The Pennsylvania State University
The Graduate School

STOCHASTIC VARIATIONAL & HIERARCHICAL PROBLEMS: MODELS, ALGORITHMS, AND APPLICATIONS TO TRANSMISSION EXPANSION PROBLEMS

A Dissertation in
Industrial Engineering
by
Shisheng Cui

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2019

The dissertation of Shisheng Cui was reviewed and approved* by the following:

Uday V. Shanbhag
Gary and Sheila Bello Chair Professor of Industrial and Manufacturing Engineering
Dissertation Advisor, Chair of Committee

Mort D. Webster
Professor of Energy Engineering

Constantino M. Lagoa
Professor of Electrical Engineering

Huanan Zhang
Assistant Professor of Industrial and Manufacturing Engineering

Benjamin F. Hobbs
Theodore M. and Kay W. Schad Professor of Environmental Health and Engineering, Johns Hopkins University
Special Member

Janis P. Terpenny
Peter and Angela Dal Pezzo Department Head and Professor of Industrial and Manufacturing Engineering
Department Head

*Signatures are on file in the Graduate School.

Abstract

This dissertation comprises of three main essays, all of which fall under the umbrella of models and algorithms for stochastic variational and hierarchical problems.

(i). Extragradient-type schemes for stochastic variational inequality problems. When solving stochastic variational inequality problems, the classical extragradient method for merely monotone problems is complicated by a key challenge: the scheme requires two projections on what could be a complicated convex set and two evaluations of the map for every major iteration. We consider two related avenues where every iteration requires only a single projection: (i) A stochastic projected reflected gradient (**SPRG**) method requiring a single evaluation of the map and a single projection; and (ii) A stochastic subgradient extragradient (**SSE**) method that requires two evaluations of the map and a single projection. We prove the almost-sure convergence of the iterates to a random point in the solution set for both schemes under suitable requirements and prove that variance-reduced counterparts achieve the canonical deterministic rates of convergence. To contend with complex feasibility sets given by an intersection of a large number of convex sets, we provide rate statements for a variant of our scheme where the projection is replaced by a random projections.

(ii). Sampling-based Proximal and splitting schemes for monotone stochastic inclusion problems. Next, we consider a stochastic generalized equation with monotone operators, a class of problems that subsumes convex stochastic optimization problems as well as subclasses of convex Nash games and variational inequality problems. A direct application of proximal and splitting schemes are complicated by the need to resolve problems with expectation-valued maps at each step, a concern that is addressed by using sampling. Accordingly, we propose two avenues for addressing uncertainty in the mapping. (i) *Stochastic proximal point methods* (**SPP**). We develop amongst the first variance-reduced stochastic proximal point scheme that achieves deterministic rates of convergence in terms of solving proximal-point problems. Notably, the presented schemes achieve deterministic rates of convergence and

oracle complexity bounds are provided; (ii) *Stochastic modified forward-backward splitting scheme* (**SMFBS**). In settings, where the map is a sum of two maps, of which the first is an expectation-valued map and the second has a cheap resolvent. In such cases, we prove that variance-reduced splitting schemes display deterministic rates of convergence under maximal monotonicity and optimal oracle complexities. We proceed to show that both schemes achieve the optimal (deterministic) rates of convergence and provide oracle complexity guarantees.

(iii). Competitive transmission expansion under uncertainty. Finally, we consider a problem of competitive transmission expansion framework under uncertainty in the context of power systems planning. This work is motivated by the need to consider transmission charges and their impact on subsequent generation expansion decisions and the resulting social welfare in a competitive environment, as captured by a market equilibrium problem. We model this hierarchical problem as a mathematical program with equilibrium constraints complicated by the presence of first-stage binary expansion decisions. Empirical studies suggest that by smoothing the problem allows for improving the efficiency of computing near-global solutions in a fraction of the time taken by standard branching schemes. Preliminary insights from the model show that using charges such as MW-miles leads to lower social welfare compared to that obtained from either the flat rate charge or no charge.

Table of Contents

List of Figures

List of Tables

List of Symbols

Sets and Indices

N	Nodes indexed by i, j
L	Transmission lines indexed by l
L_E	Existing lines
L_C	Candidate lines
Ω_l	Ordered pair of nodes connected by line l
T	Time periods indexed by t
G	Generator type indexed by g
KVL_v	Set of lines in a voltage loop indexed by v

Parameters

$CK_{g,i}$	Annualized capital cost of g at node i [$/MW/y]
CT_l	Annualized capital cost of line l [$/MW/y]
Rev	Annual grid O&M costs [$/y]
Ps	Revenue split ratio for demand
$FOR_{g,i,t}$	Force outage rate for g at node i, time t
CL_l	"Cost-adjusted" length of line l [$/mile/y]
F_l	Line capacity of line l [MW]
S_l	Line susceptance of line l [p.u.]
M	Large positive number

$A_{i,l}$	Elements of network incidence matrix A of all lines (both candidate and existing lines)
X	Diagonal matrix containing the reactances of all lines (both candidate and existing lines)
$MC_{g,i}(q)$	Marginal cost for generator g at node i
H_t	Length of time period t [Hours]
$DEM_{i,t}$	Linear inverse demand curve of i at t, defined by an intercept and slope
$GSCALE_g$	Scaling factor of transmission charge for g
INT_i	Intercept of the inverse demand curve at i
SLP_i	Slope of the inverse demand curve at i

Decision Variables

$f_{l,t}$	Power flow on line l at time t
$q_{g,i,t}$	Generation of g at i, t [MW]
$k_{g,i}$	Capacity of generator g at node i [MW]
y_l	Transmission investment decision for line l, $y_l = 1 \quad \forall l \in L_E$
$\theta_{i,t}$	Phase angle i, t [radian]
$a_{i,t}$	ISO's arbitrage (net injection) at i, t [MW]
$d_{i,t}$	Consumer's demand at i, t [MW]
$f0_l^+$	Positive part of power flow on l for transmission charge evaluation [MW]
$f0_l^-$	Negative part of power flow on l for transmission charge evaluation [MW]
c_1, c_2	Revenue recovery coefficients
kt_i	Injection at node i when generation is scaled to the annual peak demand [MW]
$p_{i,t}$	Price at i, t [\$/MWh]
tl_l	Capacity of l [MW]
ic_i	Transmission charge at i [\$/MW/y]

Acknowledgments

I would like to express my special appreciation and sincere gratitude to my academic advisor Professor Uday V. Shanbhag, who has been a tremendous mentor for me. Uday is someone you will instantly love and never forget once you meet him. He has set an example of excellence as a researcher, mentor, instructor, and role model. I would like to thank him for encouraging my research and for allowing me to grow as a researcher. His advice on both research as well as on my career have been invaluable.

I wish to thank the members of my dissertation committee: Professor Benjamin F. Hobbs, Professor Constantino M. Lagoa, Professor Mort D. Webster and Professor Huanan Zhang for generously offering their time, support, guidance and good will throughout the preparation and review of this document. I also want to thank them for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions.

I am also grateful to Pengcheng Ding from Johns Hopkins University for his collaboration and contribution in the transmission expansion work. I really appreciate his impressive modeling skills, remarkable insights and rigorous attitude.

The best outcome from these past years is finding my best friend, soul-mate, and wife. I married the best person out there for me. Chunchun is the only person who can appreciate my quirkiness and sense of humor. She has been extremely supportive of me throughout this entire process and has unconditionally loved me during my good and bad times. She has been non-judgmental of me and instrumental in instilling confidence. These past several years have not been an easy ride, both academically and personally. I truly thank Chunchun for standing by me through all my absences, my fits of pique and impatience. I feel that what we both learned a lot about life and strengthened our commitment and determination to each other and to live life to the fullest.

To my family, I would like to thank them for encouraging me in all of my pursuits and inspiring me to follow my dreams. Thank you to my father, I always knew that he believed in me and wanted the best for me. Thank you to my mother for her endless love and guiding me as a person. I am especially grateful to my parents-in-law, who provided unconditional support for us.

I would also like to acknowledge the support of NSF Grants 1408366 and 1710974 as well as DOE ARPA-E 1952-1515.

Dedication

This dissertation is dedicated to my wife Chunchun, my parents and my lovely girl Emma.

Chapter 1 Introduction

The role of optimization in planning and operation has been crucial in the context of range of settings including supply-chain management, power systems, transportation, and telecommunication. Many of these problems have been complicated by uncertainty and discreteness, fueling the development of scalable and convergent algorithms to contend with the size and complexity of such problems. Yet, two crucial concerns complicate the resolution of both problems.

(i) *Competition:* While traditional models assume the presence of a single user or a centralized planner, increasingly such problems are complicated by the presence of competition and the relevant question is one of computing an equilibrium rather than an optimum. When overlaid by uncertainty, under suitable convexity requirements, the necessary and sufficient equilibrium conditions can be compactly stated as stochastic variational inequality problems, or more generally, stochastic generalized equations.

(ii) *Hierarchy:* In may settings, there is a notion of leadership in the decision-making framework with respect to a collection of followers. Such leadership may emerge from the timing of the decision-making or may be a consequence of informational asymmetries. A popular approach for capturing such a framework lies in formulation this problem as a mathematical program with equilibrium constraints, where the equilibrium constraints capture the optimal follower decisions, contingent on leader-level decisions.

This dissertation is devoted to the development of tools and techniques for the resolution to aforementioned problems. The remainder of this chapter is partitioned into three Sections, each of which corresponds to the three main essays in the dissertation. In Section 1.1, we consider the computation of solutions to monotone stochastic variational inequality problems by developing stochastic extragradient-type schemes reliant on a *single projection* (rather than two projections onto a complicated set) or on *random projections* (when the set is defined as an intersection on a large collection of convex sets). The more general problem of resolving stochastic generalized equations with strongly monotone and maximal monotone operators is addressed in Section 1.2, which are addressed via proximal-point and splitting schemes in deterministic regimes. Relatively less is known regarding variance-reduced sampling-based counterparts of such schemes and is the focus of Section 1.2 where convergence, rate, and complexity statements are provided. Finally, in Section 1.3, we consider the competitive transmission expansion problem, a rather challenging class of mixed-binary hierarchical optimization problems in which a transmission provider determines an expansion plan subsequent to the optimal reaction of the generation providers and the associated energy market. It may be observed that while the first two chapters focus on stochastic variational inequality problems and multi-valued generalizations (in the form of stochastic generalized equations), the third chapter is a hierarchical generalization in that it considers mathematical programs with equilibrium constraints. In effect, each chapter can be viewed as an essay under the broader umbrella of stochastic variational and hierarchical problems.

1.1 Stochastic Variational Inequality Problems

1.1.1 Variational Inequality Problems

The variational inequality (VI) problem is a fundamental object in optimization and variational analysis. The variational inequality has its roots in the Signorini problem that considered the equilibrium configuration of an elastic body on a frictionless plane [1]. This problem was subsequently announced by Fichera [2] who modeled this problem as a variational inequality problem and subsequently provided existence and uniqueness conditions in [3], marking the first instance of such a problem in the literature. The earliest work in the finite-dimensional regime setting considered the traffic network equilibrium problem[1] by Dafermos [4]. Recently, the study of variational inequality problems has been supported through a series of edited volumes and monographs [5–7]. Next we state a general definition for the variational inequality problem and an illustration is shown in Figure 1.1.

Definition 1. *Given a set $X \subseteq \mathbb{R}^n$ and a single-valued map $F : \mathbb{R}^n \to \mathbb{R}^n$, the variational inequality problem VI(X, F) requires finding a point $x^* \in X$ such that*

$$F(x^*)^T(x - x^*) \geq 0, \quad \forall x \in X. \tag{VI(X,F)}$$

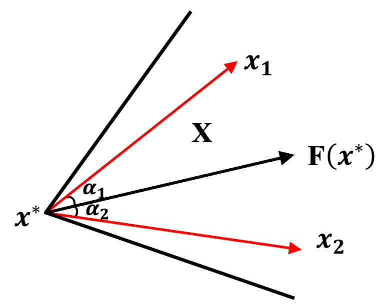


Figure 1.1: Variational inequality problem illustration with $\alpha_1, \alpha_2 \in [0, \frac{\pi}{2}]$

1.1.2 Applications

Variational inequality problems can capture a large number of problems in economics, engineering, and the applied sciences. Some of important applications include convex optimization problems, convex Nash games over continuous strategy sets, economic equilibrium problems, and complementarity problems, among others [6]. We briefly review several well-known optimization and equilibrium problems that can be captured by this object:

Convex optimization problems. Consider the convex optimization problem

[1]A pair of vectors $(f^*, d^*) \in K$ is an traffic network equilibrium pattern if and only if it satisfies the variational inequality problem

$$c(f^*)^T(f - f^*) - \lambda(d^*)^T(d - d^*) \geq 0, \quad \forall (f, d) \in K,$$

where f denotes the link flows, $c(f)$ denotes the user cost, d denotes the demand, $\lambda(d)$ denotes the travel disutility functions and K denotes the feasible set.

given by the following:

$$\min_{x \in X} f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function and $X \in \mathbb{R}^n$ is a closed and convex set. Then, by the minimum principle, any local solution of this problem is necessarily a global solution and must satisfy the following condition:

$$\nabla f(x^*)^T (x - x^*) \geq 0, \quad \forall x \in X.$$

Complementarity problems. Suppose X is a cone. Recall that X is a cone if $x \in X$ implies that $\lambda x \in X$ where λ is a nonnegative scalar. The complementarity problem, denoted by $\mathrm{CP}(X, F)$, requires an $x \in X$ such that

$$X \ni x \perp F(x) \in X^*, \qquad (\mathrm{CP}(X, F))$$

where X^* denotes the dual cone (Figure 1.2). In such an instance, $\mathrm{VI}(X, F)$ is equivalent to the complementarity problem $\mathrm{CP}(X, F)$ [8] in the sense that

$$x \text{ solves } \mathrm{VI}(X, F) \quad \Leftrightarrow \quad x \text{ solves } \mathrm{CP}(X, F).$$

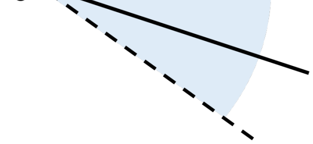


Figure 1.2: Dual cone illustration

Convex Nash games[2] Consider an N-player Nash game in which the ith solves

[2] A detailed example of stochastic Cournot game is provided in Section 2.5.1.

the following parametrized convex optimization problem:

$$\min_{x_i \in X_i} f_i(x_i; x_{-i}), \qquad (\mathrm{Player}_i(x_{-i}))$$

where $f_i(x_i; x_{-i})$ is a convex and continuously differentiable function in x_i, for every x_{-i}, and $X_i \in \mathbb{R}^{n_i}$ is a closed and convex set. A Nash equilibrium of this game is defined as a tuple $\{x_i^*\}_{i=1}^N$ such that

$$x_i^* \in \mathrm{SOL}(\mathrm{Player}_i(x_{-i}^*)), \quad i = 1, \ldots, N,$$

where $\mathrm{SOL}(\cdot)$ denotes solution to the corresponding problem. It is well known [6] that $x \triangleq \{x_i^*\}_{i=1}^N$ is a Nash equilibrium of this game if and only if x is a solution of $\mathrm{VI}(X, F)$, where

$$X \triangleq \prod_{i=1}^N X_i \quad \text{and} \quad F(x) \triangleq \begin{pmatrix} \nabla x_1 f_1(x) \\ \vdots \\ \nabla x_N f_N(x) \end{pmatrix}.$$

Finally, we comment on the relevance of equilibrium models in policy-making. Equilibrium models have played an important role in the the design of policy mechanisms. An early application of complementarity models for energy policy modeling was provided by Hogan [9] (also see [10]) while a recent tutorial [11] discussed the role of complementarity models on building policy models. Over the last several decades, complementarity-based models have assumed relevance in the context of electricity [12–19] and natural gas markets [20–22]. A review of complementarity modeling in energy markets was provided in the comprehensive monograph [22]. Such models are capable of capturing a range of competitive interactions including Nash [13, 19] and Stackelberg [14, 16], allow for modeling the impact of a range of market intricacies including sequential market clearings [16, 23] and price caps [18], amongst others. As there is a growing need to model uncertainty arising from demand and renewable supply, there has been an effort to model such problems in uncertain regimes [23–25]. Accordingly, this motivates the present efforts in the examination of stochastic variational inequality problems and their multi-valued generalizations (which may be modeled as stochastic generalized equations), the focus of the Chapters 2 and 3.

1.1.3 Stochastic Variational Inequality Problems

Uncertainty plays an increasingly significant role in modern systems and is inherent to any practical problem that relies on data. In most practical variational inequality problems, the stochastic generalization is of relevance as well. An introduction to the stochastic variational inequality problem may be found in [26] and such problems find application in a range of settings including power markets, communication networks, traffic networks, amongst others (cf. [24, 27–32]) as well. In the stochastic generalization, the components of the map F are expectation-valued; specifically $F_i(x) \triangleq \mathbb{E}[F_i(x, \xi(\omega))]$, where $\xi : \Omega \to \mathbb{R}^d$ is a random variable, $F_i : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is a single-valued function, and the $\mathbb{E}[\cdot]$ denotes the expectation and the associated probability space being denoted by $(\Omega, \mathcal{F}, \mathbb{P})$. In short, we are interested in a vector $x^* \in X$ such that

$$\mathbb{E}[F(x^*, \omega)]^T(x - x^*) \geq 0, \quad \forall x \in X, \tag{SVI(X, F)}$$

where $\mathbb{E}[F(x, \omega)] = \Big(\mathbb{E}[F_i(x, \omega)]\Big)_{i=1}^{N}$.

It may be less possible to capture certain application settings with the expected-value formulation. For instance, in many cases, the focus is on obtaining a solution to a variational inequality problem that is robust to parametric uncertainty. This leads to the almost-sure formulation of the stochastic variational inequality. Given a random mapping F, the almost-sure variational inequality, denoted by asVI, requires a vector $x^* \in X$ such that for almost every $\omega \in \Omega$,

$$F(x^*, \omega)^T(x - x^*) \geq 0, \quad \forall x \in X. \tag{asVI(X, F)}$$

It is highly unlikely that the problem admits a solution. If X is an n-dimensional cone, then asVI(X, F) reduces to asCP(X, F); this problem requires an x such that

$$0 \leq x \perp F(x; \omega) \geq 0 \quad \text{for almost every } \omega \in \Omega. \tag{asCP(X, F)}$$

One may recast this nonlinear CP as a system of equations by using the Fischer-Burmeister function. For a fixed but arbitrary realization $\omega \in \Omega$, the residual of this equation can be minimized as follows:

$$\min_{x \geq 0} \ \|\Phi(x;\omega)\|,$$

where $\Phi(x;\omega) \triangleq \left(\sqrt{x_i^2 + F_i(x;\omega)^2} - (x_i + F_i(x;\omega))\right)_{i=1}^n$. One avenue toward solving asCP(X, F) is by considering the expected residual minimization (ERM) problem [33, 34]:

$$\min_{x \geq 0} \quad \mathbb{E}[\|\Phi(x;\omega)\|]. \tag{ERM}$$

More specifically, x^* is a solution of asCP(X, F) if and only if x^* is a minimizer of (ERM) with $\mathbb{E}[\|\Phi(x;\omega)\|] = 0$.

Naturally, when the expectation of the mapping under uncertainty in instances is simple to evaluate (such as when the sample space Ω has finite cardinality), the resulting SVI(X, F) can be easily converted to its deterministic counterpart. Unfortunately, in most stochastic regimes, this evaluation is impossible to be expressed as closed form and it relies on a multidimensional integration which is quite time consuming. From a computational standpoint, two distinct ways are employed to solve SVIs. Of these, the first leverages sample-average approximation schemes [35]. In such an approach, the expectation is replaced by a mean value of a number of samples, and the effort is on the asymptotic and rate analysis of the resulting estimators, which are obtainable by solving a deterministic variational inequality problem [32, 36–38]. A natural concern is the development of confidence statements for such estimators. The other avenue was proposed first by Jiang and Xu [39], where a stochastic approximation scheme was developed for solving such stochastic variational inequality problems. Two regularized counterparts were presented by Koshal et al [40]. To overcome the shortcoming of standard stochastic approximation schemes where there is little information on the choice of step-length sequences, Yousefian et al. [41] developed distributed stochastic approximation schemes where users can independently choose a step-length rule. Our work focus on the second approach, that we develop stochastic approximation type algorithms for solving SVIs.

1.1.4 Solution Schemes

Amongst the simplest of schemes for deterministic VIs are analogs of the standard projection-based gradient schemes for convex optimization problems:

$$x_{k+1} := \Pi_X(x_k - \gamma_k F(x_k)), \tag{PG}$$

where $\Pi_X(y)$ denotes the projection of y onto X and γ_k denotes the steplength at each iteration. This method generally requires a strong monotonicity assumption on F to ensure convergence of the iterates. An extension was suggested by Antipin [42] and Korpelevich [43] that weakened the requirement on F to mere monotonicity and required 2 steps:

$$\begin{aligned} x_{k+\frac{1}{2}} &:= \Pi_X(x_k - \gamma F(x_k)), \\ x_{k+1} &:= \Pi_X\left(x_k - \gamma F\left(x_{k+\frac{1}{2}}\right)\right). \end{aligned} \tag{EG}$$

In this extragradient scheme (presumably associated with taking an "extra" gradient step), two projections were required at each iteration to obtain a new point and convergence was proved under the assumptions of Lipschitz continuity and mere monotonicity of the map F. Then we extend both schemes to be applied to SVI(X, F):

$$x_{k+1} := \Pi_X(x_k - \gamma_k F(x_k, \omega_k)). \tag{SPG}$$

Similarly, a stochastic counterpart to (EG) is (SEG) and is defined below:

$$\begin{aligned} x_{k+\frac{1}{2}} &:= \Pi_X(x_k - \gamma_k F(x_k, \omega_k)), \\ x_{k+1} &:= \Pi_X\left(x_k - \gamma_k F\left(x_{k+\frac{1}{2}}, \omega_{k+\frac{1}{2}}\right)\right). \end{aligned} \tag{SEG}$$

1.1.5 Main Work

Classical extragradient schemes and their stochastic counterpart represent a cornerstone for resolving monotone variational inequality problems. Yet, such schemes have a per-iteration complexity of two projections on a convex set and two evaluations of the map, the former of which could be relatively expensive if X is a complicated set. In Chapter 2, we consider two related avenues where the per-iteration complexity is significantly reduced: (i) A stochastic projected reflected gradient (**SPRG**) method requiring a single evaluation of the map and a single projection; and (ii) A stochastic subgradient extragradient (**SSE**) method that requires two evaluations of the map, a single projection, and a projection onto a halfspace (computable in closed form). Under suitable conditions, we prove almost sure (a.s.) convergence of the iterates to a random point in the solution set. Additionally, we show that under a variance-reduced framework, both schemes display a non-asymptotic rate of $\mathcal{O}(1/K)$, matching their deterministic counterparts. To address constraints with a complex structure, we prove that random projection variants of both schemes also display a.s. convergence while displaying a rate of $\mathcal{O}(1/\sqrt{K})$ in terms of the sub-optimality and infeasibility. Preliminary numerics support theoretical findings and the schemes outperform their standard extragradient counterparts in terms of the per-iteration complexity.

1.2 Stochastic Generalized Equations

1.2.1 Generalized Equations

In Chapter 3, we extend our work from Chapter 2 to developing solution methods to stochastic generalized equations which require finding a zero of a an expectation-valued set-valued mapping, defined next.

Definition 2. *Given a set-valued map $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$, the generalized equation (GE) requires finding a point $x^* \in \mathbb{R}^n$ such that*

$$0 \in T(x^*). \tag{GE}$$

We are interested in the stochastic version of (GE), which is finding an $x^* \in \mathbb{R}^n$ such that

$$0 \in \mathbb{E}[T(x^*, \omega)], \text{ where } \mathbb{E}[T(x^*, \omega)] \triangleq \begin{pmatrix} \mathbb{E}[T_1(x^*, \omega)] \\ \vdots \\ \mathbb{E}[T_n(x^*, \omega)] \end{pmatrix}. \tag{SGE}$$

1.2.2 Applications

Stochastic generalized equations subsume a broad class of problems, including possibly nonsmooth stochastic convex optimization problems, nonsmooth saddle-point problems, as well as a broad class of nonsmooth stochastic Nash games and stochastic variational inequality problems. We discuss two broad problem classes of interest.

(i) *Stochastic variational inequality problems with multi-valued maps.* Consider the stochastic variational inequality problem given by $\mathrm{VI}(X, \Phi)$ which requires an $x \in X$ and $u \in \Phi(x)$ such that

$$(y - x)^T u \geq 0, \qquad \forall y \in X.$$

This problem can be equivalently stated as follows.

$$0 \in \Phi(x) + \mathcal{N}_X(x),$$

where $\Phi(x) \triangleq \mathbb{E}[F(x, \omega)]$, $F(x, \omega)$ is a set-valued map, $\mathcal{N}_X(x)$ denotes the normal cone at x associated with the set X (Figure 1.3), the expectation is specified in a component-valued sense, and the integral of the set-valued map is defined in an Aumann sense. If $F(x, \omega)$ is a monotone map on X for a.e. $\omega \in \Omega$, it can be easily shown that $\mathbb{E}[F(x, \omega)]$ is also a monotone map. Furthermore, since $\mathcal{N}_X(x)$ is maximal monotone map, it follows that $\Phi(x) + \mathcal{N}_X(x)$ is a maximal monotone map [6, Prop. 12.3.6]. This problem is a generalized equation in which the mapping $T(x)$ is defined as $T(x) \triangleq \Phi(x) + \mathcal{N}_X(x)$. We observe that such problems arise commonly when modeling markets complicated by the presence of price-caps [18] and risk measures [27].

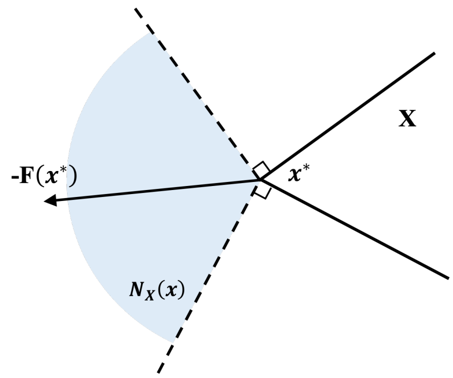


Figure 1.3: Normal cone illustration

(ii) *A subclass of stochastic multi-leader multi-follower games.* Consider a class of multi-leader multi-follower games [44, 45] in which there is a set of leaders, denoted by $\mathcal{N} \triangleq \{1, \ldots, N\}$ and a collection of leaders given by $\mathcal{M} \triangleq \{1, \ldots, M\}$. We consider the setting with a single follower where $M = 1$.

$$\begin{aligned} \min_{y_i} \quad & \tfrac{1}{2} y_i^T Q_i(\omega) y_i - d_i(x_i)^T y \\ \text{subject to} \quad & y_i \geq \ell_i(x_i), \end{aligned} \tag{$\mathrm{Follow}_i(x_i,\omega)$}$$

where $Q_i(\omega) \succ 0$ (implying that $Q_i(\omega)$ is a positive definite matrix) for every $\omega \in \Omega$. Suppose the ith leader solves the following parametrized problem

$$\begin{aligned} \min_{x_i} \quad & c_i(x_i) - p(\bar{x})x_i + \mathbb{E}[a_i(\omega)^T y_i] \\ \text{subject to} \quad & x_i \in X_i, \end{aligned} \tag{$\mathrm{Leader}_i(x_{-i})$}$$

where $c_i : X_i \to \mathbb{R}$ is a convex function, $p : \mathbb{R}^+ \to \mathbb{R}^+$ is a positive-valued strictly decreasing function of $\bar{x} \triangleq \sum_{i=1}^N x_i$, X_i is a closed and convex set in $\mathbb{R}$. We may compute the best-response of the follower by considering the necessary and sufficient conditions of optimality:

$$\begin{aligned} 0 \leq \lambda_i \perp y_i - \ell_i(x_i) \geq 0 \\ 0 = Q_i(\omega) y_i - d_i(x_i) - \lambda_i. \end{aligned}$$

This system can be equivalently stated as follows.

$$0 \leq Q_i(\omega) y_i - d_i(x_i) \perp y_i - \ell_i(x_i) \geq 0 \equiv y_i = \max\{Q_i(\omega)^{-1} d_i(x_i), \ell_i(x_i)\}.$$

Consequently, we may eliminate follower decision in the leader level problem, leading to a nonsmooth stochastic Nash game given by the following:

$$\begin{aligned} \min_{x_i} \quad & c_i(x_i) - p(\bar{x})x_i + \mathbb{E}[\underbrace{a_i(\omega)^T \max\{Q_i(\omega)^{-1}d_i(x_i), \ell_i(x_i)\}}_{\triangleq h_i(x_i,\omega)}] \\ \text{subject to} \quad & x_i \in X_i. \end{aligned} \tag{Leader$_i(x_{-i})$}$$

Under convexity of d_i and ℓ_i and suitable assumptions on $Q_i(\omega)$ and $a_i(\omega)$, the expression $a_i(\omega)^T \max\{Q_i(\omega)^{-1}d_i(x_i), \ell_i(x_i)\}$ is a convex function in x_i, a fact that follows from observing that this term is a scaling of the maximum of two convex functions. Consequently, the necessary and sufficient equilibrium conditions of this game are given by the following.

$$0 \in \partial_{x_i} c_i(x_i) - p(\bar{x}) - p'(\bar{x})x_i + \mathbb{E}[\partial_{x_i} h_i(x_i,\omega)] + \mathcal{N}_{X_i}(x_i), \quad i = 1, \ldots, N.$$

If $h_i(x_i,\omega)$ is a convex function in x_i for every ω and $p(\bar{x})$ is an affine and strictly decreasing function of x_i, then this is monotone stochastic generalized equation where the mapping T is defined as follows.

$$T(x) = \prod_{i=1}^{N} \left[\partial_{x_i} c_i(x_i) - p(\bar{x}) - p'(\bar{x})x_i + \mathbb{E}[\partial_{x_i} h_i(x_i,\omega)] + \mathcal{N}_{X_i}(x_i)\right].$$

1.2.3 Algorithms for the Solution of Generalized Equations

A range of algorithms have been developed for the resolution of generalized equations with maximal monotone operators, which are defined next.

Definition 3. *A monotone map T is maximal monotone if no monotone map $T^{'}$ exists such that gph $T \subset$ gph $T^{'}$, where gph $T = \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^n : y \in T(x)\}$.*

We now review a subset of avenues for the resolution of this problem.

1.2.3.1 Proximal-point Schemes.

A standard scheme to solve the problem $0 \in T(x)$ is the proximal point algorithm proposed by Martinet [46] and Rockafellar [47, 48]:

$$x_{k+1} \coloneqq (I + \gamma_k T)^{-1}(x_k), \tag{PP}$$

where γ_k denotes the steplength, T is required to be maximal monotone and $J_{\gamma_k}^T \triangleq (I + \gamma_k T)^{-1}$ is defined as the resolvent operator of T [47].

1.2.3.2 Splitting Schemes

In many applications, while the map T does not have a tractable resolvent operator, either the resolvent of A or B (or both) are tractable where $T \triangleq A + B$. In such instances, splitting schemes assume relevance. Moreover, if the resolvent of B is easier to evaluate, we may develop algorithms leveraging these resolvents. Assuming A and B are maximal monotone, the forward-backward splitting method was proposed by [49, 50] respectively and was applied to convex optimization by [51]:

$$x_{k+1} \coloneqq (I + \gamma_k B)^{-1}(I - \gamma_k A)(x_k). \tag{FBS}$$

A drawback of this method is that it generally requires a strong monotonicity assumption on A to ensure convergence. An extension was suggested by Tseng [52] that weakened the requirement on A to be mere monotonicity:

$$\begin{aligned} x_{k+\frac{1}{2}} &\coloneqq (I + \gamma_k B)^{-1}(I - \gamma_k A)(x_k), \\ x_{k+1} &\coloneqq \Pi_X \left(x_{k+\frac{1}{2}} - \gamma_k \left(A\left(x_{k+\frac{1}{2}}\right) - A(x_k) \right) \right). \end{aligned} \tag{MFBS}$$

1.2.4 Main Work

First, we develop a stochastic proximal point framework where we consider one of two avenues:

(A) The resolvent of the expectation-valued map, denoted by $(I + \gamma_k \mathbb{E}[T(x, \omega)])^{-1}$ is approximated with increasing accuracy via Monte-Carlo sampling. In particular, $(I + \gamma_k \mathbb{E}[T(x, \omega))^{-1}$ requires computing a solution of a strongly monotone stochastic generalized equation, which itself can be obtained via a stochastic approximation framework.

(B) The resolvent of the sample-average approximation of the map is employed, as defined by $(I + \gamma_k \bar{T}_k)^{-1}$, where $\bar{T}_k \triangleq \frac{\sum_{j=1}^{N_k} T(x,\omega_j)}{N_k}$.

Yet, the applicability of proximal-point schemes is crucially reliant on the tractability of evaluating the resolvent operator $J_{\gamma_k}^T(x_k)$. When this evaluation is problematic, splitting-based approaches have emerged as important alternatives.

Thereafter, we consider a setting where $A(x) \triangleq \mathbb{E}[A(x,\omega)]$ and $B(x)$ admits a relatively cheap resolvent. Here, we develop a stochastic modified forward-backward splitting framework in which the resolvent of A is approximated by Monte-Carlo sampling techniques. In both sets of schemes, we consider variance-reduced schemes that admit deterministic rates of convergence and near-optimal oracle complexities in some instances.

1.3 Transmission Expansion Problems

An important issue in the planning of electric energy systems is the capacity expansion of transmission lines [53–55] that enables energy flows from generation nodes to demand nodes. It requires a transmission planner to identify the optimal transmission reinforcements to be carried out with the aim of facilitating energy exchange among producers and consumers, e.g., by maximizing social welfare; this represents the essence of the transmission expansion problem.

The transmission expansion problem is generally tackled under two different frameworks: centralized and competitive. In a centralized framework, a planning entity controls both generation and transmission facilities and is in charge of performing generation and transmission expansion. In a competitive environment, an independent and regulated entity is generally in charge of operating and expanding the transmission network with the aim of maximizing social welfare over the entire market of generators and consumers. We briefly introduce the two framework in the follows.

1.3.1 Centralized Transmission Expansion Problems

Traditionally, before a competitive environment was considered in power systems planning, transmission expansion planning was modeled based on the demand forecast together with the associated generation capacity expansion plans. The mathematical model for this kind of problems is usually a single-level optimization problem with a general objective function subject to linear/nonlinear or integer constraints and is solved by standard mathematical programming methods [56–59]. Some of the past work has leveraged the use of Bender's decomposition [60] and heuristic approaches [61]. The general form of this problem is as follows:

$$\begin{aligned} \min_{x,z} \quad & h(x,z) \\ \text{subject to} \quad & f(x,z) = 0 \\ & g(x,z) \geq 0 \\ & z \in \{0,1\}^p, \end{aligned} \tag{ceTEP}$$

where z denotes the transmission expansion decision, x denotes other associated variables including generations, generation capacity expansion decisions, consumer demand, ISO arbitrage, etc. Furthermore, h denoted the objective value to be optimized while f and g represent linear/nonlinear constraints. If the size of the transmission expansion problem is too large to make it computationally intractable, a screening model [62] can be employed to reduce the number of investment variables.

1.3.2 Competitive Transmission Expansion Problems

In a competitive framework, a single regulated entity, referred to as the transmission system operator, determines the transmission expansion plans. Generation companies freely decide on investment decisions and the TSO is not directly involved in those decisions. The TSO makes expansion decisions regarding the existing transmission network with the goal of maximizing social welfare (the sum of consumer and producer surpluses) by utilizing a joint economic and engineering objective. Specifically, transmission expansion may allow for reducing overall costs of generation and improve the reliability in the supply of demand. This expansion problem is usually considered over a long-term planning horizon and accounts for (uncertain) changes in demand and generation capacity. Note that generation is not modeled as a player but is instead modeled as a random variable. Techniques from stochastic optimization [63–68] can be leveraged to cope with these uncertainties.

With competition, transmission expansion problems are generally highly complex, involving multiple optimization problems and nonlinear constraints. The problem is usually captured by two or more levels optimization problems. In general, tri-level [67, 69, 70] and bilevel [68, 71] structures are most popular in transmission expansion problems. Figure 1.4 and 1.5 illustrate both structures where y denotes transmission expansion decisions and k denotes generation capacity expansion variables, respectively. The difference between the tri-level and bilevel model lies in whether the spot market clearing condition is included in the second level (leading to a bilevel problem) or not (resulting in a tri-level problem). As a result, different approaches such as decomposition techniques [72–74] and heuristics [75–77] have been proposed.

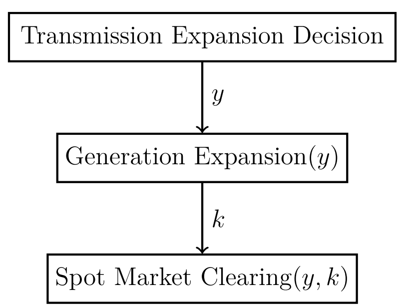


Figure 1.4: Tri-level transmission expansion model

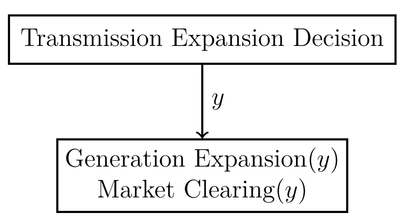


Figure 1.5: Bilevel transmission expansion model

We generally employ the following model to describe the framework:

$$\begin{aligned}
\min_{x,z} \quad & h(x,y,z) \\
\text{subject to} \quad & 0 \leq y \perp F(x,y,z) \geq 0 \\
& f(x,y,z) = 0 \\
& g(x,y,z) \geq 0 \\
& z \in \{0,1\}^p,
\end{aligned} \tag{coTEP}$$

where $0 \leq y \perp F(x,y,z) \geq 0$ are the equivalent optimal solutions to a second level.

1.3.3 Solution Methods in Transmission Expansion Problems

Transmission expansion planning problems have received attention for a long time. Many classical optimization methods have been used in analyzing transmission systems, such as linear programming [56, 78] and mixed integer programming [72, 73, 79–82]. Models with other kinds of typical techniques are also presented [75, 83–87]. [88] proposed a transmission planning framework which assumed only the generation sector is deregulated and future generation expansion of generation companies was taken into account by the transmission planner. An iterative solution algorithm was presented linking agent-based systems and search-based optimization technique.

Recently, multi-level modeling in transmission expansion problems has become popular, and some novel techniques have been invented. Jin and Ryan [69, 70] developed a tri-level model of centralized transmission and decentralized generation expansion planning. The second and third level may be collectively reformulated as a bilevel (MPEC) problem and this optimization problem can be resolved by either a diagonalization method (DM) or a complementarity problem (CP) reformulation. Then a hybrid iterative algorithm combining (CP) and (DM) was proposed. A mixed-integer linear programming model was developed in [89], consisting of three levels, transmission investment, generation investment and market operation. It integrated the three stages within a single optimization model and proposed a solution method after approximating the impedance as a function of capacity. In [90], the authors developed methodologies for capturing interactions between generation and transmission investment planning in the expansion planning problem of power networks. The first methodology formulated the problem as a mixed-integer nonlinear program (MINLP) which assuming a central regulatory organization exists while the second represents a bilevel model where the transmission operator is a leader.

Transmission planning under uncertainty also plays an important role in the literature [63–66]. In [67], the authors presented a three-level model for generation and transmission expansion planning that considered uncertainty in demand and the use of renewable sources in the model. The model included an (MPEC) and equilibrium program with equilibrium constraints (EPEC) formulations. These studies assumed that all the investments are made at the same time, which may not be realistic. The two-stage framework can also cope with the uncertainty in investment decisions better than a single stage decision. Similarly, a two-stage stochastic optimization model regarding planning electricity transmission was presented in [68] where a second decision was made based on the performance of a certain period after the first optimization solution. A robust analysis was conducted for the given practical problems. Some other multi-stage models also been developed in the literature (cf. [91, 92]).

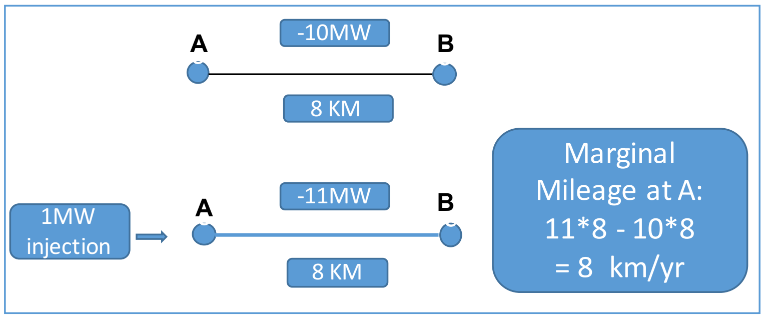
1.3.4 Main Work

In Chapter 4, we consider transmission expansion planning in competitive environment subjected to the response of a generation market with imperfect transmission pricing. In particular, we add annual transmission charges to the model for efficiency analysis purpose. Two different regimes on pricing models are analyzed and compared: one applies flat rate charge to generators, the other relies on a MW-miles based charging model. In the charging model, transmission charge to each generator is calculated by the following steps:

- Model power flows at system peak (the base case)
- Inject an additional MW at each node
- Look how power flows change compared to the base case
- Measure length of network traversed by additional MW flows.

Then this length represents the investment required for additional generation connecting at this node. The process is illustrated in a simple example shown in Figure 1.6.

Figure 1.6: MW-miles illustration



In effect, we leverage a bilevel structure to describe the power system and recast the model as a mathematical program with equilibrium constraints (MPEC). Furthermore, leveraging techniques from robust solutions to uncertain complementary problems, we incorporate robustness into the transmission expansion problem and reformulate it as a determinate model. Although the optimal solution to (MPEC)s is challenging given the discreteness and nonlinearity, we propose a direct solution method leveraging a branching scheme. To accelerate the solving process, we consider a smoothing scheme where discrete variables are relaxed by continuous functions. Preliminary results are considered for analysis.

Chapter 2
On the optimality of single projection variants of extragradient schemes for monotone stochastic variational inequality problems

2.1 Introduction

This chapter considers the solution of stochastic variational inequality problems, a stochastic generalization of the variational inequality problem. Given a set $X \subseteq \mathbb{R}^n$ and a map $F : \mathbb{R}^n \to \mathbb{R}^n$, the variational inequality problem VI(X, F) requires finding a point $x^* \in X$ such that

$$F(x^*)^T(x - x^*) \geq 0, \quad \forall x \in X. \tag{VI(X,F)}$$

In the stochastic generalization, the components of the map F are expectation-valued; specifically $F_i(x) \triangleq \mathbb{E}[F_i(x, \xi(\omega))]$, where $\xi : \Omega \to \mathbb{R}^d$ is a random variable, $F_i : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is a single-valued function, and the $\mathbb{E}[\cdot]$ denotes the expectation and the associated probability space being denoted by $(\Omega, \mathcal{F}, \mathbb{P})$. In short, we are interested in a vector $x^* \in X$ such that

$$\mathbb{E}[F(x^*, \omega)]^T(x - x^*) \geq 0, \quad \forall x \in X, \tag{SVI(X,F)}$$

where $\mathbb{E}[F(x,\omega)] = \left(\mathbb{E}[F_i(x,\omega)]\right)_{i=1}^{K}$. The variational inequality problem is an immensely relevant problem that finds application in engineering, economics, and applied sciences (cf. [6, 93–96]). Increasingly, the stochastic generalization is of relevance and has found application in the study of a broad class of equilibrium problems under uncertainty. Of these, sample average approximation (SAA) scheme solves the expected value of the stochastic mapping which is approximated via the average over a large number of samples (cf. [28–30, 32]). A counterpart to SAA schemes is the stochastic approximation (SA) methods where at each iteration, a sample of the stochastic mapping is used (cf. [24, 27, 31]). Amongst the simplest of SA schemes are analogs of the standard projection-based schemes, which we review next.

2.1.1 Projection-based Schemes and Their Variants

$$x_{k+1} := \Pi_X(x_k - \gamma_k F(x_k)), \tag{PG}$$

where $\Pi_X(y)$ denotes the projection of y onto X and γ denotes the steplength. This method generally requires a strong monotonicity assumption on F to ensure convergence. An extension, suggested by Antipin [42] and Korpelevich [43], required that F be merely monotone:

$$\begin{aligned} x_{k+\frac{1}{2}} &:= \Pi_X(x_k - \gamma F(x_k)), \\ x_{k+1} &:= \Pi_X\left(x_k - \gamma F\left(x_{k+\frac{1}{2}}\right)\right). \end{aligned} \tag{EG}$$

In (EG) however, two projections were required at each iteration to obtain a new point and convergence was proved under the assumptions of Lipschitz continuity and monotonicity of the map F. Naturally, when the set X is not necessarily a *simple* set, this projection operation by no means cheap. There have been several schemes in which merely monotone variational inequality problems can be addressed by taking a single projection operation and we consider two instances. In recent work, a *projected reflected gradient* (PRG) method was proposed by Malitsky [97], requiring a **single**, rather than **two**, projections:

$$x_{k+1} := \Pi_X(x_k - \gamma_k F(2x_k - x_{k-1})). \tag{PRG}$$

Intuitively, this scheme has a similar structure to the projected gradient scheme taking a form with the following key distinction: Rather than evaluating the map at x_k (as in (PG)), the map is evaluated at the reflection of x_{k-1} in x_k which is $x_k - (x_{k-1} - x_k) = 2x_k - x_{k-1}$. Remarkably, this simple modification allows for proving convergence of this scheme for merely monotone Lipschitz continuous maps [97]. An alternate modification of the extragradient method was proposed by Censor, Gibali and Reich and was referred to as the *subgradient extragradient method* (SE) [98]:

$$\begin{aligned} x_{k+\frac{1}{2}} &:= \Pi_X(x_k - \gamma_k F(x_k)), \\ x_{k+1} &:= \Pi_{C_k}\left(x_k - \gamma_k F\left(x_{k+\frac{1}{2}}\right)\right), \end{aligned} \tag{SE}$$

where $C_k \triangleq \{w \in \mathbb{R}^n \mid (x_k - \gamma_k F(x_k) - x_{k+\frac{1}{2}})^T(w - x_{k+\frac{1}{2}}) \leq 0\}$. In (SE), the two projections are replaced by a projection onto the set and a second onto a halfspace (computable in closed form).

2.1.2 Stochastic Variational Inequality Problems.

There have been schemes analogous to (PG) and (EG) in this regime with the key distinction that an evaluation of the map, namely $F(x_k)$, is replaced by $F(x_k, \omega_k)$, in the spirit of stochastic approximation [99]. Jiang and Xu [39] appear amongst the first who applied SA methods to solve stochastic variational inequalities. An extension to address merely monotone stochastic VIs was studied by Koshal et al. [40]. A regularized smoothing SA method to address stochastic VIs with non-Lipschitzian and merely monotone mappings was proposed in [100]. Recently, a class of prox generalization of SA methods were developed (cf. [101–104]) for solving smooth and nonsmooth stochastic convex optimization problems and variational inequalities. For instance, a simple stochastic extension of the standard projection scheme for VI(X, F) leads to a stochastic approximation scheme [99]:

$$x_{k+1} := \Pi_X(x_k - \gamma_k F(x_k, \omega_k)). \tag{SPG}$$

Similarly, an extragradient counterpart to (EG) is (SEG) and is defined below:

$$\begin{aligned} x_{k+\frac{1}{2}} &:= \Pi_X(x_k - \gamma_k F(x_k, \omega_k)), \\ x_{k+1} &:= \Pi_X\left(x_k - \gamma_k F\left(x_{k+\frac{1}{2}}, \omega_{k+\frac{1}{2}}\right)\right). \end{aligned} \tag{SEG}$$

Fig. 2.1 illustrates (SEG) scheme. Extragradient-based schemes (and their stochas-

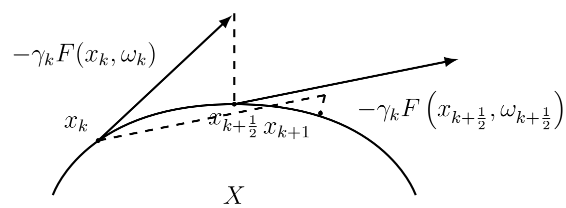


Figure 2.1: Stochastic extragradient (SEG)

tic mirror-prox counterparts) represent amongst the simplest of schemes for monotone SVIs (cf. [31, 105]). However, each iteration requires **two** projection steps, rather than one (as in (SPG)). We summarize much of the prior results in Table 3.1. Given that this class of Monte-Carlo approximation schemes routinely requires 10s or 100s of thousands of steps, our interest lies in ascertaining whether projection-based schemes can be developed requiring a single projection step per iteration, *reducing the per-iteration complexity by a factor of two.* We consider two such schemes given a random point $x_0 \in X$:

(i) **Stochastic projected reflected gradient schemes (SPRG)**;

$$x_{k+1} := \Pi_X(x_k - \gamma_k F(2x_k - x_{k-1}, \omega_k)), \tag{SPRG}$$

and (ii) **Stochastic subgradient extragradient schemes (SSE)**.

$$\begin{aligned} x_{k+\frac{1}{2}} &:= \Pi_X(x_k - \gamma_k F(x_k, \omega_k)), \\ x_{k+1} &:= \Pi_{C_k}\left(x_k - \gamma_k F\left(x_{k+\frac{1}{2}}, \omega_{k+\frac{1}{2}}\right)\right), \end{aligned} \tag{SSE}$$

where $C_k \triangleq \left\{w \in \mathbb{R}^n \mid \left(x_k - \gamma_k F(x_k, \omega_k) - x_{k+\frac{1}{2}}\right)^T \left(w - x_{k+\frac{1}{2}}\right) \leq 0\right\}$. Clearly, the second projection is a simple optimization problem. Solving for x_{k+1}, we could obtain an equivalent scheme which requires a single projection (the proof is in appendix). Fig. 2.2 illustrate the steps of these schemes.

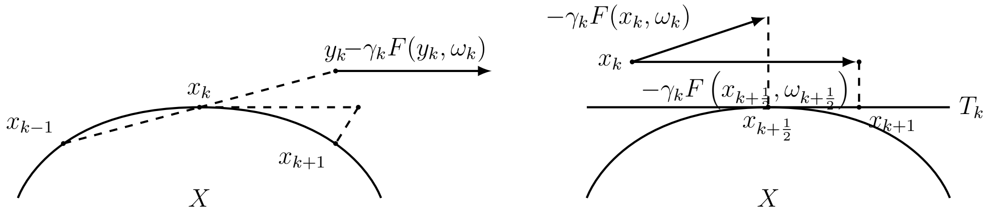


Figure 2.2: Left: (SPRG); Right: (SSE)

Table 2.1: A review of stochastic extragradient schemes

Ref.	Applicability	Avg.	Metric	Rate	A.s.	# proj.
[39]	Monotone, uniqueness	N	Iterates	-	Y	
[40]	Monotone, Lipschitz	N	Iterates	-	Y	
[100]	Monotone, non-Lip.	N	Iterates	-	Y	
[31]	Monotone, non-Lip.	Y	Gap fn.	$\mathcal{O}(1/\sqrt{K})$	N	
[102]	Monotone, non-Lip.	Y	Gap fn.	$\mathcal{O}(1/\sqrt{K})$	Y	
[106]	Strongly pseudo/monotone+weak-sharp	N	MSE	$\mathcal{O}(1/K)$	Y	
[107]	Strongly monotone, Lip., random proj.	N	Iterates	-	Y	
[108]	Pesudo monotone, Lip., var. reduction	N	Iterates	$\mathcal{O}(1/K)$	Y	
[95]	Monotone+weak-sharp, Lip., random proj.	Y	Dist. fn.	$\mathcal{O}(1/\sqrt{K})$	Y	2
v-SPRG	Monotone+weak-sharp, Lip., var. reduction	Y	Dist. fn.	$\mathcal{O}(1/K)$	Y	1
v-SSE	Monotone, Lip., var. reduction	Y	Gap fn.	$\mathcal{O}(1/K)$	Y	1
r-SPRG	Monotone+weak-sharp, Lip., random proj.	Y	Dist. fn.	$\mathcal{O}(1/\sqrt{K})$	Y	1
r-SSE	Monotone+weak-sharp, Lip., random proj.	Y	Dist. fn.	$\mathcal{O}(1/\sqrt{K})$	Y	1

2.1.3 Incorporating Variance Reduction and Random Projections.

To mitigate computational complexity, we define two variable sample-size counterparts of (SPRG) and (SEG), where N_k samples of the map are utilized at iteration k to approximate the expectation. We define (i) **Variable sample-size stochastic projected reflected gradient schemes**:

$$x_{k+1} := \Pi_X \left(x_k - \gamma_k \frac{\sum_{j=1}^{N_k} F(2x_k - x_{k-1}, \omega_{j,k})}{N_k} \right), \qquad \text{(v-SPRG)}$$

and (ii) **Variable sample-size stochastic subgradient extragradient schemes**.

$$\begin{aligned} x_{k+\frac{1}{2}} &:= \Pi_X \left(x_k - \gamma_k \frac{\sum_{j=1}^{N_k} F(x_k, \omega_{j,k})}{N_k} \right), \\ x_{k+1} &:= \Pi_{C_k} \left(x_k - \gamma_k \frac{\sum_{j=1}^{N_k} F\left(x_{k+\frac{1}{2}}, \omega_{j,k+\frac{1}{2}}\right)}{N_k} \right), \end{aligned} \qquad \text{(v-SSE)}$$

where $C_k \triangleq \left\{w \in \mathbb{R}^n \mid \left(x_k - \gamma_k F(x_k, \omega_k) - x_{k+\frac{1}{2}}\right)^T \left(w - x_{k+\frac{1}{2}}\right) \leq 0\right\}$.
A difficulty arises when implementing such schemes on a complex set X when X is defined as the intersection of a large number of convex sets. Inspired by [107], we consider extending our work to random projections when X is defined as the intersection of a finite number of sets:

$$X = \bigcap_{i \in \mathcal{I}} X_i,$$

where $\mathcal{I}$ is a finite set and $X_i \subseteq \mathbb{R}^n$ is closed and convex for all $i \in \mathcal{I}$. The key distinction is that at each iteration, we project onto a random subset X_{l_k} rather than X, where $\{l_k\}$ is a sequence of random variables in the appropriate steps of (SPRG) and (SSE). In prior work, Nedić [109, 110] considered random projection algorithms for convex optimization problems with similarly defined sets and related schemes were subsequently considered for nonsmooth convex regimes [111–113]. Wang and Bertsekas [107] extended (SPG) to allow for projecting on a subset of constraints based on either a random projection technique on either a random or deterministic (such as cyclic projection) subset. We consider analogous generalizations to (SPRG) and (SEG):

(i) **Random projected stochastic projected reflected gradient schemes** (r-SPRG);

$$x_{k+1} \coloneqq \Pi_{l_k}(x_k - \gamma_k F(2x_k - x_{k-1}, \omega_k)), \tag{r-SPRG}$$

and (ii) **Random projected stochastic subgradient extragradient schemes** (r-SSE).

$$\begin{aligned} x_{k+\frac{1}{2}} &\coloneqq \Pi_{l_k}(x_k - \gamma_k F(x_k, \omega_k)), \\ x_{k+1} &\coloneqq \Pi_{C_k}\left(x_k - \gamma_k F\left(x_{k+\frac{1}{2}}, \omega_{k+\frac{1}{2}}\right)\right), \end{aligned} \tag{r-SSE}$$

where $C_k \triangleq \left\{w \in \mathbb{R}^n \mid \left(x_k - \gamma_k F(x_k, \omega_k) - x_{k+\frac{1}{2}}\right)^T \left(w - x_{k+\frac{1}{2}}\right) \leq 0\right\}$.

2.1.4 Contributions

We summarize the key aspects of our schemes in Tables 2.2 and elaborate on these next:

Table 2.2: (SRPG), (SSE) and (SEG) schemes comparison

	Variance-reduced schemes			Random projection		
	Assump.	Rate	a.s.	Assump.	Rate, infeas.	a.s.
SPRG	monotone+weak-sharp	$\mathcal{O}(1/K)$	✔	monotone+weak-sharp	$\mathcal{O}\left(1/\sqrt{K}\right)$, $\mathcal{O}\left(1/\sqrt{K}\right)$	✔
SSE	monotone	$\mathcal{O}(1/K)$	✔	monotone+weak-sharp	$\mathcal{O}\left(1/\sqrt{K}\right)$, $\mathcal{O}\left(1/\sqrt{K}\right)$	✔
SEG	monotone	$\mathcal{O}(1/K)$	✔	monotone+weak-sharp	$\mathcal{O}\left(1/\sqrt{K}\right)$, $\mathcal{O}\left(1/\sqrt{K}\right)$	✔

(i) In Section 2.3, we prove that in monotone regimes, the iterates produced by both (SPRG) and (SSE) converge almost surely (a.s.) to the solution and the expectation of the distance function (for (SPRG)) or the gap function (for (SSE)) diminishes at $\mathcal{O}(1/K)$, matching the deterministic rate of convergence.

(ii) In Section 2.4, under merely monotone settings with a weak-sharpness requirement, random projection variants of (SPRG) and (SSE) are examined and we proceed to prove a.s. convergence of the iterates to the solution set. Additionally, we proceed to show that the expected distance to both the optimal solution set X^* and the feasible set X diminish at the rate of $\mathcal{O}(1/\sqrt{K})$.

(iii) In Section 2.5, preliminary numerics are observed support our expectations based on the theoretical findings.

2.2 Background and Assumptions

We consider the schemes (SPRG) and (SSE) where $x_0 \in X$ is a random initial point and $\{\gamma_k\}$ denotes the steplength sequence. We begin by imposing an assumption on the map F which will be valid through the remainder of this chapter.

Assumption 1. *The mapping F is L-Lipschitz continuous and monotone on $\mathbb{R}^n$; i.e. for all $x, y \in \mathbb{R}^n$,*

$$\|F(x) - F(y)\| \leq L\|x - y\| \text{ and } (F(x) - F(y))^T(x - y) \geq 0.$$

We often impose a boundedness requirement on the set X and $F(x^*)$.

Assumption 2. *The set X is bounded, i.e., there exists a scalar $B > 0$ such that $\|x - y\| \leq B$ for all $x, y \in X$.*

Assumption 3. *There exists a constant $C > 0$ such that $\|F(x^*)\| \leq C$.*

In some instances, a weak-sharpness requirement is imposed on $\mathrm{VI}(X, F)$.

Assumption 4 (**Weak sharpness**)**.** *The variational inequality problem VI(X, F) satisfies the weak sharpness property implying that there exists an $\alpha > 0$ such that for all $x \in X$, $(x - x^*)^T F(x^*) \geq \alpha \, dist\,(x, X^*)$.*

The following lemma is used in our analysis proofs and may be found in [114].

Lemma 1. *Let X be nonempty closed convex set in $\mathbb{R}^n$. Then for all $y \in X$ and for any $x \in \mathbb{R}^n$, we have that the following hold: (i) $(\Pi_X(x) - x)^T(y - \Pi_X(x)) \geq 0$; and (ii) $\|\Pi_X(x) - y\|^2 \leq \|x - y\|^2 - \|x - \Pi_X(x)\|^2$.*

We assume the presence of a stochastic oracle that can provide a conditionally unbiased estimator of $F(x)$, given by $F(x, \omega)$ such that $\mathbb{E}[F(x, \omega) \mid x] = F(x)$. Define $w_k \triangleq F(x_k, \omega_k) - F(x_k)$, $\bar{w}_k \triangleq \frac{\sum_{j=1}^{N_k} F(x_k, \omega_{j,k})}{N_k} - F(x_k)$, $w_{k+1/2} \triangleq F(x_{k+1/2}, \omega_{k+1/2}) - F(x_{k+1/2})$ and $\bar{w}_{k+1/2} \triangleq \frac{\sum_{j=1}^{N_k} F(x_{k+1/2}, \omega_{j,k})}{N_k} - F(x_{k+1/2})$, where N_k denotes the batch-size of sampled maps $F(x, \omega_{j,k})$ at iteration k. Furthermore, let $\mathcal{F}_k$ denote the history up to iteration k, i.e., $\mathcal{F}_k \triangleq \left\{x_0, \omega_0, \omega_{\frac{1}{2}}, \omega_1, \cdots, \omega_{k-1}, \omega_{k-\frac{1}{2}}\right\}$ and $\mathcal{F}_{k+\frac{1}{2}} \triangleq \mathcal{F}_k \cup \{\omega_k\}$.

Assumption 5. *At an iteration k, the following hold in an a.s. sense: (i) The conditional means $\mathbb{E}[w_k \mid \mathcal{F}_k]$ and $\mathbb{E}\left[w_{k+\frac{1}{2}} \mid \mathcal{F}_{k+\frac{1}{2}}\right]$ are zero for all k in an a.s. sense; (ii) The conditional second moment-2s are bounded in an a.s. sense or $\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] \leq \nu^2$ and $\mathbb{E}\left[\|w_{k+\frac{1}{2}}\|^2 \mid \mathcal{F}_{k+\frac{1}{2}}\right] \leq \nu^2$ for all k in an a.s. sense.*

Assumption 6. *The diminishing sequence γ_k is square-summable but non-summable: $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$, $\sum_{k=0}^{\infty} \gamma_k = \infty$.*

The following super-martingale convergence Lemma is essential to our proof [115].

Lemma 2. *Let v_k, u_k, δ_k, ψ_k be nonnegative random variables adapted to σ-algebra $\mathcal{F}_k$, and let the following relations hold almost surely:*

$$\mathbb{E}[v_{k+1} \mid \mathcal{F}_k] \leq (1+u_k)v_k - \delta_k + \psi_k, \quad \forall k; \quad \sum_{k=0}^{\infty} u_k < \infty, \text{ and } \sum_{k=0}^{\infty} \psi_k < \infty.$$

Then a.s., we have that $\lim_{k\to\infty} v_k = v$ and $\sum_{k=0}^{\infty} \delta_k < \infty$, where $v \geq 0$ is a random variable.

2.3 Convergence Analysis for (v-SPRG) and (v-SSE)

2.3.1 Stochastic Projected Reflected Gradient Schemes

In this subsection, we prove the a.s. convergence of the iterates produced by (SPRG) when F is a Lipschitz continuous and monotone map on $\mathbb{R}^n$, satisfying a weak-sharpness requirement. We begin with an intermediate lemma that relates the error in consecutive iterates.

Lemma 3. *Let Assumptions 1, 4, and 5 hold and let $0 < \gamma_k = \gamma \leq \frac{1}{8L}$ for all k. Consider a sequence generated by (v-SPRG). For any $x_0 \in X$, the following holds for all $k \geq 0$:*

$$\begin{aligned}
&\|x_{k+1} - x^*\|^2 + \frac{3}{4}\|x_{k+1} - y_k\|^2 + 2\gamma F(x^*)^T(x_k - x^*) \\
&\leq \|x_k - x^*\|^2 + \frac{3}{4}\|x_k - y_{k-1}\|^2 + 2\gamma F(x^*)^T(x_{k-1} - x^*) \\
&+ 8\gamma^2\|w_k - w_{k-1}\|^2 - \left(1 - 16\gamma^2 L^2\right)\|x_k - y_k\|^2 \\
&- 2\gamma\alpha\, dist(x_k, X^*) - 2\gamma w_k^T(y_k - x^*).
\end{aligned}$$

Proof. Define $y_k \triangleq 2x_k - x_{k-1}$ for all $k \geq 1$ and $\bar{F}(y_k) \triangleq \frac{\sum_{j=1}^{N_k} F(y_k,\omega_{k,j})}{N_k}$. By Lemma 1(ii) and noting that $x_{k+1} = \Pi_X(x_k - \gamma_k \bar{F}(y_k))$ and $\bar{F}(y_k) = F(y_k) + \bar{w}_k$, the following holds for x_{k+1} and any solution x^*.

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq \|x_k - \gamma_k\bar{F}(y_k) - x^*\|^2 - \|x_k - \gamma_k\bar{F}(y_k) - x_{k+1}\|^2 \\
&= \|x_k - x^*\|^2 - \|x_{k+1} - x_k\|^2 - 2\gamma_k(F(y_k) + \bar{w}_k)^T(x_{k+1} - x^*). \quad (2.1)
\end{aligned}$$

Since F is monotone over $\mathbb{R}^n$, by adding $2\gamma_k(F(y_k) - F(x^*))^T(y_k - x^*)$ to the rhs of (2.1), we obtain:

$$
\begin{aligned}
&\|x_{k+1}-x^*\|^2 \le \|x_k-x^*\|^2 - \|x_{k+1}-x_k\|^2 + 2\gamma_k(F(y_k)-F(x^*))^T(y_k-x^*)\\
&-2\gamma_k(F(y_k)+\bar{w}_k)^T(x_{k+1}-x^*)\\
&= \|x_k-x^*\|^2 - \|x_{k+1}-x_k\|^2 + 2\gamma_k F(y_k)^T(y_k-x_{k+1}) + 2\gamma_k F(y_k)^T(x_{k+1}-x^*)\\
&-2\gamma_k F(x^*)^T(y_k-x^*) - 2\gamma_k F(y_k)^T(x_{k+1}-x^*) + 2\gamma_k\bar{w}_k^T(y_k-x_{k+1})\\
&-2\gamma_k\bar{w}_k^T(y_k-x^*)\\
&= \|x_k-x^*\|^2 - \|x_{k+1}-x_k\|^2 + 2\gamma_k(F(y_k)+\bar{w}_k)^T(y_k-x_{k+1})\\
&-2\gamma_k(F(x^*)+\bar{w}_k)^T(y_k-x^*)\\
&= \|x_k-x^*\|^2 - \|x_{k+1}-x_k\|^2 + \underbrace{2\gamma_k(F(y_k)-F(y_{k-1}))^T(y_k-x_{k+1})}_{\text{Term 1}}\\
&+\underbrace{2\gamma_k(F(y_{k-1})+\bar{w}_k)^T(y_k-x_{k+1})}_{\text{Term 2}} - \underbrace{2\gamma_k(F(x^*)+\bar{w}_k)^T(y_k-x^*)}.
\end{aligned}
\tag{2.2}
$$

Since $x_{k+1}, x_{k-1} \in X$, by Lemma 1(i), we may conclude that

$$
\begin{aligned}
&(x_k - x_{k-1} + \gamma_{k-1}(F(y_{k-1})+\bar{w}_{k-1}))^T(x_k - x_{k+1}) \le 0 \text{ and}\\
&(x_k - x_{k-1} + \gamma_{k-1}(F(y_{k-1})+\bar{w}_{k-1}))^T(x_k - x_{k-1}) \le 0.
\end{aligned}
$$

Adding these two inequalities yields the following:

$$
(x_k - x_{k-1} + \gamma_{k-1}(F(y_{k-1})+\bar{w}_{k-1}))^T(y_k - x_{k+1}) \le 0,
$$

since $y_k = 2x_k - x_{k-1}$, leading to the following inequality:

$$
\begin{aligned}
2\gamma_{k-1}(F(y_{k-1})+\bar{w}_{k-1})^T(y_k-x_{k+1}) &\le 2(x_k-x_{k-1})^T(x_{k+1}-y_k)\\
&= 2(y_k-x_k)^T(x_{k+1}-y_k) = \|x_{k+1}-x_k\|^2 - \|x_k-y_k\|^2 - \|x_{k+1}-y_k\|^2,
\end{aligned}
\tag{2.3}
$$

where the first equality follows from recalling that $y_k = 2x_k - x_{k-1}$. Now, we may bound $2\gamma_k(F(y_{k-1})+\bar{w}_k)^T(y_k-x_{k+1})$ as follows:

$$
\begin{aligned}
&\text{Term 2} = 2\gamma_k(F(y_{k-1})+\bar{w}_k)^T(y_k-x_{k+1}) = 2\gamma_k(F(y_{k-1})+\bar{w}_k)^T(y_k-x_{k+1})\\
&-2\gamma_k(F(y_{k-1})+\bar{w}_{k-1})^T(y_k-x_{k+1}) + 2\gamma_k(F(y_{k-1})+\bar{w}_{k-1})^T(y_k-x_{k+1})\\
&= 2\gamma_k(\bar{w}_k-\bar{w}_{k-1})^T(y_k-x_{k+1}) + 2\left(\frac{\gamma_k}{\gamma_{k-1}}\right)\gamma_{k-1}(F(y_{k-1})+\bar{w}_{k-1})^T(y_k-x_{k+1})
\end{aligned}
$$

$$
\begin{aligned}
&\leq 8\gamma_k^2\|\bar{w}_k - \bar{w}_{k-1}\|^2 + \frac{1}{8}\|x_{k+1} - y_k\|^2 - \frac{\gamma_k}{\gamma_{k-1}}\|x_{k+1} - y_k\|^2 + \frac{\gamma_k}{\gamma_{k-1}}\|x_{k+1} - x_k\|^2 \\
&- \frac{\gamma_k}{\gamma_{k-1}}\|x_k - y_k\|^2 \\
&= 8\gamma_k^2\|\bar{w}_k - \bar{w}_{k-1}\|^2 + \left(\frac{1}{8} - \frac{\gamma_k}{\gamma_{k-1}}\right)\|x_{k+1} - y_k\|^2 + \frac{\gamma_k}{\gamma_{k-1}}\|x_{k+1} - x_k\|^2 \\
&- \frac{\gamma_k}{\gamma_{k-1}}\|x_k - y_k\|^2, \qquad (2.4)
\end{aligned}
$$

where $2\gamma_k(w_k - w_{k-1})^T(y_k - x_{k+1}) \leq 8\gamma_k^2\|w_k - w_{k-1}\|^2 + \frac{1}{8}\|x_{k+1} - y_k\|^2$ and inequality (2.3) allows for bounding $2\gamma_{k-1}(F(y_{k-1}) + w_{k-1})^T(y_k - x_{k+1})$. Next we estimate $(F(y_k) - F(y_{k-1})^T(y_k - x_{k+1})$. By the Cauchy-Schwarz inequality and the Lipschitz continuity of the map (Ass. 1), it follows that

$$
\begin{aligned}
&\text{Term } 1 = 2\gamma_k(F(y_k) - F(y_{k-1}))^T(y_k - x_{k+1}) \leq 2\gamma_k\|F(y_k) - F(y_{k-1})\|\|y_k - x_{k+1}\| \\
&\leq 2\gamma_k L\|y_k - y_{k-1}\|\|y_k - x_{k+1}\| \leq 8\gamma_k^2 L^2\|y_k - y_{k-1}\|^2 + \frac{1}{8}\|x_{k+1} - y_k\|^2 \\
&\leq 16\gamma_k^2 L^2\|x_k - y_{k-1}\|^2 + 16\gamma_k^2 L^2\|x_k - y_k\|^2 + \frac{1}{8}\|x_{k+1} - y_k\|^2, \qquad (2.5)
\end{aligned}
$$

where (2.5) follows from $\|u + v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$. Using (2.4) and (2.5), we deduce from (2.2) that

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 &- \left(1 - \frac{\gamma_k}{\gamma_{k-1}}\right)\|x_{k+1} - x_k\|^2 \\
&- \left(\frac{\gamma_k}{\gamma_{k-1}} - 16\gamma_k^2 L^2\right)\|x_k - y_k\|^2 - \left(\frac{\gamma_k}{\gamma_{k-1}} - \frac{1}{4}\right)\|x_{k+1} - y_k\|^2 \\
&+ 16\gamma_k^2 L^2\|x_k - y_{k-1}\|^2 + 8\gamma_k^2\|\bar{w}_k - \bar{w}_{k-1}\|^2 \\
&- 2\gamma_k(F(x^*) + \bar{w}_k)^T(y_k - x^*). \qquad (2.6)
\end{aligned}
$$

By assumption, $\gamma_k = \gamma \leq 1/(8L)$, for all k,

$$
16\gamma_k^2 L^2 \leq \frac{1}{4} \leq \left(\frac{\gamma_{k-1}}{\gamma_{k-2}} - \frac{1}{4}\right). \qquad (2.7)
$$

Consequently, from (2.6) and by invoking (2.7), we may conclude the following:

$$
\|x_{k+1} - x^*\|^2 + \left(\frac{\gamma_k}{\gamma_{k-1}} - \frac{1}{4}\right)\|x_{k+1} - y_k\|^2 \leq \|x_k - x^*\|^2
$$

$$+\left(\frac{\gamma_{k-1}}{\gamma_{k-2}}-\frac{1}{4}\right)\|x_k-y_{k-1}\|^2+8\gamma_k^2\|w_k-w_{k-1}\|^2-\left(\frac{\gamma_k}{\gamma_{k-1}}-16\gamma_k^2L^2\right)\|x_k-y_k\|^2$$
$$-2\gamma_kF(x^*)^T(y_k-x^*)+\gamma_k\bar{w}_k^T(y_k-x^*).$$

We may bound $2\gamma_kF(x^*)^T(y_k-x^*)$ as follows:

$$\begin{aligned}
&-2\gamma_kF(x^*)^T(y_k-x^*)=-2\gamma_kF(x^*)^T(x_k-x^*)-2\gamma_kF(x^*)^T(x_k-x^*)\\
&+2\gamma_kF(x^*)^T(x_{k-1}-x^*)\\
&\leq -2\gamma_kF(x^*)^T(x_k-x^*)-2\gamma_kF(x^*)^T(x_k-x^*)+2\gamma_{k-1}F(x^*)^T(x_{k-1}-x^*). \quad (2.8)
\end{aligned}$$

By the weak sharpness property, we have that $F(x^*)^T(x_k-x^*)\geq \alpha\mathrm{dist}\,(x_k,X^*)\,,$ which together with (2.8), implies that

$$\begin{aligned}
&\|x_{k+1}-x^*\|^2+\frac{3}{4}\|x_{k+1}-y_k\|^2+2\gamma F(x^*)^T(x_k-x^*)\\
&\leq \|x_k-x^*\|^2+\frac{3}{4}\|x_k-y_{k-1}\|^2+2\gamma F(x^*)^T(x_{k-1}-x^*)\\
&+8\gamma^2\|\bar{w}_k-\bar{w}_{k-1}\|^2-\left(1-16\gamma^2L^2\right)\|x_k-y_k\|^2\\
&-2\gamma\alpha\mathrm{dist}\,(x_k,X^*)-2\gamma\bar{w}_k^T(y_k-x^*). \quad (2.9)
\end{aligned}$$

□

With this lemma, we now analyze convergence of (v-SPRG).

Proposition 1 (a.s. convergence of (v-SPRG))**.** *Consider the scheme (v-SPRG). Let Assumptions 1, 4, and 5 hold. Let $0<\gamma_k=\gamma\leq\frac{1}{8L}$ for all $k\geq 0$ and $\sum_{k=1}^{\infty}\frac{1}{N_k}<\infty$. Then for any $x_0\in X$, a sequence generated by (v-SPRG) converges to a solution $x^*\in X$ in an a.s. sense.*

Proof. Using (2.9) and taking expectations conditioned on $\mathcal{F}_k$,

$$\begin{aligned}
&\mathbb{E}\left[\|x_{k+1}-x^*\|^2+\frac{3}{4}\|x_{k+1}-y_k\|^2+2\gamma F(x^*)^T(x_k-x^*)|\mathcal{F}_k\right]\\
&\leq \|x_k-x^*\|^2+\frac{3}{4}\|x_k-y_{k-1}\|^2+2\gamma F(x^*)^T(x_{k-1}-x^*)-2\alpha\gamma\mathrm{dist}\,(x_k,X^*)\\
&+8\gamma^2\mathbb{E}[\|\bar{w}_k-\bar{w}_{k-1}\|^2|\mathcal{F}_k]-\left(1-16\gamma^2L^2\right)\|x_k-y_k\|^2\\
&\leq \|x_k-x^*\|^2+\frac{3}{4}\|x_k-y_{k-1}\|^2+2\gamma F(x^*)^T(x_{k-1}-x^*)-2\alpha\gamma\mathrm{dist}\,(x_k,X^*)\\
&+32\gamma^2\frac{\nu^2}{N_k}-\left(1-16\gamma^2L^2\right)\|x_k-y_k\|^2=v_k-\delta_k+\psi_k, \quad (2.10)
\end{aligned}$$

where v_k, δ_k, and ψ_k are nonnegative random variables defined as

$$v_k \triangleq \|x_k - x^*\|^2 + \frac{3}{4}\|x_k - y_{k-1}\|^2 + 2\gamma F(x^*)^T(x_k - x^*),$$
$$\delta_k \triangleq \left(1 - 16\gamma^2 L^2\right)\|x_k - y_k\|^2 + 2\alpha\gamma \text{dist}\left(x_k, X^*\right) \text{ and } \psi_k \triangleq 32\gamma^2 \frac{\nu^2}{N_k}.$$

We note that $\sum_k \psi_k < \infty$ since $\sum_k \frac{1}{N_k} < \infty$ and $\delta_k \geq 0$ since $\text{dist}(x_k, X^*) \geq 0$ for all k and

$$\left(1 - 16\gamma^2 L^2\right) \geq \frac{1}{4}.$$

We may now invoke Lemma 2 to claim that $v_k \to \bar{v} \geq 0$ and $\sum_k \delta_k < \infty$ in an a.s. sense, implying the following holds a.s.:

$$\begin{aligned}
\infty &> \sum_k \left(\left(1 - 16\gamma^2 L^2\right)\|x_k - y_k\|^2 + 2\alpha\gamma \text{dist}\left(x_k, X^*\right)\right) \\
&\geq \sum_k \left(\left(1 - \frac{1}{4}\right)\|x_k - y_k\|^2 + 2\alpha\gamma \text{dist}\left(x_k, X^*\right)\right) \\
&= \sum_k \left(\frac{3}{4}\|x_k - y_k\|^2 + 2\alpha\gamma \text{dist}\left(x_k, X^*\right)\right),
\end{aligned}$$

where the second inequality follows from $\gamma \leq 1/(8L)$. Consequently, we have that

$$\infty > \sum_k \left(\frac{3}{4}\|x_k - y_k\|^2 + 2\alpha\gamma \text{dist}\left(y_k, X^*\right)\right).$$

It follows that in an a.s. sense,

$$\infty > \sum_k \|x_k - y_k\|^2 = \sum_k \|x_k - x_{k-1}\|^2. \tag{2.11}$$

From (2.11), $x_k - y_k \to 0$ as $k \to \infty$ in an a.s. sense. Furthermore, in an a.s. sense, $\sum_k \alpha\gamma \text{dist}\left(x_k, X^*\right) < \infty$ and in an a.s. sense, we have

$$\lim_{k\to\infty} \text{dist}(x_k, X^*) = 0.$$

This implies that the entire sequence of $\{x_k\}$ converges to a point in X^* in an a.s. sense. Since $\{x_k\}$ and $\{y_k\}$ have the same limit points almost surely, we have that $\{y_k\}$ also converges to a point in X^* in an a.s. sense. $\square$

We are now in a position to derive a rate statement for the sequence of iterates. Importantly, we attain a rate of $\mathcal{O}(1/K)$ in terms of the distance to the solution, an improvement over the rate of $\mathcal{O}(1/\sqrt{K})$ by using an increasing batch-size sequence $\{N_k\}$.

Proposition 2 (**Rate statement for (SPRG)**). *Consider the (v-SPRG) scheme. Let Assumptions 1, 2, and 5 hold. Let $0 < \gamma_k = \gamma \leq 1/8L$ for all $k \geq 0$, $\sum_{k=1}^{\infty} \frac{1}{N_k} < M$, and $\bar{x}_K \triangleq \sum_{k=1}^{K} x_k/K$.*
(1). Then for any K, $\mathbb{E}[dist(\bar{x}_K, X^*)] \leq \mathcal{O}\left(\frac{1}{K}\right)$. (2). Suppose $N_k = \lfloor k^a \rfloor$, for $a > 1$. The oracle complexity to obtain an x_K such that $\mathbb{E}[dist(x_k, X^*)] \leq \epsilon$ is bounded as follows. $\sum_{k=1}^{K} N_k \leq \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$.

Proof. (1). From (2.10), taking expectations on both sides and by summing over k from 1 to K, we have the following inequality:

$$\sum_{k=1}^{K} 2\alpha\gamma\mathbb{E}[\mathrm{dist}(\bar{x}_k, X^*)] \leq \mathbb{E}[\|x_1 - x^*\|^2] + \frac{3}{4}\mathbb{E}[\|x_1 - y_0\|^2]$$
$$+ 2\gamma F(x^*)^T(x_1 - x^*) + 32\gamma^2\nu^2 \sum_{k=1}^{K} \frac{1}{N_k}.$$

Dividing both sides by $2K\alpha\gamma$, we have the following sequence of inequalities:

$$\begin{aligned}
\frac{\sum_{k=1}^{K} 2\alpha\gamma\mathbb{E}[\mathrm{dist}(x_k, X^*)]}{2\sum_{k=1}^{K}\alpha\gamma} &\leq \frac{\mathbb{E}[\|x_1 - x^*\|^2] + \frac{3}{4}\mathbb{E}[\|x_1 - y_0\|^2] + 2\gamma F(x^*)^T(x_1 - x^*)}{2K\alpha\gamma} \\
&\quad + \frac{16\gamma\nu^2\sum_{k=1}^{K}\frac{1}{N_k}}{K\alpha} \\
&\leq \frac{\frac{7}{4}B^2 + 2\gamma BC}{2K\alpha\gamma} + \frac{16\gamma\nu^2\sum_{k=1}^{K}\frac{1}{N_k}}{K\alpha},
\end{aligned}$$

where the second inequality follows from the boundedness of X. By the convexity of the distance function, we have that

$$\mathbb{E}[\mathrm{dist}(\bar{x}_K, X^*)] \leq \frac{\sum_{k=1}^{K} 2\alpha\gamma\mathbb{E}[\mathrm{dist}(x_k, X^*)]}{2\sum_{k=1}^{K}\alpha\gamma}, \text{ where } \bar{x}_K \triangleq \frac{\sum_{k=1}^{K} x_k}{K}.$$

By choosing N_k such that $\sum_{k=1}^{K}\frac{1}{N_k} < M < \infty$, we have

$$\mathbb{E}[\mathrm{dist}(\bar{x}_K, X^*)] \leq \frac{1}{K}\underbrace{\left(\frac{\frac{7}{4}B^2 + 2\gamma BC}{2\alpha\gamma} + \frac{16\gamma\nu^2 M}{\alpha}\right)}_{\triangleq\widehat{C}} \leq \mathcal{O}\left(\frac{1}{K}\right).$$

(2). It follows from Proposition 2(1) that for ϵ sufficiently small,

$$\sum_{k=1}^{K} N_k \leq \sum_{k=1}^{\lceil(\widehat{C}/\epsilon)\rceil} N_k \leq \sum_{k=1}^{\lceil(\widehat{C}/\epsilon)\rceil} k^a \leq \int_{k=1}^{(\widehat{C}/\epsilon)+1} x^a dx$$
$$\leq \frac{((\widehat{C}/\epsilon)+1)^{a+1}}{a+1} \leq \left(\frac{\widehat{C}}{\epsilon^{a+1}}\right).$$

□

2.3.2 Stochastic Subgradient Extragradient Schemes

We begin by proving the a.s. convergence of the iterates produced by (v-SSE). Unlike (v-SPRG), this scheme does not require an assumption of weak sharpness but mere monotonicity suffices.

Proposition 3 (a.s. convergence of (v-SSE)). *Consider the scheme (v-SSE). Let Assumptions 1 and 5 hold. Suppose* $0 < \gamma_k = \gamma \leq \frac{1}{2L}$ *and* $\sum_{k=1} \frac{1}{N_k} < M$*. Then any sequence generated by (v-SSE) converges to a solution* $x^* \in X$ *in an a.s. sense.*

Proof. By Lemma 1(ii) we have

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq \left\|x_k - \gamma_k \left(F\left(x_{k+\frac{1}{2}}\right) + \bar{w}_{k+\frac{1}{2}}\right) - x^*\right\|^2 \\ &- \left\|x_k - \gamma_k \left(F\left(x_{k+\frac{1}{2}}\right) + \bar{w}_{k+\frac{1}{2}}\right) - x_{k+1}\right\|^2 \\ &= \|x_k - x^*\|^2 - \|x_k - x_{k+1}\|^2 + 2\gamma_k \left(F\left(x_{k+\frac{1}{2}}\right) + \bar{w}_{k+\frac{1}{2}}\right)^T (x^* - x_{k+1}).\end{aligned} \tag{2.12}$$

It is clear that

$$\begin{aligned}F\left(x_{k+\frac{1}{2}}\right)^T (x_{k+1} - x^*) &= F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+1} - x_{k+\frac{1}{2}} + x_{k+\frac{1}{2}} - x^*\right) \\ &= F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+1} - x_{k+\frac{1}{2}}\right) + F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right).\end{aligned} \tag{2.13}$$

Substituting (2.13) in (2.12), we obtain

$$\begin{aligned}&\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \|x_k - x_{k+1}\|^2 + 2\gamma_k F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x_{k+1}\right) \\ &- F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) + 2\gamma_k \bar{w}_{k+\frac{1}{2}}^T (x^* - x_{k+1}) \\ &= \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}} + x_{k+\frac{1}{2}} - x_{k+1}\right\|^2 + 2\gamma_k F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x_{k+1}\right)\end{aligned}$$

$$
\begin{aligned}
&- F\left(x_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}} - x^*\right) + 2\gamma_k \bar{w}_{k+\frac{1}{2}}^T(x^* - x_{k+1}) \\
&= \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - \left\|x_{k+\frac{1}{2}} - x_{k+1}\right\|^2 - 2\left(x_k - x_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}} - x_{k+1}\right) \\
&+ 2\gamma_k F\left(x_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}} - x_{k+1}\right) - F\left(x_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}} - x^*\right) + 2\gamma_k \bar{w}_{k+\frac{1}{2}}^T(x^* - x_{k+1}) \\
&= \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - \left\|x_{k+\frac{1}{2}} - x_{k+1}\right\|^2 - F\left(x_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}} - x^*\right) \\
&+ 2\left(x_{k+1} - x_{k+\frac{1}{2}}\right)^T\left(x_k - \gamma_k F\left(x_{k+\frac{1}{2}}\right) - x_{k+\frac{1}{2}}\right) + 2\gamma_k \bar{w}_{k+\frac{1}{2}}^T(x^* - x_{k+1}).
\end{aligned} \tag{2.14}
$$

By definition of C_k, we have

$$
\left(x_{k+1} - x_{k+\frac{1}{2}}\right)^T\left(x_k - \gamma_k(F(x_k) + \bar{w}_k) - x_{k+\frac{1}{2}}\right) \leq 0. \tag{2.15}
$$

Substituting (2.15) in (2.14), we deduce that

$$
\begin{aligned}
&\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - \left\|x_{k+\frac{1}{2}} - x_{k+1}\right\|^2 \\
&- F\left(x_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}} - x^*\right) + 2\gamma_k\left(x_{k+1} - x_{k+\frac{1}{2}}\right)^T\left(F(x_k) - F\left(x_{k+\frac{1}{2}}\right)\right) \\
&+ 2\gamma_k \bar{w}_k^T\left(x_{k+1} - x_{k+\frac{1}{2}}\right) + 2\gamma_k \bar{w}_{k+\frac{1}{2}}^T(x^* - x_{k+1}) \\
&\leq \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - \left\|x_{k+\frac{1}{2}} - x_{k+1}\right\|^2 - F\left(x_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}} - x^*\right) \\
&+ 2\gamma_k\left\|x_{k+1} - x_{k+\frac{1}{2}}\right\|\left\|F(x_k) - F\left(x_{k+\frac{1}{2}}\right)\right\| + 2\gamma_k\left(\bar{w}_k - \bar{w}_{k+\frac{1}{2}}\right)^T\left(x_{k+1} - x_{k+\frac{1}{2}}\right) \\
&+ 2\gamma_k \bar{w}_{k+\frac{1}{2}}^T\left(x^* - x_{k+\frac{1}{2}}\right) \\
&\leq \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - \left\|x_{k+\frac{1}{2}} - x_{k+1}\right\|^2 + \frac{1}{2}\left\|x_{k+1} - x_{k+\frac{1}{2}}\right\|^2 \\
&+ 2\gamma_k^2 L^2\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2\left\|\bar{w}_k - \bar{w}_{k+\frac{1}{2}}\right\|^2 + \frac{1}{2}\left\|x_{k+1} - x_{k+\frac{1}{2}}\right\|^2 \\
&- F\left(x_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}} - x^*\right) + 2\gamma_k \bar{w}_{k+\frac{1}{2}}^T\left(x^* - x_{k+\frac{1}{2}}\right) \\
&= \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2)\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2\left\|\bar{w}_k - \bar{w}_{k+\frac{1}{2}}\right\|^2 \\
&- F\left(x_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}} - x^*\right) + 2\gamma_k \bar{w}_{k+\frac{1}{2}}^T\left(x^* - x_{k+\frac{1}{2}}\right) \\
&= \|x_k - x^*\|^2 - (1 - 2\gamma^2 L^2)\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma^2\left\|\bar{w}_k - \bar{w}_{k+\frac{1}{2}}\right\|^2 \\
&- F\left(x_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}} - x^*\right) + 2\gamma \bar{w}_{k+\frac{1}{2}}^T\left(x^* - x_{k+\frac{1}{2}}\right),
\end{aligned} \tag{2.16}
$$

by noticing that $\gamma_k = \gamma$. Define $r_\gamma(x) \triangleq \|x - \Pi_X(x - \gamma F(x))\|$ as a residual function. We have

$$
r_\gamma^2(x_k) = \|x_k - \Pi_X(x_k - \gamma F(x_k))\|^2
$$

$$
\begin{aligned}
&= \left\| x_k - x_{k+\frac{1}{2}} + \Pi_X(x_k - \gamma F(x_k) - \gamma \bar{w}_k) - \Pi_X(x_k - \gamma F(x_k)) \right\| \\
&\leq 2 \left\| x_k - x_{k+\frac{1}{2}} \right\|^2 + 2\gamma^2 \|\bar{w}_k\|^2.
\end{aligned}
$$

It follows that

$$
-\frac{1}{2} \left\| x_k - x_{k+\frac{1}{2}} \right\|^2 \leq -\frac{1}{4} r_\gamma^2(x_k) + \frac{1}{2}\gamma^2 \|\bar{w}_k\|^2. \tag{2.17}
$$

Using (2.17) in (2.16), we obtain

$$
\begin{aligned}
&\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \left(\frac{1}{2} - 2\gamma^2 L^2\right) \left\| x_k - x_{k+\frac{1}{2}} \right\|^2 + 2\gamma^2 \left\| \bar{w}_k - \bar{w}_{k+\frac{1}{2}} \right\|^2 \\
&- F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) + 2\gamma \bar{w}_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right) - \frac{1}{2} \left\| x_k - x_{k+\frac{1}{2}} \right\|^2 \\
&\leq \|x_k - x^*\|^2 - \left(\frac{1}{2} - 2\gamma^2 L^2\right) \left\| x_k - x_{k+\frac{1}{2}} \right\|^2 + 2\gamma^2 \left\| \bar{w}_k - \bar{w}_{k+\frac{1}{2}} \right\|^2 \\
&- F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) + 2\gamma \bar{w}_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right) - \frac{1}{4} r_\gamma^2(x_k) + \frac{1}{2}\gamma^2 \|\bar{w}_k\|^2 \\
&\leq \|x_k - x^*\|^2 - \left(\frac{1}{2} - 2\gamma^2 L^2\right) \left\| x_k - x_{k+\frac{1}{2}} \right\|^2 + \frac{9}{2}\gamma^2 \|\bar{w}_k\|^2 + 4\gamma^2 \left\| \bar{w}_{k+\frac{1}{2}} \right\|^2 \\
&- F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) + 2\gamma \bar{w}_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right) - \frac{1}{4} r_\gamma^2(x_k).
\end{aligned}
$$

Taking expectations conditioned on $\mathcal{F}_k$ and leveraging $\gamma \leq \frac{1}{2L}$, we obtain the following bound:

$$
\begin{aligned}
&\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^*\|^2 + \mathbb{E}\left[\mathbb{E}\left[4\gamma^2 \left\| \bar{w}_{k+\frac{1}{2}} \right\|^2 \mid \mathcal{F}_{k+\frac{1}{2}}\right] \mid \mathcal{F}_k\right] \\
&+ \mathbb{E}\left[\frac{9}{2}\gamma^2 \|\bar{w}_k\|^2 \mid \mathcal{F}_k\right] - \mathbb{E}\left[\mathbb{E}\left[2\gamma \bar{w}_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right) \mid \mathcal{F}_{k+\frac{1}{2}}\right] \mid \mathcal{F}_k\right] - \frac{1}{4} r_\gamma^2(x_k) \\
&\leq \|x_k - x^*\|^2 + \frac{17}{2}\gamma^2 \frac{\nu^2}{N_k} - \mathbb{E}\left[2\gamma \mathbb{E}\left[\bar{w}_{k+\frac{1}{2}} \mid \mathcal{F}_{k+\frac{1}{2}}\right]^T \left(x_{k+\frac{1}{2}} - x^*\right) \mid \mathcal{F}_k\right] - \frac{1}{4} r_\gamma(x_k)^2 \\
&= \|x_k - x^*\|^2 + \frac{17}{2}\gamma^2 \frac{\nu^2}{N_k} - \frac{1}{4} r_\gamma^2(x_k).
\end{aligned}
$$

We may now apply Lemma 2 which allows us to claim that $\{\|x_k - x^*\|\}$ is convergent and $\sum_k r_\gamma(x_k)^2 < \infty$ in an a.s. sense. Therefore, in an a.s. sense, we have

$$
\lim_{k \to \infty} r_\gamma(x_k)^2 = 0.
$$

This implies that the entire sequence $\{x_k\}$ converges to a point in X^* in an a.s. sense. $\square$

Next we derive rate statements for the averaged sequence in the mere monotonicity. Unlike in stochastic convex optimization where the function value represents a metric to ascertain progress of the algorithm, a similar metric is not immediately available for variational inequality problems. Instead, the progress of the scheme can be ascertained by using the gap function, defined next.

Definition 4 (Gap function). *Given a nonempty closed set $X \subseteq \mathbb{R}^n$ and a mapping $F : \mathbb{R}^n \to \mathbb{R}^n$, then the gap function at x is denoted by $G(x)$ and is defined as follows for any $x \in X$.*

$$G(x) \triangleq \sup_{y \in X} F(y)^T (x - y).$$

The gap function is nonnegative for all $x \in X$ and is zero if and only if x is a solution of SVI (cf. [6]). We establish the convergence rate for (v-SSE) by using the gap function.

Proposition 4. *Consider the (v-SSE) scheme and let $\{\bar{x}_K\}$ be defined as $\bar{x}_K = \sum_{k=1}^{K} x_{k+\frac{1}{2}}/K$, where $0 < \gamma_k = \gamma \leq 1/(2L)$ for all $k \geq 0$ and $\sum_{k=1}^{\infty} \frac{1}{N_k} < M$. Let Assumptions 1, 2, 5 hold. (1). Then we have $\mathbb{E}[G(\bar{x}_K)] \leq \mathcal{O}\left(\frac{1}{K}\right)$ for any K. (2). Suppose $N_k = \lfloor k^a \rfloor$, for $a > 1$. Then the oracle complexity to compute an $\bar{x}_K$ such that $\mathbb{E}[G(\bar{x}_K)] \leq \epsilon$ is bounded as follows: $\sum_{k=1}^{K} N_k \leq \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$.*

Proof. (1). From (2.16) and by replacing x^* by y, we obtain

$$\begin{aligned} F(y)^T \left(x_{k+\frac{1}{2}} - y\right) &\leq \|x_k - y\|^2 - \|x_{k+1} - y\|^2 - (1 - 2\gamma^2 L^2) \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 \\ &\quad + 2\gamma^2 \left\|\bar{w}_k - \bar{w}_{k+\frac{1}{2}}\right\|^2 + 2\gamma \bar{w}_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right). \end{aligned}$$

Summing over k, we obtain the following bound:

$$\begin{aligned} \sum_{k=1}^{K} F(y)^T \left(x_{k+\frac{1}{2}} - y\right) &\leq \|x_1 - y\|^2 + 2\gamma^2 \sum_{k=1}^{K} \left\|\bar{w}_k - \bar{w}_{k+\frac{1}{2}}\right\|^2 \\ &\quad + 2\gamma \sum_{k=1}^{K} \bar{w}_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right) \\ \implies \frac{1}{K} \sum_{k=1}^{K} F(y)^T \left(x_{k+\frac{1}{2}} - y\right) &\leq \frac{1}{K} \|x_1 - y\|^2 + \frac{2\gamma^2 \sum_{k=1}^{K} \left\|\bar{w}_k - \bar{w}_{k+\frac{1}{2}}\right\|^2}{K} \\ &\quad + \frac{\sum_{k=1}^{K} 2\gamma \bar{w}_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right)}{K} \end{aligned}$$

$$\text{or } F(y)^T(\bar{x}_K - y) \leq \frac{1}{K}\|x_1 - y\|^2 + \frac{2\gamma^2 \sum_{k=1}^K \left\|\bar{w}_k - \bar{w}_{k+\frac{1}{2}}\right\|^2}{K} + \frac{\sum_{k=1}^K 2\gamma \bar{w}_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right)}{K}.$$

By taking supremum over $y \in X$, we obtain the following inequality:

$$\begin{aligned}
\sup_{y\in X} F(y)^T(\bar{x}_K - y) &\leq \frac{1}{K}\sup_{y\in X}\|x_1 - y\|^2 + \frac{2\gamma^2 \sum_{k=1}^K \left\|\bar{w}_k - \bar{w}_{k+\frac{1}{2}}\right\|^2}{K} \\
&\quad + \frac{\sum_{k=1}^K 2\gamma \bar{w}_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right)}{K} \\
\implies G(\bar{x}_K) &\leq \frac{B^2}{K} + \frac{2\gamma^2 \sum_{k=1}^K \left\|\bar{w}_k - \bar{w}_{k+\frac{1}{2}}\right\|^2}{K} + \frac{\sum_{k=1}^K 2\gamma \bar{w}_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right)}{K}.
\end{aligned}$$

Taking expectations on both sides, leads to the following inequality.

$$\begin{aligned}
\mathbb{E}[G(\bar{x}_K)] &\leq \frac{B^2}{K} + \frac{2\gamma^2 \sum_{k=1}^K \mathbb{E}\left[\left\|\bar{w}_k - \bar{w}_{k+\frac{1}{2}}\right\|^2\right]}{K} + \frac{\sum_{k=1}^K 2\gamma \mathbb{E}\left[\bar{w}_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right)\right]}{K} \\
&\leq \frac{B^2 + \sum_{k=1}^K \frac{8\gamma^2\nu^2}{N_k}}{K}.
\end{aligned}$$

It follows that $\mathbb{E}[G(\bar{x}_K)] \leq \mathcal{O}(1/K)$.
(2). We can use a same proof manner with Proposition 2(2). □

Remark: While the statements display the similar rates for these three methods, the constants are naturally quite distinct. In particular, we note that the Lipschitz constant appears in the bounds defining the complexity of (SPRG) and lead to a somewhat poorer bound. Yet, as the numerics display, these distinctions are less evident in practice suggesting that the bounds are relatively weak.

2.4 Incorporating Random Projections in (SPRG) and (SSE)

In this section, we assume that even a single projection onto the feasible set X is challenging. We assume that X is given by an intersection of a collection of closed and convex sets $\{X_i\}_{i\in I}$ where I is a finite set and consider a variants of

(**SPRG**) and (**SSE**) where the projection onto X is replaced by a projection onto a randomly selected set X_i. In Section 2.4.1, we review our main assumptions and any supporting results and proceed to derive asymptotic and rate guarantees in Sections 2.4.2 and 2.5.2 for the random projection variants of (**SPRG**) and (**SSE**), respectively.

2.4.1 Assumptions and Supporting Results

To establish the convergence, we need the following addtional assumptions on the projection set $X = \bigcap_{i\in I} X_i$ and random projection process Π_{l_k}. The following assumption is known as linear regularity discussed in [107]. It indicates that this condition is a mild restriction in practice.

Assumption 7. *There exists a positive scalar η such that for any $x \in \mathbb{R}^n$*

$$\|x - \Pi_X(x)\|^2 \leq \eta \max_{i\in I} \|x - \Pi_{X_i}(x)\|^2,$$

where I is a finite set of indexes, $I = \{1, \ldots, m\}$.

The following assumption requires that each constraint is sampled with at least some probability and the random samples are nearly independent, which refers to [107].

Assumption 8. *The random variables $l_k, k = 0, 1, \ldots,$ are such that*

$$\inf_{k\geq 0} P(l_k = X_i \mid \mathcal{F}_k) \geq \frac{\rho_i}{m}, \quad i = 1, \ldots, m,$$

with probability 1, where for $i = 1, \ldots, m$, $\rho_i \in (0, 1]$ is a scalar.

The following lemma is essential to our proofs and it leverages basic properties of projection.

Lemma 4. *Let X be a closed convex subset of $\mathbb{R}^n$. We have*

$$\|y - \Pi_X(y)\|^2 \leq 2\|x - \Pi_X(x)\|^2 + 8\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n.$$

Proof. Since $y - \Pi_X(y) = (x - \Pi_X(x)) - (x - y) + (\Pi_X(x) - \Pi_X(y))$, we have

$$\|y - \Pi_X(y)\| \leq \|x - \Pi_X(x)\| + \|x - y\| + \|\Pi_X(x) - \Pi_X(y)\|$$

$$\leq \|x - \Pi_X(x)\| + 2\|x - y\|.$$

Thus,

$$\|y - \Pi_X(y)\|^2 \leq 2\|x - \Pi_X(x)\|^2 + 8\|x - y\|^2,$$

where the last inequality leverages $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. □

The following lemma provides an inequality which is useful in deriving lower bound of $\|x_{k+1} - x^*\|^2$.

Lemma 5. *Under Assumptions 1, 3 and 4, we have*

$$F(x)^T(x - x^*) \geq \alpha\, dist\,(\Pi_X(x), X^*) - C\, dist(x, X), \quad \forall x \in \mathbb{R}^n.$$

Proof. We have

$$\begin{aligned} F(x)^T(x - x^*) = &(F(x) - F(x^*))^T(x - x^*) + F(x^*)^T(\Pi_X(x) - x^*) \\ &+ F(x^*)^T(x - \Pi_X(x)). \end{aligned} \tag{2.18}$$

From the monotonicity assumption on F, we have

$$(F(x) - F(x^*))^T(x - x^*) \geq 0. \tag{2.19}$$

Since x^* is a solution, it follows that from the weak sharpness property,

$$F(x^*)^T(\Pi_X(x) - x^*) \geq \alpha \text{dist}\,(\Pi_X(x), X^*)\,. \tag{2.20}$$

Finally, $F(x^*)^T(\Pi_X(x) - x) \leq \|F(x^*)\|\|x - \Pi_X(x)\|$ and $\|F(x^*)\| \leq C$ (by Assumption 3),

$$F(x^*)^T(x - \Pi_X(x)) \geq -\|F(x^*)\|\|x - \Pi_X(x)\| \geq -C\text{dist}(x, X). \tag{2.21}$$

By substituting (2.19) – (2.21) in (2.18), the result follows. □

Lemma 6. *Suppose Assumptions 1 and 3 hold. Then for any $x \in \mathbb{R}^n$,*

$$\|F(x)\|^2 \leq 2L^2\|x - x^*\|^2 + 2C^2.$$

Proof. The result follows by using the triangle inequality $\|F(x)\| \leq \|F(x) - F(x^*)\| + \|F(x^*)\|$. □

Lemma 7. *Suppose Assumptions 7 and 8 hold. Then for any $l_k \in I$ and any $x \in \mathbb{R}^n$,*

$$\mathbb{E}[\|x - \Pi_{l_k}(x)\|^2 \mid \mathcal{F}_k] \geq \frac{\rho}{m\eta} dist^2(x, X), \quad k \geq 0,$$

with probability 1, where $\rho \triangleq \min_{i \in I}\{\rho_i\}$.

Proof. Following from Assumption 8, we have

$$\begin{aligned}
\mathbb{E}[\|x - \Pi_{l_k}(x)\|^2 \mid \mathcal{F}_k] &= \sum_{i=1}^{m} P(l_k = i \mid \mathcal{F}_k)\|x - \Pi_i(x)\|^2 \\
&\geq \frac{\rho}{m}\|x - \Pi_j(x)\|^2, \quad \forall j = 1, \ldots, m \\
\implies \mathbb{E}[\|x - \Pi_{l_k}(x)\|^2 \mid \mathcal{F}_k] &\geq \frac{\rho}{m}\max_j \|x - \Pi_j(x)\|^2 \overset{\text{(Ass. 7)}}{\geq} \frac{\rho}{m\eta}\text{dist}^2(x, X).
\end{aligned}$$

□

2.4.2 SPRG with Random Projections

We begin with an a.s. convergence claim for (**r-SPRG**).

Proposition 5. *Let Assumptions 1, 3 – 8 hold. Then any sequence generated by (r-SPRG), where the projections are randomly generated, converges to a solution $x^* \in X$ in an a.s. sense.*

Proof. Define $y_k = 2x_k - x_{k-1}$ for all $k \geq 1$. By Lemma 1(ii) and by noting that $x_{k+1} = \Pi_X(x_k - \gamma_k F(2x_k - x_{k-1}))$ and $F(x_k, \omega_k) = F(x_k) + w_k$, we have the following inequality:

$$\begin{aligned}
&\|x_{k+1} - x^*\|^2 \leq \|x_k - \gamma_k F(y_k, \omega_k) - x^*\|^2 - \|x_k - \gamma_k F(y_k, \omega_k) - x_{k+1}\|^2 \\
&= \|x_k - x^*\|^2 - \|x_{k+1} - x_k\|^2 - 2\gamma_k (F(y_k) + w_k)^T (x_{k+1} - x^*) \\
&= \|x_k - x^*\|^2 - \|x_{k+1} - x_k\|^2 - 2\gamma_k F(y_k)^T (x_{k+1} - x^*) - 2\gamma_k w_k^T (x_{k+1} - x^*).
\end{aligned} \tag{2.22}$$

Since

$$\|y_k - x_{k+1}\|^2 = 2\|x_k - x_{k+1}\|^2 - \|x_{k-1} - x_{k+1}\|^2 + 2\|x_k - x_{k-1}\|^2,$$

We have

$$\frac{1}{4}\|x_k - x_{k+1}\|^2 = \frac{1}{8}\|y_k - x_{k+1}\|^2 + \frac{1}{8}\|x_{k-1} - x_{k+1}\|^2 - \frac{1}{4}\|x_k - x_{k-1}\|^2. \tag{2.23}$$

Using (2.23) in (2.22), we obtain

$$\begin{aligned}
&\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \frac{3}{4}\|x_{k+1} - x_k\|^2 - \frac{1}{8}\|y_k - x_{k+1}\|^2 - \frac{1}{8}\|x_{k-1} - x_{k+1}\|^2 \\
&+ \frac{1}{4}\|x_k - x_{k-1}\|^2 - 2\gamma_k F(y_k)^T(x_{k+1} - x^*) - 2\gamma_k w_k^T(x_{k+1} - x^*) \\
&= \|x_k - x^*\|^2 - \frac{3}{4}\|x_{k+1} - x_k\|^2 - \frac{1}{8}\|y_k - x_{k+1}\|^2 - \frac{1}{8}\|x_{k-1} - x_{k+1}\|^2 \\
&+ \frac{1}{4}\|x_k - x_{k-1}\|^2 - 2\gamma_k F(y_k)^T(y_k - x^*) - 2\gamma_k F(y_k)^T(x_{k+1} - y_k) \\
&- 2\gamma_k w_k^T(x_{k+1} - x^*) \\
&\leq \|x_k - x^*\|^2 - \frac{3}{4}\|x_{k+1} - x_k\|^2 - \frac{1}{8}\|y_k - x_{k+1}\|^2 - \frac{1}{8}\|x_{k-1} - x_{k+1}\|^2 \\
&+ \frac{1}{4}\|x_k - x_{k-1}\|^2 - 2\gamma_k \alpha \mathrm{dist}\,(\Pi_X(y_k), X^*) + 2\gamma_k C \mathrm{dist}(y_k, X) \\
&- 2\gamma_k F(y_k)^T(x_{k+1} - y_k) - 2\gamma_k w_k^T(x_{k+1} - x^*),
\end{aligned} \tag{2.24}$$

where the last inequality follows from Lemma 5. Since

$$-2\gamma_k F(y_k)^T(x_{k+1} - y_k) \leq 16\gamma_k^2\|F(y_k)\|^2 + \frac{1}{16}\|x_{k+1} - y_k\|^2 \tag{2.25}$$

Using (2.25) in (2.24), we have

$$\begin{aligned}
&\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \frac{3}{4}\|x_{k+1} - x_k\|^2 - \frac{1}{16}\|y_k - x_{k+1}\|^2 - \frac{1}{8}\|x_{k-1} - x_{k+1}\|^2 \\
&+ \frac{1}{4}\|x_k - x_{k-1}\|^2 - 2\gamma_k \alpha \mathrm{dist}\,(\Pi_X(y_k), X^*) + 2\gamma_k C \mathrm{dist}(y_k, X) + 16\gamma_k^2\|F(y_k)\|^2 \\
&+ 16\gamma_k^2\|w_k\|^2 - 2\gamma_k w_k^T(y_k - x^*) \\
&\leq \|x_k - x^*\|^2 - \frac{3}{4}\|x_{k+1} - x_k\|^2 - \frac{1}{16}\|y_k - x_{k+1}\|^2 - \frac{1}{8}\|x_{k-1} - x_{k+1}\|^2 \\
&+ \frac{1}{4}\|x_k - x_{k-1}\|^2 - 2\gamma_k \alpha \mathrm{dist}\,(\Pi_X(y_k), X^*) + 2\gamma_k C \mathrm{dist}(y_k, X) + 32\gamma_k^2 L^2\|y_k - x^*\|^2 \\
&+ 32\gamma_k^2 C^2 + 16\gamma_k^2\|w_k\|^2 - 2\gamma_k w_k^T(y_k - x^*).
\end{aligned}$$

Since

$$-2\gamma_k \alpha \mathrm{dist}\,(\Pi_X(y_k), X^*) \leq -2\gamma_k \alpha \mathrm{dist}\,(x_k, X^*) + 2\gamma_k \alpha\|x_k - \Pi_X(y_k)\|$$

$$
\begin{aligned}
&\leq -2\gamma_k\alpha \mathrm{dist}\,(x_k, X^*) + 2\gamma_k\alpha\|x_k - y_k\| + 2\gamma_k\alpha\|y_k - \Pi_X(y_k)\| \\
&= -2\gamma_k\alpha \mathrm{dist}\,(x_k, X^*) + 2\gamma_k\alpha\|x_k - y_k\| + 2\gamma_k\alpha \mathrm{dist}(y_k, X),
\end{aligned}
$$

we have

$$
\begin{aligned}
&\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\gamma_k\alpha \mathrm{dist}\,(x_k, X^*) - \frac{3}{4}\|x_{k+1} - x_k\|^2 \\
&- \frac{1}{16}\|y_k - x_{k+1}\|^2 - \frac{1}{8}\|x_{k-1} - x_{k+1}\|^2 + \frac{1}{4}\|x_k - x_{k-1}\|^2 + 2\gamma_k\alpha\|x_k - y_k\| \\
&+ 2\gamma_k(C+\alpha)\mathrm{dist}(y_k, X) + 64\gamma_k^2 L^2\|x_k - x^*\|^2 + 64\gamma_k^2 L^2\|x_k - x_{k-1}\|^2 \\
&+ 32\gamma_k^2 C^2 + 16\gamma_k^2\|w_k\|^2 - 2\gamma_k w_k^T(y_k - x^*). \qquad (2.26)
\end{aligned}
$$

By Lemma 7,

$$
\mathbb{E}[\|y_k - x_{k+1}\|^2 \mid \mathcal{F}_k] \geq \mathbb{E}[\|y_k - \Pi_{l_k} y_k\|^2 \mid \mathcal{F}_k] \geq \frac{\rho}{m\eta} d^2(y_k). \qquad (2.27)
$$

Taking expectations conditioned on $\mathcal{F}_k$ and using (2.27) in (2.26), we have

$$
\begin{aligned}
&\mathbb{E}[\|x_{k+1} - x^*\|^2 + \frac{3}{4}\|x_{k+1} - x_k\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^*\|^2 - 2\gamma_k\alpha \mathrm{dist}\,(x_k, X^*) \\
&- \frac{1}{16}\mathbb{E}[\|y_k - x_{k+1}\|^2 \mid \mathcal{F}_k] - \frac{1}{8}\mathbb{E}[\|x_{k-1} - x_{k+1}\|^2 \mid \mathcal{F}_k] + \frac{1}{4}\|x_k - x_{k-1}\|^2 \\
&+ 2\gamma_k\alpha\|x_k - y_k\| + 2\gamma_k(C+\alpha)\mathrm{dist}(y_k, X) + 64\gamma_k^2 L^2\|x_k - x^*\|^2 \\
&+ 64\gamma_k^2 L^2\|x_k - x_{k-1}\|^2 + 32\gamma_k^2 C^2 + 16\gamma_k^2\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] \\
&\leq \|x_k - x^*\|^2 - 2\gamma_k\alpha \mathrm{dist}\,(x_k, X^*) - \frac{1}{16}\frac{\rho}{m\eta} d^2(y_k) + \frac{1}{4}\|x_k - x_{k-1}\|^2 \\
&+ 2\gamma_k\alpha\|x_k - y_k\| + 2\gamma_k(C+\alpha)\mathrm{dist}(y_k, X) \\
&+ 64\gamma_k^2 L^2\|x_k - x^*\|^2 + 64\gamma_k^2 L^2\|x_k - x_{k-1}\|^2 + 32\gamma_k^2 C^2 + 16\gamma_k^2\nu^2 \\
&= \|x_k - x^*\|^2 + \frac{3}{4}\|x_k - x_{k-1}\|^2 - 2\gamma_k\alpha \mathrm{dist}\,(x_k, X^*) - \frac{1}{2}\|x_k - x_{k-1}\|^2 \\
&+ 2\gamma_k\alpha\|x_k - x_{k-1}\| + 2\gamma_k(C+\alpha)\mathrm{dist}(y_k, X) - \frac{1}{16}\frac{\rho}{m\eta} d^2(y_k) \\
&+ 64\gamma_k^2 L^2\|x_k - x^*\|^2 + 64\gamma_k^2 L^2\|x_k - x_{k-1}\|^2 + 32\gamma_k^2 C^2 + 16\gamma_k^2\nu^2 \\
&\leq \|x_k - x^*\|^2 + \frac{3}{4}\|x_k - x_{k-1}\|^2 - 2\gamma_k\alpha \mathrm{dist}\,(x_k, X^*) - \frac{1}{2}\|x_k - x_{k-1} - 2\gamma_k\alpha\|^2 \\
&+ 2\gamma_k^2\alpha^2 - \frac{\rho}{16m\eta}\left(\mathrm{dist}(y_k, X) - \frac{16m\eta\gamma_k(C+\alpha)}{\rho}\right)^2 + \frac{16m\eta(C+\alpha)^2}{\rho}\gamma_k^2 \\
&+ 64\gamma_k^2 L^2\|x_k - x^*\|^2 + 64\gamma_k^2 L^2\|x_k - x_{k-1}\|^2 + 32\gamma_k^2 C^2 + 16\gamma_k^2\nu^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq (1+86\gamma_k^2L^2)\left(\|x_k-x^*\|^2+\frac{3}{4}\|x_k-x_{k-1}\|^2\right) \\
&-\underbrace{\left(\tfrac{1}{2}\|x_k-x_{k-1}+2\gamma_k\alpha\|^2+2\gamma_k\alpha\text{dist}(x_k,X^*)+\tfrac{\rho}{16m\eta}\left(\text{dist}(y_k,X)-\tfrac{16m\eta\gamma_k(C+\alpha)}{\rho}\right)^2\right)}_{\beta_k} \\
&+\underbrace{\left(2\gamma_k^2\alpha^2+\frac{16m\eta(C+\alpha)^2}{\rho}\gamma_k^2+32\gamma_k^2C^2+16\gamma_k^2\nu^2\right)}_{\eta_k}. \qquad (2.28)
\end{aligned}
$$

In effect, we obtain the following recursion:

$$
\mathbb{E}[v_{k+1} \mid \mathcal{F}_k] \leq (1-u_k)v_k - \beta_k + \eta_k,
$$

where $v_k \triangleq \left(\|x_k-x^*\|^2+\frac{3}{4}\|x_k-x_{k-1}\|^2\right)$ and $u_k = 86\gamma_k^2L^2$. Since $\sum \gamma_k^2 < \infty$, it follows that u_k and β_k are summable. We may then invoke Lemma 2 and it follows that with probability one, the random sequence $\{\|x_k-x^*\|^2+\frac{3}{4}\|x_k-x_{k-1}\|^2\}$ is convergent and $\sum\{\frac{1}{2}\|x_k-x_{k-1}-2\gamma_k\alpha\|^2+2\gamma_k\alpha\text{dist}\,(x_k,X^*)\} < \infty$ with probability one. We have that $\sum_k \frac{1}{2}\|x_k-x_{k-1}-2\gamma_k\alpha\|^2 < \infty$ implying that $\|x_k-x_{k-1}-2\gamma_k\alpha\| \to 0$ in a.s. sense. It follows that $\|y_k-x_k-2\gamma_k\alpha\| \to 0$ a.s. Since $\gamma_k \to 0$, it follows that $y_k - x_k \to 0$ in an a.s. sense, which means $x_k - x_{k-1} \to 0$ in an a.s. sense. Thus $\{\|x_k-x^*\|\}$ is convergent in an a.s. sense. We may then conclude by contradiction that $\text{dist}(x_k,X^*) \to 0$ in an a.s. sense. If not, then with finite probability, every subsequence of $\{x_k\}$ satisfies $\text{dist}(x_k,X^*) \to h(\omega) \geq \bar{h} > 0$ implying that $\sum_{k=1}^{\infty}\gamma_k\alpha\text{dist}(x_k,X^*) = \infty$ with finite probability. This contradicts $\sum_k \beta_k < \infty$, implying that $x_k \xrightarrow{k\to\infty} x^*$ in an a.s. sense.

□

We now provide a rate and oracle complexity statement for this scheme.

Proposition 6. *Let Assumptions 1 – 5, 7 – 8 hold and let* $0 < \gamma_k = \gamma = \frac{\sqrt{7}B}{2\sqrt{M_1K}}$*, where* K *is the pre-defined termination number of iterations and* $M_1 = \frac{301}{2}L^2B^2 + 2\alpha^2 + \frac{16m\eta(C+\alpha)^2}{\rho} + 32C^2 + 16\nu^2$*. Then the following holds for any sequence generated by (r-SPRG) in an expected value sense, where* $\bar{x}_k = \sum_{k=0}^{K-1} x_k/K$*: (1)* $\mathbb{E}[\mathit{dist}\,(\bar{x}_K,X^*)] \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$*; (2) The oracle complexity to compute an* $\bar{x}_K$ *such that* $\mathbb{E}[\mathit{dist}(\bar{x}_k,X^*)]$ *is bounded as follows:* $\sum_{k=1}^{K} N_k \leq \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$*, where* $N_k = 1$ *for all* k*.*

Proof. (1). Taking expectations on both sides of (2.28), we have

$$
2\gamma_k\alpha\mathbb{E}[\text{dist}\,(x_k,X^*)] \leq \mathbb{E}\left[\|x_k-x^*\|^2+\frac{3}{4}\|x_k-x_{k-1}\|^2\right]
$$

$$
\begin{aligned}
&- \mathbb{E}\left[\|x_{k+1} - x^*\|^2 + \frac{3}{4}\|x_{k+1} - x_k\|^2\right] + 86\gamma_k^2 L^2 \mathbb{E}\left[\|x_k - x^*\|^2 + \frac{3}{4}\|x_k - x_{k-1}\|^2\right] \\
&+ 2\gamma_k^2\alpha^2 + \frac{16m\eta(C+\alpha)^2}{\rho}\gamma_k^2 + 32\gamma_k^2 C^2 + 16\gamma_k^2\nu^2 \\
&\leq \mathbb{E}\left[\|x_k - x^*\|^2 + \frac{3}{4}\|x_k - x_{k-1}\|^2\right] \\
&- \mathbb{E}\left[\|x_{k+1} - x^*\|^2 + \frac{3}{4}\|x_{k+1} - x_k\|^2\right] + \gamma_k^2 M_1,
\end{aligned}
$$

where $M_1 = \frac{301}{2}L^2B^2 + 2\alpha^2 + \frac{16m\eta(C+\alpha)^2}{\rho} + 32C^2 + 16\nu^2$.
Summing over k from $k = 0$ to $K - 1$, we have

$$
\begin{aligned}
2\gamma\alpha \sum_{k=0}^{K-1} \mathbb{E}[\mathrm{dist}\,(x_k, X^*)] &\leq \mathbb{E}\left[\|x_0 - x^*\|^2 + \frac{3}{4}\|x_0 - x_{-1}\|^2\right] \\
&\quad - \mathbb{E}\left[\|x_K - x^*\|^2 + \frac{3}{4}\|x_K - x_{K-1}\|^2\right] + K\gamma^2 M_1 \\
&\leq \mathbb{E}\left[\|x_0 - x^*\|^2 + \frac{3}{4}\|x_0 - x_{-1}\|^2\right] + K\gamma^2 M_1.
\end{aligned}
$$

It follows that $2\gamma\alpha\mathbb{E}[\mathrm{dist}\,(\bar{x}_K, X^*)] \leq \frac{7B^2}{4K} + \gamma^2 M_1$. Dividing both sides by $2\gamma\alpha$ and optimizing the right-hand side in γ, we obtain the following when $\gamma^* = \frac{\sqrt{7}B}{2\sqrt{M_1 K}}$.

$$
\mathbb{E}[\mathrm{dist}\,(\bar{x}_K, X^*)] \leq \frac{7B^2}{8K\gamma\alpha} + \frac{\gamma M_1}{2\alpha} = \frac{\sqrt{7M_1}B}{2\alpha\sqrt{K}} = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right),
$$

(2). From (1), we know that $K = \mathcal{O}(1/\epsilon^2)$ and it follows that

$$
\sum_{k=1}^{K} N_k = \sum_{k=1}^{K} 1 = K = \mathcal{O}\left(\frac{1}{\epsilon^2}\right).
$$

□

The feasibility error arises because the random projection algorithms cannot guarantee $\{x_k\}$ to be feasible. First we conduct almost-sure convergence analysis on the metric $\{\mathrm{dist}(x_k, X)\}$ for both randomly generated algorithms and then derive the optimal rate of convergence. To establish the rate of convergence, we need the following lemma.

Lemma 8. *Let $\{\delta_k\}$ and $\{\alpha_k\}$ be sequences of nonnegative scalars such that*

$$
\delta_{k+1} \leq (1-\beta)\delta_k + K\alpha_k^2, \quad \forall k \geq 0,
$$

where $\beta \in (0,1)$ and $K \leq 0$ are constants. If there exists $\bar{k} \geq 0$ such that $\alpha_{k+1}^2 \geq (1-\frac{\beta}{2})\alpha_k^2$ for all $k \geq \bar{k}$, we have

$$\delta_k \leq \frac{2N}{\beta}\alpha_k^2 + \delta_0(1-\beta)^k + \left(K\sum_{t=0}^{\bar{k}}\alpha_t^2\right)(1-\beta)^{k-\bar{k}}.$$

Proof. Please refer to [107]. □

Proposition 7. *Let Assumptions 1 – 3, 5 – 8 hold. Suppose $\{x_k\}$ is generated by (r-SPRG), where the projections are randomly generated. Then $\mathbb{E}[dist(\bar{x}_K, X)] \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ for any $K > 0$.*

Proof. Let $z_k = x_k - \gamma_k F(2x_k - x_{k-1}, \omega_k)$. We have

$$\begin{aligned}\operatorname{dist}^2(x_{k+1}, X) &\leq \|x_{k+1} - \Pi_X(z_k)\|^2 = \|\Pi_{l_k}(z_k) - \Pi_X(z_k)\|^2 \\ &\leq \|z_k - \Pi_X(z_k)\|^2 - \|\Pi_{l_k}(z_k) - z_k\|^2, \end{aligned} \tag{2.29}$$

where it follows from Lemma 1. By leveraging $\|a+b\|^2 \leq \left(1+\frac{4m\eta}{\rho}\right)\|a\|^2 + \left(1+\frac{\rho}{4m\eta}\right)\|b\|^2$, we obtain

$$\begin{aligned}\|z_k - \Pi_X(z_k)\|^2 &\leq \|z_k - \Pi_X(x_k)\|^2 = \|z_k - x_k + x_k - \Pi_X(x_k)\|^2 \\ &\leq \left(1+\frac{4m\eta}{\rho}\right)\|z_k - x_k\|^2 + \left(1+\frac{\rho}{4m\eta}\right)\|x_k - \Pi_X(x_k)\|^2. \end{aligned} \tag{2.30}$$

Combining (2.29) and (2.30), we get

$$\begin{aligned}\operatorname{dist}^2(x_{k+1}, X) \leq &\left(1+\frac{4m\eta}{\rho}\right)\|z_k - x_k\|^2 + \left(1+\frac{\rho}{4m\eta}\right)\operatorname{dist}^2(x_k, X) \\ &- \|\Pi_{l_k}(z_k) - z_k\|^2. \end{aligned} \tag{2.31}$$

Following from Lemma 4 and 7, we have

$$\begin{aligned}&\mathbb{E}[\|z_k - \Pi_{l_k}(z_k)\|^2 \mid \mathcal{F}_k] \geq \frac{\rho}{m\eta}d^2(z_k) \geq \frac{\rho}{m\eta}\left(\frac{1}{2}\operatorname{dist}^2(x_k, X) - 4\|z_k - x_k\|^2\right) \\ &\geq \frac{\rho}{2m\eta}\operatorname{dist}^2(x_k, X) - \frac{4\rho}{m\eta}\|z_k - x_k\|^2 \geq \frac{\rho}{2m\eta}\operatorname{dist}^2(x_k, X) - 4\|z_k - x_k\|^2. \end{aligned} \tag{2.32}$$

Applying (2.32) to (2.31), it follows that

$$\mathbb{E}[\operatorname{dist}^2(x_{k+1}, X) \mid \mathcal{F}_k] \leq \left(1-\frac{\rho}{4m\eta}\right)\operatorname{dist}^2(x_k, X) + \left(5+\frac{4m\eta}{\rho}\right)\|z_k - x_k\|^2$$

$$\leq \left(1 - \frac{\rho}{4m\eta}\right) \mathrm{dist}^2(x_k, X) + \left(5 + \frac{4m\eta}{\rho}\right) (4L^2B^2 + 4C^2 + 2\nu^2)\gamma_k^2.$$

It is clear that $\gamma_{k+1}^2 \geq \left(1 - \frac{\rho}{8m\eta}\right)\gamma_k^2$ when k is sufficiently large. Leveraging Lemma 8, we have

$$\mathbb{E}[\mathrm{dist}^2(x_k, X)] \leq \left(\frac{40m\eta}{\rho} + \frac{32m^2\eta^2}{\rho^2}\right)(4L^2B^2 + 4C^2 + 2\nu^2)\gamma_k^2 + d(x_0)\left(1 - \frac{\rho}{4m\eta}\right)^k$$
$$+ \left(\left(5 + \frac{4m\eta}{\rho}\right)(4L^2B^2 + 4C^2 + 2\nu^2)\sum_{t=0}^{\bar{k}} \gamma_t^2\right)\left(1 - \frac{\rho}{4m\eta}\right)^{k-\bar{k}}$$

When k is sufficiently large, it satisfies that

$$\mathbb{E}[\mathrm{dist}^2(x_k, X)] \leq \left(\left(\frac{40m\eta}{\rho} + \frac{32m^2\eta^2}{\rho^2}\right)(4L^2B^2 + 4C^2 + 2\nu^2) + U_1\right)\gamma_k^2,$$

where U_1 is a large number. It follows that

$$\mathbb{E}[\mathrm{dist}^2(\bar{x}_K, X)] \leq \frac{\sum_{k=0}^{K-1}\mathbb{E}[\mathrm{dist}^2(x_k, X)]}{K} \leq \frac{\mathcal{O}\left(\sum_{k=0}^{K-1}\gamma_k^2\right)}{K} = \mathcal{O}\left(\frac{1}{K}\right), \tag{2.33}$$

where we assume $\sum_{k=0}^{K-1}\gamma_k^2 < M < \infty$. By Jensen's inequality, we obtain

$$\mathbb{E}[\mathrm{dist}(\bar{x}_K, X)] \leq \sqrt{\mathbb{E}[\mathrm{dist}^2(\bar{x}_K, X)]} = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$$

□

2.4.3 SSE with Random Projections

We now proceed to provide an analogous set of statements for the SSE scheme with random projections.

Proposition 8. *Let Assumptions 1, 3 – 8 hold and let $\gamma_k \leq \frac{1}{2L}$. Then any sequence generated by (r-SSE), where the projections are randomly generated, converges to a solution $x^* \in X$ in an a.s. sense.*

Proof. By Lemma 1(ii), we have

$$\|x_{k+1} - x^*\|^2 \leq \left\|x_k - \gamma_k\left(F\left(x_{k+\frac{1}{2}}\right) + w_{k+\frac{1}{2}}\right) - x^*\right\|^2$$
$$- \left\|x_k - \gamma_k\left(F\left(x_{k+\frac{1}{2}}\right) + w_{k+\frac{1}{2}}\right) - x_{k+1}\right\|^2$$

$$= \|x_k - x^*\|^2 - \|x_k - x_{k+1}\|^2 + 2\gamma_k \left(F\left(x_{k+\frac{1}{2}}\right) + w_{k+\frac{1}{2}}\right)^T (x^* - x_{k+1}). \tag{2.34}$$

It is clear that

$$F\left(x_{k+\frac{1}{2}}\right)^T (x_{k+1} - x^*) = F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+1} - x_{k+\frac{1}{2}}\right) + F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right). \tag{2.35}$$

Using (2.35) in (2.34), we obtain

$$\begin{aligned}
&\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - \|x_k - x_{k+1}\|^2 + 2\gamma_k F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x_{k+1}\right) \\
&+ 2\gamma_k w_{k+\frac{1}{2}}^T (x^* - x_{k+1}) - 2\gamma_k F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&= \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}} + x_{k+\frac{1}{2}} - x_{k+1}\right\|^2 + 2\gamma_k F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x_{k+1}\right) \\
&+ 2\gamma_k w_{k+\frac{1}{2}}^T (x^* - x_{k+1}) - 2\gamma_k F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&= \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - \left\|x_{k+\frac{1}{2}} - x_{k+1}\right\|^2 - 2\left(x_k - x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x_{k+1}\right) \\
&+ 2\gamma_k F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x_{k+1}\right) + 2\gamma_k w_{k+\frac{1}{2}}^T (x^* - x_{k+1}) \\
&- 2\gamma_k F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&= \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - \left\|x_{k+\frac{1}{2}} - x_{k+1}\right\|^2 \\
&+ 2\left(x_{k+1} - x_{k+\frac{1}{2}}\right)^T \left(x_k - \gamma_k F\left(x_{k+\frac{1}{2}}\right) - x_{k+\frac{1}{2}}\right) \\
&+ 2\gamma_k w_{k+\frac{1}{2}}^T (x^* - x_{k+1}) - 2\gamma_k F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right).
\end{aligned}$$

With the similar approach in Proposition 3, we obtain

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2) \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2 \left\|w_k - w_{k+\frac{1}{2}}\right\|^2 \\
&+ 2\gamma_k w_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right) - 2\gamma_k F\left(x_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&\leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2) \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2 \left\|w_k - w_{k+\frac{1}{2}}\right\|^2 \\
&+ 2\gamma_k w_{k+\frac{1}{2}}^T \left(x^* - x_{k+\frac{1}{2}}\right) - 2\gamma_k \alpha \mathrm{dist}\left(\Pi_X\left(x_{k+\frac{1}{2}}\right), X^*\right) + 2\gamma_k C d\left(x_{k+\frac{1}{2}}\right).
\end{aligned} \tag{2.36}$$

Invoking weak sharpness property, we have

$$-2\gamma_k \alpha \mathrm{dist}\left(\Pi_X\left(x_{k+\frac{1}{2}}\right), X^*\right) \leq -2\gamma_k \alpha \mathrm{dist}\left(x_k, X^*\right) + 2\gamma_k \alpha \left\|x_k - x_{k+\frac{1}{2}}\right\|$$

$$+ 2\gamma_k \alpha d\left(x_{k+\frac{1}{2}}\right) \tag{2.37}$$

and

$$\begin{aligned}
2\gamma_k(C+\alpha)d\left(x_{k+\frac{1}{2}}\right) &\leq 2\gamma_k(C+\alpha)\mathrm{dist}(x_k, X) + 2\gamma_k(C+\alpha)\left\|x_k - x_{k+\frac{1}{2}}\right\| \\
&\leq 2\gamma_k(C+\alpha)\mathrm{dist}(x_k, X) + 4\gamma_k^2(C+\alpha)^2 + \frac{1}{4}\left\|x_k - x_{k+\frac{1}{2}}\right\|^2,
\end{aligned} \tag{2.38}$$

Using (2.37) and (2.38) in (2.36), we obtain

$$\begin{aligned}
&\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2)\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2\left\|w_k - w_{k+\frac{1}{2}}\right\|^2 \\
&+ 2\gamma_k w_{k+\frac{1}{2}}^T\left(x^* - x_{k+\frac{1}{2}}\right) - 2\gamma_k\alpha\mathrm{dist}\left(\Pi_X\left(x_{k+\frac{1}{2}}\right), X^*\right) + 2\gamma_k C d\left(x_{k+\frac{1}{2}}\right) \\
&\leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2)\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2\left\|w_k - w_{k+\frac{1}{2}}\right\|^2 \\
&+ 2\gamma_k w_{k+\frac{1}{2}}^T\left(x^* - x_{k+\frac{1}{2}}\right) - 2\gamma_k\alpha\mathrm{dist}\left(x_k, X^*\right) + 2\gamma_k\alpha\left\|x_k - x_{k+\frac{1}{2}}\right\| \\
&+ 2\gamma_k(C+\alpha)d\left(x_{k+\frac{1}{2}}\right) \\
&\leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2)\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2\left\|w_k - w_{k+\frac{1}{2}}\right\|^2 \\
&+ 2\gamma_k w_{k+\frac{1}{2}}^T\left(x^* - x_{k+\frac{1}{2}}\right) - 2\gamma_k\alpha\mathrm{dist}\left(x_k, X^*\right) + 2\gamma_k\alpha\left\|x_k - x_{k+\frac{1}{2}}\right\| \\
&+ 2\gamma_k(C+\alpha)\mathrm{dist}(x_k, X) + 4\gamma_k^2(C+\alpha)^2 + \frac{1}{4}\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 \\
&\leq \|x_k - x^*\|^2 - 2\gamma_k\alpha\mathrm{dist}\left(x_k, X^*\right) - \left(\frac{5}{8} - 2\gamma_k^2 L^2\right)\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 \\
&- \frac{1}{8}\left\|x_k - x_{k+\frac{1}{2}} - 8\gamma_k\alpha\right\|^2 + 8\gamma_k^2\alpha^2 + 4\gamma_k^2(C+\alpha)^2 + 2\gamma_k(C+\alpha)\mathrm{dist}(x_k, X) \\
&+ 2\gamma_k^2\left\|w_{k+\frac{1}{2}} - w_k\right\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}} - x^*\right).
\end{aligned}$$

Taking expectations conditioned on $\mathcal{F}_k$, we obtain

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] &\leq \|x_k - x^*\|^2 - 2\gamma_k\alpha\mathrm{dist}\left(x_k, X^*\right) \\
&- \left(\frac{5}{8} - 2\gamma_k^2 L^2\right)\mathbb{E}\left[\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 \mid \mathcal{F}_k\right] \\
&+ 8\gamma_k^2\alpha^2 + 4\gamma_k^2(C+\alpha)^2 + 2\gamma_k(C+\alpha)d(x_k) + 8\gamma_k^2\nu^2.
\end{aligned} \tag{2.39}$$

According to Lemma 7, we have

$$\mathbb{E}[\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 \mid \mathcal{F}_k] = \mathbb{E}[\|x_k - \Pi_{l_k}(x_k - \gamma_k F(x_k, \omega_k))\|^2 \mid \mathcal{F}_k]$$

$$\geq \mathbb{E}[\|x_k - \Pi_{l_k}(x_k)\|^2 \mid \mathcal{F}_k] \geq \frac{\rho}{m\eta}\text{dist}^2(x_k, X). \tag{2.40}$$

where the last inequality follows from Lemma 4. Multiplying (2.40) by $\frac{1}{8}$ and using it in (2.39), we have

$$\begin{aligned}
&\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^*\|^2 - 2\gamma_k\alpha\text{dist}\,(x_k, X^*) \\
&- \left(\frac{3}{4} - 2\gamma_k^2 L^2\right)\mathbb{E}\left[\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 \mid \mathcal{F}_k\right] + 8\gamma_k^2\alpha^2 + 4\gamma_k^2(C+\alpha)^2 + 2\gamma_k(C+\alpha)d(x_k) \\
&- \frac{\rho}{8m\eta}\text{dist}^2(x_k, X) + 8\gamma_k^2\nu^2 \\
&= \|x_k - x^*\|^2 - 2\gamma_k\alpha\text{dist}\,(x_k, X^*) - \left(\frac{3}{4} - 2\gamma_k^2 L^2\right)\mathbb{E}\left[\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 \mid \mathcal{F}_k\right] + 8\gamma_k^2\alpha^2 \\
&+ 4\gamma_k^2(C+\alpha)^2 - \frac{\rho}{8m\eta}\left(\text{dist}(x_k, X) - \frac{8m\eta\gamma_k(C+\alpha)}{\rho}\right)^2 \\
&+ \frac{8m\eta(C+\alpha)^2}{\rho}\gamma_k^2 + 8\gamma_k^2\nu^2 \\
&\leq \|x_k - x^*\|^2 - 2\gamma_k\alpha\text{dist}\,(x_k, X^*) + 8\gamma_k^2\alpha^2 + 4\gamma_k^2(C+\alpha)^2 \\
&+ \frac{8m\eta(C+\alpha)^2}{\rho}\gamma_k^2 + 8\gamma_k^2\nu^2
\end{aligned} \tag{2.41}$$

Now we may invoke Lemma 2. It follows that $\{\|x_k - x^*\|^2\}$ is convergent a.s. and $\sum 2\gamma_k\alpha\text{dist}\,(x_k, X^*) < \infty$. It remains to show that $\text{dist}(x_k, X^*) \xrightarrow{k\to\infty} 0$ a.s.. We proceed by contradiction and assume that with finite probability, $\text{dist}(x_k, X^*) \to h(\omega) > 0$. Since $\sum_k \gamma_k = \infty$, it follows that $\sum_k \gamma_k \text{dist}(x_k, X^*) = \infty$ with finite probability. But this contradicts $\sum 2\gamma_k\alpha\text{dist}\,(x_k, X^*) < \infty$ a.s.. Therefore, $\text{dist}\,(x_k, X^*) \to 0$ in an a.s. sense. $\square$

Proposition 9. *Let Assumptions 1 – 5, 7 – 8 hold and let* $0 < \gamma_k = \gamma = \frac{B}{\sqrt{M_2 K}}$, *where* K *is the pre-defined termination number of iterations and* $M_2 = 8\alpha^2 + 4(C+\alpha)^2 + \frac{8m\eta(C+\alpha)^2}{\rho} + 8\nu^2$. *Then the following holds for any sequence generated by (r-SSE) in an expected value sense, where* $\bar{x}_k = \sum_{k=0}^{K-1} x_k/K$*:(1)* $\mathbb{E}[\mathit{dist}\,(\bar{x}_K, X^*)] \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$; *(2) The oracle complexity to compute an* $\bar{x}_K$ *such that* $\mathbb{E}[\mathit{dist}(\bar{x}_k, X^*)]$ *is bounded as follows:* $\sum_{k=1}^{K} N_k \leq \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, *where* $N_k = 1$ *for all* k.

Proof. (1). Taking expectations on both sides of (2.41) and using a similar derivation with the proof of Proposition 6, we have

$$2\gamma\alpha\mathbb{E}[\text{dist}\,(\bar{x}_K, X^*)] \leq \frac{B^2}{K} + \gamma^2 M_2,$$

where $M_2 = 8\alpha^2 + 4(C+\alpha)^2 + \frac{8m\eta(C+\alpha)^2}{\rho} + 8\nu^2$.
Dividing both sides by $2\gamma\alpha$ and minimizing the right-hand side in γ, we obtain the following at the optimal $\gamma = \frac{B}{\sqrt{M_2 K}}$.

$$\mathbb{E}[\text{dist}\,(\bar{x}_K, X^*)] \leq \frac{B^2}{2K\gamma\alpha} + \frac{\gamma M_2}{2\alpha} = \frac{\sqrt{M_2}B}{\alpha\sqrt{K}} = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

(2). The result follows using the same avenue as Proposition 6(2). $\square$

We conclude with an analysis of the infeasibility sequence.

Proposition 10. *Let Assumptions 1 – 3, 5 – 8 hold. Let $\gamma_k \leq \frac{1}{2L}$. Suppose $\{x_k\}$ is generated by (r-SSE), where the projections are randomly generated. Then the feasibility error satisfies $\mathbb{E}[dist(\bar{x}_k, X)] \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$.*

Proof. Let $z_k = x_k - \gamma_k F\left(x_{k+\frac{1}{2}}, \omega_{k+\frac{1}{2}}\right)$. We have

$$\begin{aligned}
&\text{dist}^2(x_{k+1}, X) \leq \left\|x_{k+1} - \Pi_X\left(x_{k+\frac{1}{2}}\right)\right\|^2 = \left\|\Pi_{T_k}(z_k) - x_{k+\frac{1}{2}} + x_{k+\frac{1}{2}} - \Pi_X\left(x_{k+\frac{1}{2}}\right)\right\|^2 \\
&\leq \left(1 + \frac{4m\eta}{\rho}\right)\|\Pi_{T_k}(z_k) - x_{k+\frac{1}{2}}\|^2 + \left(1 + \frac{\rho}{4m\eta}\right)\left\|x_{k+\frac{1}{2}} - \Pi_X\left(x_{k+\frac{1}{2}}\right)\right\|^2 \\
&= \left(1 + \frac{4m\eta}{\rho}\right)\|\Pi_{T_k}(z_k) - \Pi_{l_k}(x_k)\|^2 + \left(1 + \frac{\rho}{4m\eta}\right)\left\|x_{k+\frac{1}{2}} - \Pi_X\left(x_{k+\frac{1}{2}}\right)\right\|^2 \\
&= \left(1 + \frac{4m\eta}{\rho}\right)\|\Pi_{T_k}(z_k) - \Pi_{T_k}(x_k)\|^2 + \left(1 + \frac{\rho}{4m\eta}\right)\left\|x_{k+\frac{1}{2}} - \Pi_X\left(x_{k+\frac{1}{2}}\right)\right\|^2 \\
&\leq \left(1 + \frac{4m\eta}{\rho}\right)\|z_k - x_k\|^2 + \left(1 + \frac{\rho}{4m\eta}\right)\left\|x_{k+\frac{1}{2}} - \Pi_X\left(x_{k+\frac{1}{2}}\right)\right\|^2, \qquad (2.42)
\end{aligned}$$

where we leverage $\|a+b\|^2 \leq \left(1 + \frac{4m\eta}{\rho}\right)\|a\|^2 + \left(1 + \frac{\rho}{4m\eta}\right)\|b\|^2$. We have that

$$\begin{aligned}
\mathbb{E}[d^2\left(x_{k+\frac{1}{2}}\right) \mid \mathcal{F}_k] \leq &\left(1 - \frac{\rho}{4m\eta}\right)\text{dist}^2(x_k, X) \\
&+ \left(5 + \frac{4m\eta}{\rho}\right)(4L^2B^2 + 4C^2 + 2\nu^2)\gamma_k^2. \qquad (2.43)
\end{aligned}$$

Using (2.43) in (2.42), we obtain

$$\begin{aligned}
\mathbb{E}[\text{dist}^2(x_{k+1}, X) \mid \mathcal{F}_k] \leq &\left(1 - \frac{\rho^2}{16m^2\eta^2}\right)\text{dist}^2(x_k, X) \\
&+ \left(8 + \frac{12m\eta}{\rho} + \frac{5\rho}{4m\eta}\right)(4L^2B^2 + 4C^2 + 2\nu^2)\gamma_k^2.
\end{aligned}$$

It is clear that $\gamma_{k+1}^2 \geq \left(1 - \frac{\rho^2}{32m^2\eta^2}\right)\gamma_k^2$ when k is sufficiently large. Leveraging Lemma 8, we have

$$\begin{aligned}\mathbb{E}[\mathrm{dist}^2(x_k, X)] &\leq \left(\frac{256m^2\eta^2}{\rho^2} + \frac{384m^3\eta^3}{\rho^3} + \frac{40m\eta}{\rho}\right)(4L^2B^2 + 4C^2 + 2\nu^2)\gamma_k^2 \\ &+ d(x_0)\left(1 - \frac{\rho^2}{16m^2\eta^2}\right)^k \\ &+ \left(\left(8 + \frac{12m\eta}{\rho} + \frac{5\rho}{4m\eta}\right)(4L^2B^2 + 4C^2 + 2\nu^2)\sum_{t=0}^{\bar{k}}\gamma_t^2\right)\left(1 - \frac{\rho^2}{16m^2\eta^2}\right)^{k-\bar{k}}\end{aligned}$$

When k is sufficiently large, it satisfies that

$$\mathbb{E}[\mathrm{dist}^2(x_k, X)] \leq \left(\frac{256m^2\eta^2}{\rho^2} + \frac{384m^3\eta^3}{\rho^3} + \frac{40m\eta}{\rho}\right)(4L^2B^2 + 4C^2 + 2\nu^2)\mathcal{O}(\gamma_k^2).$$

By employing the same technique used in (2.33), we have $\mathbb{E}[\mathrm{dist}(\bar{x}_K, X)] \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$. $\square$

2.5 Numerical Results

In this section, we apply the schemes on a stochastic Nash-Cournot game (Section 2.5.1) and the computation of the invariant distribution of a Markov chain (Section 2.5.2).

2.5.1 A Stochastic Nash-Cournot Game

In this section, we present and compare the computational results of applying the extragradient schemes aforementioned to a stochastic Nash-Cournot game. This game is assumed that $\mathcal{I}$ firms compete over a network of $\mathcal{J}$ nodes. Level of production and sales of firm $i \in \mathcal{I}$ at node $j \in \mathcal{J}$ are denoted by p_{ij} and s_{ij}, respectively. Furthermore, we assume the cost of production at node j is $C_{ij}(p_{ij})$ and the price at node j is denoted by $Q_j(\bar{s}_j, \xi)$, where $\bar{s}_j$ is the aggregate sales at node j. For simplicity, we assume the transportation costs are zero. Thus, each firm i will solve a profit maximization problem given by the following:

$$\max \quad \mathbb{E}[f_i(x, \xi)] = \mathbb{E}\left[\sum_{j\in\mathcal{J}}(Q_j(\bar{s}_j, \xi)s_{ij} - C_{ij}(p_{ij}))\right]$$

$$\text{subject to} \quad \sum_{j \in \mathcal{J}} p_{ij} = \sum_{j \in \mathcal{J}} s_{ij}, \quad p_{ij} \leq \text{cap}_{ij}, \quad s_{ij}, q_{ij} \geq 0, \quad \forall j \in \mathcal{J}.$$

The equilibrium conditions of this problem can be captured by a variational inequality VI(X,F), where $F = (F_1(x); ...; F_{\mathcal{I}}(x))$ with $F_i(x) = \mathbb{E}[\nabla_{x_i} f_i(x, \xi)]$. In our original setting, we assume there are $\mathcal{I} = 5$ firms and $\mathcal{J} = 4$ nodes, and the capacity $\text{cap}_{ij} = 300$, $\forall i, j$. $C_{ij}(p_{ij})) \triangleq c_{ij} p_{ij} + d_{ij}$, where $c_{ij} = 1.5$ and d_{ij} is a constant, $\forall i, j$. $Q_j(\bar{s}_j, \xi) \triangleq a_j - b_j \bar{s}_j$, where $b_j = 0.05$ and a_j is a uniformly distributed random variable sampled from $[49.5, 50.5]$, $\forall j$. With the above parameters, it can be shown that the mapping F is strictly monotone.

We assume square-summable and non-summable step sizes in our experiments and utilize gap function as our metric. Table 2.3 shows the empirical and theoretical errors at the 4000th iteration with a diminishing steplength. Parameters of this problem are $L = 0.3$, $B = 2.25\text{e}2$ and $\nu = 10/\sqrt{3}$.

Table 2.3: Empirical and theoretical errors under mere monotonicity

	SEG	SPRG	SSE
Empirical	9.8574e-3	9.2702e-3	9.1534e-3
Theoretical	2.259e3	3.373e3	2.259e3

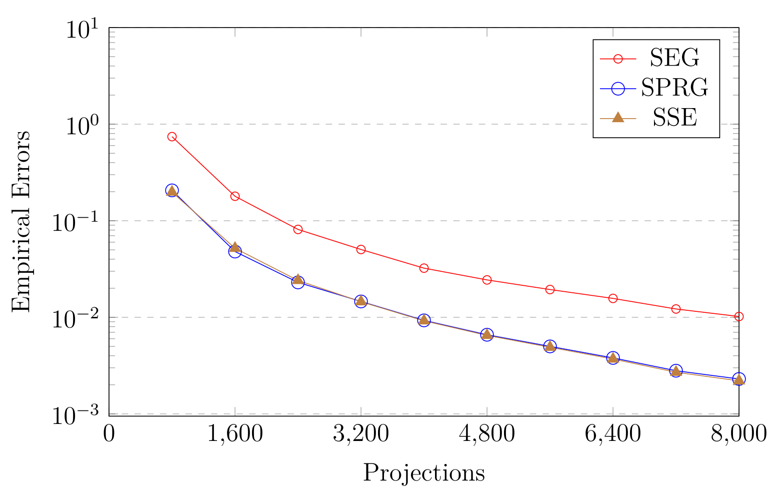


Figure 2.3: Convergence based on projections under mere monotonicity

Recall that SEG has two projections onto the set, while the other two schemes just require one. We compare their performance under the same number of projections

(Fig. 3.1). Next we change the size and parameters of the original game to ascertain parametric sensitivity. In Table 2.4 we consider test problems which are a set of 16 problems where the settings and their corresponding empirical errors and elapsed time are shown in Table 2.4. Table 2.4 shows the performance after 4000 iterations and find that while SEG has almost the same empirical error with the others but with significant computational cost. To check the performance of variance reduction, we enlarge the random set for random variable a_j to $[40, 60]$. We show the difference of convergence results between enlarged random set and the original narrow set in Fig. 2.4. Fig. 2.5 shows comparison of variance reduction schemes with original ones under the same number of iterations. Table 2.5 shows the results generated from different nodes in the system. The number of iterations used is 4000. We note that all schemes show relatively similar sensitivity to the changes introduces.

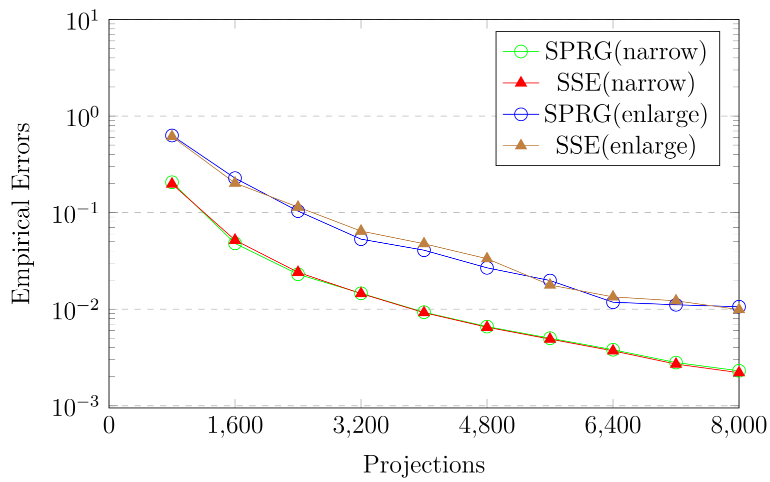


Figure 2.4: Convergence comparison between the enlarged random set and the original narrow one

Key findings. The key findings are that (**SPRG**) and (**SSE**) produce empirical errors but do so in approximately 65% of the time utilized by (**SEG**). Moreover, the presence of variance reduction allows for significant improvement in the empirical rates from the single-sample counterparts (See Table 2.5).

Table 2.4: Errors and elapsed time comparison of the three schemes with different parameters under mere monotonicity

	SEG	Time	SSE	Time	SPRG	Time
$\mathcal{I}=5, \mathcal{J}=4, c_{ij}=2, b_j=0.05$	9.1e-3	2.4e3s	9.1e-3	1.6e3s	9.2e-3	1.5e3s
$\mathcal{I}=6, \mathcal{J}=4, c_{ij}=2, b_j=0.05$	1.0e-2	2.4e3s	1.1e-2	1.6e3s	1.1e-2	1.5e3s
$\mathcal{I}=5, \mathcal{J}=5, c_{ij}=2, b_j=0.05$	1.2e-2	2.5e3s	1.2e-2	1.8e3s	1.2e-2	1.5e3s
$\mathcal{I}=6, \mathcal{J}=5, c_{ij}=2, b_j=0.05$	1.2e-2	2.5e3s	1.1e-2	1.9e3s	1.3e-2	1.5e3s
$\mathcal{I}=5, \mathcal{J}=4, c_{ij}=1, b_j=0.05$	9.1e-3	2.3e3s	9.2e-3	1.7e3s	9.3e-3	1.4e3s
$\mathcal{I}=6, \mathcal{J}=4, c_{ij}=1, b_j=0.05$	1.1e-2	2.3e3s	1.1e-2	1.8e3s	1.1e-2	1.4e3s
$\mathcal{I}=5, \mathcal{J}=5, c_{ij}=1, b_j=0.05$	1.2e-2	2.4e3s	1.3e-2	1.8e3s	1.3e-2	1.5e3s
$\mathcal{I}=6, \mathcal{J}=5, c_{ij}=1, b_j=0.05$	1.2e-2	2.4e3s	1.3e-2	1.9e3s	1.3e-2	1.5e3s
$\mathcal{I}=5, \mathcal{J}=4, c_{ij}=2, b_j=0.1$	1.1e-2	2.4e3s	1.1e-2	1.6e3s	1.2e-2	1.4e3s
$\mathcal{I}=6, \mathcal{J}=4, c_{ij}=2, b_j=0.1$	1.1e-2	2.4e3s	1.0e-2	1.6e3s	1.1e-2	1.5e3s
$\mathcal{I}=5, \mathcal{J}=5, c_{ij}=2, b_j=0.1$	1.2e-2	2.4e3s	1.1e-2	1.7e3s	1.2e-2	1.4e3s
$\mathcal{I}=6, \mathcal{J}=5, c_{ij}=2, b_j=0.1$	1.1e-2	2.5e3s	1.2e-2	1.8e3s	1.3e-2	1.4e3s
$\mathcal{I}=5, \mathcal{J}=4, c_{ij}=1, b_j=0.1$	1.0e-2	2.4e3s	1.0e-2	1.7e3s	1.1e-2	1.3e3s
$\mathcal{I}=6, \mathcal{J}=4, c_{ij}=1, b_j=0.1$	1.1e-2	2.4e3s	1.1e-2	1.6e3s	1.1e-2	1.3e3s
$\mathcal{I}=5, \mathcal{J}=5, c_{ij}=1, b_j=0.1$	1.2e-2	2.4e3s	1.2e-2	1.8e3s	1.1e-2	1.4e3s
$\mathcal{I}=6, \mathcal{J}=5, c_{ij}=1, b_j=0.1$	1.1e-2	2.4e3s	1.1e-2	1.7e3s	1.2e-3	1.0e3s

Table 2.5: Errors and elapsed time comparison of the schemes with different sizes under the same number of iterations

Network Size	SEG	Time	SSE	Time	v-SSE	Time	SPRG	Time	v-SPRG	Time
20	1.0e-1	2.4e3s	1.1e-1	1.7e3s	7.5e-3	1.9e3s	1.1e-1	1.5e3s	7.4e-3	1.6e3s
24	1.3e-1	2.4e3s	1.4e-1	1.8e3s	7.7e-3	2.0e3s	1.3e-1	1.5e3s	7.7e-3	1.7e3s
28	1.8e-1	2.7e3s	1.7e-1	1.9e3s	7.9e-3	2.1e3s	1.9e-1	1.6e3s	8.0e-3	1.7e3s
32	2.0e-1	2.8e3s	1.9e-1	1.9e3s	8.3e-3	2.2e3s	2.0e-1	1.7e3s	8.2e-3	1.8e3s
36	2.5e-1	3.1e3s	2.5e-1	2.2e3s	8.7e-3	2.4e3s	2.4e-1	2.0e3s	8.8e-3	2.1e3s
40	3.4e-1	3.2e3s	3.5e-1	2.3e3s	9.0e-3	2.5e3s	3.5e-1	2.1e3s	9.1e-3	2.2e3s

2.5.2 Markov Invariant Distribution Approximation

We test the performance of the random projection schemes on an example from [107] which requires computing a low-dimensional approximation to the invariant distribution of a Markov chain. We denote its transition matrix by P and its stationary distribution as π. The number of states is assumed to be 1000 and we want to approximate the states in a low-dimensional subspace of $\mathbb{R}^{20}$ with a transformation matrix Σ. Then we use a projection approach to approximate $\pi = P^T\pi$ as

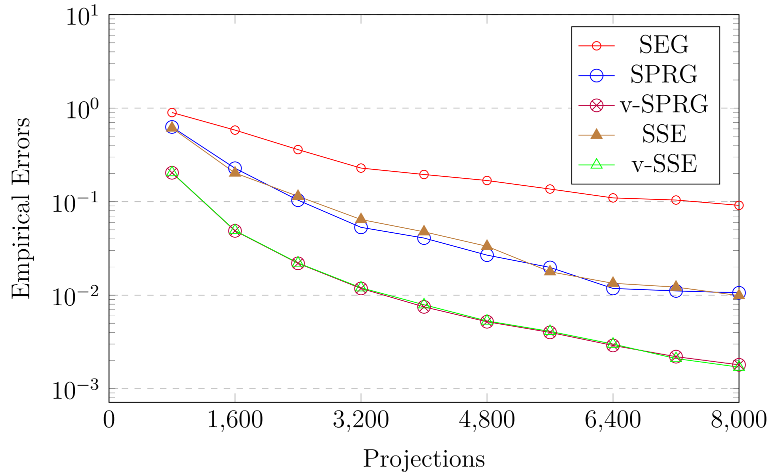


Figure 2.5: Performance comparison between variance reduction schemes and original ones

$\Sigma x = \Pi_X(P^T \Sigma x)$, where $X \triangleq \{x \mid \Sigma x \geq 0, e^T \Sigma x = 1\}$. It has been proved [107,116] that the projected equation is equivalent to the VI:

$$(x - x^*)^T S x^* \geq 0, \quad \forall x \in \mathbb{R}^{20}, \Sigma x \geq 0, e^T \Sigma x = 1,$$

where $S = \Sigma^T (I - P^T) \Sigma$. We generate the transition matrix P randomly in our experiment. The schemes are under strong monotone as well. Table 2.6 shows the empirical and theoretical errors of all extragradient-type schemes at the 10000th iteration. Figure 2.6 illustrates the convergence performance of the extragradient schemes considered.

Table 2.6: Empirical and Theoretical errors on random projections

	r-SEG	r-SPRG	r-SSE
Empirical	0.0776	0.0758	0.0657
Theoretical	2.0616	2.9183	2.0616

We record the elapsed time and empirical errors of each scheme with 10 different transition matrices, as shown in Table 2.7 while the comparison between original stochastic schemes and the random projection variants are shown in Table 2.8.

Key insights. In random projection variants, the projection onto each random constraint is cheap. Thus, the run-time benefits of (r-SSE) are not obvious when

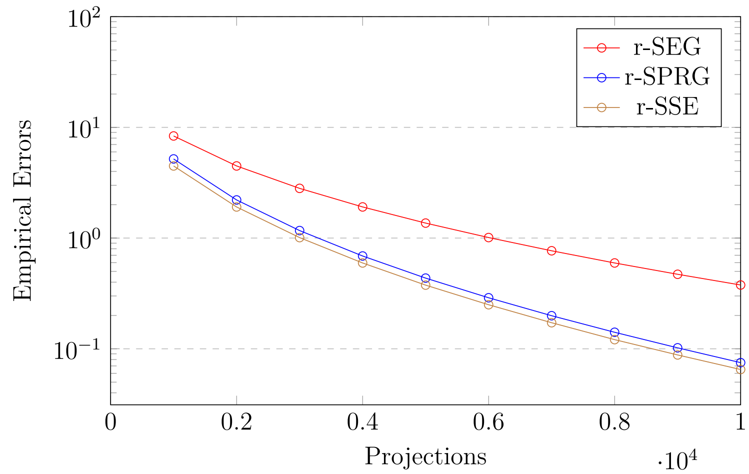


Figure 2.6: Convergence based on projections on random projections

compared with (r-SEG) while (r-SPRG) is still faster than others. This is because the second projection in (r-SSE), while computable in closed form, is almost as expensive as a (cheap) projection.

Table 2.7: Errors and elapsed time comparison of the three schemes with different transition matrices on random projections

Matrix	r-SEG	Time	r-SSE	Time	r-SPRG	Time
No.1	7.7e-2	1.4e3s	6.5e-2	1.4e3s	7.5e-2	0.7e3s
No.2	4.0e-2	1.3e3s	3.9e-2	1.4e3s	4.0e-2	0.7e3s
No.3	1.8e-2	1.3e3s	1.7e-2	1.4e3s	1.8e-2	0.7e3s
No.4	5.2e-2	1.4e3s	4.9e-2	1.4e3s	5.1e-2	0.7e3s
No.5	4.7e-2	1.3e3s	4.4e-2	1.4e3s	4.6e-2	0.7e3s
No.6	5.9e-2	1.3e3s	5.5e-2	1.4e3s	5.8e-2	0.7e3s
No.7	2.7e-2	1.4e3s	2.6e-2	1.4e3s	2.7e-2	0.7e3s
No.8	5.8e-2	1.3e3s	5.3e-2	1.4e3s	5.7e-2	0.7e3s
No.9	2.6e-2	1.4e3s	2.3e-2	1.4e3s	2.5e-2	0.7e3s
No.10	3.3e-2	1.4e3s	3.1e-2	1.4e3s	3.2e-2	0.7e3s

Table 2.8: Errors and elapsed time comparison between the three schemes with random projections and original ones

	SEG	r-SEG	SSE	r-SSE	SRPG	r-SRPG
Error	4.3e-3	7.7e-2	3.7e-3	6.5e-2	4.2e-3	7.5e-2
Time	2.8e4s	1.4e3s	1.6e4s	1.4e3s	1.5e4s	0.7e3s

2.6 Concluding Remarks

Extragradient schemes and their sampling-based counterparts represent a key cornerstone of solving monotone deterministic and stochastic variational inequality problems. Yet, the per-iteration complexity of such schemes is twice as high as their single projection counterparts. We consider two avenues in which the two projections are replaced by exactly one projection (a projected reflected scheme) or a single projection onto the set and another onto a halfpace, the second of which is computable in closed form (a subgradient extragradient scheme). In both instances, we derive a.s. convergence statements and rate statements under variance reduction. Notably, the sequences achieve a non-asymptotic rate of $\mathcal{O}(1/K)$, matching its deterministic counterpart. Furthermore, when this set is itself challenging to project onto, we develop a random projection variant for each scheme. Again, a.s. convergence and rate statements are provided. Empirical behavior of both schemes show significant benefits in terms of per-iteration complexity compared to extragradient counterparts.

Chapter 3
Stochastic proximal-point and splitting schemes for monotone stochastic generalized equations

3.1 Introduction

This chapter considers the resolution of the stochastic generalized equation, a problem that requires an $x \in \mathbb{R}^n$ such that

$$0 \in \mathbb{E}[T(x, \xi(\omega))], \tag{SGE}$$

where the components of the map T are denoted by T_i, $i = 1, \ldots, n$; $\xi : \Omega \to \mathbb{R}^d$ is a random variable, $T_i : \mathbb{R}^n \times \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ is a set-valued map, $\mathbb{E}[\cdot]$ denotes the expectation, and the associated probability space is given by $(\Omega, \mathcal{F}, \mathbb{P})$. In the remainder of this chapter, we refer to $T(x, \xi(\omega))$ by $T(x, \omega)$. The expectation of a set-valued map leverages the Aumann integral [117] and is formally defined next.

$$\mathbb{E}[T_i(x, \xi(\omega))] = \left\{ \int v_i(\omega) dP(\omega) \mid \quad v_i(\omega) \in T_i(x, \xi(\omega)) \right\}.$$

Consequently, the expectation $\mathbb{E}[T(x, \omega)]$ can be defined as a Cartesian product of the sets $\mathbb{E}[T_i(x, \omega)$.

$$\mathbb{E}[T(x^*, \omega)] \;\triangleq\; \prod_{i=1}^{n} \mathbb{E}[T_i(x^*, \omega)].$$

Generalized equations find broad applicability in a range of settings in mathematical programming, some instances of which are provided next.

(a) Convex optimization. Consider a convex programming problem defined as (C-Opt)

$$\min_{x \in X} \; f(x), \tag{C-Opt}$$

where f is a convex function defined on a closed and convex set X. Then x is a solution of (C-Opt) if and only if

$$0 \in \partial f(x) + \mathcal{N}_X(x), \tag{C-Opt-GE}$$

where $\mathcal{N}_X(x)$ denotes the normal cone of X at x and $\mathcal{N}_X(x) \triangleq \{y \in \mathbb{R}^n : y^T(u - x) \leq 0, \text{ for all } u \in X\}$. But this in effect requires solving a generalized equation $0 \in T(x) \triangleq \partial f(x) + \mathcal{N}_X(x)$.

(b) Saddle-point problems. Consider the solution of a saddle point problem (Saddle-pt)

$$\min_{x \in X} \; \max_{y \in Y} \mathcal{L}(x, y), \tag{Saddle-pt}$$

where $\mathcal{L}(x, y)$ is a convex-concave function over $X \times Y$ and X and Y are closed and convex sets. Then (x, y) is a solution of (Saddle-pt) if and only if

$$\begin{aligned} 0 &\in \partial_x \mathcal{L}(x, y) + \mathcal{N}_X(x), \\ 0 &\in -\partial_y \mathcal{L}(x, y) + \mathcal{N}_Y(y). \end{aligned} \tag{Saddle-GE}$$

where $\mathcal{N}_X(x)$ and $\mathcal{N}_Y(y)$ denotes the normal cones of X at x and Y at y, respectively. It follows that computing a saddle-point of (Saddle) is equivalent solving a generalized equation $0 \in T(x, y)$ where $T(x, y) = T_1(x, y) \times T_2(x, y)$, $T_1(x, y) \triangleq \partial_x \mathcal{L}(x, y) + \mathcal{N}_X(x)$, and $T_2(x, y) \triangleq -\partial_y \mathcal{L}(x, y) + \mathcal{N}_Y(y)$.

(c) Variational inequality problems. Consider a variational inequality problem, a class of problems that captures a range of settings including convex optimization problems, Nash equilibrium problems, traffic equilibrium problems, amongst others. Recall that a variational inequality problem VI(X, F) requires an $x \in X$ such that

$$(y - x)^T F(x) \geq 0, \quad \forall y \in X.$$

But computing such an x is equivalent to solving the following inclusion problem.

$$0 \in F(x) + \mathcal{N}_X(x).$$

In effect, computing a solution to $\mathrm{VI}(X, F)$ is equivalent to resolving $0 \in T(x)$ where $T(x) \triangleq F(x) + \mathcal{N}_X(x)$.

Generalized equations have been extensively examined since the 70s when Rockafellar [47] developed a proximal point scheme for a generalized equation characterized by monotone operators. In fact, this scheme subsumes a range of well known schemes such as the augmented Lagrangian method [118], Douglas-Rachford splitting [119], amongst others. While generalized equations represent a unifying paradigm for a broad class of decision-making problems, there is growing need to contend with the manifold uncertainty that emerges in a range of applications. To this end, stochastic linear programming through the seminal work by Dantzig [120] and Beale [121] provided a crucial foundation and much of the subsequent research in this area focused on extending the realm of stochastic programming to convex [122], nonlinear [123], and integer regimes [124]. Forays into the regime of stochastic generalized equations have been far less common and been limited to recent efforts on the analysis [125] and computation of solutions to stochastic variational inequality problems via stochastic approximation [39] and sample-average approximation (SAA) [126]. We consider the solution of generalized equations under uncertainty, a class of problems that has seen relatively little study, crucial exceptions being the SAA-based analysis of such problems [30, 127]. In this paper, we consider how the proximal point framework for stochastic generalized equations can be extended to account for uncertainty. Such, *stochastic* generalizations of the proximal point framework have significant reach since such developments will immediately allow for stating stochastic counterparts for a range of the aforementioned techniques.

In fact, there have been some recent efforts in examining stochastic counterparts of proximal point schemes [128–131] but all of these schemes are characterized by a convergence rates which do not match their deterministic counterparts. These gaps, in part, motivate the present work. The crucial challenge in developing a direct extension of the proximal-point framework to the stochastic regime arises in the evaluation of the resolvent operator (which will be subsequently defined), a computation that is complicated by the presence of an expectation-valued operator. We provide two distinct avenues for addressing this concern: (i) In our first approach, qualified as "sample-then-resolve", we utilize increasingly large sample-sizes, as a consequence of which the resolvent operator is based on a deterministic (albeit sample-average) operator; (ii) Our second scheme, referred to as "resolve-then-sample", maintains the expectation in the resolvent but utilizes a Monte-Carlo sampling scheme for computing an inexact solution to this problem. In both instances, we develop rate statements as well as oracle complexity bounds when the mappings are maximal monotone and expectation-valued.

Often the operator T is characterized by a distinct structure in that it can be cast as the sum of two operators A and B. Under suitable requirements on A and B, a range of splitting methods can be developed and has represented a vibrant area of research over the last two decades [49, 50, 119, 132]. The regime where the maps are expectation-valued has seen relatively less study [133]. We extend the modified backward-forward splitting scheme developed by Tseng [52] to the setting where the maps are expectation-valued. The key contributions of this chapter can be articulated as follows:

I. Stochastic proximal point framework. In Section 3.3, we present a stochastic proximal point framework for a subclass of stochastic generalized equations where the operator is either strongly monotone or maximal monotone. By employing a sample-then-resolve approach, denoted by (**str-SPP**), we show that when the sample-size sequences are raised at a suitable rate, we prove that the resulting sequence of iterates converges either at a linear rate (strongly monotone) or at a rate of $\mathcal{O}(1/k)$ (maximal monotone) (in terms of a suitable expectation-valued metric), leading to oracle complexities of $\mathcal{O}(1/\epsilon)$ and $\mathcal{O}(1/\epsilon^{2a+1})$ for $a > 1$ when the sample-size is raised at the rate of $\lceil (k+1)^{2a} \rceil$, respectively. In contrast, when adopting a "resolve-then-sample" approach, denoted by (**rts-SPP**), the resolvent operator is evaluated inexactly through a Monte-Carlo sampling scheme.

II. Stochastic splitting schemes. In Section 3.4, we consider structured regimes in which the map can be rewritten as the sum of two maps, facilitating the use of splitting-based framework. In this context, when one of the maps is expectation-valued while the other has a cheap resolvent, we consider a scheme in which a sample-average of the expectation-valued map is utilized in the forward step. Akin to the prior scheme, when the sample-size is increased at a suitable rate, the resulting sequence of iterates converges either at a linear rate (strongly monotone) or at a rate of $\mathcal{O}(1/k)$ (maximal monotone), leading to oracle complexities of $\mathcal{O}(1/\epsilon)$ and $\mathcal{O}(1/\epsilon^2)$, respectively. We compare the rate statements with prior work in Tables 3.1 and 3.2.

Table 3.1: A review of proximal point schemes

Ref.	Applicability	Sto.	Avg.	N_k	Metric	Rate	Oracle
[47]	Strongly monotone	N	N	1	Iterates	Linear	-
[134]	Maximal monotone	N	N	1	Yosida	$\mathcal{O}(1/K)$	-
[134]	Strongly monotone	N	N	1	Iterates	Linear	-
[128]	Maximal mono., Lip.	Y	Y	1	Gap	$\mathcal{O}(1/\sqrt{K})$	-
[128]	Strongly mono., Lip.	Y	N	1	Iterates	$\mathcal{O}(1/K)$	-
[130]	Strongly monotone	Y	N	1	Iterates	$\mathcal{O}(1/K)$	-
[129]	Maximal monotone	Y	Y	1	Iterates	A.s.	-
str-SPP	Maximal monotone	Y	N	$\lceil (k+1)^{2a} \rceil$	Yosida	$\mathcal{O}(1/K)$	$\mathcal{O}(1/\epsilon^{2a+1})$
str-SPP	Strongly monotone	Y	N	$\lfloor \rho^{-2(k+1)} \rfloor$	Iterates	Linear	$\mathcal{O}(1/\epsilon)$
rts-SPP	Maximal monotone	Y	N	$\lceil (k+1)^{2a} \rceil$	Yosida	$\mathcal{O}(1/K)$	$\mathcal{O}(1/\epsilon^{2a+1})$
rts-SPP	Strongly monotone	Y	N	$\lfloor \rho^{-2(k+1)} \rfloor$	Iterates	Linear	$\mathcal{O}(1/\epsilon)$

Table 3.2: A review of splitting schemes

Ref.	Applicability	Sto.	Avg.	N_k	Metric	Rate	Oracle
[52]	Strongly mono., Lip.	N	N	1	Iterates	Linear	-
[135]	Maximal monotone	N	N	1	Iterates	$\mathcal{O}(1/K)$	-
[136]	Strongly monotone	Y	N	1	Iterates	$\mathcal{O}(1/K)$	-
[137]	Maximal monotone	Y	N	1	Iterates	A.s.	-
SMFBS	Maximal monotone	Y	Y	k	Gap Fn.	$\mathcal{O}(1/K)$	$\mathcal{O}(1/\epsilon^2)$
SMFBS	Strongly monotone	Y	N	$\lfloor \rho^{-(k+1)} \rfloor$	Iterates	Linear	$\mathcal{O}(1/\epsilon)$

The remainder of the chapter is organized into the four sections. In Section 3.2, we provide some background on proximal and splitting schemes and outline the main assumptions. In Section 3.3, we present and analyze a stochastic proximal-point framework while in Section 3.4 examines the stochastic splitting framework. We conclude this chapter by applying the two schemes to a set of applications.

3.2 Background and Assumptions

In this section, we provide some background on proximal-point and splitting schemes which provide a foundation for the development in the sections to follow. We conclude this section with an outline of the assumptions as well as any results.

3.2.1 Generalized Equations

In this section, we provide some background on solving generalized equations of the form

$$0 \in T(x), \tag{GE}$$

where the mapping T is a set-valued maximal monotone map.

3.2.2 Algorithms for the solution of generalized equations

A range of algorithms have been developed for the resolution of (GE). We now review a subset of the important avenues for the resolution of this problem.

I. Proximal-point schemes. A standard scheme to solve (GE) is the proximal point algorithm proposed by Martinet [46] and Rockafellar [47, 48]:

$$x_{k+1} \coloneqq (I + \gamma_k T)^{-1}(x_k), \tag{PP}$$

where γ_k denotes the steplength, T is required to be maximal monotone. The map $(I + \gamma_k T)^{-1}$, referred to as the resolvent of T, is denoted by $J^T_{\gamma_k} \triangleq (I + \gamma_k T)^{-1}$ [47]. This resolvent is a single-valued, nonexpansive map for a monotone T; the domain of $J^T_{\gamma_k}$ is equal to $\mathbb{R}^n$ if T is maximal monotone [6]. In [47], Rockafellar developed a proximal-point framework for generalized equations with monotone operators, presenting a linear rate statement for strongly monotone T. More recently, in [134], Corman and Yuan proved that under maximal monotonicity, the proximal-point scheme produced sequences which diminishes to zero at the rate of $\mathcal{O}(1/k)$ under an appropriate metric while a linear rate can be proven in strongly monotone regimes. In Section 3.3, we develop a stochastic proximal point framework where we consider one of two avenues:

(A) The resolvent of the expectation-valued map, denoted by $(I+\gamma_k\mathbb{E}[T(x,\omega)])^{-1}$, is approximated with increasing accuracy via Monte-Carlo sampling. In particular, $(I+\gamma_k\mathbb{E}[T(x,\omega))^{-1}$ requires computing a solution of a stochastic generalized equation with a strongly monotone operator, which can often be obtained via a stochastic approximation framework.

(B) The resolvent of the sample-average approximation of the map is employed, as defined by $(I+\gamma_k\bar{T}_k)^{-1}$, where $\bar{T}_k \triangleq \frac{\sum_{j=1}^{N_k} T(x,\omega_j)}{N_k}$.

Yet, the applicability of proximal-point schemes is crucially reliant on the tractability of evaluating the resolvent operator $J_{\gamma_k}^T(x_k)$. When this evaluation is problemmatic, splitting-based approaches have emerged as important alternatives.

(II). Splitting schemes. In many applications, the map T may not have a tractable resolvent operator. However, suppose either the resolvent of A or B (or both) is tractable where $T \triangleq A+B$, then splitting schemes assume relevance.

(a) *Douglas-Rachford Splitting [49, 119].* In this scheme, the resolvent of A and B can be separately evaluated to generate a sequence defined as follows.

$$x_{k+1} := (I+\gamma_k B)^{-1}((I+\gamma_k A)^{-1}(I-\gamma_k B)+\gamma_k B)(x_k), \tag{DRS}$$

(b) *Peaceman-Rachford Splitting [49, 132].* In contrast, in the Peaceman-Rachford splitting method, the update rule is slightly different, given by the following.

$$x_{k+1} := (I+\gamma_k B)^{-1}(I-\gamma_k A)(I+\gamma_k A)^{-1}(I-\gamma_k B)(x_k). \tag{PRS}$$

(c) *Forward-backward splitting.* Moreover, if the resolvent of B is easier to evaluate, we may develop algorithms leveraging these resolvents. Assuming A and B are maximal monotone, the forward-backward splitting method was proposed by [49, 50] respectively and was applied to convex optimization by [51]:

$$x_{k+1} := (I+\gamma_k B)^{-1}(I-\gamma_k A)(x_k). \tag{FBS}$$

In [137], a stochastic variants of the forward-backward splitting method is developed for strongly monotone and equipped with a rate of $\mathcal{O}(\frac{1}{K})$ while in [136], maximal monotone regimes are examined and a.s. convergence statements are provided. Notably, we present schemes that provide distinct improvements over these rates.

(d) *Modified backward-forward splitting.* A drawback of (FBS) is that it generally requires a strong monotonicity assumption on A to ensure convergence. An extension was suggested by Tseng [52] that weakened the requirement on A to be mere monotonicity:

$$\begin{aligned} x_{k+\frac{1}{2}} &:= (I+\gamma_k B)^{-1}(I-\gamma_k A)(x_k), \\ x_{k+1} &:= \Pi_X\left(x_{k+\frac{1}{2}} - \gamma_k\left(A\left(x_{k+\frac{1}{2}}\right) - A(x_k)\right)\right). \end{aligned} \tag{MFBS}$$

In this scheme, convergence was proved under the assumptions of Lipschitz continuity and monotonicity of the map A.

In Section 3.4, we consider a setting where $A(x) \triangleq \mathbb{E}[A(x,\omega)]$ and develop a modified forward-backward splitting framework in which the resolvent of A is computed via either utilizing the resolvent of a single sample or the sample-average.

3.2.3 Assumptions and Supporting Results

We outline some assumptions on T which we will utilize when necessary.

Assumption 9. *The mapping T is maximal monotone.*

In some instances,T is assumed to be strongly monotone.

Assumption 10. *The mapping T is σ-strongly monotone, i.e. there exists $\sigma > 0$ such that*

$$(u-v)^T(x-y) \geq \sigma\|x-y\|^2, \quad \forall x,y \in \mathbb{R}^n, u \in T(x), v \in T(y).$$

Assumption 11. *The mapping T is L-Lipschitz continuous on $\mathbb{R}^n$, i.e. for all $x, y \in \mathbb{R}^n$,*

$$\|u-v\| \leq L\|x-y\|, \quad \forall u \in T(x),\ v \in T(y).$$

We consider schemes in which the $x_0 \in \mathbb{R}^n$ is a random initial point and $\{\gamma_k\}$ denotes the steplength sequence.

Assumption 12. *The diminishing steplength sequence γ_k is square-summable but non-summable; $\sum_{k=0}^{\infty} \gamma_k^2 < \infty, \sum_{k=0}^{\infty} \gamma_k = \infty$.*

We assume the presence of a stochastic first-order oracle that can provide an unbiased estimator of $T(x)$, given by $T(x, \omega)$, such that $\mathbb{E}[T_i(x, \omega)] = T_i(x)$. Define $v_k \in T(x_k), v_k(\omega_k) \in T(x_k, \omega_k)$ and $w_k \triangleq v_k(\omega_k) - v_k$. Furthermore we denote $\mathcal{F}_k$ as the history up to iteration k, i.e., $\mathcal{F}_k \triangleq \{x_0, \omega_0, \omega_1, \cdots, \omega_{k-1}\}$. The following requirements on the first and second moments are imposed when necessary.

Assumption 13. *At an iteration k, the following hold in an a.s. sense: (i) The conditional mean $\mathbb{E}[w_k \mid \mathcal{F}_k]$ is zero for all k; (ii) The conditional second moment is bounded or $\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] \leq \nu^2$ for all k.*

3.3 Stochastic Proximal Point Schemes

In this section, we consider the stochastic counterpart of the proximal-point method. While stochastic counterparts of the proximal gradient method (and its accelerated counterparts) have received much attention [111, 138, 139], there has been less progress in extending the proximal-point method. Recently, Asi and Duchi [130] developed model-based methods combining stochastic proximal point method which produces a convergence rate of $\mathcal{O}(\frac{1}{k})$.

$$x_{k+1} := \operatorname*{argmin}_{x \in X} \left\{ f(x, \omega_k) + \frac{1}{2\gamma_k} \|x - x_k\|_2^2 \right\}.$$

Bianchi [129] established the almost-sure weak ergodic convergence of the following scheme under a weaker assumption of maximal monotonicity.

$$x_{k+1} := (I + \gamma_k T(x_k, \omega_k))^{-1}(x_k).$$

Under a strong monotonicity requirement on the operator, Patrascu and Necoara [128] proposed a variant of stochastic proximal point method on a problem over the intersection of a large collection of convex sets,

$$\begin{aligned} y_k &:= \operatorname{argmin}_{x \in \mathbb{R}^n} \left\{ f(x, \omega_k) + \frac{1}{2\gamma_k} \|x - x_k\|_2^2 \right\} \\ x_{k+1} &:= \Pi_{X_{\omega_k}}(y_k), \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and ω is a random variable with its associated probability space (Ω, P). This leads to a convergence rate of $\mathcal{O}(\frac{1}{k})$. When implementing (PP) on an expectation-valued map $T(x_k)$, where $T_i(x) \triangleq \mathbb{E}[T_i(x, \xi(\omega))]$, the resolvent is challenging to compute since the expectation is unavailable in closed form. To this end, we consider two avenues.

I. Sample-then-resolve Stochastic Proximal-Point schemes (str-SPP). Our first scheme replaces the expectation $T(x_k)$ by a sample-average $\bar{T}_k \triangleq \frac{\sum_{j=1}^{N_k} T(x,\omega_{j,k})}{N_k}$ in computing the resolvent; in effect, we first *sample* and then *resolve*. The resulting scheme is defined as follows given a random point $x_0 \in \mathbb{R}^n$,

$$x_{k+1} \coloneqq (I + \gamma \bar{T}_k)^{-1}(x_k), \qquad \text{(str-SPP)}$$

where γ is a suitably chosen steplength. If $N_k = 1$ for all k, we obtain a counterpart that is aligned with stochastic approximation schemes [99] while $N_k = N$ for all k leads to a mini-batch framework. If however, N_k is an increasing sequence, a variance-reduced scheme is obtained with the intent of recovering deterministic rates of convergence [108,140]. We analyze the convergence of (**str-SPP**) in Section 3.3.1.

II. Resolve-then-sample Stochastic Proximal-Point schemes (rts-SPP). Our second scheme retains the expectation $T(x_k)$ in the resolvent operator but computes the resolvent operator by a Monte-Carlo sampling scheme, leading to an error; in effect, we articulate the *resolvent* problem then utilize *sampling* to get an approximation, thus the qualifier *resolve* then *sample*. The resulting scheme is defined as follows given a random point $x_0 \in \mathbb{R}^n$,

$$x_{k+1} \coloneqq (I + \gamma_k T(x_k))^{-1}(x_k) + e_k, \qquad \text{(rts-SPP)}$$

where e_k denotes the random error in computing the resolvent $(I + \gamma_k T(x_k))^{-1}$ when employing Monte-Carlo sampling schemes. We analyze the convergence of (**rts-SPP**) in Section 3.3.2.

3.3.1 Convergence Analysis of (str-SPP)

We first introduce the Yosida approximation operator [141] and recall some of its properties.

Definition 5 (**Yosida approximation).** *For a set-valued maximal monotone operator $T : \mathbb{R}^n \to \mathbb{R}^n$, the Yosida approximation operator is denoted as $T_\gamma(\bullet)$ and is defined as follows.*

$$T_\gamma \triangleq \frac{1}{\gamma}(I - J_\gamma^T).$$

The Yosida approximation has an important property (cf. [134]) in that a zero of the operator is a solution to the original generalized equation, i.e.

$$0 \in T(x) \Leftrightarrow T_\gamma(x) = 0.$$

Recall that the Yosida operator $T_\gamma(x)$ is single-valued and $\frac{1}{\gamma}$-Lipschitz continuous and has the following property [134].

Lemma 9. *Given a set-valued maximal monotone operator T: $\mathbb{R}^n \rightrightarrows \mathbb{R}^n$, let $J_{\gamma_k}^T(x_k)$ denote the resolvent operator while T_{γ_k} denotes the Yosida approximation operator of T. Then we have that $T_{\gamma_k}(x) \in T(J_{\gamma_k}^T(x))$ for all $x \in \mathbb{R}^n$.*

We begin with a result that provides a bound for the sequence of iterates produced by a deterministic proximal point scheme.

Lemma 10. *Let Assumptions 9, 13 hold and let $\gamma_k = \gamma > 0$ for all k. Assume $x^* \in T^{-1}(0)$ is a solution. Consider any sequence generated by (**SPP**). Then the following holds for all $k > 0$:*

$$\|J_{\gamma_k}^T(x_k) - x^*\|^2 = \|x_k - x^*\|^2 - \gamma_k^2\|T_{\gamma_k}(x_k)\|^2 - 2\gamma_k T_{\gamma_k}(x_k)^T(J_{\gamma_k}^T(x_k) - x^*).$$

Proof. By the definition of the Yosida approximation operator, we can write $J_{\gamma_k}^T(x_k)$ as follows.

$$J_{\gamma_k}^T(x_k) = x_k - \gamma_k T_{\gamma_k}(x_k). \tag{3.1}$$

Using (3.1), we have

$$\begin{aligned} \|J_{\gamma_k}^T(x_k) - x^*\| &= \|x_k - \gamma_k T_{\gamma_k}(x_k) - x^*\|^2 \\ &= \|x_k - x^*\|^2 + \gamma_k^2\|T_{\gamma_k}(x_k)\|^2 - 2\gamma_k T_{\gamma_k}(x_k)^T(x_k - x^*). \end{aligned} \tag{3.2}$$

Using the definition of T_{γ_k}, we have

$$x_k = \gamma_k T_{\gamma_k}(x_k) + J^T_{\gamma_k}(x_k) \text{ and } x^* = \gamma_k T_{\gamma_k}(x^*) + J^T_{\gamma_k}(x^*). \tag{3.3}$$

Using (3.3), it follows that

$$\begin{aligned} T_{\gamma_k}(x_k)^T(x_k - x^*) &= T_{\gamma_k}(x_k)^T(\gamma_k T_{\gamma_k}(x_k) - \gamma_k \underbrace{T_{\gamma_k}(x^*)}_{=0}) \\ &\quad + T_{\gamma_k}(x_k)^T(J^T_{\gamma_k}(x_k) - \underbrace{J^T_{\gamma_k}(x^*)}_{=x^*}) \\ &= T_{\gamma_k}(x_k)^T(\gamma_k T_{\gamma_k}(x_k)) + T_{\gamma_k}(x_k)^T(J^T_{\gamma_k}(x_k) - x^*) \end{aligned} \tag{3.4}$$

The conclusion follows by inserting (3.4) in (3.2). □

We are now ready to prove our main convergence statement for (**str-SPP**). Since every iterate utilizes the resolvent of the sample-average map, we require the following assumption on $\bar{w}_{k,N_k}$, defined as follows.

$$\begin{aligned} \bar{v}_{k,N_k} &= x_{k+1} - y_{k+1} \\ \text{where } y_{k+1} &= (I + T(x_k))^{-1}(x_k). \end{aligned}$$

We then impose the following assumption on the conditional second moment of v_k.

Assumption 14. *At an iteration k, the following hold in an a.s. sense: (i) The conditional second moment is bounded or $\mathbb{E}[\|v_k\|^2 \mid \mathcal{F}_k] \leq \frac{C\nu^2}{N_k}$ for all k.*

Proposition 11 (Rate statement for (**str-SPP**))**.** *Let Assumptions 9 and 14 hold and let $\gamma > 0$ and $N_k = \lceil (k+1)^{2a} \rceil$ for $k > 0$ where $a > 1$. Consider a sequence $\{x_k\}$ generated by (**str-SPP**). Then the following hold.*
(1). The following holds for any $k > 0$.

$$\mathbb{E}[\|T_\gamma(x_k)\|^2] = \mathcal{O}\left(\frac{1}{k+1}\right).$$

(2).If x_K is such that $\mathbb{E}[\|T_\gamma(x_k)\|^2 \leq \epsilon$, then the oracle complexity of the number of evaluations of $T(x,\omega)$ are bounded as follows.

$$\sum_{k=1}^{K} N_k = \mathcal{O}\left(\frac{1}{\epsilon^{2a+1}}\right).$$

Proof. (1). We begin with invoking Lemma 9, implying that

$$\begin{aligned} & (T_\gamma(x_k) - T_\gamma(x^*))^T(J^T_\gamma(x_k) - J^T_\gamma(x^*)) \geq 0, \\ \equiv \quad & (T_\gamma(x_k))^T(J^T_\gamma(x_k) - x^*) \geq 0, \end{aligned} \tag{3.5}$$

by noticing that $T_\gamma(x^*) = 0$ and $J^T_\gamma(x^*) = x^*$. Substituting (3.5) in Lemma 10, we obtain the following relation

$$\begin{aligned} \|J^T_\gamma(x_k) - x^*\|^2 &= \|x_k - x^*\|^2 - \gamma^2\|T_\gamma(x_k)\|^2 - 2\gamma T_\gamma(x_k)^T(J^T_\gamma(x_k) - x^*) \\ &\leq \|x_k - x^*\|^2 - \gamma_k^2\|T_{\gamma_k}(x_k)\|^2 \leq \|x_k - x^*\|^2. \end{aligned} \tag{3.6}$$

With the aforementioned inequalities, we may provide the following bound in expectation.

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x^*\|] &\leq \mathbb{E}[\|J^T_\gamma(x_k) - x^*\|] + \mathbb{E}[\|x_{k+1} - J^T_\gamma(x_k)\|] \\ &\leq \mathbb{E}[\|J^T_\gamma(x_k) - x^*\|] + \frac{\gamma C\nu}{\sqrt{N_{k+1}}} \\ &\overset{\text{(From (3.6))}}{\leq} \mathbb{E}[\|x_k - x^*\|] + \frac{C\gamma\nu}{\sqrt{N_{k+1}}} && (3.7) \\ &\leq \|x_0 - x^*\| + C\gamma\nu\sum_{i=0}^{\infty}\frac{1}{\sqrt{N_{i+1}}}. && (3.8) \end{aligned}$$

Recall that $\|x_{k+1} - x^*\|^2$ can be bounded as follows.

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|J^T_{\gamma_k}(x_k) - x^*\|^2 + \|x_{k+1} - J^T_{\gamma_k}(x_k)\|^2 \\ &\quad + 2(J^T_{\gamma_k}(x_k) - x^*)^T(x_{k+1} - J^T_{\gamma_k}(x_k)) \\ &\leq \|J^T_{\gamma_k}(x_k) - x^*\|^2 + \|x_{k+1} - J^T_{\gamma_k}(x_k)\|^2 \\ &\quad + 2\|J^T_{\gamma_k}(x_k) - x^*\|\|x_{k+1} - J^T_{\gamma_k}(x_k)\|. \end{aligned} \tag{3.9}$$

Taking expectations on both sides of (3.9), we get

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x^*\|^2] &\leq \mathbb{E}[\|J^T_\gamma(x_k) - x^*\|^2] + \mathbb{E}[\|x_{k+1} - J^T_\gamma(x_k)\|^2] \\ &\quad + 2\mathbb{E}[\|J^T_\gamma(x_k) - x^*\|\|x_{k+1} - J^T_\gamma(x_k)\|] \\ &\leq \mathbb{E}[\|J^T_\gamma(x_k) - x^*\|^2] + \mathbb{E}[\|x_{k+1} - J^T_\gamma(x_k)\|^2] \\ &\quad + 2\mathbb{E}[\mathbb{E}[\|J^T_\gamma(x_k) - x^*\|\|x_{k+1} - J^T_\gamma(x_k)\| \mid \mathcal{F}_k]] \\ &= \mathbb{E}[\|J^T_\gamma(x_k) - x^*\|^2] + \mathbb{E}[\|x_{k+1} - J^T_\gamma(x_k)\|^2] \end{aligned}$$

$$
\begin{aligned}
&+ 2\mathbb{E}[\mathbb{E}[\|J_\gamma^T(x_k) - x^*\|]\mathbb{E}[\|x_{k+1} - J_\gamma^T(x_k)\| \mid \mathcal{F}_k]] \\
&= \mathbb{E}[\|J_\gamma^T(x_k) - x^*\|^2] + \mathbb{E}[\|x_{k+1} - J_\gamma^T(x_k)\|^2] \\
&+ 2\mathbb{E}[\|J_\gamma^T(x_k) - x^*\|]\mathbb{E}[\mathbb{E}[\|x_{k+1} - J_\gamma^T(x_k)\| \mid \mathcal{F}_k]]. \qquad (3.10)
\end{aligned}
$$

Invoking nonexpansivity of $J_{\gamma_k}^T$, Lemma 10, and (3.8), it follows that

$$
\mathbb{E}[\|J_{\gamma_k}^T(x_k) - J_{\gamma_k}^T(x^*)\|] \leq \mathbb{E}[\|x_k - x^*\|] \leq \|x_0 - x^*\| + C\gamma\nu \sum_{i=0}^{\infty} \frac{1}{\sqrt{N_{i+1}}}. \qquad (3.11)
$$

By invoking Lemma 13 and 10 and (3.11), the expression in (3.10) can be bounded as follows by noticing $\gamma_k = \gamma$.

$$
\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^*\|^2] &\leq \mathbb{E}[\|x_k - x^*\|^2] - \gamma^2\mathbb{E}[\|T_\gamma(x_k)\|^2] + \frac{C^2\gamma^2\nu^2}{N_{k+1}} \\
&+ \frac{2C\gamma\nu}{\sqrt{N_{k+1}}}\left(\|x_0 - x^*\| + C\gamma\nu\sum_{i=0}^{\infty}\frac{1}{\sqrt{N_{i+1}}}\right).
\end{aligned}
$$

Defining $E_1 \triangleq \sum_{i=0}^{\infty} \frac{1}{\sqrt{N_i}}$,$E_2 \triangleq \sum_{i=0}^{\infty} \frac{1}{N_i}$ and summing from $i = 0, \cdots, k$, we get

$$
\begin{aligned}
\gamma^2\sum_{i=0}^{k}\mathbb{E}[\|T_\gamma(x_i)\|^2] &\leq \|x_0 - x^*\|^2 - \mathbb{E}[\|x_{k+1} - x^*\|^2] \\
&+ \sum_{i=0}^{k}\left(\frac{C^2\gamma^2\nu^2}{N_{i+1}} + \frac{2C\gamma\nu}{\sqrt{N_{i+1}}}\left(\|x_0 - x^*\| + C\gamma\nu\sum_{\ell=0}^{\infty}\frac{1}{\sqrt{N_{\ell+1}}}\right)\right) \\
&\leq \|x_0 - x^*\|^2 + C^2\gamma^2\nu^2E_2 + 2C\gamma\nu E_1\|x_0 - x^*\| + 2\left(C\gamma\nu\right)^2 E_1^2 \\
&= (\|x_0 - x^*\| + C\gamma\nu E_1)^2 + C^2\gamma^2\nu^2(E_2 + E_1^2). \qquad (3.12)
\end{aligned}
$$

We have

$$
\begin{aligned}
T_\gamma(x_k) &= \frac{1}{\gamma}(x_k - J_\gamma^T(x_k)) = \frac{1}{\gamma}(x_{k+1} - J_\gamma^T(x_k) - (x_{k+1} - x_k)) \\
&= \frac{1}{\gamma}(x_{k+1} - J_\gamma^T(x_k)) - \frac{1}{\gamma}(J_\gamma^T(x_{k+1}) - J_\gamma^T(x_k)) - (T_\gamma(x_{k+1}) - T_\gamma(x_k))
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&(T_\gamma(x_{k+1}) - T_\gamma(x_k))^T T_\gamma(x_k) = \frac{1}{\gamma}(T_\gamma(x_{k+1}) - T_\gamma(x_k))^T(x_{k+1} - J_\gamma^T(x_k)) \\
&\underbrace{-\frac{1}{\gamma}(T_\gamma(x_{k+1}) - T_\gamma(x_k))^T(J_\gamma^T(x_{k+1}) - J_\gamma^T(x_k))}_{\geq\, 0} - \|T_\gamma(x_{k+1}) - T_\gamma(x_k)\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{1}{\gamma}(T_\gamma(x_{k+1}) - T_\gamma(x_k))^T(x_{k+1} - J_\gamma^T(x_k)) \\
&- \|T_\gamma(x_{k+1}) - T_\gamma(x_k)\|^2. \qquad (3.13)
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\|T_\gamma(x_{k+1})\|^2 &= \|T_\gamma(x_k)\|^2 + \|T_\gamma(x_{k+1}) - T_\gamma(x_k)\|^2 + 2(T_\gamma(x_{k+1}) - T_\gamma(x_k))^T T_\gamma(x_k) \\
&\overset{(3.13)}{\leq} \|T_\gamma(x_k)\|^2 - \|T_\gamma(x_{k+1}) - T_\gamma(x_k)\|^2 \\
&+ \frac{2}{\gamma}\|T_\gamma(x_{k+1}) - T_\gamma(x_k)\|\|x_{k+1} - J_\gamma^T(x_k)\| \\
&\leq \|T_\gamma(x_k)\|^2 - \|T_\gamma(x_{k+1}) - T_\gamma(x_k)\|^2 + \|T_\gamma(x_{k+1}) - T_\gamma(x_k)\|^2 + \frac{\nu^2}{\gamma^2 N_{k+1}} \\
&= \|T_\gamma(x_k)\|^2 + \frac{\nu^2}{\gamma^2 N_{k+1}}. \qquad (3.14)
\end{aligned}
$$

By (3.14), we have the following relationship.

$$
\|T_\gamma(x_k)\|^2 \leq \|T_\gamma(x_i)\|^2 + \sum_{j=i}^{k-1} \frac{\nu^2}{\gamma^2 N_{j+1}}, \quad \forall i = 0, \cdots, k-1. \qquad (3.15)
$$

Thus, we have

$$
\begin{aligned}
\mathbb{E}[\|T_\gamma(x_k)\|^2] &\overset{(3.12),(3.15)}{\leq} \frac{(\|x_0 - x^*\| + C\gamma\nu E_1)^2 + \left(C^2\gamma^2\nu^2 E_2 + (C\gamma\nu E_1)^2\right)}{\gamma^2(k+1)} \\
&+ \frac{\nu^2 \sum_{i=0}^{k}\sum_{j=i}^{k-1} \frac{1}{N_{j+1}}}{(k+1)}.
\end{aligned}
$$

Recalling $N_k = \lceil (k+1)^{2a} \rceil,\ a > 1$, it follows that

$$
\begin{aligned}
\sum_{i=0}^{k}\sum_{j=i}^{k-1} \frac{1}{N_{j+1}} &= \sum_{i=0}^{k}\sum_{j=i}^{k-1} \frac{1}{\lceil (j+1)^{2a} \rceil} \leq \sum_{i=0}^{k}\sum_{j=i}^{k-1} \frac{1}{(j+1)^{2a}} \\
&\leq \int_0^{k+1}\int_y^{k+1} \frac{1}{(x+1)^{2a}}\, dx\, dy \leq \frac{1}{2(2a-1)(a-1)}.
\end{aligned}
$$

Since $\sum_{i=0}^{\infty} \frac{1}{\sqrt{N_{i+1}}} < +\infty$ and $\sum_{i=0}^{\infty} \frac{1}{N_{i+1}} < +\infty$, we have $\mathbb{E}[\|T_\gamma(x_k)\|^2] = \mathcal{O}\left(\frac{1}{k+1}\right)$.
(2). Suppose x_K is such that $\mathbb{E}[\|T_\gamma(x_K)\|^2] \leq \epsilon$. From (1), we have, for sufficiently

small ϵ,

$$\sum_{k=1}^{K} N_k \leq \sum_{k=1}^{\lceil \widehat{C}/\epsilon \rceil - 1} N_k \leq \sum_{k=1}^{\lceil \widehat{C}/\epsilon \rceil - 1} \lceil (k+1)^{2a} \rceil \leq 2 \sum_{k=1}^{\lceil \widehat{C}/\epsilon \rceil - 1} (k+1)^{2a}$$
$$\leq 2 \int_{x=1}^{\widehat{C}/\epsilon} (x+1)^{2a} \, dx \leq \frac{2(\widehat{C}/\epsilon + 1)^{2a+1}}{2a+1} \leq \left(\frac{\widehat{C}}{\epsilon^{2a+1}} \right).$$

□

Next, we derive a rate statement under a strong monotonicity assumption on T. To begin, we need the following lemmas:

Lemma 11. *Given a scalar y, we have that $\lfloor y \rfloor \geq \left\lceil \frac{1}{2} y \right\rceil, \quad \forall y \geq 1, y \in \mathbb{R}$.*

Proof. Please refer to Appendix. □

Lemma 12. *Let Assumptions 9 – 10, 13 hold and let $\gamma > 0$. Assume $x^* \in T^{-1}(0)$ is a solution. Then we have the following for all k.*

$$\|J_{\gamma}^{T}(x_k) - x^*\| \leq (1 + \sigma\gamma)^{-1} \|x_k - x^*\|.$$

Proof. Suppose $y_{k+1} = J_{\gamma}^{T}(x_k)$. From the definition of $J_{\gamma_k}^{T}(x_k)$, we have $x_k = y_{k+1} + \gamma T_k(y_{k+1})$. It follows that

$$\begin{aligned}
\|x_k - x^*\|^2 &= \|y_{k+1} - x^*\|^2 + \gamma^2 \|\bar{T}_k(y_{k+1}) - \bar{T}_k(x^*)\|^2 \\
&\quad + 2\gamma(\bar{T}_k(y_{k+1}) - \bar{T}_k(x^*))^T (y_{k+1} - x^*) \\
&\geq (1 + 2\sigma\gamma)\|y_{k+1} - x^*\|^2 + \gamma^2 \|\bar{T}_k(y_{k+1}) - \bar{T}_k(x^*)\|^2 \\
&\geq (1 + \sigma\gamma)^2 \|y_{k+1} - x^*\|^2,
\end{aligned}$$

where the first inequality follows from the strong monotonicity of T_k and the second inequality is a consequence of $\|T_k(y_{k+1}) - T_k(x^*)\|^2 \geq \sigma^2 \|y_{k+1} - x^*\|^2$. Note that $\|T_k(y_{k+1}) - T_k(x^*)\|^2 \geq \sigma \|y_{k+1} - x^*\|^2$ follows from strong monotonicity of T_k and by invoking the Cauchy-Schwarz inequality. It follows that

$$\|x_k - x^*\|^2 \geq (1 + \sigma\gamma_k)^2 \|y_{k+1} - x^*\|^2 = (1 + \sigma\gamma)^2 \|J_{\gamma_k}^{T}(x_k) - x^*\|^2.$$

□

With Lemma 11 and 12, we may derive the following rate statement under strong monotonicity.

Proposition 12 (Linear convergence of (str-SPP) under strong monotonicity). *Let Assumptions 10 and 14 hold and assume $\|x^0 - x^*\|$ is bounded by C. Let $\gamma_k = \gamma$. Suppose $N_{k+1} = \lfloor \rho^{-2(k+1)} \rfloor$, $D \triangleq \sqrt{2}\gamma\nu$, $q \triangleq (1+\sigma\gamma)^{-1}$, $\tilde{\rho} \in (q,1)$, $\hat{D} > 1/\ln(\tilde{\rho}/\rho)^e$ and $\tilde{D} \triangleq \left(C + D\frac{1}{1-\min\{(q/\rho),(\rho/q)\}}\right)$. Then the following hold.*
*(a). Then any sequence generated by (**str-SPP**) converges to a solution $x^* \in T^{-1}(0)$ at a linear rate in an expected value sense .*
(b). Suppose x_{K+1} satisfies $\mathbb{E}[\|x_{K+1} - x^*\|] \leq \epsilon$. Then $\sum_{k=1}^{K} N_k \leq \mathcal{O}\left(\frac{1}{\epsilon}\right)$.

Proof. (a). By invoking Lemma 12 and Assumption 14, we obtain the following.

$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^*\|] &\leq \mathbb{E}[\|y_{k+1} - x^*\|] + \mathbb{E}[\|x_{k+1} - y_{k+1}\|] \\ &\leq (1+\sigma\gamma)^{-1}\mathbb{E}[\|x_k - x^*\|] + \frac{\gamma\nu}{\sqrt{N_{k+1}}}. \end{aligned} \tag{3.16}$$

According to Lemma 11, we have

$$N_k = \lfloor \rho^{-2(k+1)} \rfloor \geq \left\lceil \frac{1}{2}\rho^{-2(k+1)} \right\rceil \geq \frac{1}{2}\rho^{-2(k+1)}. \tag{3.17}$$

If $q < \rho < 1$: Using (3.17) in (3.16), we deduce

$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^*\|] &\leq q\mathbb{E}[\|x_k - x^*\|] + \frac{\gamma\nu}{\sqrt{N_{k+1}}} = q\mathbb{E}[\|x_k - x^*\|] + D\rho^{k+1} \\ &\leq q^{k+1}\mathbb{E}[\|x_0 - x^*\|] + D\sum_{j=1}^{k+1} q^{k+1-j}\rho^j \\ &\leq q^{k+1}C + D\rho^{k+1}\sum_{j=1}^{k+1}(q/\rho)^{k+1-j} \leq \tilde{D}\rho^{k+1}.\end{aligned}$$

If $\rho < q < 1$: With a similar approach, we obtain $\mathbb{E}[\|x_{k+1} - x^*\|] \leq \tilde{D}q^{k+1}$.
If $\rho = q$: We have that

$$\mathbb{E}[\|x_{k+1} - x^*\|] \leq q^{k+1}\mathbb{E}[\|x_0 - x^*\|] + Dq^{k+1} \leq q^{k+1}C + D\sum_{j=1}^{k+1} q^{j+1}. \tag{3.18}$$

According to [142, Lemma 4], we can deduce from (3.18) that

$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^*\|] &\leq q^{k+1}C + D\sum_{j=1}^{k+1} q^{j+1} \leq Cq^{k+1} + Dq^{k+1}(k+1) \\ &\leq Cq^{k+1} + \hat{D}\tilde{\rho}^{k+1} \leq (C+\hat{D})\tilde{\rho}^{k+1}.\end{aligned}$$

Thus, we conclude that any sequence generated by (**str-SPP**) converges linearly in an expected-value sense.

(b). If $q < \rho < 1$: From (a), it follows that

$$\mathbb{E}[\|x_{K+1} - x^*\|] \leq \tilde{D}\rho^{K+1} \leq \epsilon \Longrightarrow K = \lceil \log_{1/\rho}(\tilde{D}/\epsilon) \rceil - 1.$$

For the optimal oracle complexity, we require $\sum_{k=1}^{K} N_k$ gradients. Since $N_k = \lfloor \rho^{-2(k+1)} \rfloor \leq \rho^{-2(k+1)}$, we have

$$\sum_{k=1}^{\log_{1/\rho}(\tilde{D}/\epsilon)} \rho^{-2(k+1)} \leq \frac{1}{\left(\frac{1}{\rho} - 1\right)} \left(\frac{1}{\rho}\right)^{2+2\log_{1/\rho}(\tilde{D}/\epsilon)} \leq \left(\frac{\tilde{D}}{\epsilon}\right)^2 \frac{1}{\rho(1-\rho)}.$$

If $\rho < q < 1$: Similarly, we have $\sum_{k=1}^{K} N_k \leq \left(\frac{\tilde{D}}{\epsilon}\right)^2 \frac{1}{q(1-q)}$.

If $q = \rho$: From (a), we have that $\mathbb{E}[\|x_{K+1} - x^*\|] \leq \tilde{C}\tilde{\rho}^{K+1}$, where $\tilde{C} = C + \hat{D}$. It follows that $\sum_{k=1}^{K} N_k \leq \left(\frac{\tilde{C}}{\epsilon}\right) \frac{1}{\rho(1-\rho)}$. □

Crucial to the rate and complexity statements is Assumption 14. While there are many settings under which this Assumption holds we now show that when Assumption 11 holds, Assumption 13 implies that Assumption 14 is satisfied.

Lemma 13. *Let Assumptions 9, 11, 13 hold. Consider a sequence generated by* (**str-SPP**)*. For any $x_0 \in \mathbb{R}^n$, the following holds for all $k > 0$:*

$$\mathbb{E}[\|x_{k+1} - J_\gamma^T(x_k)\|^2] \leq \frac{\gamma^2\nu^2}{(1-2\gamma L)^2 N_{k+1}}.$$

Proof. Suppose $y_{k+1} = J_\gamma^T(x_k)$. From the definition of $J_\gamma^T(x_k)$ and $J_\gamma^{T_k}(x_k)$,

$$\begin{aligned}
\left[y_{k+1} = (I+\gamma T)^{-1}(x_k)\right] &\equiv [y_{k+1} + \gamma T(y_{k+1}) = x_k] \\
\left[x_{k+1} = (I+\gamma \bar{T}_k)^{-1}(x_k)\right] &\equiv \left[x_{k+1} + \gamma \bar{T}_k(x_{k+1}) = x_k\right].
\end{aligned} \tag{3.19}$$

Recall from the monotonicity of T, we have the following.

$$\begin{aligned}
&((I+\gamma T)(x_{k+1}) - (I+\gamma T)(y_{k+1}))^T(x_{k+1} - y_{k+1}) \\
&= ((x_{k+1} - y_{k+1}) + \gamma(T(x_{k+1}) - T(y_{k+1})))^T(x_{k+1} - y_{k+1}) \\
&= (x_{k+1} - y_{k+1})^T(x_{k+1} - y_{k+1}) + \gamma(T(x_{k+1}) - T(y_{k+1}))^T(x_{k+1} - y_{k+1}) \\
&\geq \|x_{k+1} - y_{k+1}\|^2.
\end{aligned}$$

It follows that

$$
\begin{aligned}
&\|x_{k+1}-y_{k+1}\|^2 \leq ((I+\gamma T)(x_{k+1})-(I+\gamma T)(y_{k+1}))^T(x_{k+1}-y_{k+1}) \\
&= ((I+\gamma T)(x_{k+1})-x_k-(I+\gamma T)(y_{k+1})+x_k)^T(x_{k+1}-y_{k+1}) \\
&\overset{(3.19)}{=} ((I+\gamma T)(x_{k+1})-x_k)^T(x_{k+1}-y_{k+1}) \\
&= ((I+\gamma T)(x_{k+1})-(I+\gamma \bar{T}_k)(x_{k+1})+(I+\gamma \bar{T}_k)(x_{k+1})-x_k)^T(x_{k+1}-y_{k+1}) \\
&\overset{(3.19)}{=} \gamma(T(x_{k+1})-\bar{T}_k(x_{k+1}))^T(x_{k+1}-y_{k+1}).
\end{aligned} \tag{3.20}
$$

We have the following equality

$$
\begin{aligned}
T(x_{k+1})-\bar{T}_k(x_{k+1}) &= T(x_{k+1})-\bar{T}_k(x_{k+1})+\bar{T}_k(y_{k+1})-\bar{T}_k(y_{k+1}) \\
&= T(x_{k+1})-T(y_{k+1})-\bar{T}_k(x_{k+1})+\bar{T}_k(y_{k+1})+T(y_{k+1}) \\
&\quad -\bar{T}_k(y_{k+1}).
\end{aligned} \tag{3.21}
$$

Substituting (3.21) to (3.20), it follows that

$$
\begin{aligned}
\|x_{k+1}-y_{k+1}\|^2 &\leq \gamma(T(x_{k+1})-\bar{T}_k(x_{k+1}))^T(x_{k+1}-y_{k+1}) \\
&= \gamma(T(x_{k+1})-T(y_{k+1}))^T(x_{k+1}-y_{k+1}) \\
&\quad -\gamma(\bar{T}_k(x_{k+1})-\bar{T}_k(y_{k+1}))^T(x_{k+1}-y_{k+1}) \\
&\quad +\gamma(T(y_{k+1})-\bar{T}_k(y_{k+1}))^T(x_{k+1}-y_{k+1}) \\
&\overset{\text{Lip.}}{\leq} 2\gamma L\|x_{k+1}-y_{k+1}\|^2 \\
&\quad +\gamma\|T(y_{k+1})-\bar{T}_k(y_{k+1})\|\|x_{k+1}-y_{k+1}\|.
\end{aligned} \tag{3.22}
$$

Following (3.22), we have

$$
(1-2\gamma L)\|x_{k+1}-y_{k+1}\|^2 \leq \gamma\|T(y_{k+1})-\bar{T}_k(y_{k+1})\|.
$$

Therefore, by leveraging Assumption 13, we have that

$$
\begin{aligned}
\mathbb{E}[\|x_{k+1}-J_\gamma^T(x_k)\|^2] = \mathbb{E}[\|x_{k+1}-y_{k+1}\|^2] &\leq \frac{\gamma^2}{(1-2\gamma L)^2}\mathbb{E}[\|T(y_{k+1})-\bar{T}_k(y_{k+1})\|^2] \\
&= \frac{\gamma^2}{(1-2\gamma L)^2}\mathbb{E}[\mathbb{E}[\|T(y_{k+1})-\bar{T}_k(y_{k+1})\|^2 \mid \mathcal{F}_k]] \\
&\leq \frac{\gamma^2\nu^2}{(1-2\gamma L)^2 N_{k+1}}.
\end{aligned}
$$

$\square$

3.3.2 Convergence Analysis of (rts-SPP)

We now prove convergence of the (**rts-SPP**) scheme and being by making an assumption on the random error e_k that emerges in the (**rts-SPP**) update rule. Notably, this error arises from solving the resolvent problem inexactly by using a Monte-Carlo sampling scheme.

Assumption 15. *At an iteration k, the following hold in an a.s. sense: (i) The conditional second moment is bounded or $\mathbb{E}[\|e_k\|^2 \mid \mathcal{F}_k] \leq \frac{C\nu^2}{N_k}$ for all k.*

Proposition 13 (Rate statement for (**rts-SPP**))**.** *Let Assumptions 9 and 15 hold, $\gamma_k = \gamma > 0$, and $N_k \triangleq \lceil (k+1)^{2a} \rceil$ for $k > 0$, where $a > 1$. Consider a sequence generated by (**rts-SPP**).*

*(1). Any sequence generated by (**rts-SPP**) converges to a solution $x^* \in T^{-1}(0)$ as specified by the following. $\mathbb{E}[\|T_\gamma(x_k)\|^2] = \mathcal{O}\left(\frac{1}{k+1}\right)$.*

(2). Suppose x_K satisfies $\mathbb{E}[\|T(x_K)\|^2] \leq \epsilon$. Then the oracle complexity of computing x_K is bounded as follows

$$\sum_{k=1}^{K} N_k = \mathcal{O}\left(\frac{1}{\epsilon^{2a+1}}\right).$$

Proof. The proof proceed in a similar manner to that provided for Proposition 11, the only difference being that $\mathbb{E}[\|x_{k+1} - J^T_{\gamma_k}(x_k)\|]^2 = \frac{C^2\nu^2}{N_{k+1}}$. Thus, we have

$$\begin{aligned}\sum_{k=1}^{K} N_k &\leq \sum_{k=1}^{\lceil \widehat{C}/\epsilon \rceil - 1} N_k \leq \sum_{k=1}^{\lceil \widehat{C}/\epsilon \rceil - 1} \lceil (k+1)^{2a} \rceil \leq 2 \sum_{k=1}^{\lceil \widehat{C}/\epsilon \rceil - 1} (k+1)^{2a} \\ &\leq 2 \int_{x=1}^{\widehat{C}/\epsilon} (x+1)^{2a} \, dx \leq \frac{2(\widehat{C}/\epsilon + 1)^{2a+1}}{2a+1} \leq \left(\frac{\widehat{C}}{\epsilon^{2a+1}}\right).\end{aligned}$$

$\square$

In the (**rts-SPP**) scheme, we employ a Monte-Carlo sampling-based scheme for approximating the resolvent. Based on the nature of this problem, we may derive an overall iteration complexity in terms of “minor” iterations for solving the resolvent. So one thing we benefit from using (**rts-SPP**) is that we can obtain a superior iteration complexity for approximating the resolvent by leveraging the structure of different problems.

Proposition 14 (**Linear convergence of (rts-SPP) under strong monotonicity**). *Let Assumptions 9 – 10, 13 hold and assume $\|x^0 - x^*\|$ is bounded by C.*
*(a). Let $\gamma_k = \gamma$. Suppose $N_{k+1} = \lfloor \rho^{-2(k+1)} \rfloor$, $D \triangleq \sqrt{2}\gamma\nu$, $q \triangleq (1+\sigma\gamma)^{-1}$, $\tilde{\rho} \in (q,1)$, $\hat{D} > 1/\ln(\tilde{\rho}/\rho)^e$ and $\tilde{D} \triangleq \left(C + D\frac{1}{1-\min\{(q/\rho),(\rho/q)\}}\right)$. Then any sequence generated by (**rts-SPP**) converges to a solution $x^* \in T^{-1}(0)$ at a linear rate in an expected value sense .*
(b). To x_{K+1} such that $\mathbb{E}[\|x_{K+1} - x^*\|] \leq \epsilon$, we have $\sum_{k=1}^{K} N_k \leq \mathcal{O}\left(\frac{1}{\epsilon}\right)$.

Proof. Please refer to the proof of Proposition 12 by noticing that $\mathbb{E}[\|x_{k+1} - J^T_{\gamma_k}(x_k)\|] = \frac{\nu}{\sqrt{N_{k+1}}}$ instead of $\mathbb{E}[\|x_{k+1} - J^T_{\gamma_k}(x_k)\|] \leq \frac{\gamma\nu}{\sqrt{N_{k+1}}}$. □

3.4 Stochastic Modified Forward-Backward Schemes

In this section, we consider the development of stochastic splitting schemes when $T = A + B$ with our interest lying in computing an $x \in X \subseteq \mathbb{R}^n$ such that

$$0 \in T(x) \triangleq A(x) + B(x),$$

where $A(x) \triangleq \mathbb{E}[A(x,\omega)]$ and B has a relatively cheap resolvent. In such instances, we may consider develop a stochastic generalization of a splitting-based framework. We first provide a short summary of our approach and relate it to prior work and then proceed to derive rate statements for the proposed scheme.

3.4.1 Overview of proposed framework

While there are a range of splitting schemes that may be adopted, we consider a stochastic modified forward-backward scheme that utilizes $\frac{\sum_{j=1}^{N_k} A(x_k,\omega_{j,k})}{N_k}$ to approximate $\mathbb{E}[A(x_k,\omega)]$ at iteration k. We formally defined such a scheme next, given a random point $x_0 \in \mathbb{R}^n$.

$$\begin{aligned} x_{k+\frac{1}{2}} &:= (I + \gamma_k B)^{-1}(x_k - \gamma_k A_k), \\ x_{k+1} &:= \Pi_X\left(x_{k+\frac{1}{2}} - \gamma_k\left(A_{k+\frac{1}{2}} - A_k\right)\right), \end{aligned} \qquad \text{(SMFBS)}$$

where $A_k \triangleq \frac{\sum_{j=1}^{N_k} A(x_k,\omega_{j,k})}{N_k}$, $A_{k+\frac{1}{2}} \triangleq \frac{\sum_{j=1}^{N_k} A(x_{k+1/2},\omega_{j,k+1/2})}{N_k}$ are unbiased estimators of $A(x_k)$ and $A\left(x_{k+\frac{1}{2}}\right)$, respectively. We provide rate statements for settings where $N_k = 1$ for all k (a stochastic approximation counterpart) and $N_k = K$ for all k (a mini-batch stochastic approximation framework). When N_k is an increasing sequence, a variance-reduced scheme is obtained and allows for recovering deterministic rates of convergence under some assumptions. Throughout the remainder of this section, x^* refers to a point in $T^{-1}(0) \cap X$. We analyze the rate of convergence of (**SMFBS**) in Section 3.4.2

3.4.2 Convergence Analysis

We assume the following on the map A and B.

Assumption 16. *The mapping $A(x)$ is L-Lipschitz continuous, monotone, and single-valued on $\mathbb{R}^n$ and the mapping B is maximal monotone on $\mathbb{R}^n$, i.e.*

$$\begin{aligned} \|A(x) - A(y)\| \leq L\|x-y\|, \quad \forall x, y \in \mathbb{R}^n \\ (A(x) - A(y))^T(x-y) \geq 0, \quad \forall x, y \in \mathbb{R}^n, \\ \textit{and } (u-v)^T(x-y) \geq 0, \quad \forall x, y \in \mathbb{R}^n, \quad \forall u \in B(x), \quad \forall v \in B(y). \end{aligned}$$

Assumption 17. *The single-valued mapping A is co-coercive on $\mathbb{R}^n$; there exists a constant c_1 such that $(A(x) - A(y))^T(x-y) \geq c_1\|A(x) - A(y)\|^2$ for all $x, y \in \mathbb{R}^n$.*

We assume the presence of a stochastic oracle that can provide an unbiased estimator of $A(x)$, given by $A(x,\omega)$ such that $\mathbb{E}[A(x,\omega)] = A(x)$. Define $w_k \triangleq A(x_k,\omega_k) - A(x_k)$ and $w_{k+\frac{1}{2}} \triangleq A\left(x_{k+\frac{1}{2}}, \omega_{k+\frac{1}{2}}\right) - A\left(x_{k+\frac{1}{2}}\right)$. Furthermore we denote $\mathcal{F}_k$ as the history up to iteration k, i.e., $\mathcal{F}_k \triangleq \left\{x_0, \omega_0, \omega_{\frac{1}{2}}, \omega_1, \cdots, \omega_{k-1}, \omega_{k-\frac{1}{2}}\right\}$ and $\mathcal{F}_{k+\frac{1}{2}} \triangleq \mathcal{F}_k \cup \{\omega_k\}$.

Assumption 18. *At an iteration k, the following hold in an a.s. sense: (i) The conditional means $\mathbb{E}[w_k \mid \mathcal{F}_k]$ and $\mathbb{E}\left[w_{k+\frac{1}{2}} \mid \mathcal{F}_{k+\frac{1}{2}}\right]$ are zero for all k; (ii) The conditional second moments are bounded or $\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] \leq \frac{\nu^2}{N_k}$ and $\mathbb{E}\left[\left\|w_{k+\frac{1}{2}}\right\|^2 \mid \mathcal{F}_{k+\frac{1}{2}}\right] \leq \frac{\nu^2}{N_k}$ for all k.*

Lemma 14. *Let Assumptions 12, 16 – 18 hold and assume A is strictly monotone. Let $\gamma_k \leq \frac{1}{4L}$ and $c \leq \frac{1}{2L}$. Then for any k, we have the following bound.*

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq \ \|x_k - x^*\|^2 - \gamma_k c_1 \|u_k - u^*\|^2 + 8\gamma_k^2 \tfrac{\nu^2}{N_k}.$$

Proof. From the definition of $x_{k+\frac{1}{2}}$ and x_{k+1}, we have

$$\begin{aligned} x_{k+1/2} &= (I + \gamma_k B)^{-1}(x_k - \gamma_k A_k) \text{ or } x_{k+\frac{1}{2}} + \gamma_k v_{k+\frac{1}{2}} = x_k - \gamma_k(u_k + w_k), \\ x_{k+1} &= \Pi_X \left(x_{k+\frac{1}{2}} - \gamma_k \left(u_{k+\frac{1}{2}} + w_{k+\frac{1}{2}} - u_k - w_k \right) \right). \end{aligned}$$

where $u_k = A(x_k), u_{k+\frac{1}{2}} = A\left(x_{k+\frac{1}{2}}\right), v_{k+\frac{1}{2}} \in B\left(x_{k+\frac{1}{2}}\right)$. From $0 \in A(x^*) + B(x^*)$,

$$u^* + v^* = 0, \text{ where } u^* = A(x^*), \quad v^* \in B(x^*) \tag{3.23}$$

If $z_k = x_{k+\frac{1}{2}} - \gamma_k \left(u_{k+\frac{1}{2}} + w_{k+\frac{1}{2}} - u_k - w_k \right)$, we have the following equality:

$$\begin{aligned} &\|x_k - x^*\|^2 = \left\| x_k - x_{k+\frac{1}{2}} + x_{k+\frac{1}{2}} - z_k + z_k - x^* \right\|^2 \\ &= \left\| x_k - x_{k+\frac{1}{2}} \right\|^2 + \left\| x_{k+\frac{1}{2}} - z_k \right\|^2 + \|z_k - x^*\|^2 + 2\left(x_k - x_{k+\frac{1}{2}} \right)^T \left(x_{k+\frac{1}{2}} - x^* \right) \\ &+ 2\left(x_{k+\frac{1}{2}} - z_k \right)^T (z_k - x^*) \\ &= \left\| x_k - x_{k+\frac{1}{2}} \right\|^2 + \left\| x_{k+\frac{1}{2}} - z_k \right\|^2 + \|z_k - x^*\|^2 + 2\left(x_k - x_{k+\frac{1}{2}} \right)^T \left(x_{k+\frac{1}{2}} - x^* \right) \\ &- 2\left\| x_{k+\frac{1}{2}} - z_k \right\|^2 + 2\left(x_{k+\frac{1}{2}} - z_k \right)^T \left(x_{k+\frac{1}{2}} - x^* \right) \\ &= \left\| x_k - x_{k+\frac{1}{2}} \right\|^2 - \left\| x_{k+\frac{1}{2}} - z_k \right\|^2 + \|z_k - x^*\|^2 + 2(x_k - z_k)^T \left(x_{k+\frac{1}{2}} - x^* \right) \\ &= \left\| x_k - x_{k+\frac{1}{2}} \right\|^2 - \gamma_k^2 \left\| u_{k+\frac{1}{2}} + w_{k+\frac{1}{2}} - u_k - w_k \right\|^2 + \|z_k - x^*\|^2 \\ &+ 2\gamma_k \left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}} + w_{k+\frac{1}{2}} \right)^T \left(x_{k+\frac{1}{2}} - x^* \right). \end{aligned} \tag{3.24}$$

Since A is co-coercive and B is monotone on $\mathbb{R}^n$, it follows that

$$\begin{aligned} &\left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}} - u^* - v^* \right)^T \left(x_{k+\frac{1}{2}} - x^* \right) \geq c_1 \left\| u_{k+\frac{1}{2}} - u^* \right\|^2 \\ &\implies \left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}} \right)^T \left(x_{k+\frac{1}{2}} - x^* \right) \geq c_1 \left\| u_{k+\frac{1}{2}} - u^* \right\|^2. \qquad \text{(From (3.23))} \end{aligned} \tag{3.25}$$

By (3.25), $\gamma_k \leq 1/2L$, and moving $\|x_{k+1} - x^*\|^2$ to the left hand side, from (3.24),

$$\begin{aligned} &\|x_{k+1} - x^*\|^2 \leq \|z_k - x^*\|^2 \\ &= \|x_k - x^*\|^2 - \left\| x_k - x_{k+\frac{1}{2}} \right\|^2 \end{aligned}$$

$$
\begin{aligned}
&+ \gamma_k^2 \left\|u_{k+\frac{1}{2}} + w_{k+\frac{1}{2}} - u_k - w_k\right\|^2 - 2\gamma_k \left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}} + w_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&= \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + \gamma_k^2 \left\|u_{k+\frac{1}{2}} + w_{k+\frac{1}{2}} - u_k - w_k\right\|^2 \\
&- 2\gamma_k \left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) - 2\gamma_k w_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&\leq \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2 \left\|u_{k+\frac{1}{2}} - u_k\right\|^2 + 2\gamma_k^2 \left\|w_{k+\frac{1}{2}} - w_k\right\|^2 \\
&- 2\gamma_k \left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) - 2\gamma_k w_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&\leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2) \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - 2\gamma_k \left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&+ 2\gamma_k^2 \left\|w_{k+\frac{1}{2}} - w_k\right\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right) \qquad (3.26) \\
&\overset{(3.25)}{\leq} \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2) \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - 2\gamma_k c_1 \left\|u_{k+\frac{1}{2}} - u^*\right\|^2 \\
&+ 2\gamma_k^2 \left\|w_{k+\frac{1}{2}} - w_k\right\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right). \qquad (3.27)
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2) \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - 2\gamma_k c_1 \left\|u_{k+\frac{1}{2}} - u^*\right\|^2 \\
&+ 2\gamma_k^2 \left\|w_{k+\frac{1}{2}} - w_k\right\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&\leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2) \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - \gamma_k c_1 \|u_k - u^*\|^2 \\
&+ 2\gamma_k c_1 \left\|u_k - u_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2 \left\|w_{k+\frac{1}{2}} - w_k\right\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&= \|x_k - x^*\|^2 - \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - \gamma_k c_1 \|u_k - u^*\|^2 + \frac{1}{2}(2\gamma_k + c_1)^2 L^2 \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 \\
&- \frac{1}{2} c_1^2 L^2 \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2 \left\|w_{k+\frac{1}{2}} - w_k\right\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&\overset{(2\gamma_k + c_1)^2 L^2 < 1)}{\leq} \|x_k - x^*\|^2 - \frac{1}{2} \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 - \gamma_k c_1 \|u_k - u^*\|^2 \\
&- \frac{1}{2} c_1^2 L^2 \left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2 \left\|w_{k+\frac{1}{2}} - w_k\right\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&\leq \|x_k - x^*\|^2 - \gamma_k c_1 \|u_k - u^*\|^2 + 2\gamma_k^2 \left\|w_{k+\frac{1}{2}} - w_k\right\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right) \\
&\leq \|x_k - x^*\|^2 - \gamma_k c_1 \|u_k - u^*\|^2 + 4\gamma_k^2 \left\|w_{k+\frac{1}{2}}\right\|^2 + 4\gamma_k^2 \|w_k\|^2 \\
&- 2\gamma_k w_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right),
\end{aligned}
$$

where we leverage $\|u_k - u^*\|^2 \leq 2 \left\|u_k - u_{k+\frac{1}{2}}\right\|^2 + 2 \left\|u_{k+\frac{1}{2}} - u^*\right\|^2$ and $\left\|w_{k+\frac{1}{2}} - w_k\right\|^2 \leq 2 \left\|w_{k+\frac{1}{2}}\right\|^2 + 2\|w_k\|^2$ in the last inequality. Taking expectations conditioned on $\mathcal{F}_k$, we obtain the following bound:

$$\begin{aligned}\mathbb{E}[\|x_{k+1}-x^*\|^2 \mid \mathcal{F}_k] &\leq \|x_k - x^*\|^2 - \gamma_k c_1 \|u_k - u^*\|^2 \\ &\quad + 4\gamma_k^2 \mathbb{E}\left[\mathbb{E}\left[\left\|w_{k+\frac{1}{2}}\right\|^2 \mid \mathcal{F}_{k+\frac{1}{2}}\right] \mid \mathcal{F}_k\right] + 4\gamma_k^2 \mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] \\ &\quad - 2\gamma_k \mathbb{E}\left[\mathbb{E}\left[w_{k+\frac{1}{2}}^T \left(x_{k+\frac{1}{2}} - x^*\right) \mid \mathcal{F}_{k+\frac{1}{2}}\right] \mid \mathcal{F}_k\right] \\ &\leq \|x_k - x^*\|^2 - \gamma_k c_1 \|u_k - u^*\|^2 + 8\gamma_k^2 \frac{\nu^2}{N_k}.\end{aligned}$$

□

We are now ready to prove the a.s. convergence of the sequence generated by this scheme.

Proposition 15. *Let Assumptions 12, 16 – 18 hold and assume A is strictly monotone. Let $\gamma_k \leq \frac{1}{4L}$ and $c \leq \frac{1}{2L}$. Suppose $N_k \geq 1$ for all k. Then for any $x_0 \in X$, a sequence generated by (SMFBS) converges to a solution x^* in an a.s. sense.*

Proof. We may now apply Lemma 2 which allows us to claim that $\{\|x_k - x^*\|\}$ is convergent and $\sum_k \gamma_k c_1 \|u_k - u^*\|^2 < \infty$ in an a.s. sense. Since $\sum_k \gamma_k = \infty$, in an a.s. sense, we have

$$\liminf_{k\to\infty} \|u_k - u^*\|^2 = 0.$$

Consequently, a subsequence of $\{u_k\}$ converges to u^*. Furthermore, since $\|x_k - x^*\| \to r \geq 0$ in an a.s. sense implying that every limit point of $\{x_k\}$ satisfies $\|\bar{x} - x^*\| = r$. Furthermore, for every limit point $\bar{x}$, we have that $A(\bar{x}) = u^*$. Since A is strictly monotone, thus $(A(\bar{x}) - u^*)^T(\bar{x} - x^*) > 0$ holds at x^*. But for every $\bar{x}$, we have $A(\bar{x}) - u^* = 0$, implying that $\bar{x} = x^*$. Consequently, every limit point of $\{x_k\}$ is x^* in an a.s. sense and we may claim that the entire sequence $\{x_k\}$ converges to x^* in an a.s. sense. □

It is worth emphasizing that the aforementioned result holds for both constant and increasing sequences $\{N_k\}$.

3.4.3 Rate under Strong Monotonicity of A

Next, we derive the rate of convergence of the iterates generated by (SMFBS). The following bound is used in deriving the rate of convergence for (SMFBS).

Lemma 15. *Suppose $\{x_k\}$ and $\{x_{k+\frac{1}{2}}\}$ denote the sequences generated by (SMFBS). Then $\left\|x_k - x_{k+\frac{1}{2}}\right\|$ is bounded for all k, i.e. $\left\|x_k - x_{k+\frac{1}{2}}\right\| \leq C_1$.*

Proof. Taking limits on both sides of the first step of (SMFBS), we obtain

$$\lim_{k\to\infty}\left(x_{k+\frac{1}{2}} + \gamma_k B\left(x_{k+\frac{1}{2}}\right)\right) = \lim_{k\to\infty}(x_k - \gamma_k A(x_k)).$$

Since $\lim_{k\to\infty}\gamma_k = 0$, we have $\lim_{k\to\infty}\left(x_{k+\frac{1}{2}} - x_k\right) = 0$ and the result follows. □

Assumption 19. *The solution set X^* is bounded as specified by $\sup_{x^*\in X^*}\|x^*\| \leq C$.*

Assumption 20. *The mapping A is σ-strongly monotone, i.e. $(A(x)-A(y))^T(x-y) \geq \sigma\|x-y\|^2, \quad \forall x,y \in \mathbb{R}^n$.*

We now consider two sets of cases.

(i). $N_k = 1$ for every k. In this setting, we assume that the steplength sequence is given by

$$\gamma_k = \frac{\gamma_0}{k}. \tag{3.28}$$

We need the following lemma to establish the convergence rate [29].

Lemma 16. *Consider the following recursion: $a_{k+1} \leq \left(1 - \frac{2c\theta}{k}\right)a_k + \frac{\theta^2 M^2}{2k^2}$, where θ, M are positive constants and $1 - 2c\theta < 0$. Then for $k \geq 1$, we have that $2a_k \leq \frac{1}{k}\max\left(\frac{\theta^2 M^2}{2c\theta - 1}, 2a_1\right)$.*

We now proceed to prove our rate statement under strong monotonicity.

Lemma 17. *Let Assumptions 16, 18, 20 hold. Then we have that the following holds for every k.*

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq (1-\sigma\gamma_k)\|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2 - 2\sigma\gamma_k)\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 \\ &\quad + 4\gamma_k^2\left\|w_{k+\frac{1}{2}}\right\|^2 + 4\gamma_k^2\|w_k\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}} - x^*\right).\end{aligned}$$

Proof. According to Assumption 20, we have

$$\begin{aligned}-2\gamma_k\left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}} - x^*\right) &\leq -2\gamma_k\sigma\left\|x_{k+\frac{1}{2}} - x^*\right\|^2 \\ &\leq 2\gamma_k\sigma\left\|x_{k+\frac{1}{2}} - x_k\right\|^2 - \gamma_k\sigma\|x_k - x^*\|^2.\end{aligned} \tag{3.29}$$

Using (3.29) in (3.27), we deduce

$$
\begin{aligned}
&\|x_{k+1}-x^*\|^2 \le \|x_k-x^*\|^2 - (1-2\gamma_k^2L^2)\left\|x_k-x_{k+\frac{1}{2}}\right\|^2 - 2\gamma_k\sigma\left\|x_{k+\frac{1}{2}}-x^*\right\|^2 \\
&+ 2\gamma_k^2\left\|w_{k+\frac{1}{2}}-w_k\right\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}}-x^*\right) \\
&\le \|x_k-x^*\|^2 - (1-2\gamma_k^2L^2)\left\|x_k-x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k\sigma\left\|x_{k+\frac{1}{2}}-x_k\right\|^2 \\
&- \gamma_k\sigma\|x_k-x^*\|^2 + 2\gamma_k^2\left\|w_{k+\frac{1}{2}}-w_k\right\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}}-x^*\right) \\
&\le (1-\sigma\gamma_k)\|x_k-x^*\|^2 - (1-2\gamma_k^2L^2-2\sigma\gamma_k)\left\|x_k-x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2\left\|w_{k+\frac{1}{2}}-w_k\right\|^2 \\
&- 2\gamma_k w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}}-x^*\right) \\
&\le (1-\sigma\gamma_k)\|x_k-x^*\|^2 - (1-2\gamma_k^2L^2-2\sigma\gamma_k)\left\|x_k-x_{k+\frac{1}{2}}\right\|^2 + 4\gamma_k^2\left\|w_{k+\frac{1}{2}}\right\|^2 \\
&+ 4\gamma_k^2\|w_k\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}}-x^*\right).
\end{aligned}
$$

□

Proposition 16 (Rate of convergence under $N_k = 1$)**.** *Let Assumptions 16, 18, 20 hold, let $\{\gamma_k\}$ be given by* (3.28) *and $N_k = 1$ for every k. Then any sequence generated by (SMFBS) converges to a solution x^* in an expected value sense:*

$$\mathbb{E}[\|x_k-x^*\|^2] = \mathcal{O}\left(\frac{1}{k}\right).$$

Proof. According to Lemma 17, and taking expectations conditioned on $\mathcal{F}_k$, we obtain

$$
\begin{aligned}
\mathbb{E}[\|x_{k+1}-x^*\|^2 \mid \mathcal{F}_k] &\le (1-\sigma\gamma_k)\|x_k-x^*\|^2 \\
&- (1-2\gamma_k^2L^2-2\sigma\gamma_k)\mathbb{E}\left[\left\|x_k-x_{k+\frac{1}{2}}\right\|^2 \mid \mathcal{F}_k\right] \\
&+ 4\gamma_k^2\mathbb{E}\left[\mathbb{E}\left[\left\|w_{k+\frac{1}{2}}\right\|^2 \mid \mathcal{F}_{k+\frac{1}{2}}\right] \mid \mathcal{F}_k\right] + 4\gamma_k^2\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] \\
&- 2\gamma_k\mathbb{E}\left[\mathbb{E}\left[w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}}-x^*\right) \mid \mathcal{F}_{k+\frac{1}{2}}\right] \mid \mathcal{F}_k\right] \\
&\le (1-\sigma\gamma_k)\|x_k-x^*\|^2 + (2\gamma_k^2L^2+\gamma_k^2\sigma^2)\mathbb{E}\left[\left\|x_k-x_{k+\frac{1}{2}}\right\|^2 \mathcal{F}_k\right] + 8\gamma_k^2\nu^2,
\end{aligned}
\tag{3.30}
$$

where the second inequality leverages $1-2\gamma_k\sigma \ge -\gamma_k^2\sigma^2$. Invoking Lemma 15, we have $\left\|x_k - x_{k+\frac{1}{2}}\right\| \leq C_1$. Then taking expectations on both sides of (3.30), we get

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq (1 - \sigma\gamma_k)\mathbb{E}[\|x_k - x^*\|^2] + (2L^2 + \sigma^2)\gamma_k^2 C_1^2 + 8\gamma_k^2\nu^2. \tag{3.31}$$

By assuming $1 < \sigma\gamma_0$ and invoking Lemma 16, we get

$$\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{M_1}{k},$$

where $M_1 \triangleq \max\left(\frac{\gamma_0^2((2L^2+\sigma^2)C_1^2+8\nu^2)}{\sigma\gamma_0-1}, \mathbb{E}[\|x_0 - x^*\|^2]\right)$. This means $\{x_k\}$ converges to x^* in an expected value sense. □

Next we provide rate statements involving (SMFBS) with increasing sample-size. We are now to provide a proposition that deriving rate statements and oracle complexity bounds.

Proposition 17 (Rate and oracle complexity under $N_k = \lfloor \rho^{-(k+1)} \rfloor$)**.** *Let Assumptions 16, 18, and 20 hold. Let $\gamma_k = \gamma \leq \min\left\{\frac{1}{\sigma}, \frac{-\sigma+\sqrt{\sigma^2+2L^2}}{2L^2}\right\}$. Suppose $N_k = \lfloor \rho^{-(k+1)} \rfloor$, $C_2 \triangleq \|x_0\|+C$, $D \triangleq 16\gamma^2\nu^2$, $q \triangleq 1-\sigma\gamma$, $\tilde{\rho} \in (q, 1)$, $\hat{D} > 1/\ln(\tilde{\rho}/\rho)^e$ and $\tilde{D} \triangleq \left(C_2^2 + D\frac{1}{1-\min\{(q/\rho),(\rho/q)\}}\right)$. Then the following hold.*
(a) Any sequence generated by (SMFBS) converges at a linear rate to a solution x^* in an expected value sense.
(b). Let x_{K+1} be such that $\mathbb{E}[\|x_{K+1} - x^*\|^2] \leq \epsilon$. Then we have $\sum_{k=1}^{K} N_k \leq \mathcal{O}\left(\frac{1}{\epsilon}\right)$.

Proof. (a). Invoking Lemma 17 and since $\gamma \leq \frac{-\sigma+\sqrt{\sigma^2+2L^2}}{2L^2}$, we obtain

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq (1 - \sigma\gamma)\|x_k - x^*\|^2 - (1 - 2\gamma^2L^2 - 2\sigma\gamma)\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 \\
&\quad + 4\gamma^2\left\|w_{k+\frac{1}{2}}\right\|^2 + 4\gamma^2\|w_k\|^2 - 2\gamma w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}} - x^*\right) \\
&\leq (1 - \sigma\gamma)\|x_k - x^*\|^2 + 4\gamma^2\left\|w_{k+\frac{1}{2}}\right\|^2 + 4\gamma^2\|w_k\|^2 \\
&\quad - 2\gamma w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}} - x^*\right).
\end{aligned}$$

Taking expectations conditioned on $\mathcal{F}_k$, we obtain

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] &\leq (1 - \sigma\gamma)\|x_k - x^*\|^2 + 4\gamma^2\mathbb{E}\left[\mathbb{E}\left[\left\|w_{k+\frac{1}{2}}\right\|^2 \mid \mathcal{F}_{k+\frac{1}{2}}\right] \mid \mathcal{F}_k\right] \\
&\quad + 4\gamma^2\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] - 2\gamma\mathbb{E}\left[\mathbb{E}\left[w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}} - x^*\right) \mid \mathcal{F}_{k+\frac{1}{2}}\right] \mid \mathcal{F}_k\right] \\
&\leq (1 - \sigma\gamma)\|x_k - x^*\|^2 + \frac{8\gamma^2\nu^2}{N_k}.
\end{aligned} \tag{3.32}$$

By taking conditions on both sides of (3.32), we obtain

$$
\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^*\|^2] &\leq (1-\sigma\gamma)\mathbb{E}[\|x_k - x^*\|^2] + \frac{8\gamma^2\nu^2}{N_k} \\
&= q\mathbb{E}[\|x_k - x^*\|^2] + \frac{8\gamma^2\nu^2}{N_k}.
\end{aligned} \tag{3.33}
$$

According to Lemma 11, we have

$$
N_k = \lfloor \rho^{-(k+1)} \rfloor \geq \left\lceil \frac{1}{2}\rho^{-(k+1)} \right\rceil \geq \frac{1}{2}\rho^{-(k+1)}. \tag{3.34}
$$

We now consider three cases.
Case (i). If $q < \rho < 1$: Using (3.34) in (3.33), we deduce that

$$
\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq q\mathbb{E}[\|x_k - x^*\|^2] + \frac{8\gamma^2\nu^2}{N_k} &= q\mathbb{E}[\|x_k - x^*\|^2] + D\rho^{k+1} \\
&\leq q^{k+1}\mathbb{E}[\|x_0 - x^*\|^2] + D\sum_{j=1}^{k+1} q^{k+1-j}\rho^j \\
&\leq q^{k+1}C_2^2 + D\rho^{k+1}\sum_{j=1}^{k+1}(q/\rho)^{k+1-j} \leq \tilde{D}\rho^{k+1}.
\end{aligned}
$$

Case (ii). $\rho < q < 1$. If $\rho < q < 1$: Employing a similar approach, we obtain

$$
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \tilde{D}q^{k+1}.
$$

Case (iii). $\rho = q < 1$. If $\rho = q$: We have that

$$
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq q^{k+1}\mathbb{E}[\|x_0 - x^*\|^2] + Dq^{k+1} \leq q^{k+1}C_2^2 + D\sum_{j=1}^{k+1} q^{j+1}. \tag{3.35}
$$

According to [142, Lemma 4], we can deduce from (3.35) that

$$
\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^*\|^2] &\leq q^{k+1}C_2^2 + D\sum_{j=1}^{k+1} q^{j+1} \leq C_2^2 q^{k+1} + Dq^{k+1}(k+1) \\
&\leq C_2^2 q^{k+1} + \hat{D}\tilde{\rho}^{k+1} \leq (C_2^2 + \hat{D})\tilde{\rho}^{k+1}.
\end{aligned}
$$

Therefore, we conclude that any sequence generated by (SMFBS) converges linearly in an expected sense.
(b). Again, we consider three cases.

Case (i) $q < \rho < 1$. If $q < \rho < 1$: From part 17(a), it follows that

$$\mathbb{E}[\|x_{K+1} - x^*\|^2] \leq \tilde{D}\rho^{K+1} \leq \epsilon \Longrightarrow K = \lceil \log_{1/\rho}(\tilde{D}/\epsilon) \rceil - 1.$$

For the optimal oracle complexity, we require $\sum_{k=1}^{K} N_k$ gradients. Since $N_k = \lfloor \rho^{-(k+1)} \rfloor \leq \rho^{-(k+1)}$, we have

$$\sum_{k=1}^{\log_{1/\rho}(\tilde{D}/\epsilon)} \rho^{-(k+1)} \leq \frac{1}{\left(\frac{1}{\rho} - 1\right)} \left(\frac{1}{\rho}\right)^{2+\log_{1/\rho}(\tilde{D}/\epsilon)} \leq \left(\frac{\tilde{D}}{\epsilon}\right) \frac{1}{\rho(1-\rho)}.$$

Case (ii) $\rho < q < 1$. If $\rho < q < 1$: With a similar manner, we have

$$\sum_{k=1}^{K} N_k \leq \left(\frac{\tilde{D}}{\epsilon}\right) \frac{1}{q(1-q)}.$$

Case (iii) $\rho = q < 1$. If $q = \rho$: From Proposition 17(a), we have that $\mathbb{E}[\|x_{K+1} - x^*\|^2] \leq \tilde{C}\tilde{\rho}^{K+1}$, where $\tilde{C} = C_2^2 + \hat{D}$. It follows that

$$\sum_{k=1}^{K} N_k \leq \left(\frac{\tilde{C}}{\epsilon}\right) \frac{1}{\rho(1-\rho)}.$$

□

3.4.4 Rate under Maximal Monotonicity of A

To establish rate of convergence under maximal monotonicity, we need introduce a metric for ascertaining progress. In strongly monotone regimes, the mean-squared error serves this roole. However, when the solution set is multi-valued, this avenue is no longer available unless one can derive a bound on the distance to X^*, the solution set of the stochastic generalized equation. In stochastic convex optimization, the function value represents such a metric. Instead, the progress of the scheme can be ascertained by using a suitably defined gap function [143], which is inspired by a Fitzpatrick function [144, 145].

Definition 6 (**Gap function**). *Given a nonempty closed set $X \subseteq \mathbb{R}^n$ and a set-valued mapping $T : \mathbb{R}^n \to \mathbb{R}^n$, then the gap function at x is denoted by $G(x)$ and is* *defined as follows:*

$$G(x) \triangleq \sup_{y\in X} \sup_{z\in T(y)} z^T(x-y), \quad \forall x \in X. \quad \square$$

The gap function is nonnegative for all $x \in X$ and is zero if and only if x is a solution of the monotone inclusion problem. To derive the convergence rate under maximal monotonicity, we require boundedness of X as captured by the following assumption.

Assumption 21. *The set X is bounded as specified by $\|x-y\| \leq B \quad \forall x,y \in X$.*

We establish the convergence rate for (SMFBS) by using the gap function.

Lemma 18. *Let Assumptions 16, 18 hold and $0 < \gamma_k \leq 1/\sqrt{2}L$ for all k. Assume $\{x_k\}$ and $\{x_{k+\frac{1}{2}}\}$ be sequences generated by (SMBFS). It follows that the following inequality holds for all $y \in X$, $z \in T(y)$ and all $k \geq 0$:*

$$2\gamma_k \mathbb{E}\left[z^T\left(x_{k+\frac{1}{2}}-y\right)\right] \leq \mathbb{E}[\|x_k-y\|^2] - \mathbb{E}[\|x_{k+1}-y\|^2] + 8\gamma_k^2\frac{\nu^2}{N_k}. \tag{3.36}$$

Proof. According to (3.26),we have that

$$\begin{aligned}
2\gamma_k\left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}}-y\right) &\leq \|x_k-x^*\|^2 - \|x_{k+1}-x^*\|^2 \\
&- (1-2\gamma_k^2L^2)\left\|x_k - x_{k+\frac{1}{2}}\right\|^2 + 2\gamma_k^2\left\|w_{k+\frac{1}{2}}-w_k\right\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}}-x^*\right) \\
&\leq \|x_k-x^*\|^2 - \|x_{k+1}-x^*\|^2 + 4\gamma_k^2\left\|w_{k+\frac{1}{2}}\right\|^2 + 4\gamma_k^2\|w_k\|^2 \\
&- 2\gamma_k w_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}}-x^*\right).
\end{aligned}$$

Taking expectations conditioned on $\mathcal{F}_k$, we obtain

$$\begin{aligned}
2\gamma_k\mathbb{E}\left[\left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}}\right)^T\left(x_{k+\frac{1}{2}}-y\right) \mid \mathcal{F}_k\right] &\leq \|x_k-x^*\|^2 - \mathbb{E}[\|x_{k+1}-x^*\|^2 \mid \mathcal{F}_k] \\
&+ 4\gamma_k^2\mathbb{E}\left[\mathbb{E}\left[\left\|w_{k+\frac{1}{2}}\right\|^2 \mid \mathcal{F}_{k+\frac{1}{2}}\right] \mid \mathcal{F}_k\right] + 4\gamma_k^2\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] \\
&- 2\gamma\mathbb{E}\left[\mathbb{E}\left[\tilde{w}_{k+\frac{1}{2}}^T\left(x_{k+\frac{1}{2}}-x^*\right) \mid \mathcal{F}_{k+\frac{1}{2}}\right] \mid \mathcal{F}_k\right] \\
&\leq \|x_k-x^*\|^2 - \mathbb{E}[\|x_{k+1}-x^*\|^2 \mid \mathcal{F}_k] + 8\gamma_k^2\frac{\nu^2}{N_k}.
\end{aligned} \tag{3.37}$$

By taking conditions on both sides of (3.37), it deduce that

$$2\gamma_k \mathbb{E}\left[\left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - y\right)\right] \leq \mathbb{E}[\|x_k - x^*\|^2] - \mathbb{E}[\|x_{k+1} - x^*\|^2] + 8\gamma_k^2 \frac{\nu^2}{N_k}.$$

According to monotonicity property of T, we conclude that

$$\begin{aligned} 2\gamma_k \mathbb{E}\left[z^T \left(x_{k+\frac{1}{2}} - y\right)\right] &\leq 2\gamma_k \mathbb{E}\left[\left(u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}}\right)^T \left(x_{k+\frac{1}{2}} - y\right)\right] \\ &\leq \mathbb{E}[\|x_k - x^*\|^2] - \mathbb{E}[\|x_{k+1} - x^*\|^2] + 8\gamma_k^2 \frac{\nu^2}{N_k}. \end{aligned}$$

□

Using Lemma 18, we may derive the following rate statement for a diminishing steplength sequence $\{\gamma_k\}$. The result is provided for the sequence $\bar{x}_K$, an average of the iterates $\{x_{k+1/2}\}$ generated by (SMFBS) over the window constructed from K_l to K where $K_l \triangleq \lfloor K/2 \rfloor$ and $K \geq 2$:

$$\bar{x}_K \triangleq \frac{\sum_{k=K_l}^{K} \gamma_k x_{k+\frac{1}{2}}}{\sum_{k=K_l}^{K} \gamma_k}. \tag{3.38}$$

Proposition 18. *Consider the (SMFBS) scheme and let $\{\bar{x}_K\}$ be defined in (3.38). Let Assumptions 16, 18, 21 hold.*
(a). Let $0 < \gamma_k \leq 1/\sqrt{2}L$ for all $k \geq 0$ and $\gamma_k = \gamma_0/\sqrt{k}$. Then, we have

$$\mathbb{E}[G(\bar{x}_K)] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

(b). Let $0 < \gamma_k = \gamma \leq 1/\sqrt{2}L$ for all $k \geq 0$. Then, we have

$$\mathbb{E}[G(\bar{x}_K)] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

Consider the (SMFBS) scheme and let $\{\bar{x}_K\}$ be defined as (3.38), where $0 < \gamma_k = \gamma \leq 1/\sqrt{2}L$ for all $k \geq 0$. Let $N_k = K, \forall k$, where K is a pre-defined termination number.
(c1). Then, we have

$$\mathbb{E}[G(\bar{x}_K)] = \mathcal{O}\left(\frac{1}{K}\right).$$

(c2). The oracle complexity holds for any sequence generated by (SMFBS):

$$\sum_{k=1}^{K} N_k \leq \mathcal{O}\left(\frac{1}{\epsilon^2}\right).$$

Consider the (SMFBS) scheme and let $\{\bar{x}_K\}$ *be defined as* (A.7)*, where* $0 < \gamma_k = \gamma \leq 1/\sqrt{2}L$ *for all* $k \geq 0$*. Let* $\sum_{k=1}^{\infty} N_k < M$*.*
(d1). Then, we have

$$\mathbb{E}[G(\bar{x}_K)] = \mathcal{O}\left(\frac{1}{K}\right).$$

(d2). Suppose $N_k = \lfloor k^a \rfloor$*, for* $a > 1$*. Then the oracle complexity to compute an* $\bar{x}_K$ *such that* $\mathbb{E}[G(\bar{x}_K)] \leq \epsilon$ *is bounded as follows:*

$$\sum_{k=1}^{K} N_k \leq \mathcal{O}\left(\frac{1}{\epsilon^{a+1}}\right).$$

Proof. (a). From (3.36), by summing over k from K_l to K, we have the following for all $y \in X$, $z \in T(y)$:

$$2\sum_{k=K_l}^{K} \gamma_k \mathbb{E}\left[z^T\left(x_{k+\frac{1}{2}} - y\right)\right] \leq \mathbb{E}[\|x_{K_l} - y\|^2] - \mathbb{E}[\|x_{K+1} - y\|^2] + 8\sum_{k=K_l}^{K} \gamma_k^2 \nu^2.$$

Consequently, we have the following sequence of inequalities:

$$\begin{aligned} 2\left(\sum_{k=K_l}^{K} \gamma_k\right) \mathbb{E}[z^T(\bar{x}_K - y)] &\leq \mathbb{E}[\|x_{K_l} - y\|^2] - \mathbb{E}[\|x_{K+1} - y\|^2] + 8\sum_{k=K_l}^{K} \gamma_k^2 \nu^2 \\ &\leq \mathbb{E}[\|x_{K_l} - y\|^2] + 8\sum_{k=K_l}^{K} \gamma_k^2 \nu^2 \leq B^2 + 8\sum_{k=K_l}^{K} \gamma_k^2 \nu^2, \end{aligned} \tag{3.39}$$

where the second inequality follows from the boundedness of X. Since $\gamma_k = \gamma_0/\sqrt{k}$, it follows that for all $y \in X$, $z \in T(y)$:

$$\mathbb{E}[z^T(\bar{x}_K - y)] \leq \frac{B^2 + 8\sum_{k=K_l}^{K} \gamma_k^2 \nu^2}{2\sum_{k=K_l}^{K} \gamma_k} = \frac{B^2}{2\gamma_0} \frac{1}{\sum_{k=K_l}^{K} k^{-\frac{1}{2}}} + 4\gamma_0 \nu^2 \frac{\sum_{k=K_l}^{K} k^{-1}}{\sum_{k=K_l}^{K} k^{-\frac{1}{2}}}. \tag{3.40}$$

We now utilize the following lower bound on the denominator for $K \geq 1$:

$$\sum_{k=K_l}^{K} k^{-\frac{1}{2}} \geq \int_{\frac{K}{2}}^{K} (x+1)^{-\frac{1}{2}} dx = 2\sqrt{(K+1)} - 2\sqrt{K/2+1)} \geq 2\sqrt{K/40}. \quad (3.41)$$

Similarly an upper bound may be constructed:

$$\sum_{k=K_l}^{K} k^{-1} \leq \int_{\frac{K}{2}}^{K} x^{-1} dx + \frac{1}{\lfloor \frac{K}{2} \rfloor} \leq \log 2 + 1. \quad (3.42)$$

By substituting (3.41) and (3.42) in (3.40), we obtain that the following holds:

$$\mathbb{E}[z^T(\bar{x}_K - y)] \leq \frac{S_1}{\sqrt{K}} \text{ for all } y \in X, z \in T(y)$$
$$\text{where } S_1 \triangleq \left(\frac{\sqrt{40}B^2}{4\gamma_0} + 2\sqrt{40}(\log 2 + 1)\gamma_0 \nu^2 \right).$$

The result follows by taking supremum over $y \in X$, $z \in T(y)$.

(b). Proceeding similarly as in the prior proof, an analogous inequality to (3.39) can be derived for all $y \in X$, $z \in T(y)$:

$$2\left(\sum_{k=K_l}^{K} \gamma_k\right) \mathbb{E}[z^T(\bar{x}_K - y)] \leq B^2 + 8 \sum_{k=K_l}^{K} \gamma_k^2 \nu^2. \quad (3.43)$$

Since $\gamma_k \equiv \gamma$, we can rewrite (3.43) as follows for all $y \in X$:

$$2\left\lceil \frac{K+2}{2} \right\rceil \gamma \mathbb{E}[z^T(\bar{x}_K - y)] \leq B^2 + 8\left\lceil \frac{K+2}{2} \right\rceil \gamma^2 \nu^2,$$

leading to the following inequality for all $y \in X$, $z \in T(y)$:

$$\begin{aligned}\mathbb{E}[z^T(\bar{x}_K - y)] &\leq \frac{B^2}{2\left\lceil \frac{K+2}{2} \right\rceil \gamma} + \frac{8\left\lceil \frac{K+2}{2} \right\rceil \gamma^2 \nu^2}{2\left\lceil \frac{K+2}{2} \right\rceil \gamma} \\ &= \frac{B^2}{2\left\lceil \frac{K+2}{2} \right\rceil \gamma} + 4\gamma\nu^2 \leq \frac{B^2}{K\gamma} + 4\gamma\nu^2.\end{aligned}$$

By optimizing in γ, we have that $\gamma^* = C/2\sqrt{K}\nu$, we may deduce that

$$\mathbb{E}[z^T(\bar{x}_K - y)] \leq \frac{2C\nu}{\sqrt{K}}, \forall y \in X.$$

The result follows.

(c1). Similar to (3.43), we have

$$2\left(\sum_{k=K_l}^{K}\gamma_k\right)\mathbb{E}[z^T(\bar{x}_K-y)]\le B^2+8\sum_{k=K_l}^{K}\frac{\gamma_k^2\nu^2}{K}. \tag{3.44}$$

Since $\gamma_k\equiv\gamma$, we rewrite (3.44) as follows for all $y\in X$:

$$2\left\lceil\frac{K+2}{2}\right\rceil\gamma\mathbb{E}[z^T(\bar{x}_K-y)]\le B^2+8\left\lceil\frac{K+2}{2}\right\rceil\frac{\gamma^2\nu^2}{K},$$

leading to the following inequality for all $y\in X$, $z\in T(y)$:

$$\begin{aligned}\mathbb{E}[z^T(\bar{x}_K-y)]&\le\frac{B^2}{2\left\lceil\frac{K+2}{2}\right\rceil\gamma}+\frac{8\left\lceil\frac{K+2}{2}\right\rceil\gamma^2\nu^2}{2\left\lceil\frac{K+2}{2}\right\rceil K\gamma}\\&=\frac{B^2}{2\left\lceil\frac{K+2}{2}\right\rceil\gamma}+\frac{4\gamma\nu^2}{K}\le\frac{B^2}{K\gamma}+\frac{4\gamma\nu^2}{K}.\end{aligned} \tag{3.45}$$

By optimizing in γ, we have that $\gamma^*=B/2\nu$, we may deduce that

$$\mathbb{E}[z^T(\bar{x}_K-y)]\le\frac{2B\nu}{K},\forall y\in X.$$

The result follows.

(c2). Following (c1) and $\mathbb{E}[\|x_{K+1}-x^*\|^2]\le\epsilon$, it follows that

$$\mathbb{E}[z^T(\bar{x}_{K+1}-x^*)]\le\frac{2B\nu}{K+1}\Longrightarrow K=\lceil 2B\nu/\epsilon\rceil-1\le\frac{2B\nu}{\epsilon}.$$

Thus,

$$\sum_{k=1}^{K}N_k=KK\le\frac{4B^2\nu^2}{\epsilon^2}.$$

(d1). Similarly to (3.44), we have

$$2\left(\sum_{k=K_l}^{K}\gamma_k\right)\mathbb{E}[z^T(\bar{x}_K-y)]\le B^2+8\sum_{k=K_l}^{K}\frac{\gamma_k^2\nu^2}{N_k}.$$

Similarly to (3.45), it leads to

$$\mathbb{E}[z^T(\bar{x}_K-y)]\le\frac{B^2}{2\left\lceil\frac{K+2}{2}\right\rceil\gamma}+\frac{8\left(\sum_{k=K_l}^{K}\frac{1}{N_k}\right)\gamma^2\nu^2}{2\left\lceil\frac{K+2}{2}\right\rceil\gamma}$$

$$\leq \frac{B^2}{K\gamma} + \frac{8M\gamma\nu^2}{K}.$$

By optimizing in γ, we have that $\gamma^* = B/2\sqrt{2M}\nu$, we may deduce that

$$\mathbb{E}[z^T(\bar{x}_K - y)] \leq \frac{2\sqrt{2M}B\nu}{K}, \forall y \in X.$$

The result follows.

(d2). it follows that for ϵ sufficiently small,

$$\sum_{k=K_l}^{K} N_k \leq \sum_{k=1}^{\lceil(\widehat{C}/\epsilon)\rceil} N_k \leq \sum_{k=1}^{\lceil(\widehat{C}/\epsilon)\rceil} k^a \leq \int_{k=1}^{(\widehat{C}/\epsilon)+1} x^a dx \leq \frac{((\widehat{C}/\epsilon)+1)^{a+1}}{a+1} \leq \left(\frac{\widehat{C}}{\epsilon^{a+1}}\right).$$

□

3.5 Numerical Results

In this section, we apply the proximal point scheme to a generalized equation problem (Section 3.5.1) and the splitting schmes on a stochastic Nash-Cournot game (Section 3.5.2).

3.5.1 A Generalized Equation Problem

We test the performance of (SPP) on an example from [134] which requires finding the root of a stochastic mapping. The mapping T is defined as $T = \mathbb{E}[T(x,\xi)]$ where $T(x,\xi) = \left(\frac{x_1^3}{1+x_1^2} + \xi_1, \frac{x_2^3}{1+\|x_2\|^3} + \xi_2\right)^T$. We define ξ_1 and ξ_2 are uniformly distributed random variables sampled from $[-1,1]$ and it's clear that $(0,0)$ is a root of T. We assume a constant step size of (SPP) with $\gamma = 1$. The result is shown in 3.1.

3.5.2 A Stochastic Nash-Cournot Game

In this section, we present and compare the computational results of applying the splitting scheme aforementioned to a stochastic Nash-Cournot game. This game is assumed that $\mathcal{I}$ firms compete over a network of $\mathcal{J}$ nodes. Level of production and sales of firm $i \in \mathcal{I}$ at node $j \in \mathcal{J}$ are denoted by p_{ij} and s_{ij}, respectively. Furthermore, we assume the cost of production at node j is $C_{ij}(p_{ij})$ and the price

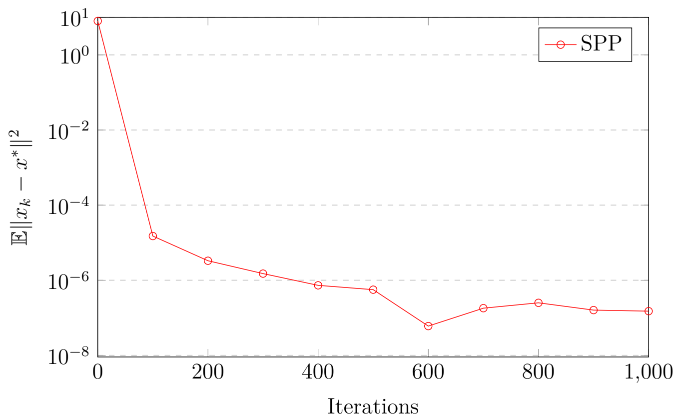


Figure 3.1: (SPP) convergence based on iterations

at node j is denoted by $Q_j(\bar{s}_j, \xi)$, where $\bar{s}_j$ is the aggregate sales at node j. For simplicity, we assume the transportation costs are zero. Thus, each firm i will solve a profit maximization problem given by the following:

$$\begin{aligned} \max \quad & \mathbb{E}[f_i(x,\xi)] = \mathbb{E}[\sum_{j\in\mathcal{J}}(Q_j(\bar{s}_j,\xi)s_{ij} - C_{ij}(p_{ij}))] \\ \text{subject to} \quad & \sum_{j\in\mathcal{J}} p_{ij} = \sum_{j\in\mathcal{J}} s_{ij}, \quad p_{ij} \leq \mathrm{cap}_{ij}, \quad s_{ij}, q_{ij} \geq 0, \quad \forall j \in \mathcal{J}. \end{aligned}$$

The equilibrium conditions of this problem can be captured by a variational inequality VI(X,F), where $F = (F_1(x); ...; F_{\mathcal{I}}(x))$ with $F_i(x) = \mathbb{E}[\nabla_{x_i} f_i(x,\xi)]$. Recall that the variational inequality problem is finding an $x \in X$ satisfying $0 \in F(x) + \mathcal{N}_X(x)$ where X is a nonempty closed convex set in $\mathbb{R}^n$ and F is a single-valued monotone mapping that is continuous on X. Thus (SMFBS) is converted to the following scheme:

$$\begin{aligned} x_{k+\frac{1}{2}} &:= \Pi_X(x_k - \gamma_k F(x_k,\omega_k)), \\ x_{k+1} &:= x_{k+\frac{1}{2}} - \gamma_k \left(F\left(x_{k+\frac{1}{2}}, \omega_{k+\frac{1}{2}}\right) - F(x_k,\omega_k)\right). \end{aligned}$$

In our original setting, we assume there are $\mathcal{I} = 5$ firms and $\mathcal{J} = 4$ nodes, and the capacity $\mathrm{cap}_{ij} = 300$, $\forall i,j$. $C_{ij}(p_{ij})) \triangleq c_{ij}p_{ij} + d_{ij}$, where $c_{ij} = 1.5$ and d_{ij} is a constant, $\forall i,j$. $Q_j(\bar{s}_j,\xi) \triangleq a_j - b_j\bar{s}_j$, where $b_j = 0.05$ and a_j is a uniformly distributed random variable sampled from $[49.5, 50.5]$, $\forall j$. With the above parameters, it can be shown that the mapping F is strictly monotone. We assume suitable step sizes in our experiments and leverage gap function as our metric. We compare the performance of (SMFBS) with the stochastic extragradient

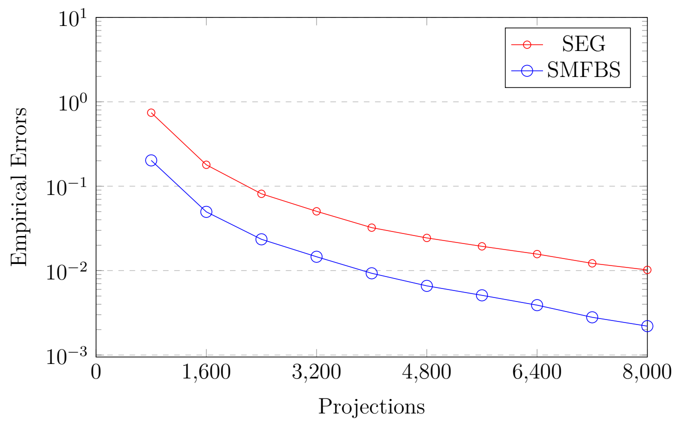


Figure 3.2: Convergence based on projections under maximal monotonicity

method (SEG) under the same number of projections (Fig. 3.2). The convergence difference here is because that (SEG) has two projections onto the set, while (SMFBS) just require one.

$$\begin{aligned} x_{k+\frac{1}{2}} &:= \Pi_X(x_k - \gamma_k F(x_k, \omega_k)), \\ x_{k+1} &:= \Pi_X\left(x_k - \gamma_k F\left(x_{k+\frac{1}{2}}, \omega_{k+\frac{1}{2}}\right)\right). \end{aligned}$$

To check the performance of variance reduction, we enlarge the random set for random variable a_j to $[40, 60]$. Fig. 3.3 shows comparison of (SMFBS) with (SMFBS) under the same number of iterations. The presence of variance reduction allows for significant improvement in the empirical rates from the single-sample counterparts.

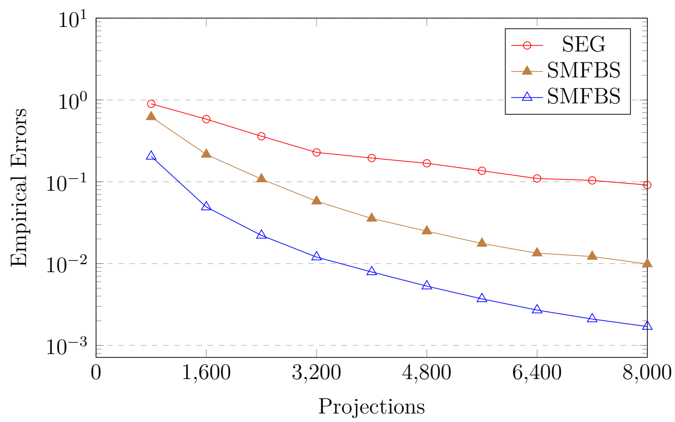


Figure 3.3: Performance comparison between variance reduction schemes and original ones

Chapter 4
Optimal proactive transmission expansion with transmission charging systems[1]

4.1 Introduction[2]

In an electric energy system, we rely on transmission networks to allow for power flow from generation nodes to demand centers. In a competitive environment the transmission system operator (TSO), an independent and regulated entity, is generally in charge of operating and expanding the transmission network with the aim of maximizing energy trade opportunities among producers and consumers. The TSO is also in charge of recovering long run network maintenance cost as well as transmission networks costs included. The use of system charges reflects the cost of installing, operating and maintaining the transmission system for the transmission owner activity. Charging mechanisms for recovering transmission network costs can significantly impact new generation investment. However, interactions between these charges and generation decisions are usually neglected in transmission planning. In this chapter we consider the proactive planning of transmission subject to the subsequent response of a generation market with imperfect transmission pricing. Prior work has emphasized short term pricing, either locational marginal pricing (LMP, which refers to the marginal cost of delivering an additional unit of energy to

[1]This chapter is joint work with Pengcheng Ding from Johns Hopkins University and has been jointly supervised by Prof. B. Hobbs from Johns Hopkins along with my advisor.

[2]This section has been jointly developed by Pengcheng and Shisheng

a node in the network), or zonal pricing systems. Here, we consider an equilibrium modeling framework with inefficient long-term pricing and the incentives provided for generation mix and siting, a framework that has not received consideration thus far. We investigate how generation decisions may be affected by alternative network cost recovery methods and how a proactive transmission planner can take those reactions into account while deciding how to plan new transmission. We model two distinct types of potentially inefficient cost recovery schemes. The first is a postage-stamp type charge, while the second is a marginal MW-miles based charge akin to that used in the UK. We develop a hierachical framework in which the system operator determines transmission investments (binary) at the upper level subject to the resulting transmission charging and the resulting market equilibrium, both of which are defined at at the lower level and are contingent on transmission level decisions. The resulting problem is a mathematical program with equilibrium constraints (MPEC) and binary variables at the first-stage, a relatively challenging class of nonconvex programs. We conduct a case study of a 5-node nodal network by resolving the mixed-binary (MPEC) via a branching scheme combined with a smoothing-based approach to further improve the efficiency of the scheme. We find that the efficiency impacts of the different cost recovery schemes (relative to the LMP ideal) vary significantly depending on the network and the exact structure of the charge.

4.1.1 Motivation and Approach[3]

Market operators across the world have designed different charging schemes [146,147] to recover both the investment and the operation and maintenance costs of the transmission grid. In addition to full cost recovery, a design goal for some charges is to approximate the short-run and long-run marginal costs of transmission assets investment, operation, and maintenance in order to incentivize efficient generation dispatch as well efficient location and mix of new supply development. The underlying philosophy is that by incorporating better representations of system marginal costs as part of the incentives to the market participants, higher economic efficiency may be achieved [148].

In Schweppe et al.'s vision [149], LMP, the most popular short-run congestion

[3]This subsection has been contributed by Pengcheng and is presented here for purposes of completeness.

management method may result in optimal short-run operations; in addition, under some assumptions, LMP also provides optimal long-run incentives for generation and transmission investment. However, the total congestion revenue resulting form LMP may only recover 20% of the total costs for the transmission system [150] and consequently system operators have to supplement LMP with additional transmission charges to recover the entirety of the costs from either the consumers and/or the generator. While LMP represents an efficient short-run incentive for dispatch, these additional charges are expected to incorporate efficient long-run incentives for investment. These may be volumetric (per MWh) or based on installed capacity and may or may not be locationally differentiated. We investigate the distortions in short and long-run decisions that could result from alternative designs of such charge considering transmission network expansions, which will heavily affect the generation expansion decisions [151]. We examine the interactions between transmission planning, generation expansion, and transmission costs recovery. These distortions can be gauged by the decrease in total social welfare, which is the sum of generation profit, consumer surplus, and transmission surplus.

In general, transmission assets require much longer lead time [86] than generation assets. It has been shown that, in theory, proactive transmission planning that anticipates how network expansions affect generation investment and siting decisions yield a higher total market efficiency, defined as the sum of all economic surpluses (welfare) of all market participants. We model transmission cost recovery charges in a proactive transmission planning framework to find how a proactive transmission plan might differ depending on the choice of transmission charging schemes. We model two potentially inefficient cost recovery schemes: (i) a postage-stamp type charge; and (ii) a marginal MW-miles based charge similar to that used by UK's Office of Gas and Electricity Markets [152]. We formulate the relationship between a proactive grid planner and the generation market [86] through a bilevel framework in which at the upper level, the system operator determines transmission investments; at the lower level, transmission charging schemes are incorporated into the lower-level generation expansion problem and the market clearing equilibrium problem.

4.1.2 Literature Review[4]

Next, we first discuss several types of transmission pricing systems and provide some detail on the Marginal MW-miles approach that is the focus of this chapter. We then review some related work that compares the economic efficiency of different pricing systems and their impact on transmission planning. We then discuss the methods used to solve a grid planning problem subject to anticipating the response of generators to prices, which is in general a non-convex problem. Finally, we review several formulations of capacity expansion problems and their solution schemes.

4.1.2.1 Transmission Pricing Systems[5]

As mentioned, LMPs are generally inadequate in terms of recovering total transmission related costs, requiring system operators to supplement the LMP with other transmission charges to recover all the costs and collect adequate revenue. In addition to revenue adequacy, a transmission charge is expected to promote efficient operation and investment in the power market while being fair and practically implementable [153, 154]. However, these goals are often in conflict with each other and different transmission charging schemes emphasize distinct goals.

We may categorize transmission charges in a variety of ways: they could be based on power flows or not; they could be based on average total costs or incremental costs; they could be location-dependent or independent. We list some popular charging methods here. The *postage-stamp* type of charge is the most widely used non-power-flow-based method and allocates costs without considering the location of the network users. Users may be charged by their energy injection or peak power. *Postage-stamp*, though simple to implement, does not distinguish between users based on how their decisions affect short- or long-run transmission costs [155] and may not encourage the most efficient use of the grid. The *contract path* method specifies a path between different nodes and the costs associated with the transmission facilities on the paths are allocated to the implied users of those paths. This method ignores how a system operates in reality and may send incorrect economic signals [153]. There is also the power flow-based *MW-miles* charging method, which considers both the flows on the lines and the length of

[4] This subsection has been jointly developed by Pengcheng and Shisheng

[5] This part is contributed by Pengcheng

the lines. It attributes flows back to the users using either distribution factors, power flow comparison methods or power flow tracing methods [156]. This kind of MW-miles method considers each user's unique contribution to network costs based on better approximations of the power flow than the contract path method and assign costs accordingly.

More recently, in [155], the authors proposed a MW-miles based charge incorporating reliability concerns by using the maximum flows under all N-1 contingencies. All the methods mentioned assign average rolled-in total costs [156]; in contrast, some researchers have argued that charges based on incremental costs may lead to higher efficiency [157].

Notable incremental cost methods include the *Investment Cost-Related Pricing* (ICRP) [152] model, the *Long-Run Incremental Cost Pricing* (LRIC) [158] model, and models based on LMP [159, 160]. The first two models approximate long-run marginal costs of the grid while the last model modifies the short-run LMP signals to encourage more efficient investment. ICRP derives the charge by evaluating the marginal MW-miles in changing incremental injection at a node while LRIC assumes that line investment would be made when the flow on the line reaches its thermal capacity and uses the change in the present value of the lines due to incremental injections in constituting a charge. A more recent model is proposed in [161] where the charge is based on the marginal value of the system's capital costs including that of the transmission lines and generators when an incremental generation investment is made. We choose to study a marginal MW-miles method based on *Investment Cost-Related Pricing* (ICRP) from OFGEM [152] given that it can be easily incorporated in an equilibrium framework.

4.1.2.2 Economic Efficiency of Alternative Systems[6]

Some prior work considers the influence of transmission charges on transmission planning [161–166]. In both [162] and [163], the authors conclude that LMP-based models are better in achieving market efficiency than non-LMP based systems. The methods considered include the postage stamp charge and a per-power-flow payment charge. Tohidi et al. [161] observe that their proposed transmission charge may distort efficient coordination between generation and transmission expansion while affecting market efficiency. Effects of ICRP charges on transmission planning has

[6]This part is contributed by Pengcheng

also been studied by Strbac et al. [164] where it has been observed that significant benefit is transferred from consumers to generators under inefficient transmission charges. ICRP is further compared with the postage-stamp charging model and LRIC in [166] where the authors build an iterative model that models different players in the market and find that the LRIC model provides the best efficiency amongst the three. However, there appears to be no convergence criterion in the scheme and it remains unclear if the resulting solution is an equilibrium.

However, the models used in the majority of the works mentioned above are single-level models [161–163] where a single entity plans everything including both transmission and generation. These models cannot represent how a proactive planner would anticipate the investment and operational decisions in response to the specific transmission charging method used. There are also models that iterate between between transmission, generation and transmission charging models [164, 166]. Solutions to the latter model can be viewed as equilibria between transmission planners who attempt to maximize net market surplus and generation investors who behave competitively facing a transmission charge, although the resulting equilibria may be worse from a welfare standpoint than the solution generated from proactive planning model.

4.1.2.3 Proactive Models[7]

Distinct from single-level models like the ones mentioned previously, in the proactive or anticipative transmission planning paradigm, decisions are made sequentially. At the upper level, the network owner or ISO makes decisions concerning which lines to build and which existing lines to upgrade, anticipating the response from the generation market. Then, given how transmission is priced in both the short-run (e.g., LMP) and long-run (e.g., transmission cost recovery charges), generators make investment and dispatch decisions in a market framework. Furthermore, a multi-level framework could model players that make decisions sequentially and may have objectives that do not align with maximizing total surplus. The latter ability is particularly useful to model market imperfections [69, 70, 86, 90, 167], such as generator market power, inefficient zonal transmission prices, extra capacity payment rules, or if reliability goals are embedded in the planning problem.

[7]This part is contributed by Pengcheng

In [165] and [168], the authors use multi-level models to consider the transmission planning problem with cost recovery. Wang et al. [165] employ a tri-level model to consider transmission planning and distributed energy resource capacity expansion with a transmission charge based on a power tracing method. However, this scheme has no convergence theory nor does the transmission charge guarantee full recovery of the transmission costs. Weibelzahl et al. [168] build a bilevel model to consider transmission planning and storage investment planning. Here, transmission costs will be fully recovered, but only one kind of transmission charge is considered and the charge is assumed to be invariant among different lines. Neither of these works emphasize the efficiency impacts of choice of transmission charging model. Here we extend the paradigm to the case of inefficient long-run transmission pricing, with the goal of investigating whether network planning can counter the distorted siting incentives arising from MW-mile or postage-stamp based pricing.

However, these multilevel models are known to be nonconvex and NP-hard in general [169] and various methods, including reformulation, penalization and regularization, have been developed to cope with them. A mixed integer linear programming model was developed in [89] which consisted of three levels, transmission investment, generation investment and market operation. It utilized disjunctive constraints to recast equilibrium constraints as mixed-integer linear constraints. Depending on the scale of the problem, it is also possible to apply a nonlinear programming to directly resolve the problem. However, most nonlinear programming solvers tend to have less predictable behavior in the resolution of large-scale (MPEC)s. Generally, mixed-integer linear/convex/nonconvex problems can in principle be solved via global optimization solvers such as `baron` and `couenne`. However, this is generally possible for small problem instances. If the problem is a mixed-integer linear/quadratic program, then `cplex` and `gurobi` can also be employed with the relaxations are convex. In this chapter, since the scale of our problem is relatively small, we attempt to compute near-global solutions by utilizing a branching scheme and solving via `baron`. By utilizing smoothing techniques, we garner further efficiencies.

4.1.2.4 Capacity Expansion Problems[8]

- **Deterministic generation capacity expansion problem:** Suppose the

[8] This part is contributed by Shisheng

generation capacity of firm j is denoted by x_j, the cost of each firm j is f_j, the discounted operating cost during time period t is $o_t(x)$ and the constraint on capacity of firm j is c_j. Then the formulation of this problem is

$$\begin{aligned} \min \quad & f^T x + \sum_t o_t(x) \\ \text{subject to} \quad & c_j \geq x_j \geq 0, \quad \forall j. \end{aligned}$$

In this formulation, the solution vector x^* describes the entire optimal plan of capacities over all time periods.

- **Generation capacity expansion problem under uncertainty:** Borison, Morris and Oren [170] suggested a static probabilistic model, where technologies are purchased and operated under uncertain conditions. It is similar to the deterministic problem can be converted to a single linear/nonlinear problem which is analogous to the deterministic case. Eager, Hobbs and Bialek [171] presented a dynamic simulation model of a certain generation investment market. A dynamic investment model was presented in which the Mix of Normals distribution (MOND) technique was embedded. The generation companies used a Value at Risk (VaR) criterion for investment decisions. A stochastic multiscale model [172] for electricity generation capacity expansion has also been considered where an efficient method for coupling multiple temporal scales in the capacity expansion problem was utilized. The solution technique relied on a finite difference approximation of the Hamilton-Jacobi-Bellman equation.

- **Generation capacity expansion problem in a competitive environment:** Let the capacity variables be denoted by x and generation decisions be denoted by y. The parameters include the price of power seen by generator p, generator's investment K and operational cost v. The ith generator i solves the following optimization problem

$$\begin{aligned} \min \quad & -(p(X) - v_i)y_i + K_i x_i \\ \text{subject to} \quad & x_i - y_i \geq 0 \\ & y_i \geq 0. \end{aligned}$$

Ehrenmann and Smeers [173] proposed a model inherited from the generation capacity expansion developed for monopolistic regimes, (adapted for a competitive environment with risk-aversion). de Frutos and Fabra [174] analyzed role of demand uncertainty in markets of fixed size, in which firms take long-run capacity decisions prior to competing in prices. In the first stage, firms simultaneously choose their capacities. Capacity decisions are assumed to be irreversible and become publicly known. In the second stage, firms simultaneously choose prices. It was analyzed as a two-stage Nash game. Three models of investments in generation capacity expansion in imperfectly competitive restructured electricity markets were considered by Murphy and Smeers [175]. These models differ in their underlying economic assumptions: the perfectly competitive model, the open-loop Cournot model and the closed-loop Cournot model. Existence and uniqueness of the solutions analyzed. A complementarity-based simulation model [176] was developed for providing strategies for risk-averse firms in competitive settings where the main uncertainty considered was the future regulation of carbon dioxide emissions.

- **Transmission capacity expansion problem:** Jin and Ryan [69] considered a tri-level model of generation and transmission expansion problem. Transmission expansion is considered at the upper level as a centralized decision while at the second level, multiple decentralized GENCOs make their own capacity expansion decisions. The third level represents the electricity market equilibrium problem. A hybrid iterative solution algorithm was proposed that combined a complementarity problem (CP) reformulation and diagonalization method (DM) solutions of the EPEC sub-problem. A similar tri-level model was considered in [67]. The solution method required approximating the line impedance values as a function of the installed transmission capacity. Bilevel models also employed for the transmission problem [68]. Transmission decisions were modeled via a two-stage, bilevel game-theoretic problem. Naturally, a two-stage framework may better cope with uncertainty than a one-time decision. The equilibrium of the lower level are imposed as constraints in the upper level and this framework may be modeled and analyzed as a mixed integer linear program if suitable bounds are available on the variables in the equilibrium problem. Alternate bilevel formulations were proposed for transmission and generation capacity expansion by [71] in which the upper level problem represents the investment problem faced by the transmission operator (TO) while the lower level captures market outcomes obtained from clearing the market. This can be reduced to a mixed-integer linear/nonlinear program. Motamedi, Zareipour, Buygi and Rosehart [88] proposed a transmission planning framework which assumed only the generation sector was deregulated and future generation expansion of GENCOs was taken into account. The four levels included transmission expansion planning, generation providers' optimal expansion problems, the market clearing problem, and the generation firms' optimal bidding strategy. The solution method was based on agent-based techniques and search-based optimization techniques which is efficient and minimal assumptions needed for market players but lacks rigorous convergence theory.

4.1.3 Contributions

1. Via an equilibrium modeling framework, we consider an inefficient long-term transmission pricing and the incentives provided for generation mix and siting. We incorporate this framework with a bilevel model for optimal transmission planning, anticipating the response to generation to the network reinforcements and inefficient financial incentives.

2. We propose a novel linear programming model to approximate the Marginal MW-miles transmission charging system. This model can be embedded in a transmission planning model to endogenously calculate the resulting transmission charge.

3. We solve the bilevel model as a Mixed Integer Mathematical Program with Equilibrium Constraints which we solve by combining a branching method with a smoothing method and allows for obtaining near global solutions.

4. We examine the economic efficiency impacts from different transmission charging schemes and illustrate their significance in the context of transmission planning. In particular, we observe that if generator investments are affected by transmission charging systems, transmission planning that ignore such interactions will result in plans with lower net market surplus.

5. Uncertainties are included in the model and we use a robust approach to address the uncertainty. Leveraging techniques from robust optimization, we incorporate robustness into the transmission expansion problems and provide some preliminary numerics.

4.2 Model Description[9]

In the proactive or anticipative transmission planning paradigm, decisions are made sequentially. At the upper level, the network owner or transmission system operator (TSO) makes decisions concerning where lines should be built and which existing lines should be upgraded, anticipating the response from the generation market. Then, given how transmission is priced in both the short (e.g., LMP) and long-run (e.g., interconnection charges), generators make investment and dispatch decisions in a market framework. Multi-level models can be used to implement this paradigm. Here we extend the paradigm to the case of inefficient long-run transmission pricing, with the goal of investigating whether network planning can counter the distorted siting incentives arising from MW-mile based pricing.

Each generation firm determines its dispatch and capacity investments given energy and transmission prices. Its objective lies in maximizing profit (revenue minus the expense of operations, capacity investments, and interconnection and transmission charges). Through this formulation, we consider investment incentives introduced by transmission charges.

Energy prices and LMPs arise from the short-run market clearing constraints, while long-run (annual) transmission charges are based on a MW-miles model that results from a separate optimization in the equilibrium model. The combination of the market clearing and the first-order conditions of generators and the long-run transmission pricing model constitutes an equilibrium model formulated as a complementarity problem.

We assume that the market is perfectly competitive; generators make investment decisions given the transmission charge but believe that they are incapable of either changing this charge or the energy price. Since the transmission planning decisions and the subsequent generation investments are made sequentially, the transmission planner (leader) is anticipating generator responses in a proactive manner. We

[9]This section is contributed by Pengcheng

employ this model to investigate the influence of transmission charging schemes on proactive transmission planning decisions. The structure of our framework is described in Fig. 4.1.

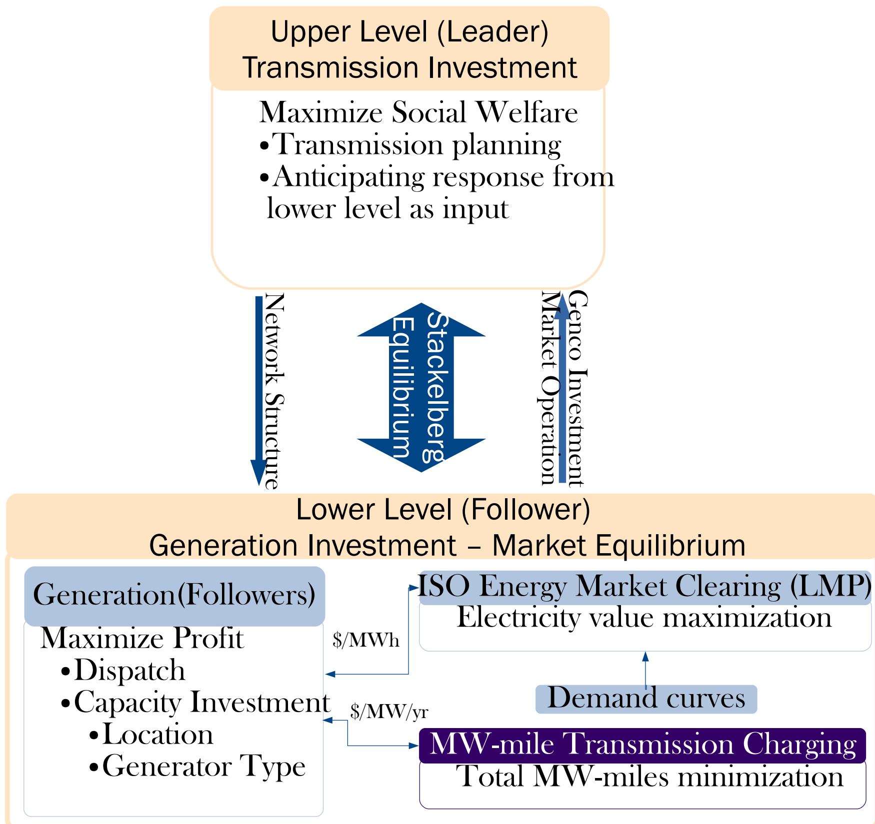


Figure 4.1: Model structure

4.2.1 Marginal MW-miles Transmission Charging Model

The transmission charge may take the form of one-time entry fee which could either be based on peak capacity during the year or the volume of energy sold. This charge could also be set to be location-dependent. In this subsection, we first describe the marginal MW-miles transmission charging model, which allocates transmission costs to network users in a location-dependent way. We then show when some of the parameters of the model are set to zero, the model reduces to a location-independent model.

Our marginal MW-miles based charging system is inspired by UK's Transmission Network Use of System Charges [152], used to collect the annual revenue of the transmission owner from the demand and generation. The UK model starts by designating a hub node, which is assumed to be the main load center in the network. It then computes the power flow in the system (based on DC load flow equations) during the peak hour assuming that transmission capacity is not constrained. A base MW-mile quantity for the system is then computed by first multiplying the absolute value of the load flow on each line by its cost-adjusted length and a summation is conducted over all lines. The cost-adjusted length is the length of the line times a factor representing the relative costs among different type of lines. The nodal MW-miles for each network location is then computed, by raising the generation capacity of that location by 1 MW and increasing the consumption at the hub node by 1 MW; the difference between the base MW-miles and the nodal MW-miles is the Nodal Marginal MW-miles. Since the UK has combined nodes into different charging zones, the obtained nodal MW-miles are then weighted by the ratio of peak nodal capacity versus overall peak capacity in the zone and summed over the zone to form Zonal Marginal MW-miles. Finally, the zonal marginal MW-miles are multiplied by a constant in an attempt to better reflect costs and then modified by a scaling variable to ensure that the payments from the demand and generation maintain a certain ratio. With the Zonal Marginal MW-miles based charge alone, it may not be possible to collect all the revenue the network owner needs, and therefore a location-independent Residual Recovery Tariff is used to recover the remaining costs.

We have adapted this model to a nodal energy market, where similar to the UK's model, we calculate the DC power flow during the "system peak" with all the nodes except the hub node producing at full capacity and the hub node withdrawing all the power. To calculate the nodal marginal MW-miles, we solve a linear program with the goal of minimizing total MW-miles. The dual variables of the nodal balance constraint represent the change in total MW-miles when an extra MW is injected at the corresponding node ($\frac{\partial \text{Total\_MW−miles}}{\partial \text{Capacity\_change}}$), and is thus equivalent to the nodal marginal MW-miles. We use these dual values as the "Base Charges" and scale them to ensure that all the revenue is collected.

We may generalize this to each node i in the system where we may represent the final cost measured in \$/MW/y as $ic_i = c_1 + c_2 \cdot \lambda_i$. The constant c_2 is applied to scale the location-dependent part akin to the constant multiplied by the zonal marginal MW-miles, while c_1 represents the residual charge that is used to make sure revenue is collected. λ_i is the base charge found as the Lagrange multiplier associated with the flow constraints.

Furthermore, c_1 and c_2 are determined by setting the sum of collected charges equal to the annual revenue required from generation. c_1 and c_2 could be uniquely found by requiring that total collected charge equals annual revenue and that the charge collected from the generator and demand maintains a fixed ratio. This charge includes both location-independent(c_1) and dependent parts ($c_2\lambda_i$). If we set $c_2 = 0$, we find that the charge would no longer depend on the location.

As discussed in the previous paragraph, we may obtain the base charge from the Lagrange multiplier associated with the flow constraints in the following optimization problem:

$$\min \quad TMM = \sum_{l\in L}(f0_l^+ + f0_l^-)CL_l \tag{TMM(y)}$$

$$\left\{ \begin{array}{l} \sum\limits_{l\in L} A_{il}(f0_l^+ - f0_l^-) = \sum\limits_{g\in G} k_{g,i} \quad (\lambda_i)\, \forall i \in N \setminus \left\{h\right\} \\ \sum\limits_{l\in KVL(v)} X_{ll}(f0_l^+ - f0_l^-) = 0 \quad \forall v \in V \\ f0_l^+, f0_l^- \geq 0, \quad \forall l \end{array} \right\}, \tag{4.1}$$

where h denotes the index of the hub node, $k_{g,i}$ denotes the capacity of generator g at i, which are the decision variables of the generation expansion problem. $f0_l^+$ and $f0_l^-$ denote positive and negative power flow on l for transmission charge evaluation; CL_l denotes cost-adjusted length of line l; A_{il} denotes elements of network incidence matrix while X denotes diagonal matrix containing the reactances of all lines. The objective lies in to maximizing the total MW-miles given by the sum of the absolute value of the flow times the cost-adjusted length over all lines. The first constraint is the nodal balance constraint, or Kirchhoff's Current Law, while the second constraint is the Kirchhoff's Voltage Law. When upper level transmission decisions y_l are given, the system (4.1) changes as follows:

$$\left\{\begin{array}{r} \sum_{g\in G} k_{g,i} = \sum_{l\in L} A_{il}(f_l^+ - f_l^-) \quad (\lambda_i) \quad \forall i \in N \setminus \left\{h\right\} \\ \sum_{l\in KVL(v)} M(y_l - 1) \leq \sum_{l\in KVL(v)} X_l(f_l^+ - f_l^-) \quad \forall v \in V \\ \sum_{l\in KVL(v)} M(1 - y_l) \geq \sum_{l\in KVL(v)} X_l(f_l^+ - f_l^-) \quad \forall v \in V \\ f_l^+, f_l^- \geq 0 \quad \forall l \end{array}\right\}.$$

We make the assumption that the reactance of the lines does not change with the thermal capacity of the lines to keep the model linear. In other words, X_l is not a variable in this model. Note that y_l and tl_l are both transmission decisions, they are related to each other following the constraint:

$$tl_l \leq y_l M.$$

To simplify the problem, we only consider charges to generators, and set them to recover the generators' share of fixed transmission costs in the below revenue sufficiency constraint. Thus, we set c_2 to a predetermined value with this change and the revenue sufficiency constraint can then be written as follows:

$$\begin{aligned}(Rev - \sum_{l\in L} tl_l CT_l)(1 - Ps) - \sum_{t\in T}\sum_{i\in N} H(t)p_{i,t} \cdot a_{i,t} \\ = \sum_{i\in N}(c_1 + c_2\lambda_i) \cdot \sum_{g\in G} k_{g,i},\end{aligned} \tag{REV}$$

where the left hand side comprises of fixed annual revenue and capital costs of constructing new lines from generation less congestion revenue (sum of LMPs times nodal injections), while the right hand side represents the total collected charge from the generation.

4.2.1.1 Transmission Charging: Generalization and Variants

Finally, with different generators using transmission grid in widely varying ways, it may be fair to define another generation-type-dependent scaling factor on the transmission charge ic, so that different types of generation pay differentiated transmission charges. Following sec 4.2.1, we redefine ic multiplying this scaling factor $Gscale_g$:

$$ic_i = Gscale_g(c_1 + c_2\lambda_i).$$

With different values for the parameters/variables, we define the following charges:

1. Marginal MW-miles Based Transmission Charge (MTC): c_2 is set to a certain nonzero value while c_1 is found through REV; $Gscale_{g_1} = Gscale_{g_2},\ \forall g_1, g_2 \in G$; cost-recovery transmission charge is based on marginal MW-miles; charges for different generators are the same

2. Generator-dependent Marginal MW-miles Based Transmission Charge (MTCg): c_2 is set to a certain nonzero value while c_1 is found through REV; $\exists g_1, g_2 \in G$: $Gscale_{g_1} \neq Gscale_{g_2}$; cost-recovery transmission charge is based on marginal MW-miles; charges for different generators are not the same

3. Flat Rate Charge (FR): Set $c_2 = 0$ and c_1 is found through REV; $Gscale_{g_1} = Gscale_{g_2},\ \forall g_1, g_2 \in G$; cost-recovery transmission charge is no longer based on marginal MW-miles; it is the same across all network nodes;

4. LMP: $c_1 = 0$ and $c_2 = 0$; locational marginal price (LMP) is the only price signal.

4.2.2 Bilevel Framework

We build a bilevel model to capture the decision making of a proactive transmission planner. In this model, the transmission planner determines transmission investment y_l, tl_l at the upper level while at the lower level values of generation investment and operation as well as prices are determined contingent on transmission decisions, given y_l, tl_l. To account for the transmission charge, different types of transmission charging frameworks are also in the lower level. At the lower level, each generation firm determines its dispatch and capacity investments given energy and transmission prices, with the objective of maximizing profit (revenue minus the expense of operations, capacity investments, and interconnection and transmission charges).

4.2.2.1 Lower-level Generation Expansion and Operation Problem

In summary, at the lower level, we model the interaction between the generators, consumers, arbitrager and the transmission charging player as a perfectly competitive equilibrium problem. Each generator determines its dispatch and capacity investments given energy and transmission prices with the overall objective of maximizing profits/benefits. Consumers react to the energy price based on their demand curves, while system operator acts as an arbitrageur in maximizing profits from the buying and selling of electricity at different nodes. The transmission charging player calculates the nodal MW-miles by solving Problem TMM(y), resulting in the transmission charges paid by generating capacity minimizes the total MW-miles. Energy prices arise from the short-run market clearing constraints. The combination of the market clearing and the first-order conditions of generators and the long-run transmission pricing model constitute a complementarity problem parameterized by transmission expansion decisions.

Next, we define the optimization problem corresponding to each market participant and assume that each node has a single generation firm. More general assumptions are readily accommodated. The ith generator solves the following problem:

$$\begin{aligned} \max \quad & \sum_{t\in T}\sum_{g\in G} H_t(p_{i,t}\cdot q_{g,i,t} - MC_{g,i}\cdot q_{g,i,t}) \\ & - \sum_{g\in G}(IC_i + CK_{g,i})k_{g,i} \qquad \text{(Problem Gen}_i(y,tl)) \\ & \left\{ \begin{array}{l} q_{g,i,t} \le k_{g,i}\cdot(1 - FOR_{g,i,t}) \quad (\mu_{g,i,t}) \quad \forall g\in G, t\in T \\ q_{g,i,t} \ge 0 \quad \forall g \in G, t\in T \\ k_{g,i} \ge 0 \quad \forall g \end{array} \right\}. \end{aligned}$$

The ith consumer solves the following problem. The resulting price $p_{i,t}$ will support decisions in generators' and ISO's problems.

$$\text{Dem}i,t(d_{i,t}) = p_{i,t} = \text{INT}_i + \text{SLP}_i d_{i,t}. \qquad \text{(Relationship Con}_i)$$

The ISO solves the following parameterized problem during each time period t to obtain short run prices and dispatch

$$\max_{\{a_{i,t}\}} \quad \sum_{i\in N} p_{i,t}a_{i,t} \qquad \text{(Problem ISO}_t(y,tl))$$

$$
\begin{aligned}
a_{i,t} &= \sum_{l=1}^{M} An_{i,l} f_{l,t} \\
\sum_{i \in N} a_{i,t} &= 0 && (\eta_t) \\
f_{l,t} &= S_l(\theta_{i,t} - \theta_{j,t}) \quad \forall (i,j) \in \Omega_l, l \in L_E && (\xi_{l,t}) \\
M_l(1 - y_l) &\geq f_{l,t} - S_l(\theta_{i,t} - \theta_{j,t}) \quad \forall (i,j) \in \Omega_l, \\
& \qquad\qquad\qquad\qquad \forall l \in L_C && (\sigma^+_{l,t}) \\
M_l(1 - y_l) &\leq -f_{l,t} + S_l(\theta_{i,t} - \theta_{j,t}) \quad \forall (i,j) \in \Omega_l, \\
& \qquad\qquad\qquad\qquad \forall l \in L_C && (\sigma^-_{l,t}) \\
f_{l,t} &\leq tl_l \quad \forall l \in L_E && (\lambda^+_{l,t}) \\
f_{l,t} &\geq -tl_l \quad \forall l \in L_E && (\lambda^-_{l,t}) \\
f_{l,t} &\leq M \cdot y_l \quad \forall l \in L_C && (\pi^+_{l,t}) \\
f_{l,t} &\geq -M \cdot y_l \quad \forall l \in L_C && (\pi^-_{l,t}) \\
\theta_{i,t} &\leq \pi \quad \forall i && (\rho^+_{i,t}) \\
\theta_{i,t} &\geq -\pi \quad \forall i. && (\rho^-_{i,t})
\end{aligned}
$$

Note that, we have linearized the constraint $f_{l,t} = S_l \cdot y_l(\theta_{i,t} - \theta_{j,t})$ using Big M method here. Finally, the market clearing conditions for each node is listed below.

$$
d_{i,t} = a_{i,t} + \sum_{g=1}^{G} q_{g,i,t} \quad \forall i \in N, t \in T. \tag{MK(y, tl)}
$$

Combining with the transmission charging model TMM(y), the four models of this section constitute the lower level problem Follower(y, tl):

$$
\boxed{\text{Follower}(y, tl)}
$$

$$
\boxed{\text{TMM}(y)}\,\boxed{\text{REV}}\,\boxed{\text{Gen}_i(y, tl)}\,\boxed{\text{Con}_i}\,\boxed{\text{ISO}_t(y, tl)}\,\boxed{\text{MK}(y, tl)}
$$

With the assumption of perfect competitiveness, we reformulate the lower level by leveraging the sufficient KKT conditions of each problem. We state the equilibrium conditions as: For generators:

$$\left\{\begin{array}{c} 0 \le q_{g,i,t} \perp H_t p_{i,t} - H_t MC_{g,i} - \mu_{g,i,t} \le 0 \\ 0 \le \mu_{g,i,t} \perp q_{g,i,t} - k_{g,i} \cdot (1 - FOR_{g,i,t}) \le 0 \\ 0 \le k_{g,i} \perp -(IC_i + CK_{g,i}) + \sum_{t=1}^{T} \mu_{g,i,t}(1 - FOR_{g,i,t}) \le 0 \end{array}\right\} \quad \forall g, i, t. \quad (CP_i^G(y))$$

For ISO: First we substitute $a_i = \sum_{l=1}^{M} A_{2i,l} f_l$ into the objective and constraints. The decision variables change to f_l and θ_l.

For $l \in L_E$,

$$\left\{\begin{array}{l} \qquad 0 = \sum_{i=1}^{N} p_{i,t} An_{i,l} - \eta_t \sum_{i=1}^{N} An_{i,l} - \xi_{l,t} - \lambda_{l,t}^{+} + \lambda_{l,t}^{-} \\ \quad f_{l,t} = S_l(\theta_{i,t} - \theta_{j,t}) \quad \forall (i < j) \in \Omega_l \\ 0 \le \lambda_{l,t}^{+} \perp f_{l,t} - F_{l,t} \le 0 \\ 0 \le \lambda_{l,t}^{-} \perp -f_{l,t} - F_{l,t} \le 0. \end{array}\right\}. \qquad (CP_i^{le}(y))$$

For $l \in L_C$,

$$\left\{\begin{array}{l} \qquad 0 = \sum_{i=1}^{N} p_{i,t} An_{i,l} - \eta_t \sum_{i=1}^{N} An_{i,l} - \sigma_{l,t}^{+} + \sigma_{l,t}^{-} - \pi_{l,t}^{+} + \pi_{l,t}^{-} \\ 0 \le \sigma_{l,t}^{+} \perp f_{l,t} - S_l(\theta_{i,t} - \theta_{j,t}) - M_l(1 - y_l) \le 0 \quad \forall (i < j) \in \Omega_l \\ 0 \le \sigma_{l,t}^{-} \perp -f_{l,t} + S_l(\theta_{i,t} - \theta_{j,t}) - M_l(1 - y_l) \le 0 \quad \forall (i < j) \in \Omega_l \\ 0 \le \pi_{l,t}^{+} \perp f_{l,t} - F_{l,t} \le 0 \\ 0 \le \pi_{l,t}^{-} \perp -f_{l,t} - F_{l,t} \le 0 \end{array}\right\}. \quad (CP_i^{lc}(y))$$

For both settings,

$$\left\{\begin{array}{l} \qquad 0 = \xi_{l,t} S_l + \sigma_{l,t}^{+} S_l - \sigma_{l,t}^{-} S_l - \rho_{i,t}^{+} + \rho_{i,t}^{-} \\ 0 \le \rho_{i,t}^{+} \perp \theta_{i,t} - \pi \le 0 \\ 0 \le \rho_{i,t}^{-} \perp -\theta_{i,t} - \pi \le 0 \end{array}\right\}. \qquad (CP_i^{l}(y))$$

4.2.2.2 Upper-level Transmission Capacity Expansion

If we denote the decision variable of the lower level equilibrium problem as x, we may write the proactive transmission planning problem as follows:

$$\max_{x,y,tl} \quad F(x,y,tl)$$
$$\left\{ \begin{array}{c} y, tl \in \text{Feasible Transmission Plans} \\ x \in \text{Follower}(y,tl) \end{array} \right\}.$$

Assuming perfect competition, every player is modeled as a price taker and the follower-level problem is recast as a complementarity problem by concatenating the player-specific necessary and sufficient optimality conditions. Specifically, we denote the complementarity problem of the associated player problem as $(\text{CP}^{\text{Prob}})$, e.g. the complementarity conditions of $\text{Gen}_i(y,tl)$ are denoted by $\text{CP}_i^{\text{Gen}}(y,tl)$. The whole follower-level complementarity problem is then denoted by $\text{CP}^{\text{Follower}(y,tl)}$, which contains $\text{CP}^{\text{TMM}}(y)$, $\text{CP}_i^{\text{Gen}}(y,tl)$, CP_i^{Con}, $\text{CP}_t^{\text{ISO}}(y,tl)$, and $\text{MK}(y,tl)$. We then introduce $\text{CP}^{\text{Follower}(y,tl)}$ into the feasible level of the transmission planner's problem. We denote λ as the tuple of Lagrange multipliers arising from $\text{CP}^{\text{Follower}(y,tl)}$. The ISO makes planning decisions in this model subject to the reaction of the generation market, as captured by the following equilibrium model:

$$\begin{aligned} \max \quad & \sum_{t\in T}\Bigg[\sum_{i\in N} H_t\Bigg(\int_0^{a_{i,t}+\sum_{g\in G} q_{g,i,t}} Dem_{i,t}(x)dx - \\ & \sum_{g\in G}\int_0^{q_{g,i,t}} MC_{g,i}(q)dq\Bigg) - \sum_{i\in N}\sum_{g\in G} CK_{g,i}k_{g,i} - \sum_{l\in L} CT_l tl_l\Bigg] \end{aligned} \tag{UP}$$
$$\text{subject to} \left\{ \begin{array}{l} a_{i,t}, q_{g,i,t}, k_{g,i}, \lambda \quad \text{solve} \quad \text{CP}^{\text{Follower}}(y,tl) \\ tl_l \geq 0, \quad y_l \in \{0,1\}, \quad \forall l \end{array} \right\}. \tag{LO}$$

More generally, the problem is given by the following compact model:

$$\begin{aligned} \min_{x,y,z} \quad & h(x,y,z) \\ \text{subject to} \quad & 0 \leq y \perp F(x,y,z) \geq 0 \\ & f(x,y,z) \geq 0 \\ & g(x,z) \geq 0 \\ & z \in \{0,1\}^p, \end{aligned}$$

where the first (complementarity) constraint captures the market equilibrium conditions, the second and third conditions represent feasibility requirements, while the final set of integrality requirements pertain to transmission expansion decisions.

4.3 Uncertainty Modeling[10]

We now consider an uncertain generalization where the inverse demand function $P_i(x_i)$ and generation marginal cost $MC_i(q_i)$ are assumed to be uncertain. We assume that these uncertain parameters have the following forms:

$$\begin{aligned} P_{i,t}(x, u_{1,i,t}) &\triangleq P_{i,t,0}(x) + u_{1,i,t} P_{1,i,t}(x) \\ MC_{g,i}(q, u_{2,g,i}) &\triangleq MC_{g,i,0}(q) + u_{2,g,i} MC_{2,g,i}(q), \end{aligned}$$

where $u_{1,i,t}$ and $u_{2,g,i}$ are stochastic components and $\|u_1\|_\infty \leq 1, \|u_2\|_\infty \leq 1$. We consider a robust approach to transform the stochastic model to a deterministic (MPEC) with binary variables.

The field of robust optimization [177] have grown immensely over the last two decades in an effort and are guided by the desire to provide solutions robust to parametric uncertainty. In general, it is interested in the convex optimization problem:

$$\min_{x \in X} f(x; u),$$

where $X \in \mathbb{R}^n$, $u \in \mathcal{U} \subseteq \mathbb{R}^L$, $f : X \times U \to \mathbb{R}$ is a convex function in x for every $u \in \mathcal{U}$. Given such a problem, one avenue for defining a robust solution to this collection of uncertain problems is given by the solution to the following worst case problem:

$$\min_{x \in X} \max_{u \in \mathcal{U}} f(x; u).$$

Inspired by the idea of robust optimization, we can derive the following robust

[10] This section is contributed by Shisheng

counterpart of our model:

$$\begin{aligned}
\min_{x,y,z} \max_{u \in \mathcal{U}} \quad & h(x,y,z;u) \\
\text{subject to} \quad & 0 \le y \perp F(x,y,z;u) \ge 0, \quad \forall u \in \mathcal{U} \\
& f(x,y,z) \ge 0, \quad \forall u \in \mathcal{U} \\
& g(x,z) \ge 0 \\
& z \in \{0,1\}^p.
\end{aligned} \tag{RC}$$

Unfortunately, the $y \perp F(x,y,z;u)$ constraint need not admit a solution for all $u \in \mathcal{U}$. Instead, we recast the uncertain complementarity constraint as follows:

$$\begin{aligned}
\min_{x,y,z} \max_{u \in \mathcal{U}} \quad & h(x,y,z;u) \\
\text{subject to} \quad & \min_{x,y,z} \max_{u \in \mathcal{U}} y^T F(x,y,z;u) \\
& y \ge 0 \\
& F(x,y,z;u) \ge 0, \quad \forall u \in \mathcal{U} \\
& f(x,y,z) \ge 0, \quad \forall u \in \mathcal{U} \\
& g(x,z) \ge 0 \\
& z \in \{0,1\}^p,
\end{aligned} \tag{Approx-RC}$$

where $y \perp F(x,y,z;u),\ \forall u \in \mathcal{U}$ is approximated by $\min_{x,y,z}\ \max_{u \in \mathcal{U}} y^T F(x,y,z;u)$.

Although (Approx-RC) may be solved computationally, it can be seen that for some u, it may not hold that $0 \le y \perp F(x,y,z;u) \ge 0$. Consequently, (x,y,z) can be viewed as not satisfying equilibrium conditions for this set of u. For example, if we assume that demand is uncertain. Thus, $p_{i,t}$ is uncertain in the complementarity problem $CP_i^G(y)$. This means the first complementarity constraint in this problem may not be satisfied for all u and operational decisions of generator i are not necessarily optimal for all u. Thus, we consider an adjustable robust technique first suggested in [178]. The idea is that some parts of uncertain variables can be chosen after the realization (akin to a recourse-based model). In our model, investment decisions are made before uncertainty is realized while operational decisions can wait until afterwards. Based on this method, we can develop an adjustable robust counterpart of our model as follows.

$$\begin{aligned}
\min_{x,z,q_1,\ldots,q_n} \max_{u\in\{1,\ldots,n\}} \quad & h(x,z,q_u) \\
\text{subject to} \quad & 0 \le q_u \perp F(x,z,q_u) \ge 0, \quad \forall u \in \{1,\ldots,n\} \\
& f(x,z,q_u) \ge 0, \quad \forall u \in \{1,\ldots,n\} \\
& g(x,z) \ge 0 \\
& z \in \{0,1\}^p,
\end{aligned} \tag{ARC}$$

where q_u are adjustable variables and we assume u has discrete realizations.

4.4 Algorithmic Framework[11]

We briefly review relevant prior research in Section 4.4.1 and describe our proposed scheme for the deterministic problem in Sections 4.4.2- 4.4.3.

4.4.1 Literature Review on Non-convex Mixed-integer Nonlinear Programming (MINLP)

The transmission planning model is a mixed integer mathematical program with equilibrium constraints (MIMPEC), a challenging problem since it involves both binary variables and as well as nonconvexit (due to the inclusion of complementarity conditions). The resulting (MIMPEC) lies in a broad class of mathematical programs referred to as mixed-integer nonlinear programs (MINLPs). Generally, (MINLP)s are categorized as one of the following: (i) The convex MINLP is easier to solve because its continuous relaxation is computationally tractable. (ii) However, the continuous relaxation of the nonconvex (MINLP) is NP-hard [179]. Our problem is an instance of (ii), namely a nonconvex (MINLP). Several techniques have been proposed to handle the difficulty, including convex under- and over-estimators [180]. If a nonconvex (MINLP) has a suitable separability structure, it can be approximated by an (MILP) [181]. Another approximation technique lies in applying a factorization [182] to a (MINLP), leading to a simpler form and allowing for generating good solutions. However if an exact solution is desired, the branching schemes are appropriate.

[11]This section is contributed by Shisheng

4.4.2 Branching Schemes

It has been observed that the resulting (MIMPEC) is difficult to solve by available global optimization solvers directly. To overcome this difficulty, we develop a branch and bound scheme and employ `baron` [183] to solve each subproblem. First, we assume all integer variables are continuous within a range. In our model, since the integer variables $y_i's$ are all binary in nature, we may relax them as $y_i \in [0, 1]$. The model with all binary variables relaxed is treated as the root node. Then we iteratively do the following: at each node, we branch on a relaxed binary variable and convert this variable to be a parameter with a value of 0 or 1 and solve this model to obtain the optimal value. After solving the model, we compare the current optimal value with a maintained lower and an upper bound to determine whether to continue along this branch or prune this branch. After this branch and bound approach terminates, we end up with an optimal solution to the original (MIMPEC). Since we split the node into two each time, the whole process is actually a binary tree. We explain this scheme with a simple example. Assume we have 2 candidate lines to upgrade and the binary decision variables are y_1, y_2. Algorithm 3 and Figure 4.2 summarize the whole process.

Algorithm 1 Branch and Bound Scheme

1: Relax y_l as $y_l \in [0, 1]$, $\forall l$. Set $k = 1$, $z_l = 0$ and $z_u = 1 \times 10^{10}$. Set $\epsilon = 10$.
2: Split y_k as $y_k = 0$ and $y_k = 1$. Optimize the corresponding problems using `baron`.
3: If $z_u - z_l < \epsilon$, then stop and report the solution as optimal to branch. Else, if y_l is binary $\forall l$ and $z > z_l$, set $z_l = z$.
4: If the optimal value $z < z_l$, prune the branch; Else, if $z > z_u$, set $z_u = z$. $k = k + 1$. Return to 2.

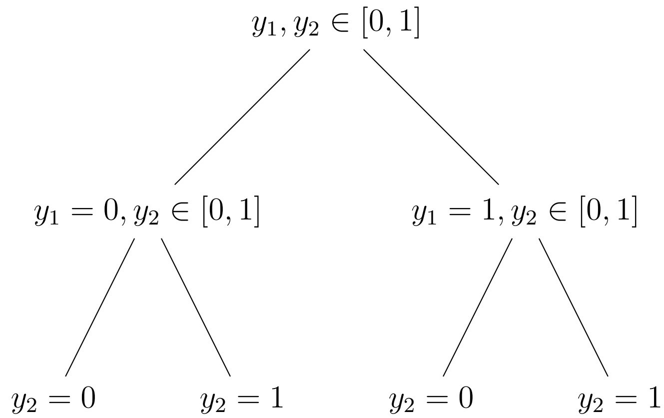


Figure 4.2: Brief explanation for the branching scheme

4.4.3 Global Smoothing

To reduce the computational effort, we leverage a smoothing scheme to provide a reasonable lower bound to the maximization problem (UP). The idea is that we smooth each y_l in the objective function by a continuous function $1 - e^{-\mu y_l}$, where μ is a large positive value. In theory, we have

$$1 - e^{-\mu y_l} \longrightarrow \begin{cases} 1, & y_l = 1 \\ 0, & y_l = 0 \end{cases}, \qquad \text{as} \qquad \mu \to \infty.$$

Thus, we may approximate the original program from a (MIMPEC) by a continuous (MPEC) and a global solution of the smoothed (MPEC) can be obtained via `baron` relatively fast. After rounding the corresponding (relaxed) values to their binary values and recovering feasibility, we resolve the resulting (MPEC) by fixing these binary values. This provides a reasonable lower bound to the original (MIMPEC) (which is being maximized). In fact, this lower bound aids significantly in pruning the tree in the branch-and-bound process, significantly improving the performance. Figure 4.3 shows plots of functions with different value of μ. It improves the efficiency dramatically. We use this smoothing-branching scheme to solve every mixed-binary (MPEC) problem in the case study.

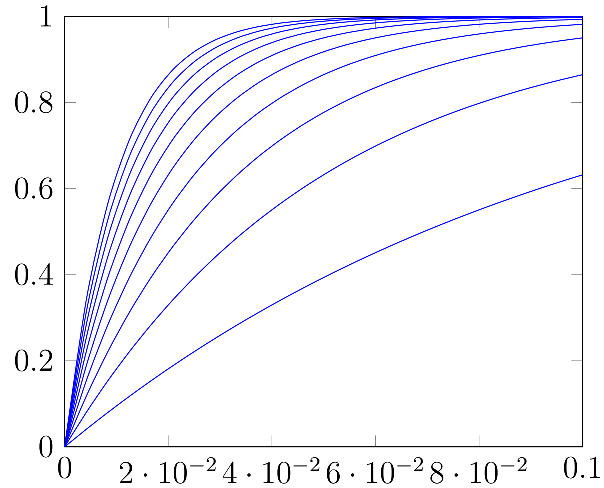


Figure 4.3: Smoothing function with μ from 10 to 100

4.5 Case Study[12]

In this section, we investigate the economic impacts of different transmission charging systems in a 5-node network (Figure 4.4). We first show how generation investment changes with transmission charges and we then show how transmission planning decisions change under different transmission charging systems. Total economic surplus is the major metric we use to compare the solutions. All programs were run on the NEOS Server [184–186] with CPU - 2x Intel Xeon X5660 @ 2.8GHz (12 cores total) CPU and 64GB RAM memories. All models are solved with the proposed branching scheme by employing the smoothing approach.

Since we study a power system with a short-temp nodal market, LMP as a kind of transmission charge always exists in the system. In addition to LMP, we study another 3 kinds of transmission charging system mentioned already in Sect 4.2.1.1, namely the "Marginal MW-miles Based Transmission Charge" (MTC), the "Flat Rate Charge" (FR) and "Generator-dependent Marginal MW-miles Based Transmission Charge" (MTCg). We set $c_2 = 29$ \$/MW/y for both MTC and MTCg. $c1$ is found through REV. With different transmission charges, we construct six different models based on the proposed framework. We then apply these models to a 5-node system for the analysis. We list them below:

[12]This section is joint work by Pengcheng and Shisheng, where Pengcheng designed the test case and analyzed the results while Shisheng developed the algorithm, solved the models, and conducted sensitivity and uncertainty analysis

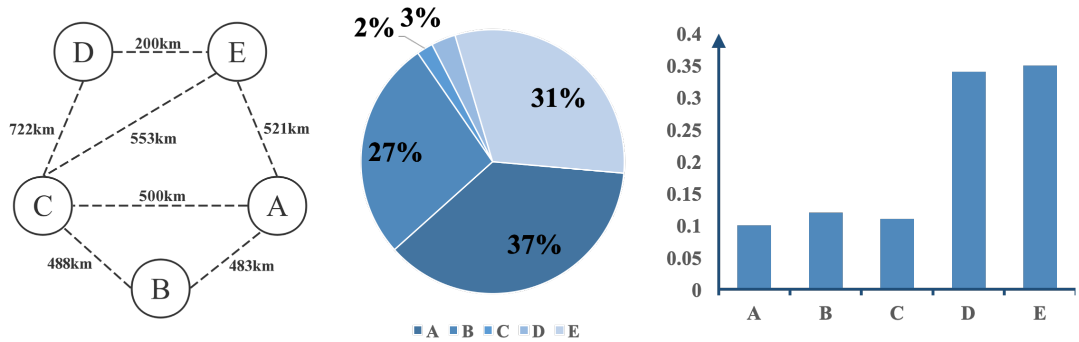


Figure 4.4: L:Five-node network M:Load distribution R:Wind capacity factor

1. Model LMP: A bilevel transmission planning model (as in section 2.2.2) with only LMP as the transmission charge in the lower level.

2. Model MTC: A bilevel transmission planning model with both LMP and MTC charge as the transmission charge.

3. Model $MTCg$: A bilevel transmission planning model with both LMP and MTCg charge as the transmission charge.

4. Model FR: A bilevel transmission planning model with both LMP and a FR charge as the transmission charge.

5. Model $LMP(MTC)$: A single level generation expansion model (as in section 2.2.1) with both LMP and MTC charge as the transmission charge. However, the network is the same as the optimal network from the solution of Model LMP.

6. Model $LMP(FR)$: A single level generation expansion model with both LMP and FR charge as the transmission charge. The network is the same as the optimal network from the solution of Model LMP.

Other fixed parameters include: the demand curve and wind capacity factors calibrated to the ERCOT system data published online of the year 2015, by clustering one year into 22 time periods, generation costs taken from EIA database [187, 188], and transmission line expansion costs are taken from WECC 2014 report [189]. The values are documented in the online Appendix.

Table 4.1: Transmission charges under all models in \$/MW/Year

	A	B	C	D	E
LMP	0	0	0	0	0
LMP(MTC)	24748	10740	24893	45657	39857
LMP(FR)	16629	16629	16629	16629	16629

Table 4.2: Generation capacity under $LMP(MTC)$ in MW

LMP(MTC)	A	B	C	D	E
CC	2690	0	1260	131	0
CT	389	525	401	30	0
Wind	0	0	0	0	0

4.5.1 Transmission Charges' effects on Generation Equilibrium

We show the effects of transmission charges on generation equilibrium by comparing results from Models LMP, $LMP(MTC)$, and $LMP(FR)$, which all share the same transmission network. Although LMP is a bilevel model, the results of the model can be seen as the lower level's response to the shared optimal network, and thus can be compared to the other two models. We first list the equilibrium charge ic_i in the Table 4.1. We can see that the MTC penalizes the nodes far away from the hub, in this case, nodes D and E, which have cheap wind resources. This could incentivize generators to site new generation in places where the local demand is high, especially the hub node B.

Next, we consider the resulting generation capacity investment at different nodes for diverse models in Table 4.2 (LMP(MTC)), Table 4.3 (LMP(FR)), and Table 4.4 (LMP). We can see that the wind generation has been completely pushed out of the system with the MTC charge. Although the wind generation capacity also decreases with the FR charge, the overall generation mix does not change much.

Table 4.3: Generation capacity under $LMP(FR)$ in MW

LMP(FR)	A	B	C	D	E
CC	2347	0	1260	134	0
CT	506	516	417	33	0
Wind	0	0	0	0	878

Table 4.4: Generation capacity under LMP in MW

LMP	A	B	C	D	E
CC	2344	0	1260	54	0
CT	535	539	450	78	0
Wind	0	0	0	195	898

Table 4.5: Social welfare decomposition in M$/Year

Model	Social Welfare Change	Consumer Benefits Change	GenCost Change	Transmission Capital Cost Change
LMP	0	0	0	0
MTC	-5.77	-11.40	7.11	-12.74
FR	-1.70	-10.10	0.28	-8.67
LMP(MTC)	-20.47	-12.30	8.17	0
LMP(FR)	-3.12	-9.06	-6.41	0

Social Welfare Change = Consumer Benefits Change - Gen Cost Change - Transmission Capital Cost Change

Reducing the capacity of wind generators would result in loss of efficiency for the whole system, especially since they have practically negative variable costs due to tax credits. The changes in the generation capacity also induces change in the energy output. 11.4% of energy is generated by wind in LMP, while there's only 9.2% in $LMP(FR)$ and the value drops to zero in case of $LMP(MTC)$.

Social surplus drops more with the MTC charge than with the FR charge, resulting in a loss of 0.674$/MWh for the $LMP(MTC)$ model and 0.103$/MWh for the $LMP(FR)$ model. With the decrease in the capacity of wind generation,

Table 4.6: Solution description

Model	WelfareLoss $/MWh/Year	Gen Mix as % of Demand (Wind;CC;CT)	Average Demand MW/hr/Year
LMP	0	11.4; 83.1; 5.5	3465.83
MTC	0.190	0.0, 95.2; 4.8	3464.56
FR	0.056	4.2; 90.7; 5.1	3465.19
LMP(MTC)	0.674	0.0; 95.2; 4.8	3462.33
LMP(FR)	0.103	9.2; 85.5; 5.3	3462.28
MTCg	0.029	17.8; 76.1; 6.1	3458.67

Table 4.7: Social welfare decomposition in M$/Year

Model	Consumer Surplus	Transmission Surplus	Generation Revenue	Consumer Payment
LMP	8610.81	13.62	1557.08	1633.77
MTC	8509.60	109.07	1695.00	1723.58
FR	8514.68	108.05	1668.33	1719.80
LMP(MTC)	8498.08	105.88	1695.77	1734.20
LMP(FR)	8515.43	105.88	1651.96	1719.62

Social Welfare = Consumer Surplus + Generator Surplus + Transmission Surplus. Since the generator surplus is always zero, it is not listed here. The transmission surplus should also include the revenue collected from consumers which is assumed to be collected by a non-distorting tax and is not included here.

Table 4.8: Comparison of Congestion Revenue, Transmission Charge, and Average Generation Costs (transmission charge excluded)

Model	Congestion Revenue M$/Year	Transmission Charge M$/Year	GenerationCosts-TransmissionCharge $/MWh/Year
LMP	76.69	0	51.29
MTC	28.58	130.82	51.54
FR	51.47	110.97	51.30
LMP(MTC)	38.43	130.52	51.61
LMP(FR)	67.66	101.29	51.13

the per MWh generation revenue also increases in both the $LMP(MTC)$ and $LMP(FR)$ models. However, the price also increased with decrease of cheap wind generation and thus the overall demand decreases. If we look at the average per MWh generation costs (with transmission charge omitted), we could see that it increases with $LMP(MTC)$ model, but decreases with $LMP(FR)$ model. We know the demand drops and lowers system costs and that the generation costs rise with less wind capacity. Overall since there is still significant wind generation in $LMP(FR)$, the reduction of costs from lower demand offsets the rise of costs from more expensive generation.

If we look at the capacity of generators at each node, the deciding factors of the system efficiency is the capacity of wind generation. In the model, the wind generators, though having a much lower capacity factor than fossil fuels generators, would be charged the same way. Additionally, since wind resources are located far away from load center, their locational transmission charge would be higher than generators at other places. These two factors also contribute to the diminishing of wind generators with location dependent transmission charge. We could also see that the more efficient way to collect transmission costs is to collect as much as possible from the congestion charge.

Comparing results from Models LMP, $LMP(MTC)$, and $LMP(FR)$, we may also illustrate the effects of transmission planning without considering actual transmission charges. The planner assumes the transmission charge is only LMP and builds the network according to the optimal plan of the LMP model, however, in reality, the equilibrium outcomes are actually from models $LMP(MTC)$ and $LMP(FR)$, where significant losses have incurred. The cost of transmission is \$63063700, if we calculate the welfare loss of $LMP(MTC)$ and $LMP(FR)$ models relative to this cost, we will find them to be 32.5% and 4.9%. The ratio from $LMP(MTC)$ model is significant and it indicates that more efficient transmission might be found through a proactive transmission planning model.

4.5.2 Transmission Charge effects on Transmission Planning

Table 4.9: Network differences

Model	OptimalNetwork	Transmission Capital M$/Year
LMP	AB=1491.725, AE=412.112, BC=5.693, DE=9.624	63.06
MTC	AB=1553	50.33
FR	AB=1553, AE=86, DE=79	54.39

Comparing the transmission expansion plans in the models LMP, MTC and FR, we observe that the optimal transmission plan changes with different transmission systems. This indicates that the proactive planner may take actions to correct distortions from the transmission charge.

If we compare the per MWh welfare loss from all the models, we see that the proactive planning models MTC and FR do better than $LMP(MTC)$ and

$LMP(FR)$. More than 50% of the losses while using a transmission charging model can be mitigated by planning proactively. However, the degree of this correction differs significantly with different transmission charging systems. The decrease in surplus of MTC is much larger than that of FR, which is caused by less transmission capacity investment due to wind's inability to enter the system. While the proactive planning can mitigate some of the inefficiencies, the welfare drop is still significant for the MTC model, which has no wind generation. We attribute these losses to the nature of the distance-based charge penalizing remote generation.

Although a proactive plan cannot correct all the distortions from the cost recovery charge, can modifying the cost recovery charge encourage wind adoption to reduce the distortion. Furthermore, the remote wind generation, while leading to larger marginal MW-miles, may not employ the transmission system in the same way as baseload or peaking plants. Building on the MTC, we scale the charge IC_i by a scaling factor, so that the new charge $SIC_{g,i}$ could be defined as:

$$SIC_{g,i} = gscale_g \cdot IC_i$$

As can be seen, it now becomes possible to charge different types of generators in a distinct fashion. The charges to the renewable, base load and peaking generators are expected to change and through some testing, we find that in our system, when the $gscale_g$ is chosen in a certain way ($gscale_{wind} = 0.1, gscale_{CC} = 1.0, gscale_{CT} = 0.5$), the system efficiency loss may be even lower than that from FR. We define the model with this charge as $MTCg$ and provide a comparison table in 4.6. With much less cost recovery charge, wind generators have an advantage in investment costs. Thus, we see the wind generation ratio is even higher than the LMP case, although at the cost of lowering system efficiency. We believe a careful design of $gscale_g$ could indeed lessen the distortion from this charge, however, its values are highly dependent on the system configuration and determining them is out of the scope of this study.

We also see that in this system, the transmission planner has to make a tradeoff between building expensive transmission capacity to accommodate remote cheap wind resources and building less transmission capacity while using more expensive but local generation. As can be seen by comparing FR and $LMP(FR)$, the proactive transmission planner improves the system efficiency by building less transmission capacity and thus limits wind generation. Since wind generation is dropping out of system and the system can no longer make use of the significantly cheaper wind generation and fossil fuel generation prices do not significantly differ across different nodes, it makes less economic sense to build expensive transmission to replace local generation.

We can see that the generation revenue also rises for both MTC and FR models and that the revenues are higher than that in $LMP(MTC)$ and $LMP(FR)$. This is caused by both the slightly higher demand and the reduced transmission capacity. However, in contrast with what is observed in section 4.1, the average cost of generation now rises for FR. This is due to the fact that although demand is lower due to higher prices, the generation costs also rise due to lower wind capacity. Overall the demand is not low enough to reduce the system costs while the depressed wind capacity causes too large an increase in costs, and thus the average generation costs (without considering transmission charges) increase.

Again, consistent with what we observe in the prior subsection, another observation is that the social surpluses rise with the percentage of transmission costs recovered from congestion charge. This could create an incentive for the transmission planner to build insufficient transmission capacity to raise the congestion revenue to recover all the costs, even when the objective to maximize social surplus.

4.5.3 Smoothing Approach Results

We compare the objective values and running time of smoothing-branching scheme with classical branching scheme in Figure 4.10. We test MTC, FR and LMP models over two different settings of network. It indicates smoothing techniques lead to significant computational savings with modest loss in solution quality.

Table 4.10: Objective value comparisons($\mu = 10$)

	Branching(global)	Smoothing
MTC(set.1)	8.7452e+9	8.7451e+9
FR(set.1)	9.2141e+9	9.2140e+9
LMP(set.1)	9.2153e+9	9.2151e+9
MTC(set.2)	8.7697e+9	8.7695e+9
FR(set.2)	9.0799e+9	9.0798e+9
LMP(set.2)	9.0799e+9	9.0797e+9
	4-5 hours	10-20 minutes

4.5.4 Sensitivity Analysis

In this section, we conduct a sensitivity analysis to examine the relation of system performance to key parameters. First, the solutions of deterministic models (Table 4.11) and the resulting networks (Figure 4.5) are provided for reference. Table 4.12 shows total welfare and transmission plan for different value of revenue to be recovered. When we increase revenue requirements from 1.6×10^8 to 2.5×10^8, the planning decisions do not change at all, implying that this parameter has little impact on transmission lines expansion. An interesting finding in Table 4.12 and Figure 4.6 is that social welfare sees a significant drop when we increase the value from 2.0×10^8 to 2.5×10^8, which indicates that the whole transmission network is redistributed although the transmission expansion decision is unchanged.

	SW	y	k(Wind)
MTC	8.6186e+9	y(AB)=1	0
FR	8.6227e+9	y(AB)=1, y(AE)=1, y(DE)=1	k(E)=401
LMP	8.6244e+9	y(AB)=1, y(AE)=1, y(BC)=1, y(DE)=1	k(D)=195,k(E)=858

Table 4.11: Social welfare comparison

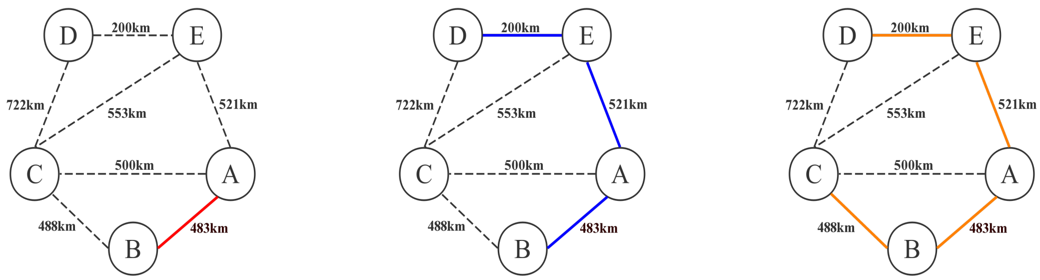


Figure 4.5: L: MTC; M: FR; R: LMP

Table 4.12: (MTC) solutions with different Revenue

Rev	*SW*	y	k(CT)
1.6e+8	8.6186e+9	y(AB)=1	A=372, B=458
1.7e+8	8.6183e+9	y(AB)=1	A=369, B=456
1.8e+8	8.6180e+9	y(AB)=1	A=367, B=454
1.9e+8	8.6177e+9	y(AB)=1	A=365, B=452
2.0e+8	8.6174e+9	y(AB)=1	A=362, B=450
2.1e+8	8.6147e+9	y(AB)=1	A=807, B=0
2.2e+8	8.6144e+9	y(AB)=1	A=803, B=0
2.3e+8	8.6140e+9	y(AB)=1	A=800, B=0
2.4e+8	8.6036e+9	y(AB)=1	A=796, B=0
2.5e+8	8.6032e+9	y(AB)=1	A=792, B=0

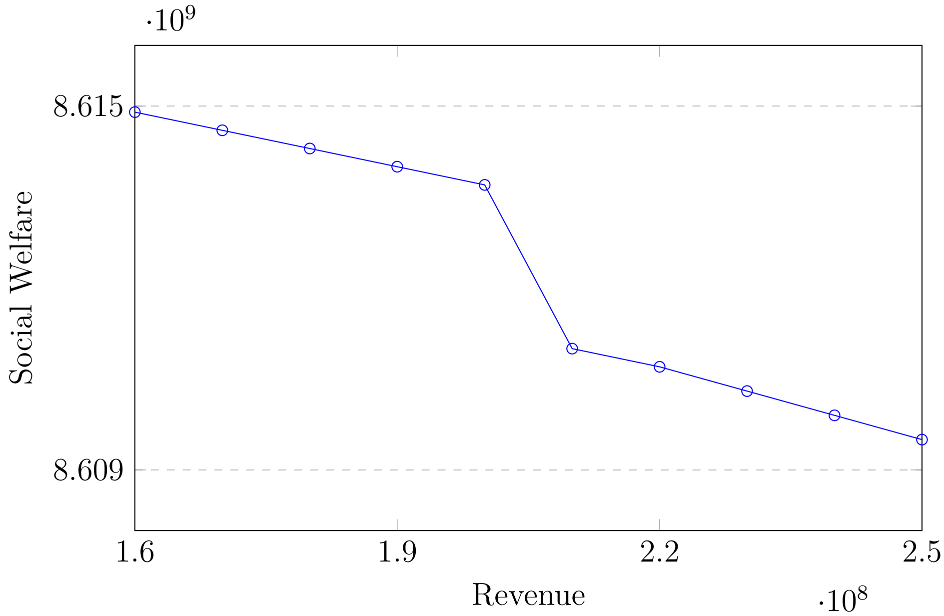


Figure 4.6: Solutions with different Revenue

We rerun the models by varying values of marginal cost for wind at each generation. Table 4.13 shows that the transmission expansion plan changes with decreasing marginal cost for wind. We assume the marginal cost is the same at each generation. Also, the social welfare has significant increase when the marginal cost is set below -35. Figure 4.7 shows the two optimal transmission networks generated by the model.

Table 4.13: (MTC) solutions with different Marginal Cost

MC(Wind)	SW	y	k(Wind)
-15	8.6186e+9	y(AB)=1	0
-20	8.6186e+9	y(AB)=1	0
-25	8.6186e+9	y(AB)=1	0
-30	8.6186e+9	y(AB)=1	0
-35	8.6186e+9	y(AB)=1	0
-40	8.7928e+9	y(AB)=1, y(AE)=1, y(CE)=1, y(DE)=1	E=4406
-45	8.8951e+9	y(AB)=1, y(AE)=1, y(CE)=1, y(DE)=1	E=5373
-50	8.9607e+9	y(AB)=1, y(AE)=1, y(BC)=1, y(DE)=1	E=5714
-55	9.0607e+9	y(AB)=1, y(AE)=1, y(BC)=1, y(DE)=1	E=5983
-60	9.2373e+9	y(AB)=1, y(AE)=1, y(CE)=1, y(DE)=1	D=403, E=5791

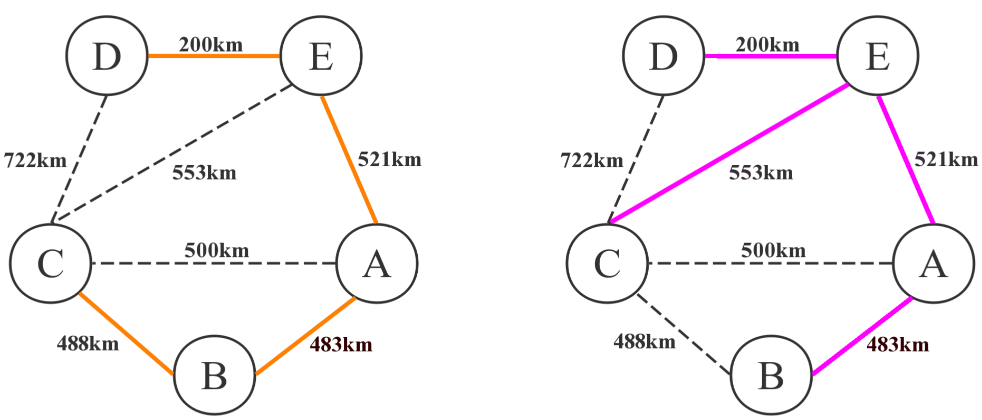


Figure 4.7: Transmission network with different MC

4.5.5 Robust Solutions

Next, we consider the role of uncertainty and examine the adjustable robust counterpart. First, we assume that the parameter a in the inverse demand function $a - bX$ is uncertain and takes on realizations denoted by $a(u)$. Next, we randomly choose 10 samples $\{a_1, \ldots, a_{10}\}$, where $a_1 < a_2 < \cdots < a_{10}$. We also generate results of another 7 random samples $a_{11}, \ldots, a_{17}$ as well as the expected-valued objective for comparison. The result is shown in Table 4.14. From the table, it is clear that (ARC) provides a lower social welfare since it represents a worst-case solution. However, the robust objective value is better than worst case (Sample 11). Another finding is that the expected-valued model generates a different transmission plan compared with (ARC). Capacity of wind remains zero throughout all scenarios, which indicates that in (MTC) we may not expect power supply from wind. Figure 4.8 displays networks generated by solving (ARC) and expected objective model. It means that the (ARC) formulation generates a different network from an expected objective model. The reason is that the expected objective model tries to maximize the social welfare of all scenarios in an averaged value sense which (ARC) ensures normal operation of the worst case scenario.

4.6 Concluding Remarks

In this chapter, we build a bi-level transmission planning model with consideration of transmission cost recovery schemes. We have the transmission planner in the upper level and the generation expansion and transmission cost recovery market equilibrium in the lower level. We are able to illustrate the behavior of a proactive planner, who plans transmission considering the response from the lower level market equilibrium with or without transmission cost recovery. We analyze the economic consequences of imposing different cost recovery schemes when the transmission planner does transmission planning considering these charges. We show that different charges would result in different social efficiency loss. The flat rate charge induces less distortion than the standard MW-miles based charge. We would like to note here that a different configuration of MW-miles based charge that differentiates among generator types results in even less efficiency loss comparing to flat rate charge, however, designing an optimal configuration and making sure it complies with the principles of pricing designs is complicated and out of the scope of this study. We also illustrate that higher efficiency loss could be observed if the transmission planner did not take into account of the cost recovery scheme while planning. This illustrates that the proactive planning could indeed correct some distortions from the cost recovery charge. With uncertainty considered in the model and leveraging techniques from robust optimization, we incorporate robustness into the transmission expansion problem. We develop an adjustable robust counterpart to the uncertain model which admits solutions. A branching scheme is developed to solve the mixed-binary (MPEC) directly and a smoothing approach is employed to accelerate the solving process.

Table 4.14: Solutions with different sample size (MTC)

MTC	Sample	SW	Planning
ARC	$a_1, \ldots, a_{10}$	**2.5572e+9**	**AB, AE, CE, DE**
Sample 11	a_{11}	1.7476e+9	AB, AE, CE, DE
Sample 12	a_{12}	2.7133e+9	AB, AE, CE, DE
Sample 13	a_{13}	3.8805e+9	AB, AE, CD, DE
Sample 14	a_{14}	5.2296e+9	AB, AE, CD, DE
Sample 15	a_{15}	6.8372e+9	AB
Sample 16	a_{16}	8.5305e+9	AB
Sample 17	a_{17}	9.5816e+9	AB
Expected	$a_1, \ldots, a_{10}$	**5.6293e+9**	**AB, AE, BC, DE**

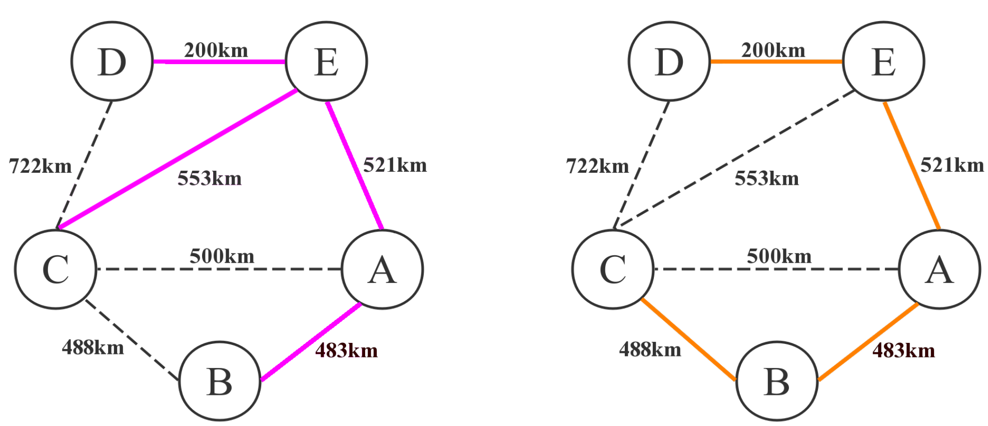


Figure 4.8: Transmission network generated by ARC and expected objective

Chapter 5
Concluding remarks and future work

In this chapter, we recap our main findings from each essay and briefly discuss future work.

5.1 Stochastic Extragradient-type Methods

In this dissertation, we first develop single projection variants of extragradient schemes for monotone stochastic variational inequality problems. Extragradient schemes and their sampling-based counterparts represent a key cornerstone of solving monotone deterministic and stochastic variational inequality problems. Yet, the per-iteration complexity of such schemes is twice as high as their single projection counterparts. We consider two avenues in which the two projections are replaced by exactly one projection (a projected reflected scheme) or a single projection onto the set and another onto a halfpace, the second of which is computable in closed form (a subgradient extragradient scheme). In both instances, we derive a.s. convergence statements and rate statements under variance reduction. Notably, the sequences achieve a non-asymptotic rate of $\mathcal{O}(1/K)$, matching its deterministic counterpart. Furthermore, when this set is itself challenging to project onto, we develop a random projection variant for each scheme. Again, a.s. convergence and rate statements are provided. Empirical behavior of both schemes show significant benefits in terms of per-iteration complexity compared to extragradient counterparts.

Yet, there is much to be gained from understanding how the schemes can be extended to pseudomonotone and non-monotone regimes. While it is plausible that such avenues could allow for rate statements under strong pseudomonotonicity, addressing more general mappings (non-monotone) requires the use of a line search. This is particularly problematic in that naive implementations lead to correlation between the steplength and the iterate, significantly complicating the analysis. A second question pertains to the use of random projections where the rate statements are poorer than the deterministic counterparts ($\mathcal{O}(1/K)$) under mere monotonicity. Can increasing number of simple projections at each step reduce this gap? Such questions remain of interest in future work.

5.2 Stochastic Proximal and Splitting Schemes

Next, we extend our work to solving the stochastic generalized equation with monotone operators, a class of problems that subsumes convex stochastic optimization problems as well as subclasses of convex Nash games and variational inequality problem. In this context, we propose two avenues where the first avenue develops a stochastic proximal point framework for a subclass of stochastic generalized equations in which the operator is either strongly monotone or maximal monotone. By employing a sample-average of the map, we proceed to show that when the sample-size sequences are raised at a suitable rate, we prove that the resulting sequence of iterates converges either at a linear rate (strongly monotone) or at a rate of $\mathcal{O}(1/k)$ (maximal monotone), leading to oracle complexities of $\mathcal{O}(1/\epsilon)$ and $\mathcal{O}(1/\epsilon^{2a+1})$, $\forall a > 1$, respectively. The second one consider structured regimes in which the map can be rewritten as the sum of two maps, facilitating the use of splitting-based framework. In this context, when one of the maps is expectation-valued while the other has a cheap resolvent, we consider a scheme in which a sample-average of the expectation-valued map. Akin to the prior scheme, when the sample-size is increased at a suitable rate, the resulting sequence of iterates converges either at a linear rate (strongly monotone) or at a rate of $\mathcal{O}(1/k)$ (maximal monotone), leading to oracle complexities of $\mathcal{O}(1/\epsilon)$ and $\mathcal{O}(1/\epsilon^2)$, respectively.

Several questions remain in developing a comprehensive study of stochastic generalized equations. A crucial open question is contending with non-monotone operators complicated by uncertainty. The resolution of such problems is challenging, even in deterministic regimes, and remains a broad goal of study. Second, one of the stochastic proximal-point schemes relies on settings where the resolvent is equivalent to solving a tractable optimization/equilibrium problem. Instead, can one develop an improved scheme for solving strongly monotone generalized equations which display a muted dependence on the condition number.

5.3 Competitive Transmission Expansion Planning

Finally, we build a bilevel transmission planning model with consideration of transmission cost recovery schemes where the transmission planner's problem is at the upper level while the generation expansion and transmission cost recovery market equilibrium are at the lower level. We are able to illustrate the behavior of a proactive planner, who plans transmission by incorporating the response from the lower level market equilibrium with or without transmission cost recovery. We analyze the economic consequences of imposing different cost recovery schemes when the transmission planner does transmission planning considering these charges. We show that diverse charges would result in widely differing social welfare loss. The flat rate charge induces less distortion than the standard MW-Miles based charge. We would like to note here that a different configuration of MW-miles based charge that differentiates among generator types results in even less efficiency loss comparing to flat rate charge, however, designing an optimal configuration and making sure it complies with the principles of pricing designs is complicated and out of the scope of this study. We also illustrate that higher efficiency loss could be observed if the transmission planner does not take into account of the cost recovery scheme while planning. This illustrates that the proactive planning could indeed correct some distortions from the cost recovery charge. Also, uncertainty is included in the model and we use a robust approach to reformulate the stochastic model. Leveraging techniques from robust solutions to uncertain complementary problems, we incorporate robustness into the transmission expansion problem and some preliminary numerics are provided.

Much remains to be understood about how such models can be extended to account for uncertainty. Currently, we employ a sampled approach to approximate the adjustable robust counterpart. However, under what conditions can the original adjustable robust counterpart be recast as a single low-dimensional nonlinear program. Can such avenues provide better expansion plans from a robustness standpoint? This remains the focus of future work.

Appendix A Convergence analysis of stochastic extragradient method

Proposition 19. *Let Assumptions 1, 4 – 6 hold and let $\gamma_k \leq \frac{1}{2L}$. Then for any $x_0 \in X$, a sequence generated by (SEG) converges to a solution $x^* \in X$ in an a.s. sense.*

Proof. By Lemma 1(ii), we have

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq \|x_k - \gamma_k F(x_{k+\frac{1}{2}}, \omega_{k+\frac{1}{2}}) - x^*\|^2 - \|x_k - \gamma_k F(x_{k+\frac{1}{2}}, \omega_{k+\frac{1}{2}}) - x_{k+1}\|^2 \\ &= \|x_k - x^*\|^2 - \|x_{k+1} - x_k\|^2 - 2\gamma_k (F(x_{k+\frac{1}{2}}) + w_{k+\frac{1}{2}})^T (x_{k+1} - x^*).\end{aligned} \tag{A.1}$$

We have

$$-F(x_{k+\frac{1}{2}})^T (x_{k+1} - x^*) = -F(x_{k+\frac{1}{2}})^T (x_{k+1} - x_{k+\frac{1}{2}}) - F(x_{k+\frac{1}{2}})^T (x_{k+\frac{1}{2}} - x^*). \tag{A.2}$$

Applying (A.2) to (A.1) yields

$$\begin{aligned}&\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \|x_{k+1} - x_k\|^2 - 2\gamma_k F(x_{k+\frac{1}{2}})^T (x_{k+1} - x_{k+\frac{1}{2}}) \\ &- 2\gamma_k F(x_{k+\frac{1}{2}})^T (x_{k+\frac{1}{2}} - x^*) - 2w_{k+\frac{1}{2}}^T (x_{k+1} - x^*) \\ &= \|x_k - x^*\|^2 - \|x_{k+\frac{1}{2}} - x_k\|^2 - \|x_{k+\frac{1}{2}} - x_{k+1}\|^2 - 2\gamma_k F(x_{k+\frac{1}{2}})^T (x_{k+\frac{1}{2}} - x^*) \\ &- 2w_{k+\frac{1}{2}}^T (x_{k+1} - x^*) + 2(x_{k+1} - x_{k+\frac{1}{2}})^T (x_k - \gamma_k F(x_{k+\frac{1}{2}}) - x_{k+\frac{1}{2}}).\end{aligned} \tag{A.3}$$

By Lemma 1(i), we have

$$(x_{k+\frac{1}{2}} - x_{k+1})^T (x_{k+\frac{1}{2}} - x_k + \gamma_k F(x_k, \omega_k)) \leq 0. \tag{A.4}$$

Using (A.4) in (A.3), we obtain

$$\begin{aligned}
&\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \|x_{k+\frac{1}{2}} - x_k\|^2 - \|x_{k+\frac{1}{2}} - x_{k+1}\|^2 \\
&- 2\gamma_k F(x_{k+\frac{1}{2}})^T (x_{k+\frac{1}{2}} - x^*) - 2w_{k+\frac{1}{2}}^T (x_{k+1} - x^*) \\
&+ 2\gamma_k (x_{k+1} - x_{k+\frac{1}{2}})^T (F(x_k) - F(x_{k+\frac{1}{2}})) + 2\gamma_k \omega_k^T (x_{k+1} - x_{k+\frac{1}{2}}) \\
&\leq \|x_k - x^*\|^2 - \|x_{k+\frac{1}{2}} - x_k\|^2 - \|x_{k+\frac{1}{2}} - x_{k+1}\|^2 - 2\gamma_k F(x_{k+\frac{1}{2}})^T (x_{k+\frac{1}{2}} - x^*) \\
&- 2w_{k+\frac{1}{2}}^T (x_{k+1} - x^*) + 2\gamma_k L \|x_{k+1} - x_{k+\frac{1}{2}}\| \|x_k - x_{k+\frac{1}{2}}\| + 2\gamma_k \omega_k^T (x_{k+1} - x_{k+\frac{1}{2}}) \\
&\leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2) \|x_k - x_{k+\frac{1}{2}}\|^2 - \frac{1}{2} \|x_{k+\frac{1}{2}} - x_{k+1}\|^2 \\
&- 2\gamma_k F(x_{k+\frac{1}{2}})^T (x_{k+\frac{1}{2}} - x^*) - 2w_{k+\frac{1}{2}}^T (x_{k+\frac{1}{2}} - x^*) + 2\gamma_k (\omega_k - \omega_{k+\frac{1}{2}})^T (x_{k+1} - x_{k+\frac{1}{2}}) \\
&\leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2) \|x_k - x_{k+\frac{1}{2}}\|^2 - 2\gamma_k F(x_{k+\frac{1}{2}})^T (x_{k+\frac{1}{2}} - x^*) \\
&- 2w_{k+\frac{1}{2}}^T (x_{k+\frac{1}{2}} - x^*) + 2\gamma_k^2 \|\omega_k - \omega_{k+\frac{1}{2}}\|^2.
\end{aligned}$$

Invoking the weak sharpness property, we have that
$F(x^*)^T (x_{k+\frac{1}{2}} - x^*) \geq \alpha \text{dist}\left(x_{k+\frac{1}{2}}, X^*\right)$, implying that

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2) \|x_k - x_{k+\frac{1}{2}}\|^2 - 2\gamma_k \alpha \text{dist}\left(x_{k+\frac{1}{2}}, X^*\right) \\
+ 2\gamma_k^2 \|w_{k+\frac{1}{2}} - w_k\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T (x_{k+\frac{1}{2}} - x^*).
\end{aligned} \tag{A.5}$$

We have the following inequality:

$$\begin{aligned}
\text{dist}\left(x_k, X^*\right) &= \|x_k - \Pi_{X^*}(x_k)\| \\
&\leq \|x_k - \Pi_{X^*}(x_{k+\frac{1}{2}})\| \\
&\leq \|x_k - x_{k+\frac{1}{2}}\| + \|x_{k+\frac{1}{2}} - \Pi_{X^*}(x_{k+\frac{1}{2}})\| \\
&= \|x_k - x_{k+\frac{1}{2}}\| + \text{dist}\left(x_{k+\frac{1}{2}}, X^*\right).
\end{aligned} \tag{A.6}$$

Using (A.6) in (A.5), we obtain

$$\begin{aligned}
&\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2) \|x_k - x_{k+\frac{1}{2}}\|^2 - 2\gamma_k \alpha \text{dist}\left(x_{k+\frac{1}{2}}, X^*\right) \\
&+ 2\gamma_k^2 \|w_{k+\frac{1}{2}} - w_k\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T (x_{k+\frac{1}{2}} - x^*)
\end{aligned}$$

$$
\begin{aligned}
&\leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2)\|x_k - x_{k+\frac{1}{2}}\|^2 - 2\gamma_k \alpha \mathrm{dist}\,(x_k, X^*) \\
&+ 2\gamma_k \alpha \|x_k - x_{k+\frac{1}{2}}\| + 2\gamma_k^2 \|w_{k+\frac{1}{2}} - w_k\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T (x_{k+\frac{1}{2}} - x^*) \\
&= \|x_k - x^*\|^2 - \left(\frac{1}{2} - 2\gamma_k^2 L^2\right) \|x_k - x_{k+\frac{1}{2}}\|^2 - 2\gamma_k \alpha \mathrm{dist}\,(x_k, X^*) \\
&- \frac{1}{2}\|x_k - x_{k+\frac{1}{2}}\|^2 + 2\gamma_k \alpha \|x_k - x_{k+\frac{1}{2}}\| + 2\gamma_k^2 \|w_{k+\frac{1}{2}} - w_k\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T (x_{k+\frac{1}{2}} - x^*) \\
&= \|x_k - x^*\|^2 - \left(\frac{1}{2} - 2\gamma_k^2 L^2\right) \|x_k - x_{k+\frac{1}{2}}\|^2 - 2\gamma_k \alpha \mathrm{dist}\,(x_k, X^*) \\
&- \frac{1}{2}\|x_k - x_{k+\frac{1}{2}} - 2\gamma_k \alpha\|^2 + 2\gamma_k^2 \alpha^2 + 2\gamma_k^2 \|w_{k+\frac{1}{2}} - w_k\|^2 - 2\gamma_k w_{k+\frac{1}{2}}^T (x_{k+\frac{1}{2}} - x^*) \\
&\leq \|x_k - x^*\|^2 - 2\gamma_k \alpha \mathrm{dist}\,(x_k, X^*) + 2\gamma_k^2 \alpha^2 + 2\gamma_k^2 \|w_{k+\frac{1}{2}} - w_k\|^2 \\
&- 2\gamma_k w_{k+\frac{1}{2}}^T (x_{k+\frac{1}{2}} - x^*)
\end{aligned}
$$

Taking expectations conditioned on $\mathcal{F}_k$, we obtain the following bound:

$$
\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^*\|^2 - 2\gamma_k \alpha \mathrm{dist}\,(x_k, X^*) + 2\gamma_k^2 \alpha^2 + 8\gamma_k^2 \nu^2,
$$

where we leverage $\|w_{k+\frac{1}{2}} - w_k\|^2 \leq 2\|w_{k+\frac{1}{2}}\|^2 + 2\|w_k\|^2$. We may now apply Lemma 2 which allows us to claim that $\{\|x_k - x^*\|\}$ is convergent and $\sum_k \gamma_k \alpha \mathrm{dist}\,(x_k, X^*) < \infty$ in an a.s. sense. Since $\sum_k \gamma_k = \infty$, in an a.s. sense, we have

$$
\liminf_{k \to \infty} \mathrm{dist}(x_k, X^*) = 0.
$$

This implies that some subsequence of $\{x_k\}$ converges to a point in X^* in an a.s. sense. Since we have known that $\{\|x_k - x^*\|\}$ is a convergent sequence in an a.s. sense, we can claim that the entire sequence $\{x_k\}$ converges to a point in X^* in an a.s. sense.

$\square$

The following rate statements are provided for the sequence $\bar{x}_N$, an average of the iterates $\{x_{k+1/2}\}$ generated by (SEG) over the window constructed from N_l to N where $N_l \triangleq \lfloor N/2 \rfloor$ and $N \geq 2$:

$$
\bar{x}_N \triangleq \frac{\sum_{k=N_l}^N \gamma_k x_{k+\frac{1}{2}}}{\sum_{k=N_l}^N \gamma_k}. \tag{A.7}
$$

Proposition 20 (Dim. steplength: SEG). *Consider the (SEG) scheme and let $\{\bar{x}_N\}$ be defined in* (A.7)*, where $0 < \gamma_k \leq 1/L$ for all $k \geq 0$ and $\gamma_k = \gamma_0/\sqrt{k}$. Let*

Assumptions 1 – 2, 5 hold. Then, we have

$$\mathbb{E}[G(\bar{x}_N)] = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

Proof. Adding $2\gamma_k F(x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}} - x^*)$ to both sides of (A.1), we can deduce

$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_{k+\frac{1}{2}}] \leq \|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2)\|x_k - x_{k+\frac{1}{2}}\|^2 \\ - 2\gamma_k F(x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}} - x^*) + 8\gamma_k^2\nu^2.\end{aligned} \tag{A.8}$$

Taking expectations on both sides of (A.8), we obtain

$$2\gamma_k \mathbb{E}[F(y)^T(x_{k+\frac{1}{2}} - y)] \leq \mathbb{E}[\|x_k - y\|^2] - \mathbb{E}[\|x_{k+1} - y\|^2] + 8\gamma_k^2\nu^2, \quad \forall y \in X. \tag{A.9}$$

From (A.9), by summing over k from N_l to N, we have the following for all $y \in X$:

$$2\sum_{k=N_l}^{N} \gamma_k \mathbb{E}[F(y)^T(x_{k+\frac{1}{2}} - y)] \leq \mathbb{E}[\|x_{N_l} - y\|^2] - \mathbb{E}[\|x_{N+1} - y\|^2] + 8\sum_{k=N_l}^{N} \gamma_k^2\nu^2.$$

Consequently, we have the following sequence of inequalities:

$$\begin{aligned}2\left(\sum_{k=N_l}^{N} \gamma_k\right) \mathbb{E}[F(y)^T(\bar{x}_N - y)] &\leq \mathbb{E}[\|x_{N_l} - y\|^2] - \mathbb{E}[\|x_{N+1} - y\|^2] \\ &\leq B_2^2 + 8\sum_{k=N_l}^{N} \gamma_k^2\nu^2,\end{aligned} \tag{A.10}$$

where the second inequality follows from the boundedness of X. Since $\gamma_k = \gamma_0/\sqrt{k}$, it follows that for all $y \in X$:

$$\mathbb{E}[F(y)^T(\bar{x}_N - y)] \leq \frac{B_2^2 + 8\sum_{k=N_l}^{N} \gamma_k^2\nu^2}{2\sum_{k=N_l}^{N} \gamma_k} = \frac{B_2^2}{2\gamma_0} \frac{1}{\sum_{k=N_l}^{N} k^{-\frac{1}{2}}} + 4\gamma_0\nu^2 \frac{\sum_{k=N_l}^{N} k^{-1}}{\sum_{k=N_l}^{N} k^{-\frac{1}{2}}}. \tag{A.11}$$

We now utilize the following lower bound on the denominator for $N \geq 1$:

$$\sum_{k=N_l}^{N} k^{-\frac{1}{2}} \geq \int_{\frac{N}{2}}^{N} (x+1)^{-\frac{1}{2}} dx = 2\sqrt{(N+1)} - 2\sqrt{N/2+1} \geq 2\sqrt{N/40}. \tag{A.12}$$

Similarly an upper bound may be constructed:

$$\sum_{k=N_l}^{N} k^{-1} \leq \int_{\frac{N}{2}}^{N} x^{-1} dx + \frac{1}{\lfloor \frac{N}{2} \rfloor} \leq \log 2 + 1. \tag{A.13}$$

By substituting (A.12) and (A.13) in (A.11), we obtain that the following holds:

$$\begin{aligned} &\mathbb{E}[F(y)^T(\bar{x}_N - y)] \leq \frac{C_4}{\sqrt{N}} \text{ for all } y \in X \\ &\text{where } C_4 \triangleq \left(\frac{\sqrt{40} B_2^2}{4\gamma_0} + 2\sqrt{40}(\log 2 + 1)\gamma_0 \nu^2 \right). \end{aligned}$$

The result follows by taking supremum over $y \in X$. $\square$

Proposition 21 (**Constant steplength: SEG).** *Consider the (SEG) scheme and let* $\{\bar{x}_N\}$ *be defined as* (A.7), *where* $0 < \gamma_k = \gamma \leq 1/L$ *for all* $k \geq 0$. *Let Assumptions 1 – 2, 5 hold. Then, we have*

$$\mathbb{E}[G(\bar{x}_N)] = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

Proof. Proceeding similarly as in the prior proof, an analogous inequality to (A.10) can be derived for all $y \in X$:

$$2\left(\sum_{k=0}^{N} \gamma_k\right) \mathbb{E}[F(y)^T(\bar{x}_N - y)] \leq B_2^2 + 8\sum_{k=0}^{N} \gamma_k^2 \nu^2. \tag{A.14}$$

Since $\gamma_k \equiv \gamma$, we can rewrite (A.14) as follows for all $y \in X$:

$$2(N+1)\gamma \mathbb{E}[F(y)^T(\bar{x}_N - y)] \leq B_2^2 + 8(N+1)\gamma^2\nu^2,$$

leading to the following inequality for all $y \in X$:

$$\begin{aligned} \mathbb{E}[F(y)^T(\bar{x}_N - y)] &\leq \frac{B_2^2}{2(N+1)\gamma} + \frac{8(N+1)\gamma^2\nu^2}{2(N+1)\gamma} = \frac{B_2^2}{2(N+1)\gamma} + 4\gamma\nu^2 \\ &\leq \frac{B_2^2}{2N\gamma} + 4\gamma\nu^2. \end{aligned}$$

Letting $\gamma = B_2/2\sqrt{2N}\nu$, we may deduce that

$$\mathbb{E}[F(y)^T(\bar{x}_N - y)] \leq \frac{\sqrt{2}B_2\nu}{\sqrt{N}}, \forall y \in X.$$

The result follows. $\square$

Proposition 22. *Let Assumptions 1 – 2, 5 hold and assume the mapping F is strongly monotone. Let $\{\gamma_k\}$ be given by $\gamma_k = \gamma_0/k$. Then any sequence generated by (SEG) converges to a solution $x^* \in X$ in an expected value sense:*

$$\mathbb{E}[\|x_k - x^*\|^2] = \mathcal{O}\left(\frac{1}{N}\right).$$

Proof. According to strong monotonicity assumption of F, we get

$$\begin{aligned}-2\gamma_k F(x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}} - x^*) &\leq -2\gamma_k\sigma\|x_{k+\frac{1}{2}} - x^*\|^2 \\ &\leq 2\gamma_k\sigma\|x_{k+\frac{1}{2}} - x_k\|^2 - \gamma_k\sigma\|x_k - x^*\|^2.\end{aligned}$$

Using it in (A.3), we deduce

$$\begin{aligned}\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 &- \|x_{k+\frac{1}{2}} - x_k\|^2 - \|x_{k+\frac{1}{2}} - x_{k+1}\|^2 - w_{k+\frac{1}{2}}^T(x_{k+1} - x^*) \\ &+ 2(x_{k+1} - x_{k+\frac{1}{2}})^T(x_k - \gamma_k F(x_{k+\frac{1}{2}}) - x_{k+\frac{1}{2}}) \\ &+ 2\gamma_k\sigma\|x_{k+\frac{1}{2}} - x_k\|^2 - \gamma_k\sigma\|x_k - x^*\|^2.\end{aligned}$$

Modifying the subsequent deduction accordingly, we obtain

$$\begin{aligned}&\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_{k+\frac{1}{2}}] \leq (1 - \sigma\gamma_k)\|x_k - x^*\|^2 - (1 - 2\gamma_k^2 L^2 - 2\sigma\gamma_k)\|x_k - x_{k+\frac{1}{2}}\|^2 \\ &+ 4\gamma_k^2\nu^2 \\ &\leq (1 - \sigma\gamma_k)\|x_k - x^*\|^2 + (2\gamma_k^2 L^2 + \gamma_k^2\sigma^2)\|x_k - x_{k+\frac{1}{2}}\|^2 + 4\gamma_k^2\nu^2, \qquad \text{(A.15)}\end{aligned}$$

where the second inequality leverages $1 - 2\gamma_k\sigma \geq -\gamma_k^2\sigma^2$. Since $\|x_k - x_{k+\frac{1}{2}}\|$ is bounded, we have $\|x_k - x_{k+\frac{1}{2}}\| \leq B_2$. Then taking expectations on both sides of (A.15), we get

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq (1 - \sigma\gamma_k)\mathbb{E}[\|x_k - x^*\|^2] + (2L^2 + \sigma^2)\gamma_k^2 B_2^2 + 4\gamma_k^2\nu^2.$$

By assuming $1 < \sigma\gamma_0$ and invoking Lemma 16, we get

$$\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{M}{k},$$

where $M = \max\left(\frac{\gamma_0^2((2L^2+\sigma^2)B_2^2+4\nu^2)}{\sigma\gamma_0-1}, \mathbb{E}[\|x_0 - x^*\|^2]\right)$. This means $\{x_k\}$ converges to x^* in an expected value sense. $\square$

Proposition 23. *Let Assumptions 1, 3 – 8 hold and let $\gamma_k \leq \frac{1}{2L}$. Then any sequence generated by (r-SEG), where the projections are random generated, converges to a solution $x^* \in X$ in an a.s. sense.*

Proof. By Lemma 1(ii), we have

$$\begin{aligned}\|x_{k+1}-x^*\|^2 &\leq \|x_k-\gamma_k(F(x_{k+\frac{1}{2}})+w_{k+\frac{1}{2}})-x^*\|^2 \\ &- \|x_k-\gamma_k(F(x_{k+\frac{1}{2}})+w_{k+\frac{1}{2}})-x_{k+1}\|^2 \\ &= \|x_k-x^*\|^2-\|x_k-x_{k+1}\|^2+2\gamma_k(F(x_{k+\frac{1}{2}})+w_{k+\frac{1}{2}})^T(x^*-x_{k+1}).\end{aligned} \tag{A.16}$$

It is clear that

$$F(x_{k+\frac{1}{2}})^T(x_{k+1}-x^*) = F(x_{k+\frac{1}{2}})^T(x_{k+1}-x_{k+\frac{1}{2}})+F(x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}}-x^*). \tag{A.17}$$

Using (A.17) in (A.16), we obtain

$$\begin{aligned}&\|x_{k+1}-x^*\|^2 = \|x_k-x^*\|^2-\|x_k-x_{k+1}\|^2+2\gamma_k F(x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}}-x_{k+1}) \\ &+2\gamma_k w_{k+\frac{1}{2}}^T(x^*-x_{k+1})-2\gamma_k F(x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}}-x^*) \\ &= \|x_k-x^*\|^2-\|x_k-x_{k+\frac{1}{2}}+x_{k+\frac{1}{2}}-x_{k+1}\|^2+2\gamma_k F(x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}}-x_{k+1}) \\ &+2\gamma_k w_{k+\frac{1}{2}}^T(x^*-x_{k+1})-2\gamma_k F(x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}}-x^*) \\ &= \|x_k-x^*\|^2-\|x_k-x_{k+\frac{1}{2}}\|^2-\|x_{k+\frac{1}{2}}-x_{k+1}\|^2-2(x_k-x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}}-x_{k+1}) \\ &+2\gamma_k F(x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}}-x_{k+1})+2\gamma_k w_{k+\frac{1}{2}}^T(x^*-x_{k+1})-2\gamma_k F(x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}}-x^*) \\ &= \|x_k-x^*\|^2-\|x_k-x_{k+\frac{1}{2}}\|^2-\|x_{k+\frac{1}{2}}-x_{k+1}\|^2 \\ &+2(x_{k+1}-x_{k+\frac{1}{2}})^T(x_k-\gamma_k F(x_{k+\frac{1}{2}})-x_{k+\frac{1}{2}})+2\gamma_k w_{k+\frac{1}{2}}^T(x^*-x_{k+1}) \\ &-2\gamma_k F(x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}}-x^*).\end{aligned}$$

With a similar approach in Proposition 8, we obtain

$$\begin{aligned}&\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2-2\gamma_k\alpha\mathrm{dist}\,(x_k,X^*)-\left(\frac{5}{8}-2\gamma_k^2L^2\right)\|x_k-x_{k+\frac{1}{2}}\|^2 \\ &-\frac{1}{8}\|x_k-x_{k+\frac{1}{2}}-8\gamma_k\alpha\|^2+8\gamma_k^2\alpha^2+4\gamma_k^2(C+\alpha)^2+2\gamma_k(C+\alpha)d(x_k) \\ &+2\gamma_k^2\|w_{k+\frac{1}{2}}-w_k\|^2-2\gamma_k w_{k+\frac{1}{2}}^T(x_{k+\frac{1}{2}}-x^*).\end{aligned}$$

Finally, we can obtain

$$\mathbb{E}[\|x_{k+1}-x^*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^*\|^2 - 2\gamma_k \alpha \mathrm{dist}\,(x_k, X^*) + 8\gamma_k^2\alpha^2 + 4\gamma_k^2(C+\alpha)^2 + \frac{8m\eta(C)+\alpha^2}{\rho}\gamma_k^2 + 8\gamma_k^2\nu^2$$

Now we can invoke Lemma 2. It follows that $\{\|x_k - x^*\|^2\}$ is convergent and $\sum 2\gamma_k\alpha\mathrm{dist}\,(x_k, X^*) < \infty$. Therefore, it is clear that $x_k \to x^*$ in an a.s. sense. □

Proposition 24. *Let Assumptions 1, 3, 5 – 8 hold and assume the mapping F is strongly monotone. Let $\{\gamma_k\}$ be given by $\gamma_k = \gamma_0/k$. Then any sequence generated by (r-SEG) converges to a solution $x^* \in X$ in an expected value sense.*

Proof. With a similar approach in Proposition 22, we obtain

$$\mathbb{E}[\|x_{k+1}-x^*\|^2] \leq (1-\gamma_k\sigma)\mathbb{E}[\|x_k - x^*\|^2] + \left(\frac{4}{3}\sigma^2 + 2L^2\right)B_2^2\gamma_k^2 + \gamma_k^2\left(32C^2 + \frac{8m\eta C^2}{\rho} + 8\nu^2\right)$$

By assuming $1 < \sigma\gamma_0$ and invoking Lemma 16, we get

$$\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{M}{k},$$

where $M = \max\left(\frac{\gamma_0^2\left(\left(\frac{4}{3}\sigma^2+2L^2\right)B_2^2+32C^2+\frac{8m\eta C^2}{\rho}+8\nu^2\right)}{\sigma\gamma_0-1}, \mathbb{E}[\|x_0-x^*\|^2]\right)$. This means $\{x_k\}$ converges to x^* in an expected value sense. □

Appendix B
Supporting mathematical results

Proposition 25. *The solution to*

$$\min_{x \in X} \quad \|y - x\|^2,$$

where $X = \{a^T x + b \geq 0\}$ *and* $y \notin X$, *is*

$$x^* = y - \frac{b + a^T y}{a^T a} a.$$

Proof. Since $y \notin X$, we know that $a^T x^* + b = 0$. This implies that optimality conditions are given by

$$\begin{aligned} 2(x^* - y) - \lambda a &= 0 \\ a^T x^* + b &= 0 \end{aligned}$$

This implies

$$x^* = y - \frac{b + a^T y}{a^T a} a,$$

which completes the proof. $\square$

Lemma 19.

$$\lfloor y \rfloor \geq \left\lceil \frac{1}{2} y \right\rceil, \quad \forall y \geq 1, y \in \mathbb{R}.$$

Proof. (i) Suppose $y = 2n$, $n \in \mathbb{Z}_+$. It is clear that

$$\lfloor y \rfloor = 2n \geq n = \left\lceil \frac{1}{2} y \right\rceil.$$

(ii) Suppose $2n < y < 2n+1$, $n \in \mathbb{Z}_+$, we have

$$\lfloor y \rfloor = 2n \geq n+1 = \left\lceil \frac{1}{2} y \right\rceil .$$

(iii) Suppose $2n+1 \leq y < 2n+2$, $n \in \mathbb{Z}_+$ and it follows that

$$\lfloor y \rfloor = 2n+1 \geq n+1 = \left\lceil \frac{1}{2} y \right\rceil .$$

The proof is complete. $\square$

Bibliography

[1] Signorini, A. (1959) "Questioni di elasticità non linearizzata e semilinearizzata," **Rendiconti di Matematica e delle sue Applicazioni**, **5**(18), pp. 95–139.

[2] Fichera, G. (1963) "Sul problema elastostatico di Signorini con ambigue condizioni al contorno," **Rendiconti della Accademia Nazionale dei Lincei, Classe di Scienze Fisiche, Matematiche e Naturali 8**, **34**(2), pp. 138–142.

[3] ——— (1964) "Problemi elastostatici con vincoli unilaterali: il problema di Signorini con ambigue condizioni al contorno," **Memorie della Accademia Nazionale dei Lincei, Classe di Scienze Fisiche, Matematiche e Naturali, 8**, **7**(2), pp. 91–140.

[4] Dafermos, S. (1980) "Traffic equilibrium and variational inequalities," **Transportation Science**, **14**(1), pp. 42–54.

[5] Ferris, M. C., O. L. Mangasarian, and J.-S. Pang (2013) **Complementarity: Applications, algorithms and extensions**, vol. 50, Springer Science & Business Media.

[6] Facchinei, F. and J.-S. Pang (2007) **Finite-dimensional variational inequalities and complementarity problems**, Springer Science & Business Media.

[7] Konnov, I. (2007) **Equilibrium models and variational inequalities**, vol. 210, Elsevier.

[8] Karamardian, S. (1972) "The complementarity problem," **Mathematical Programming**, **2**(1), pp. 107–129.

[9] Hogan, W. W. (1975) "Energy policy models for project independence," **Computers & Operations Research**, **2**(3-4), pp. 251–271.

[10] Greenberg, H. J. and F. H. Murphy (1980) "Modelling the National Energy Plan," **Journal of the Operational Research Society**, **31**(11), pp. 965–973.

[11] Murphy, F., A. Pierru, and Y. Smeers (2016) "A Tutorial on Building Policy Models as Mixed-Complementarity Problems," **INFORMS Journal on Applied Analytics**, **46**(6), pp. 465–481.

[12] Hobbs, B. F. (1986) "Network Models of Spatial Oligopoly with an Application to Deregulation of Electricity Generation," **Operations Research**, **34**(3), pp. 395–409.

[13] ——— (2001) "Linear Complementarity Models of Nash-Cournot Competition in Bilateral and POOLCO Power Markets," **IEEE Transactions on Power Systems**, **16**(2), pp. 194–202.

[14] Hobbs, B. F., C. B. Metzler, and J.-S. Pang (2000) "Strategic Gaming Analysis for Electric Power Systems: An MPEC Approach," **IEEE Transactions on Power Systems**, **15**, pp. 638–645.

[15] Metzler, C., B. F. Hobbs, and J. S. Pang (2003) "Nash-Cournot Equilibria in Power Markets on a Linearized DC Network with Arbitrage: Formulations and Properties," **Networks and Spatial Theory**, **3**(2), pp. 123–150.

[16] Kamat, R. and S. Oren (2004) "Two-settlement Systems for Electricity Markets under Network Uncertainty and Market Power," **Journal of Regulatory Economics**, **25**(1), pp. 5–37.

[17] Yao, J., S. Oren, and I. Adler (2004) "Computing Cournot Equilibria in Two Settlement Electricity Markets with Transmission Constraints." in **Proceeding of the 38th Hawaii International Conference on Systems Sciences (HICSS 38). Big Island, Hawaii**.

[18] Hobbs, B. F. and J. S. Pang (2007) "Nash-Cournot equilibria in electric power markets with piecewise linear demand functions and joint constraints," **Operations Research**, **55**(1), pp. 113–127.

[19] Hobbs, B. F. and U. Helman (2003) "Complementarity-based equilibrium modeling for electric power markets," in **Modeling Prices in Competitive Electricity Markets**, D. W. Bunn, Ed. New York: Wiley, 2004, ch. 3.

[20] Gabriel, S. A., S. Kiet, and J. Zhuang (2005) "A mixed complementarity-based equilibrium model of natural gas markets," **Operations Research**, **53**(5), pp. 799–818.

[21] Holz, F., C. Von Hirschhausen, and C. Kemfert (2008) "A strategic model of European gas supply (GASMOD)," **Energy Economics**, **30**(3), pp. 766–788.

[22] Gabriel, S. A., A. J. Conejo, J. D. Fuller, B. F. Hobbs, and C. Ruiz (2012) **Complementarity modeling in energy markets**, vol. 180, Springer Science & Business Media.

[23] Shanbhag, U. V., G. Infanger, and P. W. Glynn (2011) "A Complementarity Framework for Forward Contracting Under Uncertainty," **Operations Research**, **59**(4), pp. 810–834.

[24] Kannan, A., U. V. Shanbhag, and H. M. Kim (2013) "Addressing supply-side risk in uncertain power markets: stochastic Nash models, scalable algorithms and error analysis," **Optimization Methods and Software**, **28**(5), pp. 1095–1138.

[25] ——— (2011) "Strategic Behavior in Power Markets under Uncertainty," **Energy Systems**, **2**(2), pp. 115 – 141.

[26] Shanbhag, U. V. (2013) "Stochastic Variational Inequality Problems: Applications, Analysis, and Algorithms," **Informs Tutorials**, pp. 71 – 107.

[27] Ravat, U. and U. V. Shanbhag (2011) "On the characterization of solution sets of smooth and nonsmooth convex stochastic Nash games," **SIAM Journal on Optimization**, **21**(3), pp. 1168–1199.

[28] Birge, J. R. and F. Louveaux (2011) **Introduction to stochastic programming**, Springer Science & Business Media.

[29] Shapiro, A., D. Dentcheva, et al. (2014) **Lectures on stochastic programming: modeling and theory**, vol. 16, SIAM.

[30] Chen, X., R. J.-B. Wets, and Y. Zhang (2012) "Stochastic variational inequalities: residual minimization smoothing sample average approximations," **SIAM Journal on Optimization**, **22**(2), pp. 649–673.

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

[32] Xu, H. (2010) "Sample average approximation methods for a class of stochastic variational inequality problems," **Asia-Pacific Journal of Operational Research**, **27**(01), pp. 103–119.

[33] Chen, X. and M. Fukushima (2005) "Expected residual minimization method for stochastic linear complementarity problems," **Mathematics of Operations Research**, **30**(4), pp. 1022–1038.

[34] Fang, H., X. Chen, and M. Fukushima (2007) "Stochastic r_0 matrix linear complementarity problems," **SIAM Journal on Optimization**, **18**(2), pp. 482–506.

[35] Shapiro, A. (2003) "Monte Carlo sampling methods," **Handbooks in operations research and management science**, **10**, pp. 353–425.

[36] Gürkan, G., A. Yonca Özge, and S. M. Robinson (1999) "Sample-path solution of stochastic variational inequalities," **Mathematical Programming**, **84**(2), pp. 313–333.

[37] Ralph, D. and H. Xu (2011) "Convergence of stationary points of sample average two-stage stochastic programs: A generalized equation approach," **Mathematics of Operations Research**, **36**(3), pp. 568–592.

[38] Xu, H. and D. Zhang (2013) "Stochastic Nash equilibrium problems: sample average approximation and applications," **Computational Optimization and Applications**, **55**(3), pp. 597–645.

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

[40] Koshal, J., A. Nedić, and U. V. Shanbhag (2012) "Regularized iterative stochastic approximation methods for stochastic variational inequality problems," **IEEE Transactions on Automatic Control**, **58**(3), pp. 594–609.

[41] Yousefian, F., A. Nedić, and U. V. Shanbhag (2013) "Distributed adaptive steplength stochastic approximation schemes for Cartesian stochastic variational inequality problems," **arXiv preprint arXiv:1301.1711**.

[42] Antipin, A. (1978) "Method of Convex Programming using a symmetric modification of Lagrange function," **Matekon**, **14**(2), pp. 23–38.

[43] Korpelevich, G. (1976) "The extragradient method for finding saddle points and other problems," **Matecon**, **12**, pp. 747–756.

[44] Sherali, H. D. (1984) "A multiple leader Stackelberg model and analysis," **Operations Research**, **32**(2), pp. 390–404.

[45] Su, C.-L. (2007) "Analysis on the forward market equilibrium model," **Operations Research Letters**, **35**(1), pp. 74–82.

[46] Martinet, B. (1972) “Détermination approchée d’un point fixe d’une application pseudo-contractante,” **CR Acad. Sci. Paris**, **274**(2), pp. 163–165.

[47] Rockafellar, R. T. (1976) “Monotone operators and the proximal point algorithm,” **SIAM Journal on Control and Optimization**, **14**(5), pp. 877–898.

[48] ——— (1976) “Augmented Lagrangians and applications of the proximal point algorithm in convex programming,” **Mathematics of Operations Research**, **1**(2), pp. 97–116.

[49] Lions, P.-L. and B. Mercier (1979) “Splitting algorithms for the sum of two nonlinear operators,” **SIAM Journal on Numerical Analysis**, **16**(6), pp. 964–979.

[50] Passty, G. B. (1979) “Ergodic convergence to a zero of the sum of monotone operators in Hilbert space,” **Journal of Mathematical Analysis and Applications**, **72**(2), pp. 383–390.

[51] Han, S.-P. and G. Lou (1988) “A parallel algorithm for a class of convex programs,” **SIAM Journal on Control and Optimization**, **26**(2), pp. 345–355.

[52] Tseng, P. (2000) “A modified forward-backward splitting method for maximal monotone mappings,” **SIAM Journal on Control and Optimization**, **38**(2), pp. 431–446.

[53] Hogan, W. W. (1998) “Independent system operator: Pricing and flexibility in a competitive electricity market,” **Center for Business and Government, JF Kennedy School of Government, Harvard University, MA**.
URL `https://pdfs.semanticscholar.org/a332/ebda2c4e2dfa49ec278424da99ec019e2001.pdf`

[54] Interconnection, P. (2015) “PJM Interconnection Economic Analysis of EPA Clean Power Plan Proposal, March 2015. Nicholas Institute, Duke University,” **Assessing Impacts of the Clean Power Plan on Southeast States**.
URL `https://www.pjm.com/-/media/library/reports-notices/clean-power-plan/20150302-pjm-interconnection-economic-analysis-of-the-epa-clean-power-plan-proposal.ashx?la=en`

[55] Crampes, C. and N. Fabra (2005) “The Spanish electricity industry: plus ca change...” **The Energy Journal**, pp. 127–153.

[56] Garver, L. L. (1970) "Transmission network estimation using linear programming," **IEEE Transactions on Power Apparatus and Systems**, **PAS-89**(7), pp. 1688–1697.

[57] Youssef, H. and R. Hackam (1989) "New transmission planning model," **IEEE Transactions on Power Systems**, **4**(1), pp. 9–18.

[58] Gilles, M. L. (1986) "Optimum HVAC-Transmission Expansion Planning-A New Formulation," **IEEE Transactions on Power Systems**, **1**(1), pp. 48–56.

[59] Dos Santos, A., P. Franca, and A. Said (1989) "An optimization model for long-range transmission expansion planning," **IEEE Transactions on Power Systems**, **4**(1), pp. 94–101.

[60] Pereira, M., L. Pinto, S. Cunha, and G. Oliveira (1985) "A decomposition approach to automated generation/transmission expansion planning," **IEEE Transactions on Power Apparatus and Systems**, **PAS-104**(11), pp. 3074–3083.

[61] Bayona, G. L. (1993) **Modelos estáticos para la planificación a largo plazo de la red de transporte de energía eléctrica**, Ph.D. thesis, Universidad Pontificia Comillas.

[62] Donohoo, P., M. Webster, and I. Pérez-Arriaga (2013) "Algorithmic investment screening for wide-area Transmission Network Expansion Planning," in **2013 IEEE Power & Energy Society General Meeting**, IEEE, pp. 1–5.

[63] De la Torre, T., J. W. Feltes, T. Gomez San Roman, and H. M. Merrill (1999) "Deregulation, privatization, and competition: transmission planning under uncertainty," **IEEE Transactions on Power Systems**, **14**(2), pp. 460–465.

[64] Buygi, M. O., M. Shahidehpour, H. M. Shanechi, and G. Balzer (2004) "Market based transmission planning under uncertainties," in **2004 International Conference on Probabilistic Methods Applied to Power Systems**, IEEE, pp. 563–568.

[65] Silva, I. J., M. J. Rider, R. Romero, and C. A. Murari (2006) "Transmission network expansion planning considering uncertainty in demand," **IEEE Transactions on Power Systems**, **21**(4), pp. 1565–1573.

[66] Cedeño, E. B. and S. Arora (2011) "Performance comparison of transmission network expansion planning under deterministic and uncertain conditions," **International Journal of Electrical Power & Energy Systems**, **33**(7), pp. 1288–1295.

[67] Pozo, D., J. Contreras, and E. Sauma (2013) "If you build it, he will come: Anticipative power transmission planning," **Energy Economics**, **36**, pp. 135–146.

[68] van der Weijde, A. H. and B. F. Hobbs (2012) "The economics of planning electricity transmission to accommodate renewables: Using two-stage optimisation to evaluate flexibility and the cost of disregarding uncertainty," **Energy Economics**, **34**(6), pp. 2089–2101.

[69] Jin, S. and S. M. Ryan (2013) "A tri-level model of centralized transmission and decentralized generation expansion planning for an electricity market-Part I," **IEEE Transactions on Power Systems**, **29**(1), pp. 132–141.

[70] ——— (2014) "A tri-level model of centralized transmission and decentralized generation expansion planning for an electricity market-part II," **IEEE Transactions on Power Systems**, **29**(1), pp. 142–148.

[71] Wogrin, S., E. Centeno, and J. Barquín (2011) "Generation capacity expansion in liberalized electricity markets: A stochastic MPEC approach," **IEEE Transactions on Power Systems**, **26**(4), pp. 2526–2532.

[72] Binato, S., M. V. F. Pereira, and S. Granville (2001) "A new Benders decomposition approach to solve power transmission network design problems," **IEEE Transactions on Power Systems**, **16**(2), pp. 235–240.

[73] Romero, R. and A. Monticelli (1994) "A hierarchical decomposition approach for transmission network expansion planning," **IEEE Transactions on Power Systems**, **9**(1), pp. 373–380.

[74] ——— (1994) "A zero-one implicit enumeration method for optimizing investments in transmission expansion planning," **IEEE Transactions on Power Systems**, **9**(3), pp. 1385–1391.

[75] Contreras, J. and F. F. Wu (2000) "A kernel-oriented algorithm for transmission expansion planning," **IEEE Transactions on Power Systems**, **15**(4), pp. 1434–1440.

[76] Oliveira, G., A. Costa, and S. Binato (1995) "Large scale transmission network planning using optimization and heuristic techniques," **IEEE Transactions on Power Systems**, **10**(4), pp. 1828–1834.

[77] Romero, R., R. Gallego, and A. Monticelli (1995) "Transmission system expansion planning by simulated annealing," in **Power Industry Computer Application Conference, 1995. Conference Proceedings., 1995 IEEE**, IEEE, pp. 278–283.

[78] Villasana, R., L. Garver, and S. Salon (1985) "Transmission network planning using linear programming," **IEEE Transactions on Power Apparatus and Systems**, **PAS-104**(2), pp. 349–356.

[79] Joskow, P. and J. Tirole (2005) "Merchant transmission investment*," **The Journal of Industrial Economics**, **53**(2), pp. 233–264.

[80] Alguacil, N., A. L. Motto, and A. J. Conejo (2003) "Transmission expansion planning: A mixed-integer LP approach," **IEEE Transactions on Power Systems**, **18**(3), pp. 1070–1077.

[81] Samarakoon, H., R. Shrestha, and O. Fujiwara (2001) "A mixed integer linear programming model for transmission expansion planning with generation location selection," **International Journal of Electrical Power & Energy Systems**, **23**(4), pp. 285–293.

[82] Bahiense, L., G. C. Oliveira, M. Pereira, and S. Granville (2001) "A mixed integer disjunctive model for transmission network expansion," **IEEE Transactions on Power Systems**, **16**(3), pp. 560–565.

[83] Gallego, R., A. Monticelli, and R. Romero (1998) "Transmision system expansion planning by an extended genetic algorithm," **IEE Proceedings-Generation, Transmission and Distribution**, **145**(3), pp. 329–335.

[84] Romero, R., R. Gallego, and A. Monticelli (1995) "Transmission system expansion planning by simulated annealing," in **Proceedings of Power Industry Computer Applications Conference**, IEEE, pp. 278–283.

[85] Contreras, J. and F. F. Wu (1999) "Coalition formation in transmission expansion planning," **IEEE Transactions on Power Systems**, **14**(3), pp. 1144–1152.

[86] Sauma, E. E. and S. S. Oren (2006) "Proactive planning and valuation of transmission investments in restructured electricity markets," **Journal of Regulatory Economics**, **30**(3), pp. 261–290.

[87] ——— (2007) "Economic criteria for planning transmission investment in restructured electricity markets," **IEEE Transactions on Power Systems**, **22**(4), pp. 1394–1405.

[88] Motamedi, A., H. Zareipour, M. O. Buygi, and W. D. Rosehart (2010) "A transmission planning framework considering future generation expansions in electricity markets," **IEEE Transactions on Power Systems**, **25**(4), pp. 1987–1995.

[89] Pozo, D., E. E. Sauma, and J. Contreras (2013) "A three-level static MILP model for generation and transmission expansion planning," **IEEE Transactions on Power Systems**, **28**(1), pp. 202–210.

[90] Jenabi, M., S. M. T. Fatemi Ghomi, and Y. Smeers (2013) "Bi-level game approaches for coordination of generation and transmission expansion planning within a market environment," **IEEE Transactions on Power Systems**, **28**(3), pp. 2639–2650.

[91] Akbari, T., S. Zolfaghari, and A. Kazemi (2009) "Multi-stage stochastic transmission expansion planning under load uncertainty using benders decomposition," **International Review of Electrical Engineering**, **4**(5).

[92] Dehghan, S., A. Kazemi, and N. Neyestani (2011) "Multistage transmission expansion planning alleviating the level of transmission congestion," in **2011 IEEE Trondheim PowerTech**, IEEE, pp. 1–8.

[93] Rockafellar, R. T. and R. J.-B. Wets (2009) **Variational analysis**, vol. 317, Springer Science & Business Media.

[94] Chen, Y., G. Lan, and Y. Ouyang (2014) "Accelerated schemes for a class of variational inequalities," **Mathematical Programming**, pp. 1–37.

[95] Iusem, A. N., A. Jofré, and P. Thompson (2018) "Incremental constraint projection methods for monotone stochastic variational inequalities," **Mathematics of Operations Research**.

[96] Chen, Y., G. Lan, and Y. Ouyang (2017) "Accelerated schemes for a class of variational inequalities," **Mathematical Programming**, **165**(1), pp. 113–149.

[97] Malitsky, Y. (2015) "Projected Reflected Gradient Methods for Monotone Variational Inequalities," **SIAM Journal on Optimization**, **25**(1), pp. 502–520.

[98] Censor, Y., A. Gibali, and S. Reich (2011) "The subgradient extragradient method for solving variational inequalities in Hilbert space," **Journal of Optimization Theory and Applications**, **148**(2), pp. 318–335.

[99] Robbins, H. and S. Monro (1951) "A stochastic approximation method," **The Annals of Mathematical Statistics**, pp. 400–407.

[100] Yousefian, F., A. Nedić, and U. V. Shanbhag (2013) "A regularized smoothing stochastic approximation (RSSA) algorithm for stochastic variational inequality problems," in **Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World**, IEEE Press, pp. 933–944.

[101] Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009) "Robust stochastic approximation approach to stochastic programming," **SIAM Journal on Optimization**, **19**(4), pp. 1574–1609.

[102] Yousefian, F., A. Nedić, and U. V. Shanbhag (2014) "Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems," in **Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on**, IEEE, pp. 5831–5836.

[103] ——— (2015) "Self-tuned stochastic approximation schemes for non-Lipschitzian stochastic multi-user optimization and Nash games," **IEEE Transactions on Automatic Control**, **61**(7), pp. 1753–1766.

[104] Nemirovski, A. (2004) "Prox-method with rate of convergence O (1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems," **SIAM Journal on Optimization**, **15**(1), pp. 229–251.

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

[106] Kannan, A. and U. V. Shanbhag (2014) "Pseudomonotone Stochastic Variational Inequality Problems: Analysis and Stochastic Approximation Schemes," **arXiv preprint arXiv:1410.1628**.

[107] Wang, M. and D. P. Bertsekas (2015) "Incremental constraint projection methods for variational inequalities," **Mathematical Programming**, **150**(2), pp. 321–363.

[108] Iusem, A., A. Jofré, R. I. Oliveira, and P. Thompson (2017) "Extragradient method with variance reduction for stochastic variational inequalities," **SIAM Journal on Optimization**, **27**(2), pp. 686–724.

[109] Nedić, A. (2010) "Random projection algorithms for convex set intersection problems," in **Decision and Control (CDC), 2010 49th IEEE Conference on**, IEEE, pp. 7655–7660.

[110] ——— (2011) "Random algorithms for convex minimization problems," **Mathematical Programming**, **129**(2), pp. 225–253.

[111] Bertsekas, D. P. (2011) "Incremental proximal methods for large scale convex optimization," **Mathematical Programming**, **129**(2), pp. 163–195.

[112] Wang, M. and D. P. Bertsekas (2013) "Incremental constraint projection-proximal methods for nonsmooth convex optimization," **SIAM Journal on Optimization (to appear)**.

[113] Wang, M., Y. Chen, J. Liu, and Y. Gu (2015) "Random Multi-Constraint Projection: Stochastic Gradient Methods for Convex Optimization with Many Constraints," **arXiv preprint arXiv:1511.03760**.

[114] Bauschke, H. H. and P. L. Combettes (2011) **Convex analysis and monotone operator theory in Hilbert spaces**, Springer Science & Business Media.

[115] Polyak, B. T. (1987) **Introduction to optimization**, Optimization Software New York.

[116] Bertsekas, D. P. (2011) "Temporal difference methods for general projected equations," **IEEE Transactions on Automatic Control**, **56**(9), pp. 2128–2139.

[117] Aumann, R. J. (1965) "Integrals of set-valued functions," **Journal of Mathematical Analysis and Applications**, **12**(1), pp. 1–12.

[118] Glowinski, R. and P. Le Tallec (1989) **Augmented Lagrangian and operator-splitting methods in nonlinear mechanics**, vol. 9, SIAM.

[119] Douglas, J. and H. H. Rachford (1956) "On the numerical solution of heat conduction problems in two and three space variables," **Transactions of the American Mathematical Society**, **82**(2), pp. 421–439.

[120] Dantzig, G. B., A. Orden, P. Wolfe, et al. (1955) "The generalized simplex method for minimizing a linear form under linear inequality restraints," **Pacific Journal of Mathematics**, **5**(2), pp. 183–195.

[121] Beale, E. M. (1955) "On minimizing a convex function subject to linear inequalities," **Journal of the Royal Statistical Society: Series B (Methodological)**, **17**(2), pp. 173–184.

[122] Rockafellar, R. and R. Wets (1976) "Stochastic convex programming: basic duality," **Pacific Journal of Mathematics**, **62**(1), pp. 173–195.

[123] Ben-Tal, A. and M. Teboulle (1986) "Expected utility, penalty functions, and duality in stochastic nonlinear programming," **Management Science**, **32**(11), pp. 1445–1466.

[124] Klein Haneveld, W. K. and M. H. van der Vlerk (1999) "Stochastic integer programming: General models and algorithms," **Annals of Operations Research**, **85**, pp. 39–57.

[125] Ravat, U. and U. V. Shanbhag (2013) "On the existence of solutions to stochastic variational inequality and complementarity problems," **arXiv preprint arXiv:1306.0586**.

[126] Ruszczynski, A. and A. Shapiro (2003) **Stochastic programming (handbooks in operations research and management science)**, vol. 10, Elsevier.

[127] Shapiro, A. and H. Xu (2008) "Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation," **Optimization**, **57**(3), pp. 395–418.

[128] Patrascu, A. and I. Necoara (2017) "Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization," **The Journal of Machine Learning Research**, **18**(1), pp. 7204–7245.

[129] Bianchi, P. (2016) "Ergodic convergence of a stochastic proximal point algorithm," **SIAM Journal on Optimization**, **26**(4), pp. 2235–2260.

[130] Asi, H. and J. C. Duchi (2018) "Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity," **arXiv preprint arXiv:1810.05633**.

[131] Nedić, A. and T. Tatarenko (2017) "On stochastic proximal-point method for convex-composite optimization," in **2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)**, IEEE, pp. 198–203.

[132] Peaceman, D. W. and H. H. Rachford, Jr (1955) "The numerical solution of parabolic and elliptic differential equations," **Journal of the Society for Industrial and Applied Mathematics**, **3**(1), pp. 28–41.

[133] Ruszczyński, A. (1997) "Decomposition methods in stochastic programming," **Mathematical Programming**, **79**(1-3), pp. 333–353.

[134] Corman, E. and X. Yuan (2014) "A generalized proximal point algorithm and its convergence rate," **SIAM Journal on Optimization**, **24**(4), pp. 1614–1638.

[135] He, B. and X. Yuan (2012) "On the O(1/n) Convergence Rate of the Douglas–Rachford Alternating Direction Method," **SIAM Journal on Numerical Analysis**, **50**(2), pp. 700–709.

[136] Rosasco, L., S. Villa, and B. C. Vũ (2016) "Stochastic forward–backward splitting for monotone inclusions," **Journal of Optimization Theory and Applications**, **169**(2), pp. 388–406.

[137] Combettes, P. L. and J.-C. Pesquet (2015) "Stochastic approximations and perturbations in forward-backward splitting for monotone operators," **arXiv preprint arXiv:1507.07095**.

[138] Karampatziakis, N. and J. Langford (2010) "Online importance weight aware updates," **arXiv preprint arXiv:1011.1576**.

[139] Kulis, B. and P. L. Bartlett (2010) "Implicit online learning," in **Proceedings of the 27th International Conference on Machine Learning (ICML-10)**, pp. 575–582.

[140] Cui, S. and U. V. Shanbhag (2019) "On the optimality of single projection variants of extragradient schemes for monotone stochastic variational inequality problems," **arXiv preprint arXiv:1904.11076**.

[141] Brezis, H. (1973) "Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert. Number 5 in North Holland Math," **Studies. North-Holland, Amsterdam**.

[142] Ahmadi, H. (2016) **On the analysis of data-driven and distributed algorithms for convex optimization problems**, The Pennsylvania State University.

[143] Borwein, J. M. and J. Dutta (2016) "Maximal monotone inclusions and Fitzpatrick functions," **Journal of Optimization Theory and Applications**, **171**(3), pp. 757–784.

[144] Borwein, J. M., J. D. Vanderwerff, et al. (2010) **Convex functions: constructions, characterizations and counterexamples**, vol. 109, Cambridge University Press Cambridge.

[145] Fitzpatrick, S. et al. (1988) "Representing monotone operators by convex functions," in **Workshop/Miniconference on Functional Analysis and Optimization**, Centre for Mathematics and its Applications, Mathematical Sciences Institute, The Australian National University, pp. 59–65.

[146] Orfanos, G. A., P. S. Georgilakis, and N. D. Hatziargyriou (2013) "A more fair power flow based transmission cost allocation scheme considering maximum line loading for N-1 security," **IEEE Transactions on Power Systems**, **28**(3), pp. 3344–3352.

[147] Weibelzahl, M. (2017) "Nodal, zonal, or uniform electricity pricing: how to deal with network congestion," **Frontiers in Energy**, **11**(2), pp. 210–232.

[148] Turvey, R. (2017) **Optimal Pricing and Investment in Electricity Supply: An Esay in Applied Welfare Economics**, Routledge.

[149] Schweppe, F., M. Caramanis, R. Tabors, and R. Bohn (2013) **Spot Pricing of Electricity**, Power Electronics and Power Systems, Springer US. URL `https://books.google.com/books?id=QZ8ACAAAQBAJ`

[150] Rubio-Odériz, F. J. and I. J. Perez-Arriaga (2000) "Marginal pricing of transmission services: A comparative analysis of network cost allocation methods," **IEEE Transactions on Power systems**, **15**(1), pp. 448–454.

[151] Krishnan, V., J. Ho, B. F. Hobbs, A. L. Liu, J. D. McCalley, M. Shahidehpour, and Q. P. Zheng (2016) "Co-optimization of electricity transmission and generation resources for planning and policy analysis: review of concepts and modeling approaches," **Energy Systems**, **7**(2), pp. 297–332.

[152] Grid, N. (2015) **Connection and Use of System Code (CUSC) v1.1.10**, **Tech. rep.**, National Grid ESO. URL `https://www.nationalgrideso.com/codes/connection-and-use-system-code-cusc?code-documents`

[153] Shahidehpour, M., H. Yamin, and Z. Li (2002) **Market Operations in Electric Power Systems**, John Wiley & Sons, Inc., New York, USA.

[154] Shirmohammadi, D., B. Gorenstin, M. V. Pereira, et al. (1996) "Some fundamental, technical concepts about cost based transmission pricing," **IEEE Transactions on Power Systems**, **11**(2), pp. 1002–1008.

[155] Orfanos, G. A., P. S. Georgilakis, and N. D. Hatziargyriou (2013) "A more fair power flow based transmission cost allocation scheme considering maximum line loading for N-1 security," **IEEE Transactions on Power Systems**, **28**(3), pp. 3344–3352.

[156] Pan, J., Y. Teklu, S. Rahman, and K. Jun (2000) "Review of usage-based transmission cost allocation methods under open access," **IEEE Transactions on Power Systems**, **15**(4), pp. 1218–1224.

[157] Turvey, R. (1968) **Optimal Pricing and Investment in Electricity Supply: An Essay in Applied Welfare Economics**, MIT Press.

[158] Li, F. and D. L. Tolley (2007) "Long-run incremental cost pricing based on unused capacity," **IEEE Transactions on Power Systems**, **22**(4), pp. 1683–1689.

[159] Kim, B. H. and M. L. Baughman (1997) "The economic efficiency impacts of alternatives for revenue reconciliation," **IEEE Transactions on Power Systems**, **12**(3), pp. 1129–1135.

[160] Cervantes, J. and F. F. Choobineh (2017) "A new allocation methodology for recovering transmission investment cost," in **2017 IEEE Power & Energy Society General Meeting**, IEEE, pp. 1–5.

[161] Tohidi, Y., L. Olmos, M. Rivier, and M. R. Hesamzadeh (2016) "Coordination of generation and transmission development through generation transmission charges—A game theoretical approach," **IEEE Transactions on Power Systems**, **32**(2), pp. 1103–1114.

[162] Bertsch, J., S. Hagspiel, and L. Just (2015) **Congestion management in power systems: Long-term modeling framework and large-scale application**, **Tech. rep.**, EWI Working Paper.

[163] Bravo, D., E. Sauma, J. Contreras, S. de la Torre, J. A. Aguado, and D. Pozo (2016) "Impact of network payment schemes on transmission expansion planning with variable renewable generation," **Energy Economics**, **56**, pp. 410–421.

[164] Strbac, G., C. V. Konstantinidis, R. Moreno, I. Konstantelos, and D. Papadaskalopoulos (2015) "It's all about grids: The importance of transmission pricing and investment coordination in integrating renewables," **IEEE Power and Energy Magazine**, **13**(4), pp. 61–75.

[165] Wang, J., H. Zhong, W. Tang, R. Rajagopal, Q. Xia, and C. Kang (2018) "Tri-Level Expansion Planning for Transmission Networks and Distributed Energy Resources Considering Transmission Cost Allocation," **IEEE Transactions on Sustainable Energy**, **9**(4), pp. 1844–1856.

[166] Li, F., D. Tolley, N. P. Padhy, and J. Wang (2009) "Framework for assessing the economic efficiencies of long-run network pricing models," **IEEE Transactions on Power Systems**, **24**(4), pp. 1641–1648.

[167] Krishnan, V., J. Ho, B. F. Hobbs, A. L. Liu, J. D. McCalley, M. Shahidehpour, and Q. P. Zheng (2016) "Co-optimization of electricity transmission and generation resources for planning and policy analysis: review of concepts and modeling approaches," **Energy Systems**, **7**(2), pp. 297–332.

[168] Weibelzahl, M. and A. Märtz (2018) "Optimal storage and transmission investments in a bilevel electricity market model," **Annals of Operations Research**.

[169] Pozo, D., E. Sauma, and J. Contreras (2017) "Basic theoretical foundations and insights on bilevel models and their applications to power systems," **Annals of Operations Research**, **254**(1-2), pp. 303–334.

[170] Borison, A. B., P. A. Morris, and S. S. Oren (1984) "A state-of-the-world decomposition approach to dynamics and uncertainty in electric utility generation expansion planning," **Operations Research**, **32**(5), pp. 1052–1068.

[171] Eager, D., B. F. Hobbs, and J. W. Bialek (2012) "Dynamic modeling of thermal generation capacity investment: Application to markets with high wind penetration," **IEEE Transactions on Power Systems**, **27**(4), pp. 2127–2137.

[172] Parpas, P. and M. Webster (2014) "A stochastic multiscale model for electricity generation capacity expansion," **European Journal of Operational Research**, **232**(2), pp. 359–374.

[173] Ehrenmann, A. and Y. Smeers (2011) "Generation capacity expansion in a risky environment: a stochastic equilibrium analysis," **Operations Research**, **59**(6), pp. 1332–1346.

[174] de Frutos, M.-A. and N. Fabra (2011) "Endogenous capacities and price competition: The role of demand uncertainty," **International Journal of Industrial Organization**, **29**(4), pp. 399–411.

[175] Murphy, F. H. and Y. Smeers (2005) "Generation capacity expansion in imperfectly competitive restructured electricity markets," **Operations Research**, **53**(4), pp. 646–661.

[176] Fan, L., B. F. Hobbs, and C. S. Norman (2010) "Risk aversion and CO2 regulatory uncertainty in power generation investment: Policy and modeling implications," **Journal of Environmental Economics and Management**, **60**(3), pp. 193–208.

[177] Ben-Tal, A., L. El Ghaoui, and A. Nemirovski (2009) **Robust optimization**, Princeton University Press.

[178] Ben-Tal, A., A. Goryashko, E. Guslitzer, and A. Nemirovski (2004) "Adjustable robust solutions of uncertain linear programs," **Mathematical Programming**, **99**(2), pp. 351–376.

[179] Burer, S. and A. N. Letchford (2012) "Non-convex mixed-integer nonlinear programming: A survey," **Surveys in Operations Research and Management Science**, **17**(2), pp. 97–106.

[180] Tawarmalani, M., N. V. Sahinidis, and N. Sahinidis (2002) **Convexification and global optimization in continuous and mixed-integer nonlinear programming: theory, algorithms, software, and applications**, vol. 65, Springer Science & Business Media.

[181] Beale, E. M. (1968) **Mathematical programming in practice**, Pitman.

[182] McCormick, G. P. (1976) "Computability of global solutions to factorable nonconvex programs: Part I - Convex underestimating problems," **Mathematical Programming**, **10**(1), pp. 147–175.

[183] Sahinidis, N. V. (1996) "BARON: A general purpose global optimization software package," **Journal of Global Optimization**, **8**(2), pp. 201–205.

[184] Czyzyk, J., M. P. Mesnier, and J. J. Moré (1998) "The NEOS server," **IEEE Computational Science and Engineering**, **5**(3), pp. 68–75.

[185] Dolan, E. D. (2001) **The NEOS Server 4.0 Administrative Guide**, **Technical Memorandum ANL/MCS-TM-250**, Mathematics and Computer Science Division, Argonne National Laboratory.

[186] Gropp, W. and J. J. Moré (1997) "Optimization Environments and the NEOS Server," in **Approximation Theory and Optimization** (M. D. Buhman and A. Iserles, eds.), Cambridge University Press, pp. 167 – 182.

[187] EIA (2016) **Levelized cost and levelized avoided cost of new generation resources in the annual energy outlook**, **Tech. rep.**, EIA.

[188] ——— (2017) **Electric Power Annual 2016**, **Tech. rep.**, EIA.

[189] Pletka, R., J. Khangura, A. Rawlins, E. Waldren, and D. Wilson (2014) **Capital costs for transmission and substations**, **Tech. rep.**, Western Electricity Coordinating Council.

Vita

Shisheng Cui

Shisheng Cui was born and raised in Harbin, the provincial capital of Heilongjiang, China. He attended Fuhua primary school and later attended Songlei junior middle school. Thereafter, he was chosen to attend Harbin No. 3 high school. Shisheng continued his education at Tsinghua University of where he pursued Control Engineering in 2005. At Tsinghua University, he worked under the mentorship of Professor Mingguo Zhao studying the passive dynamical walking. He graduated in 2009, receiving a Bachelor of Engineering in Control Engineering. Shortly after graduation, Shisheng began graduate school in September of 2009 at Stanford University, majoring in Computer Science with specialization in Artificial Intelligence. He received a Master of Science in 2011. After working in industry for about two years, Shisheng headed to The Pennsylvania State University and began working under the tutelage of Professor Chia-Jung Chang in August of 2013 developing schemes for regression problems. From 2015, he has been supervised by Professor Uday V. Shanbhag working on stochastic optimization algorithms and applications.