathscr{R}(z) \geq 0$. The inequality (5.21) satisfies the requirements of Supermartingale Convergence Theorem 13. Therefore the nonnegative random sequence of $\{\|x_k - z\|\}$ converges almost surely to a random variable $v \geq 0$ for any $z \in X^*$. In other words the sequence created by the algorithm (5.2) converges to a point $z \in X^*$. As a result of the supermartingale convergence theorem

$$\sum_{k=1}^{\infty} \alpha_k (\mathscr{R}(x_{k-1}) - \mathscr{R}(z)) < \infty \quad a.s.,$$

meaning $\liminf_{k\to\infty} (\mathscr{R}(x_{k-1}) - \mathscr{R}(z)) = 0$ almost surely since the stepsize we choose is not summable, $\sum_{k=1}^{\infty} \alpha_k = \infty$. Since z is any point in the solution set $\mathscr{R}(z) = \mathscr{R}^*$. The function $\mathscr{R}(x)$ is a continous function. Since the sequence $\{\|x_k - z\|\}$ is convergent almost surely, the sequence created by the algorithm (5.2) $\{x_k\}$ is bounded. Therefore we can claim that $\lim_{k\to\infty} \{x_k\} \to z$ and z is a random point in the set X^*. As a result, the sequence generated by (5.2) $\{x_k\}$ converges and its limit point lies in X^* almost surely. □

5.3.3 Optimal One-Step Stepsize

In this section we want to find an optimal relaxation parameter at iteration k in order to bring x_k closer to any point $z \in X^*$. The stepsize α_k in algorithm (5.2) can be chosen in the interval of $[\varepsilon, 2-\varepsilon]$

$$\begin{aligned}\|x_k - z\|^2 &= \|(x_k - x_{k-1}) + (x_{k-1} - z)\|^2 = \|x_k - x_{k-1}\|^2 + \|x_{k-1} - z\|^2 + (x_k - x_{k-1})^T (x_{k-1} - z) \\ &= \alpha_k^2 \left\| -x_{k-1} + \sum_{i=1}^{m} w_i \Pi_{X_{\omega_{ik}}} [x_{k-1}] \right\|^2 + \|x_{k-1} - z\|^2 - 2\alpha_k \left(x_{k-1} - \sum_{i=1}^{m} w_i \Pi_{X_{\omega_{ik}}} [x_{k-1}] \right)^T (x_{k-1} - z).\end{aligned} \tag{5.22}$$

The equation above is quadratic in α_k. If any point $z \in X^*$ was known in advance we would be able to find an optimal stepsize. Since the target of convex feasibility algorithm (5.2) is to find the elements of X^* if it

is nonempty, the best possible option for stepsize can not be pinpointed. But we are going to try to find a range in $\varepsilon \leq \alpha_k \leq 2-\varepsilon$ for $\varepsilon > 0$.

Proposition 16. *The optimal stepsize for Algorithm* (5.2) *is* $\alpha_k^* \geq 1$.

Proof. The equality in (5.22) is minimized for

$$\alpha_k^* = \frac{\left(x_{k-1} - \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}]\right)^T (x_{k-1} - z)}{\left\| -x_{k-1} + \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] \right\|^2}. \tag{5.23}$$

We can write the numerator term of above equality as follows

$$\begin{aligned}
&\left(x_{k-1} - \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}]\right)^T (x_{k-1} - z) = \\
&\left(x_{k-1} - \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}]\right)^T \left(x_{k-1} - \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] + \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] - z\right) = \\
&\left\| x_{k-1} - \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] \right\|^2 + \left(x_{k-1} - \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}]\right)^T \left(\sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] - z\right).
\end{aligned}$$

Therefore the optimal stepsize is

$$\alpha_k^* = 1 + \frac{\left(x_{k-1} - \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}]\right)^T \left(\sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] - z\right)}{\left\| -x_{k-1} + \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] \right\|^2}.$$

Due to non-expansive property of Euclidean projection operation (5.8) and convexity of euclidean norm we have

$$\left(\sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] - z\right)^T (x_{k-1} - z) \geq \left\| \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] - z \right\|^2.$$

When we open up the inequality above we get

$$\begin{aligned}
&\left(\sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] - z\right)^T (x_{k-1} - z) \left\| \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] - z \right\|^2 = \\
&\left(x_{k-1} - \sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}]\right)^T \left(\sum_{i=1}^m w_i \Pi_{X_{\omega_{i_k}}}[x_{k-1}] - z\right) \geq 0.
\end{aligned}$$

Therefore the optimal stepsize for a single iteration is

$$\alpha_k^* \geq 1.$$

□

We have shown that the algorithm converges to the solution set if we use not summable but square summable stepsize. Based on result above the best stepsize option to ensure bringing a single iteration x_k point closest to any point z in the solution set X^* needs to be greater than 1.

5.3.4 Convergence Rate for Constant Stepsize

In this subsection we are going to explore the convergence rate of the algorithm for constant stepsize. We are going to start by building an auxiliary result that shows a bound on the function value of weighted averages.

Lemma 14. *Assume that the solution set X^* is nonempty. Let the sequence $\{x_k\}$ generated by the algorithm* (5.2)*. Lets define the weighted average as*

$$\hat{x}_t = \frac{1}{S_t}\sum_{k=1}^{t} \alpha_k x_{k-1} \quad with \quad S_t = \sum_{k=1}^{t} \alpha_k \quad for\ any\ t \geq 1.$$

Then the bound is

$$\mathsf{E}\left[\mathscr{R}\left(\hat{x}_t\right)\right] \leq \frac{1}{S_t}\mathsf{E}\left[dist^2\left(x_0, X\right)\right] + \sum_{k=1}^{t} \frac{\alpha_k^2}{S_t}\mathsf{E}\left[\mathscr{R}\left(x_{k-1}\right)\right].$$

Proof. We assumed that the solution set X is nonempty. So for any $z \in X$ and $\mathscr{R}(z) = 0$. We start by taking expectation of relation (5.20) conditioned on initial iteration. Then we obtain almost surely for any $z \in X$ and $k \geq 1$,

$$\mathsf{E}\left[\|x_k - z\|^2 \mid \mathscr{F}_0\right] \leq \mathsf{E}\left[\|x_{k-1} - z\|^2 \mid \mathscr{F}_0\right] + 2\left(\alpha_k^2 - \alpha_k\right)\mathsf{E}\left[\mathscr{R}\left(x_{k-1}\right) \mid \mathscr{F}_0\right].$$

We rearrange the terms and sum it from $k = 1$ to $k = t$ for some $t \geq 1$. So we have for all $z \in X$ and $t \geq 1$,

$$2\sum_{k=1}^{t}\left(\alpha_k - \alpha_k^2\right)\mathsf{E}\left[\mathscr{R}\left(x_{k-1}\right) \mid \mathscr{F}_0\right] \leq \|x_0 - z\|^2 - \mathsf{E}\left[\|x_t - z\|^2 \mid \mathscr{F}_0\right].$$

We already showed in Proposition 15 that for not summable but square summable stepsize the algorithm (5.2) converges to a point in solution set. Therefore if we choose t large enough x_t converges to a point

$z \in X$. And we can discard the last term in the preceding inequality. Then we take total expectation and we get

$$2\sum_{k=1}^{t}\left(\alpha_k - \alpha_k^2\right)\mathsf{E}\left[\mathscr{R}\left(x_{k-1}\right)\right] \leq \mathsf{E}\left[\mathrm{dist}^2\left(x_0, X\right)\right] \quad \text{for all } t \geq 1.$$

For $S_t = \sum_{k=1}^{t} \alpha_k$ and dividing the preceding inequality by $2S_t$, we further reach

$$\sum_{k=1}^{t}\frac{\alpha_k}{S_t}\mathsf{E}\left[\mathscr{R}\left(x_{k-1}\right)\right] - \sum_{k=1}^{t}\frac{\alpha_k^2}{S_t}\mathsf{E}\left[\mathscr{R}\left(x_{k-1}\right)\right] \leq \frac{1}{S_t}\mathsf{E}\left[\mathrm{dist}^2\left(x_0, X\right)\right] \quad \text{for all } t \geq 1.$$

As it can be observed that the terms $\dfrac{\alpha_k}{\sum_{k=1}^{t}\alpha_k}$, $k = 1, \ldots, t$ are convex weights while $\mathscr{R}(x)$ is convex. If we use average value $\hat{x} = \dfrac{1}{S_t}\sum_{k=1}^{t}\alpha_k x_{k-1}$ then for any $t \geq 1$,

$$\mathsf{E}\left[\mathscr{R}\left(\hat{x}_t\right)\right] \leq \frac{1}{S_t}\mathsf{E}\left[\mathrm{dist}^2\left(x_0, X\right)\right] + \sum_{k=1}^{t}\frac{\alpha_k^2}{S_t}\mathsf{E}\left[\mathscr{R}\left(x_{k-1}\right)\right].$$

□

The next proposition is going to provide error bounds on the performance of the algorithm (5.2) for a constant stepsize using Lemma 14.

Proposition 17. *Assume that problem* (5.1) *has a nonempty optimal set* X^*. *Let* $\{x_k\}$ *be the iterate sequence generated by the algorithm* (5.2). *Also let average of iterates to be* $\bar{x}_t = \dfrac{1}{t}\sum_{k=1}^{t} x_{k-1}$ *and the average of weighted iterates to be* $\hat{x}_t = \dfrac{1}{S_t}\sum_{k=1}^{t}\alpha_k x_{k-1}$. *If the stepsize is constant, i.e.,* $\alpha_k = \bar{\alpha}$ *and* $z \in X^*$ *the we have the following error bound for all* $t \geq 1$

$$\mathsf{E}\left[\mathscr{R}\left(\bar{x}_t\right)\right] \leq \frac{1}{\left(1 - \bar{\alpha} + \bar{\alpha}t\right)}\mathsf{E}\left[dist^2\left(x_0, X\right)\right].$$

Proof. Firstly we should note that for constant $\bar{\alpha}$ the weighted average and average of iterates are equal, $\hat{x}_t = \bar{x}_t$ with $S_t = \bar{\alpha}t$ for $t \geq 1$. Let $\alpha_k = \bar{\alpha}$ in Lemma 14 for all $t \geq 1$, and we get

$$\mathsf{E}\left[\mathscr{R}\left(\bar{x}_t\right)\right] \leq \frac{1}{\bar{\alpha}t}\mathsf{E}\left[\mathrm{dist}^2\left(x_0, X\right)\right] + \frac{\bar{\alpha}}{t}\sum_{k=1}^{t}\mathsf{E}\left[\mathscr{R}\left(x_{k-1}\right)\right]. \tag{5.24}$$

From the convexity of the function $\mathscr{R}$ we have

$$\mathsf{E}\left[\mathscr{R}\left(\frac{1}{t}\sum_{k=1}^{t}(x_{k-1})\right)\right] \leq \frac{1}{t}\sum_{k=1}^{t}\mathsf{E}\left[\mathscr{R}(x_{k-1})\right].$$

Therefore

$$\mathsf{E}\left[\mathscr{R}(\bar{x}_t)\right] \leq \frac{1}{(1-\bar{\alpha}+\bar{\alpha}t)}\mathsf{E}\left[\mathrm{dist}^2(x_0,X)\right]\frac{1}{(1-\bar{\alpha}+\bar{\alpha}t)}\mathsf{E}\left[\mathrm{dist}^2(x_0,X)\right].$$

□

From Proposition 17, for a fixed stepsize it can be deduced that the expected average iterate does not scatter away from the solution set. Although it looks like initial point affects the bound on the function value of expected average iterate and solution, as the number of iterations increases its contribution diminishes.

Using Proposition 17, we can provide a bound on the sum of the distances from the averages to the sets X_{ω_i}. The expected proximity function can be written as

$$\mathscr{R}(x) = \frac{1}{2}\sum_{i=1}^{M} w_i \mathtt{Pr}\{\omega = \omega_i\}\mathrm{dist}^2(x, X_{\omega_i}).$$

Lets define the minimum possibility of choosing a set i as follows

$$p_{\min} = \min_{1\leq i\leq m} \mathtt{Pr}\{\omega = \omega_i\}.$$

For any $x \in \mathbb{R}^n$ we have

$$\max_{1\leq i\leq m} \mathrm{dist}^2(x,X_i) \leq \sum_{i=1}^{M} w_i\, \mathrm{dist}^2(x,X_i) \leq \frac{2}{p_{\min}}\mathscr{R}(x).$$

Due to Markov's inequality we have

$$\mathtt{Pr}\{\max_{i\in\mathscr{M}} \mathrm{dist}^2(\hat{x}_t,X_i) \geq \varepsilon\} \leq \frac{1}{\varepsilon}\mathsf{E}\left[\max_i \mathrm{dist}^2(\hat{x}_t,X_i)\right]$$

and

$$\mathtt{Pr}\{\max_{i\in\mathscr{M}} \mathrm{dist}^2(\hat{x}_t,X_i) \geq \varepsilon\} \leq \frac{1}{\varepsilon}\frac{2}{p_{\min}}\frac{1}{(1-\bar{\alpha}+\bar{\alpha}t)}\mathsf{E}\left[\mathrm{dist}^2(x_0,X)\right]$$

For $t \geq 1$ and $x = \hat{x}$, we get

$$\mathsf{E}\left[\max_{i \in \mathscr{M}} \operatorname{dist}^2(\hat{x}_t, X_i)\right] \leq \frac{2}{p_{\min}} \mathscr{R}(\hat{x}_t).$$

We showed that when solution set X^* is nonempty the function value monotically decreases and converges to zero. In most basic sense maximum expected distance of average iterate to set i is bounded.

Chapter 6

RANDOM FEASIBILITY PROJECTION ALGORITHM: SIGNAL FEASIBILITY PROBLEM

We demonstrate the numerical behavior of our algorithm on a signal deconvolution problem, which is a signal estimation problem. Convolution is a mathematical operation on two functions f and g, producing a third function that is typically viewed as a modified version of one of the original functions, giving the area overlap between the two functions as a function of the amount that one of the original functions is translated. Computing the inverse of the convolution operation is called deconvolution.

6.1 Signal Deconvolution

The goal of this section is to find least-squares solution to a consistent/inconsistent convex set theoretic signal estimation problem. Signal estimation is to find a signal a^* that is feasible for a collection of $(S_i)_{1\leq i\leq \mathscr{M}}$ of sets such that

$$\text{Find} \quad a^* \in \cap_{i=1}^{\mathscr{M}} S_i \quad a^* \in \mathbb{R}^n, \tag{6.1}$$

where S_is are closed and convex.

Convex feasibility representation has been applied to a wide range of signal processing problems such as image enhancement Oh et al. (1993), image restoration Trussell and Civanlar (1984), image reconstruction Herman (1979), signal deconvolution Combettes (1994) and signal recovery from bispectrum information Cetin (1991). The survey work by Combettes (1993) presents an overview of theory and applications in this field.

The conventional approach of set theoretic estimation problems provide solution that confirm with constraint sets known a priori or observed. But it fails to take into account that sets built on a priori or observed data may carry disturbances or have erroneously predicted statistical information, which may result in inconsistent sets, where $\cap_{i=1}^{\mathscr{M}} S_i = \emptyset$. The attributes of original signal such as amplitude bound, region of support,

band-limitedness that are used to built sets in estimation problems may not be accurate. Additionally sets that are built using moments, spectral properties, distribution and bounds information are based on predicted stochastic estimations. The overly conservative confidence bounds or statistical assumptions may cause inconsistencies. Also noise pertubations in measurements or random variations in the impulse response of a system can cause inconsistencies. For signal design problems such as Goldburg et al. (1985) the definitions of constraint sets are left to users and this makes signal design problems more prone to inconsistent sets. Although the algorithm we propose (5.2) looks like it is based on finite number of sets, actually when you take into account the additive noise, u, accompanying the sets you have infinitely many options to choose from for each set.

We solve a signal deconvolution problem, which is set theoretic signal estimation to produce a signal h^* that satisfies a collection of constraints. We restore blurred and noise corrupted one dimesional discrete signal of $N = 64$. Essentially it is finding a signal that minimizes a weighted average of squares of the distances to constraint sets. This problem arises in a wide range of applications in medical imaging field. The set of signals consistent with a particular piece of information is called a property set. The set of feasible signals is the intersection of all property sets. In some cases there is no signal h^* that satisfies all the sets. Then we are looking for best possible signal that minimizes the expected proximity function (5.1). When a probabilistic description of uncertainty is not available but only the bounds on them are available then our framework for deconvolution is useful.

The model of an linear-shift invariant (LSI) imaging system can be defined as

$$x = T(h) + u, \tag{6.2}$$

where T is the blurring process and u is an additive noise component. This basic representation is suitable for most of the practical applications such as low-passed Fourier transform in band-limited extrapolation and Radon transform in tomography. A Gaussian blur is convolving an image/signal by a Gaussian function. Gaussian blur reduces the image's high frequency components hence it is a low pass filter. For a linear-shift invariant (LSI) system transfer function/system response function completely describes system behavior. Systems are linear if superposition holds and shift invariant if an input is delayed τ seconds then the output is delayed τ seconds, but the shape of the output depends only on the shape of the input and does not change with time.

We solve the discrete signal deconvolution/estimation problem of a noisy discrete-time $N = 64$ point signal. We estimate the original form of the input signal, $h \in \cap_i^m X_i$, $\{X_i \mid i \in \mathscr{M}\}$, which was passed through a linear shift-invariant system and further degraded by addition of noise. We do not have a single estimate but rather a set of possible inputs which are consistent with all the information available if such a solution set is nonempty. Output carries noise component. The additive measurement noise is bounded. The convolution kernel has unknown variations. No statistical apriori is needed for noise sets. The problem is

$$x = L * h + u, \tag{6.3}$$

where L is $N \times N$ matrix models a shift-invariant linear blur and u is a vector of bounded noise samples with $|u_i| \leq 0.15$. h is the input signal and x is the output signal. There is uncertainity in the kernel of the linear system. And there is additive measurement noise. The noise signal is bounded by a known function. The blurring kernel is a Gaussian function with a variance of 2 samples^2. We do not necessarily require $\cap_{i=1}^m X_i, \neq \emptyset$. When incompatible constraints are present then $\cap_{i=1}^m X_i, = \emptyset$. Then the weighted average of the squares of distances to constraint sets does not take the value of zero. The expected proximity function is defined as weighted least squares feasibility problem, which indicates the degree of unfeasibility of a signal, a. The solution set is defined as $X_0 = \cap_{i=1}^m X_i$.

The pixel size of a typical digital radiography or X-ray image is $2048 \times 2048 \times 12$, digital mammography is $4000 \times 5000 \times 12$ that require in the order of 200 million sets. Therefore projecting on complete collection of constraint set cyclicly or simultanously over and over again might be computationaly overwhelming. But projecting on a random subset of constraints and estimating the close enough signal might be a favorable approach. Even if the signal feasibility problem is consistent, minimizing the proximity function closely enough to zero function value is adequate in most cases. Randomly choosing a subcollection of sets and projecting on them for convex feasibility problems eases the high dimensionality of data that leads to burdensome computations. That is why we would like to test our algorithm against a noisy image/signal recovery problem.

The nth sample of the degraded signal is given by

$$x_n = \sum_{k=-l}^{l} k * h_{n-k} + u_n, \tag{6.4}$$

where original signal $h = (h_l, \ldots, h_q)$ with some blurring kernel $h = (t\_l, \ldots, t_l)$

A discrete-time signal may be a finitelength or an infinite-length sequence. The length of the signal is $N = 64$ as in Combettes (1994).

The problem consists of $m = 66$ closed and convex sets, which have been used in various theoretic signal processing applications by Combettes and Trussell (1991), Youla and Webb (1982), Levi and Stark (1983), Trussell and Civanlar (1984). The first 64 of them, $(S_i)_{1 \leq i \leq N}$ are constructed based on the knowledge of the blurring operator, L and the information that the components of the noise vector u are bounded by $|u_i| \leq 0.15$. They are hyperslabs defined as

$$S_i = \{a \in \mathbb{R}^N \mid x_i - u_i \leq \langle L_i \mid a \rangle \leq x_i + u_i\} \quad 1 \leq i \leq N, \tag{6.5}$$

where L_i is the ith row of L.

The projection of a signal a onto S_i is given by Trussell and Civanlar (1984)

$$\Pi_i = \begin{cases} a + \left[(x_i + u_i - \langle L_i \mid a \rangle) / \|L_i\|^2 \right] L_i{}^T, & \text{if } \langle L_i \mid a \rangle > x_i + u_i \\ a + \left[(x_i - u_i - \langle L_i \mid a \rangle) / \|L_i\|^2 \right] L_i{}^T, & \text{if } \langle L_i \mid a \rangle < x_i - u_i \\ a, & \text{otherwise.} \end{cases} \tag{6.6}$$

The next set is constructed by using the phase angle information of original signal, h. In practical applications phase information of h is not known but it is assumed. If A denotes the discrete Fourier transform of the signal a, the set number $i = 65$ is

$$S_{m-1} = \{a \in \mathbb{R}^N \mid (\forall k \in \{1, \ldots, N\}) \quad \angle A(k) = \angle H(k)\} \tag{6.7}$$

The projection $\Pi_{m-1}(a) = b$ of a signal a onto S_{m-1} for every k in $\{1, \ldots, N\}$ is as follows

$$B(k) = \begin{cases} |B(k)| \cos(\angle A(k) - \angle H(k)) \exp(i \angle H(k)), & \text{if } \cos(\angle A(k) - \angle H(k)) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{6.8}$$

This closed-form projection operation was given in Youla and Webb (1982).

The last set, $i = 66$ depends on amplitude information of original signal, h. The components of h are

nonnegative and bounded by 12. This leads to the bounded set

$$S_m = \{a \in \mathbb{R}^N \mid (\forall k \in \{1,\ldots,N\}) \quad 0 \leq a_i \leq 12\}. \tag{6.9}$$

The projection of a signal a onto S_m is given by $\Pi_m(a) = b$ where for every i in $\{1,\ldots,N\}$

$$b_i = \begin{cases} 0, & \text{if } a_i < 0 \\ 12, & \text{if } a_i > 0 \\ a_i & \text{otherwise.} \end{cases} \tag{6.10}$$

We concluded introducing the problem and the sets. Hereafter we are going to explain the details of implementation.

6.2 Random Projection Feasibility Algorithm Implementation

1. **Initialization**

   (a) Read original $N = 64$ point discrete signal, h which is shown in Figure 6.1.

   (b) Create the degraded signal by first blurring the original signal. The first step is constructing a $N = 64$ point Gaussian window in the column vector w. The coefficients of a Gaussian window are computed from the following equation.

$$w(n) = e^{-\frac{1}{2}\left(\alpha\frac{n}{N/2}\right)^2}$$

   where $-\frac{N-1}{2} \leq n \leq \frac{N-1}{2}$. α is inversely proportional to the standard deviation, σ of a Gaussian random variable where

$$\alpha = \frac{N}{2\sigma}.$$

   The Gaussian window vector, w is normalized and then the resulting vector is convolved with the original signal, h. When you slide the window along the original signal, the resulting vector is longer than the original signal by the length of the sliding window. This is called the "edge effect". The center N elements are the blurred vector. Edges are discarded.

   The same blurred vector can be calculated by using convolution kernel built with the gaussian function with $N = 64$ discrete points and a variance of 2 samples^2. We need to build the $N \times N$ matrix L because the first $i = 64$ sets are based on information about this shift-invariant linear blur matrix and bounded noise. At this point we would like to remind the definition of shift-invariance and explain how we built the L matrix.

   The *Kronecker delta* function is defined as follows

$$\delta(i) = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{otherwise.} \end{cases}$$

   And a *shift matrix* is defined as $S^k = \delta(i - j - k \bmod N)$. If for matrices, A and B it is the case that $AB = BA$ then we say that the matrix product commutes.

Any $N \times N$ circulant matrix product with each S^n for $1 \leq n \leq N$ commutes. A circulant matrix is defined as $A(i,j) = A(i+n \bmod N, j+n \bmod N)$. A circulant matrix, $N \times N$ is

$$\begin{bmatrix} a_0 & a_1 & \dots & a_{N-1} \\ a_{N-1} & a_0 & \dots & a_{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ a_1 & a_2 & \dots & a_0 \end{bmatrix}$$

When $n = -j$ a circulant matrix is $A(i,j) = A(i-j+ \bmod N, 0)$. Therefore the discrete convolution $\mathbf{y} = \mathbf{Ax}$ can be written as

$$y_i = \sum_{j=0}^{N-1} A(i,j)x_j = \sum_{j=0}^{N-1} A(i-j \bmod N, 0)x_j.$$

Then let

$$h(i-j \bmod N) = A(i-j \bmod N, 0).$$

and

$$y_i = \sum_{j=0}^{N-1} h(i-j \bmod N)x_j.$$

We say that $\mathbf{y}$ is the *discrete periodic convolution* of $\mathbf{h}$ and $\mathbf{x}$;

$$y = h * x.$$

For example the weight matrix for a *1D* artificial retina filter is

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The convolution kernel is the first column of the weight matrix. The convolution/blurring kernel

of L matrix is gaussian function with $N = 64$ discrete points and a variance of 2 samples^2. And by circularly shifting one element for each column we created the $L_{64\times 64}$ matrix. By multiplying h by L you can also get the blurred vector.

The blurred vector is further degraded by additon of random bounded noise of $|u_i| \leq 0.15$. The resulting degraded signal is shown in Figure 6.2.

2. **Monte Carlo Simulations**

Computational algorithms that rely on repeated random sampling to obtain numerical results should have enough samples to inspect the performance of the method. The algorithm we propose has also two random sampling in two segments of the process. First the sets X_i, $i = 1, \ldots, 64$ contain random noise elements. While projecting on them at each iteration there is going to be one random realization of u_i for each set. Second source of randomness is choosing a random subcollection of m sets at each iteration. We want to mention that for comparison purposes we also project on entire collection of sets as if it is like a parallel projection method, PPM. We gradually decreased the number of sets chosen. Our goal is to measure the trade off between iteration number and convergence rate while we decrease the number of chosen sets. In order to have 95% confidence interval on our results we followed a Monte-Carlo approach. And for each scenario we rely 101 different convergence paths on repeated random sampling.

3. **Algorithm Convergence Process**

For each Monte-Carlo iteration the algorithm runs its course until there is negligible improvement in the decrease of $\mathscr{R}$ is observed, i.e. whenever the stopping criterion is met.

(a) Convex feasibility algorithm (5.2) loop is initialized with the degraded signal, $a_0 = x$.

(b) Strictly convex equal weights (w_i) for $1 \leq i \leq m$ are chosen. Initially projection on complete collection of sets are done. For further simulation cases the number of sets chosen is decreased. Weights (w_i) are calculated accordingly.

(c) Various different stepsize options are tested for the algorithm convergence loop. We started with diminishing stepsizes, $\alpha_k = a/k^b$. Later we used stepsize, which is used by Combettes (1994) in order to roughly compare his Parallel Projection Algorithm (PPM) results with our algorithm

(5.2) performance. Combettes (1994) updates the stepsize according to rule below.

$$\text{If } \mathscr{R}(x_k) - \mathscr{R}(x_{k+1}) < \alpha_k \frac{\|\nabla \mathscr{R}(x_k)\|^2}{2}, \text{then set } \alpha_k = 0.75\alpha_k, \tag{6.11}$$

where $\alpha_0 = 1.999$. We should mention a significant difference between the implementation of PPM in our setup and Combettes (1994). Even if we choose m sets all at once, our results are not comparable to Combettes (1994) work exactly because the sets defined by Combettes (1994) are in deterministic nature as opposed to ours being stochastic.

(d) We project the iterate x_k onto randomly chosen sets. Initially projection onto complete collection of $m = 66$ sets is done as if it is a parallel projection algorithm. Combettes (1994) implements the parallel projection algorithm for the same problem with some important differences. Firstly the projection onto the sets $i = 1, \ldots, 64$, which are built on the knowledge of the blurring operator L and bounded noise vector u are deterministic in Combettes (1994). In reality bounded noise of u means that signal can vibrate between $[x_i - u_i, x_i + u_i]$. That is why we applied at each iteration random realizations of noise to sets $i = 1$ through 64. Later subcollection of sets of 44 and 33 were randomly chosen. The resulting function values $\mathscr{R}(x_k)$ and recovered signals h^* were recorded.

(e) The goal of the our algorithm (5.2) is to obtain an approximate minimum of the function $\mathscr{R}$ in a finite number of steps. The near convergence of the algorithm is measured by the stopping criteria below.

$$\mathscr{R}(x_k) - \mathscr{R}(x_{k+1}) \leq \varepsilon$$

for a small positive value ε.

6.2.1 Figures and Tables

From figures below it is seen that most feastures of h have been fairly well recovered. The limiting function values in Figure 6.4 and 6.6 are fairly close to results of Combettes (1994) which is also replicated here. Since Combettes (1994) projects onto larger sets by setting $u_i = 0.15$ he was able to achieve slightly lower limiting function values. Another important point that we need to mention is that even if the required number of iterations increases slightly by decreasing the number of sets chosen, we were still able to recover the

original signal adequately.

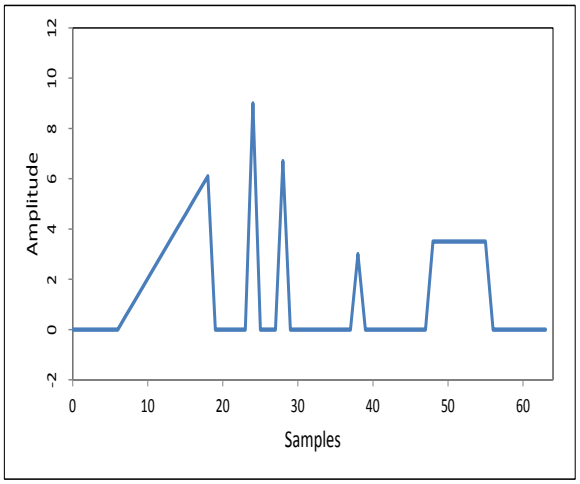

Figure 6.1: Original Signal, h

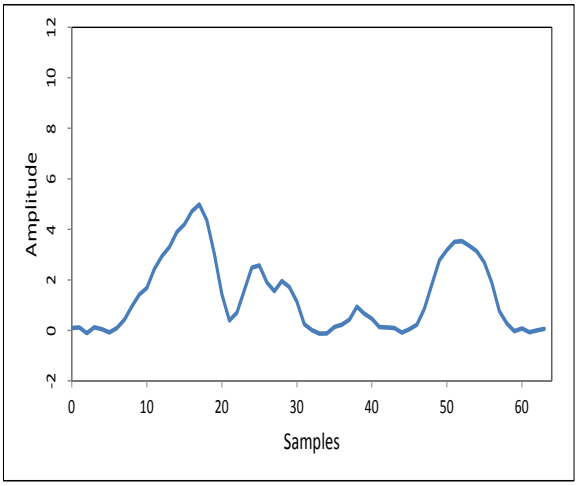

Figure 6.2: Degraded Signal, x=L*h+u

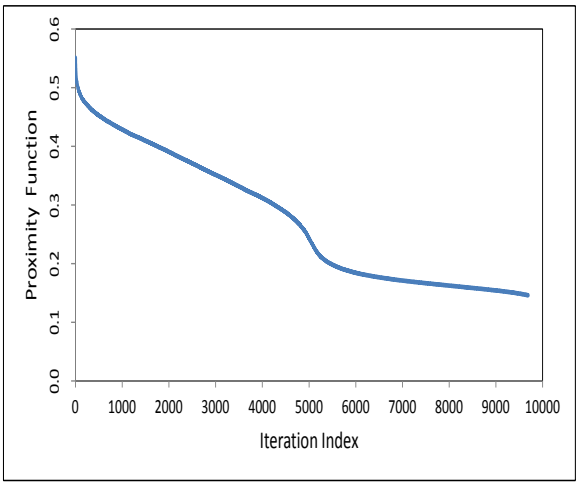

Figure 6.3: Proximity Function, $\mathscr{R}$,
$\alpha_k = 1/k^{0.75}$, X_i, $1 \leq i \leq 66$

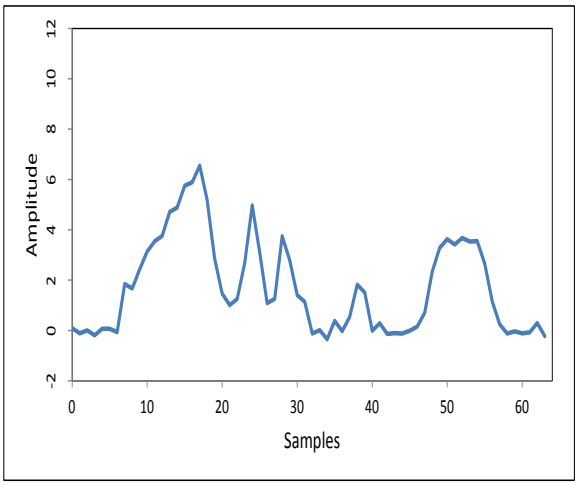

Figure 6.4: Recovered Signal, h^*,
$\alpha_k = 1/k^{0.75}$, X_i, $1 \leq i \leq 66$

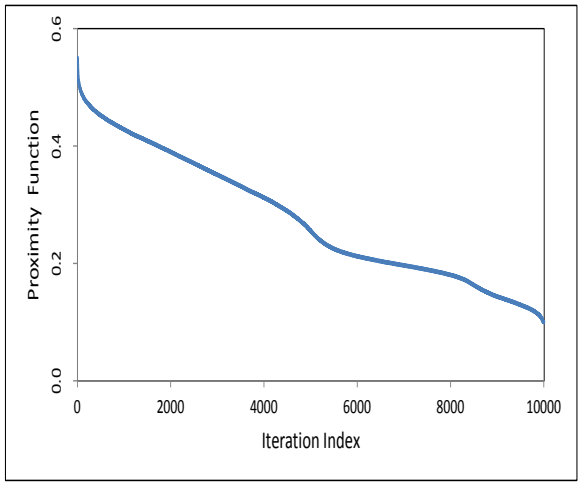

Figure 6.5: Proximity Function, $\mathscr{R}$,
$\alpha_k = 1/k^{0.75}$, X_i, $1 \leq i \leq 44$

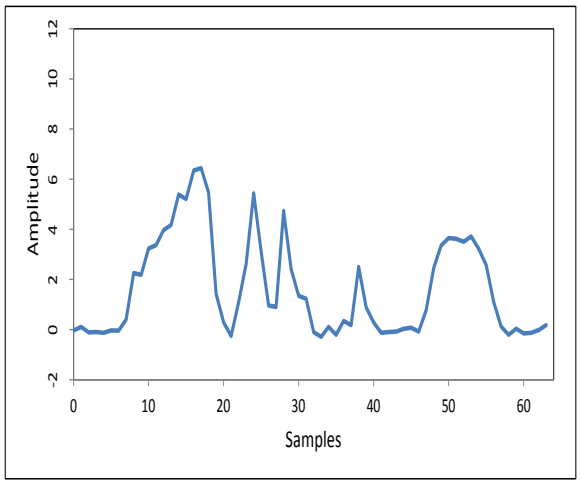

Figure 6.6: Recovered Signal, h^*,
$\alpha_k = 1/k^{0.75}$, X_i, $1 \leq i \leq 44$

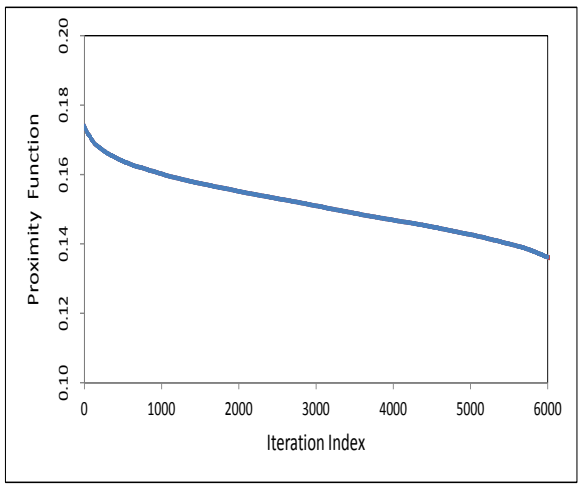

Figure 6.7: Combettes' PPM,
Proximity Function, $\mathscr{R}$,
Stepsize (6.11), X_i, $1 \leq i \leq 66$

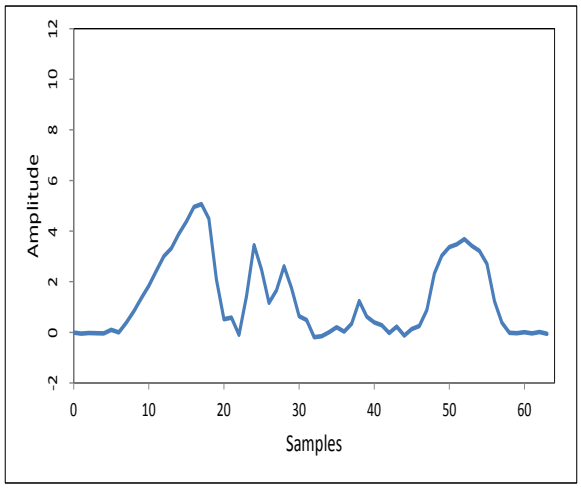

Figure 6.8: Combettes' PPM,
Recovered Signal, h^*,
Stepsize (6.11), X_i, $1 \leq i \leq 66$

REFERENCES

S. Agmon. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6(3):382–392, 1954.

R. Aharoni and Y. Censor. Block-iterative projection methods for parallel computation of solutions to convex feasibility problems. *Linear Algebra and Its Applications*, 120:165–175, 1989.

T. Alamo, R. Tempo, and E. F. Camacho. Randomized strategies for probabilistic solutions of uncertain feasibility and optimization problems. *Automatic Control, IEEE Transactions on*, 54(11):2545–2559, 2009.

H. C. Andrews and B. R. Hunt. Digital image restoration. *Prentice-Hall Signal Processing Series*, 1, 1977.

R. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.

J. Aubin. *Optima and equilibria: an introduction to nonlinear analysis*. Springer Verlag, 1998.

H. Bauschke and J. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM review*, 38(3): 367–426, 1996.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton Univ Pr, 2009.

D. Bertsekas and I. Rhodes. Recursive state estimation for a set-membership description of uncertainty. *Automatic Control, IEEE Transactions on*, 16(2):117 – 128, 1971.

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

D. Bertsekas, A. Nedić, and A. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, Cambridge, Massachusetts, 2003.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge Univ Pr, 2004.

L. Bregman. Determination of a common point of convex sets by the method of successive projection. In *Akademia Nauk, SSSR, Doklady*, volume 162, pages 487–490, 1965.

L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *Zh. Vychisl. Mat. & Mat. Fiz.*, 7:620–631, 1967.

F. E. Browder. Convergence theorems for sequences of nonlinear operators in banach spaces. *Mathematische Zeitschrift*, 100(3):201–225, 1967.

G. C. Calafiore. Random convex programs. *SIAM Journal on Optimization*, 20(6):3427–3464, 2010.

T. Capricelli and P. Combettes. A convex programming algorithm for noisy discrete tomography. In *Advances in Discrete Tomography and Its Applications*, pages 207–226. Springer, 2007.

Y. Censor. Iterative methods for the convex feasibility problem. *Ann. Discrete Math*, 20:83–91, 1984.

Y. Censor. Parallel application of block-iterative methods in medical imaging and radiation therapy. *Mathematical programming*, 42(1):307–325, 1988.

Y. Censor and G. T. Herman. On some optimization techniques in image reconstruction from projections. *Applied Numerical Mathematics*, 3(5):365–391, 1987.

Y. Censor and A. Lent. Cyclic subgradient projections. *Mathematical Programming*, 24(1):233–235, 1982.

A. E. Cetin. An iterative algorithm for signal reconstruction from bispectrum. *Signal Processing, IEEE Transactions on*, 39(12):2621–2628, 1991.

G. Cimmino. *Approximate calculation for the solutions of systems of linear equations*. Institute for application of the calculation, 1938.

P. Combettes. The convex feasibility problem in image recovery. *Advances in Imaging and Electron Physics*, 95: 155–270, 1996.

P. L. Combettes. The foundations of set theoretic estimation. *Proceedings of the IEEE*, 81(2):182–208, 1993.

P. L. Combettes. Inconsistent signal feasibility problems: Least-squares solutions in a product space. *Signal Processing, IEEE Transactions on*, 42(11):2955–2966, 1994.

P. L. Combettes. Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections. *Image Processing, IEEE Transactions on*, 6(4):493–506, 1997.

P. L. Combettes and H. J. Trussell. The use of noise properties in set theoretic estimation. *Signal Processing, IEEE Transactions on*, 39(7):1630–1641, 1991.

J. Crockett and H. Chernoff. Gradient methods of maximization. *Pacific J. Math*, 5(1), 1955.

H. Curry. The method of steepest descent for nonlinear minimization problems. *Quart. Appl. Math*, 2(3):250–261, 1944.

V. Dem'yanov. Minimization of functions on convex bounded sets. *Cybernetics and Systems Analysis*, 1(6):77–87, 1965.

V. F. Dem'Yanov, L. V. Vasil'Ev, and T. Sasagawa. *Nondifferentiable optimization*. Springer, 1985.

F. Deutsch. *Approximation Theory, Spline Functions and Applications*, volume 356 of *NATO ASI Series*. Springer Netherlands, 1992.

P. P. B. Eggermont, G. T. Herman, and A. Lent. Iterative algorithms for large partitioned linear systems, with applications to image reconstruction. *Linear algebra and its applications*, 40:37–67, 1981.

L. El Ghaoui and G. Calafiore. Robust filtering for discrete-time systems with bounded noise and parametric uncertainty. *Automatic Control, IEEE Transactions on*, 46(7):1084–1089, 2002.

L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18:1035–1064, 1997.

I. Eremin. Fejér mappings and convex programming. *Siberian Mathematical Journal*, 10(5):762–772, 1969.

F. Facchinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*, volume I and II. Springer-Verlag New York, 2003.

A. Ferguson and G. Dantzig. The allocation of aircraft to routes-an example of linear programming under uncertain demand. *Management Science*, 3(1):45–73, 1956.

S. D. Flåm and J. Zowe. Relaxed outer projections, weighted averages and convex feasibility. *Bit*, 30(2):289–300, 1990.

L. Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, 1997.

K. Goebel and W. A. Kirk. *Topics in metric fixed point theory*, volume 28. Cambridge University Press, 1990.

M. Goldburg, R. Marks, et al. Signal synthesis in the presence of an inconsistent set of constraints. *Circuits and Systems, IEEE Transactions on*, 32(7):647–663, 1985.

A. Goldstein. Convex programming in Hilbert space. *Bull. Amer. Math. Soc*, 70(5):709–710, 1964.

G. Golub and C. Van Loan. An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, 17 (6):883–893, 1980.

L. Gubin, B. Polyak, and E. Raik. The method of projections for finding the common point of convex sets. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 7(6):1211–1228, 1967.

G. Herman. Image reconstruction from projections. *Real-Time Imaging*, 1(1):3–18, 1995.

G. T. Herman. Image reconstruction from projections. *Image Reconstruction from Projections: Implementation and Applications*, 1, 1979.

G. T. Hermann. Image reconstruction from projections. *The Fundamentals of Computerized Tomography*, 1980.

V. Jakubowich. Finite convergent iterative algorithm for solving system of inequalities. In *Dokl. Akad. Nauk. SSSR*, volume 166, pages 1308–1311, 1966.

W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26:189–206, 1984.

S. Kaczmarz. Approximate solution of systems of linear equations. *International Journal of Control*, 57(6):1269–1271, 1993.

L. Kantorovich. Approximate solution of functional equations. *Uspekhi Matematicheskikh Nauk*, 11(6):99–116, 1956.

A. Lent and H. Tuy. An iterative method for the extrapolation of band-limited functions. *Journal of Mathematical Analysis and Applications*, 83(2):554 – 565, 1981.

A. Levi and H. Stark. Signal restoration from phase by projections onto convex sets. *JOSA*, 73(6):810–822, 1983.

E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1 – 50, 1966.

O. Mangasarian. Error bounds for nondifferentiable convex inequalities under strong slater constraint qualification. Technical report, University of Wisconsin, 1996.

T. Motzkin and I. Schoenberg. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6(3): 393–404, 1954.

A. Nedić. Random projection algorithms for convex set intersection problems. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 7655–7660. IEEE, 2010.

S. Oh, C. Ramon, R. Marks, A. Nelson, M. Meyer, et al. Resolution enhancement of biomagnetic images using the method of alternating protections. *Biomedical Engineering, IEEE Transactions on*, 40(4):323–328, 1993.

P. Oskoui-Fard and H. Stark. Tomographic image reconstruction using the theory of convex projections. *Medical Imaging, IEEE Transactions on*, 7(1):45–58, 1988.

B. Polyak. *Introduction to optimization*. Optimization Software, Inc., New York, 1987.

B. Polyak. Random algorithms for solving convex inequalities. In D. Butnariu, Y. Censor, and S. Reich, editors, *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, pages 409–422. Elsevier, Amsterdam, Netherlands, 2001.

B. T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864 – 878, 1963.

H. Robbins and D. Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. *Optimizing methods in statistics*, pages 233–257, 1971.

N. Shor. *Nondifferentiable optimization and polynomial problems*, volume 24. Springer, 1998.

N. Z. Shor, K. C. Kiwiel, and A. Ruszczynski. *Minimization methods for non-differentiable functions*, volume 3. Springer-Verlag Berlin, 1985.

G. Stewart and J. Sun. *Matrix perturbation theory*, volume 175. Academic press New York, 1990.

H. Trussell and M. Civanlar. The feasible solution in signal restoration. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(2):201–212, 1984.

M. Vainberg. On the convergence of the method of steepest descent for nonlinear equations, Sibirsk. *Math. Z*, 2: 201–220, 1961.

S. Van Huffel and J. Vandewalle. *The total least squares problem: computational aspects and analysis*, volume 9. Society for Industrial and Applied Mathematics, 1987.

J.-P. Vial. Strong convexity of sets and functions. *Journal of Mathematical Economics*, 9(1-2):187 – 205, 1982.

R. Vinter. *Optimal control*. Springer, 2010.

D. C. Youla and H. Webb. Image restoration by the method of convex projections: Part 1 theory. *Medical Imaging, IEEE Transactions on*, 1(2):81–94, 1982.

K. Zhou, J. Doyle, and K. Glover. *Robust and optimal control*. Prentice Hall Englewood Cliffs, NJ, 1995.

APPENDIX

Lemma 15. *For the stepsize that is not summable but square summable such that* $\sum_{k=1}^{\infty}\alpha_k = \infty$ *and* $\sum_{k=1}^{\infty}\alpha_k^2 < \infty$*, if stepsize and variance of error term satisfies* $lim_{k\to\infty}\alpha_k = \hat{\alpha} \geq 0$ *and* $lim_{k\to\infty}\nu_k = \hat{\nu} \geq 0$ *with* $\bar{\alpha} = max_k\alpha_k$ *and* $\bar{\nu} = max_k\nu_k$ *then for any* $M > K$ $limsup_{M\to\infty}\dfrac{\sum_{k=1}^{M}\nu_k\alpha_k^2}{\sum_{k=1}^{M}\alpha_k} \leq (\bar{\nu}+\varepsilon)\hat{\alpha}$*, where* $\bar{\alpha} > 0$ *and* $\bar{\nu} > 0$ *are some scalars for all* $k \geq 1$.

Proof. Let the stepsize be not summable but square summable such that $\sum_{k=1}^{\infty}\alpha_k = \infty$ and $\sum_{k=1}^{\infty}\alpha_k^2 < \infty$ for some scalar $\bar{\alpha} > 0$ and all $k \geq 1$. If the stepsize satisfies $\lim_{k\to\infty}\alpha_k = \hat{\alpha} \geq 0$ and $\sum_{k=1}^{\infty}\alpha_k = \infty$, and $\lim_{k\to\infty}\nu_k = \hat{\nu} \geq 0$ with $\bar{\alpha} = \max_k\alpha_k$ and $\bar{\nu} = max_k\nu_k$. Since $\lim_{t\to\infty}S_t = \infty$ and $\lim_{t\to\infty}\dfrac{\sum_{k=1}^{t}\alpha_k^2}{S_t} = \hat{\alpha}$, For every $\varepsilon > 0$ there is a large enough K such that $\nu_k - \bar{\nu} \leq \varepsilon$

For any $M > K$,

$$\frac{\sum_{k=1}^{M}\nu_k\alpha_k^2}{\sum_{k=1}^{M}\alpha_k} = \frac{\sum_{k=1}^{K}\nu_k\alpha_k^2}{\sum_{k=1}^{M}\alpha_k} + \frac{\sum_{k=K+1}^{M}\nu_k\alpha_k^2}{\sum_{k=1}^{M}\alpha_k} \leq \frac{\sum_{k=1}^{K}\nu_k\alpha_k^2}{\sum_{k=1}^{M}\alpha_k} + (\bar{\nu}+\varepsilon)\frac{\sum_{k=K+1}^{M}\alpha_k^2}{\sum_{k=1}^{M}\alpha_k}$$

□