Asymptotic results of Stochastic Decomposition for Two-stage Stochastic Quadratic Programming

Junyi Liu * Suvrajeet Sen *

October 15, 2019

Abstract

This paper presents stochastic decomposition (SD) algorithms for two classes of stochastic programming problems: 1) two-stage stochastic quadratic-linear programming (SQLP) in which a quadratic program defines the objective function in the first stage and a linear program defines the value function in the second stage; 2) two-stage stochastic quadratic-quadratic programming (SQQP) which has quadratic programming problems in both stages. Similar to their stochastic linear programming (SLP) predecessor, these iterative schemes in SD approximate the objective function using piecewise affine/quadratic minorants and then apply a stochastic proximal mapping to obtain the next iterate. In this paper we show that under some assumptions, the proximal mapping applied in SD obeys a contraction mapping property even though the approximations are based on sequential random samples. Following that, we demonstrate that under those assumptions, SD can provide a sequence of solutions converging to the optimal solution with the sublinear convergence rate in both SQLP and SQQP problems. Finally, we present an "in-sample" stopping rule to assess the optimality gap by constructing consistent bootstrap estimators.

1 Introduction

Stochastic programming (SP) deals with a class of optimization models and algorithms in which the distribution of the uncertainty plays a significant role. One of the standard SP models is the two-stage SLP in which the first-stage decision is made prior to observing the uncertainty, and the second-stage recourse decision is undertaken so as to adapt to the observation in an optimal manner. For instance, in a power grid with significant variable energy resources (e.g., wind and solar), thermal generators are scheduled ahead of time. As the wind and solar generation become available, some fast-ramping gas generators are deployed to accommodate wind and solar variability. Similar models arise in many other applications including financial planning, stochastic routing and others.

In general, SP provides a formal approach to measure the performance of decisions under uncertainty. Measures of the performance in SP includes the most standard measure of the optimization in expectation and the currently popular measure of distributionally robust optimization which accommodates not only the uncertainty but also errors in estimating its distribution. In some sense, the choice of measure implies the attitude of decision makers towards risk in the area of decision theory. For example, risk-sensitive decision makers adopt certain "risk/return" trade-off models (also known as mean-risk models), whereas, highly risk-averse decision makers are more likely to control the worst-case performance.

*University of Southern California, Los Angeles, CA(junyiliu@usc.edu, s.sen@usc.edu).

1.1 Stochastic methods

In this paper, we consider the standard SP objective of minimizing the expectation of a random cost function which is defined as the sum of a quadratic function and a value function of the second-stage optimization problem. Since the random variable might have a large or even continuous probability space, it may be impossible to identify a deterministic optimal solution. To address this issue, there are at least two classical methods based on Monte-Carlo sampling techniques, namely, Sample Average Approximation (SAA) and Stochastic Approximation (SA). These sampling-based methods are also the mainstay algorithms to estimate the parameters in machine learning (ML) models. Similar to Empirical Risk Minimization (ERM) in machine learning, SAA generates an approximating mathematical model using sample average. SAA methods have been studied extensively for both for convex and nonconvex stochastic programs ([1], [12], [21], [29] ,[33]). However, SAA methods require a numerical algorithm to solve the deterministic optimization problem for a finite sample size. So in large-scale problems, the SAA approach can be computationally expensive when a sequence of sample sizes need to be explored with restarts of the numerical algorithm for each sample size.

On the other hand, an SA method is computationally implementable and tractable ([5]). It has a long history tracing back to the pioneering work of Robbins and Monroe in [22]. In the context of a non-smooth optimization problem, SA is a stochastic subgradient method which updates a solution in a direction opposite the subgradient at each iteration. Early versions of the SA method required a fair amount of parameter tuning, and even so it was not particularly reliable. Modern extensions of SA methods are based on incorporating a "distance generating" function as in the mirror-descent method ([18]). More recently, the work by A. Nemirovski et al. in [17] presents a robust SA method which is less sensitive to parameter choices.

Unlike SA and SAA methods for general convex problems, the SD algorithm was first developed in [8] by taking the advantage of piecewise linear structures of two-stage stochastic linear programming(SLP). SD inherits the convergence result to the optimum [27] while maintaining an iteration complexity which is slightly greater than SA. As reported in the empirical experiments in [27], there has been significant progress of SD in terms of computational efficiency and statistical accuracy for a fairly large battery of test instances. In this paper, we focus on two classes of problems: 1. two-stage stochastic quadratic-linear programming (SQLP) problems with quadratic programming in the first stage and linear programming in the second stage; 2. two-stage stochastic quadratic-quadratic programming (SQQP) problems with quadratic programming in both stages. Through straightforward analysis, the SD algorithm as well as the convergence result will be extended to two-stage SQLP and SQQP problems. This paper is intended to provide the mathematical support via a study of convergence rates and stopping rules for SD algorithms.

1.2 Convergence properties and our contributions

There is a vast literature on convergence rate analysis pertaining to SAA and SA methods. When SAA methods are applied for a standard stochastic program with a sample set of size N corresponding to i.i.d observations of some random variable, it can provide ε-optimal solution with probability greater than $1 - O(C_\varepsilon e^{-\beta_\varepsilon N})$ where $\varepsilon > 0$ (see [30]). Recently, the almost-sure convergence rate is shown in [1] for SAA methods. For convex stochastic programming, under the assumption of the sharpness of the solution set, the solution provided by SAA converges to an optimal solution with probability $1 - O(C_0 e^{-\beta_0 N})$ (see [28] and [29]). In addition, the modified robust SA method can provide a solution sequence of which the objective value in a convex stochastic program converges in expectation at $O(N^{-1/2})$ ([17]). With assumptions of strong convexity and Lipschitz gradient,

the classical SA method provides a solution sequence which asymptotically converges in expectation at $O(N^{-1/2})$ and the objective value of a solution sequence converges in expectation at $O(N^{-1})$. It has been shown that in two-stage SLP, SD gives rise to convergence to an optimal solution with statistically verifiable bounds almost surely (see [8],[9],[27]). The lack of study on the asymptotic convergence rate of SD motivates our work in this paper. Specifically, we analyze the contraction property of the stochastic proximal mapping with constraints and demonstrate that for two-stage SQLP/SQQP problems, SD with a sequence of diminishing step sizes can provide a solution sequence which converges to the optimum in expectation at rate of $O(N^{-1})$ under several assumptions at the optimum solution. However, the SA method does not highlight the contraction property in the proximal iteration of SD, which distinguishes the convergence analysis of SD in the present paper from SA methods in [17].

One should notice that convergence rates of these stochastic methods are analyzed in different but not equivalent scales, such as expectation of the distance to the optimal solution set or the optimal value. For convex problems, the convergence of the expected distance to the optimal solution set implies the convergence in expectation in the objective value, thus the former type of convergence is stronger and more stable than the latter. section 1.2 is a summary of currently known results on the convergence rate for SAA, SA methods and our result for the SD method. The second column in section 1.2 refers to the broad class of problems addressed by each method. Let x^*, S^*, f^* denote the optimal solution, the optimal solution set, the optimal value respectively and let ε be a positive constant. The last two columns present different types of convergence rates. The rates in section 1.2 are not comparable since each method requires additional assumptions (eg, SAA needs the existence of several moments of the random objective function). For SD methods, we require additional assumptions pertaining to the structure of two-stage SP, e.g., a linear growth condition which originates from the SAA approximation in the SD method.

Table 1: The summary of convergence results of the stochastic methods for SP problems

Method	SP problem class	Convergence type	Rate
SAA[30]	Non-convex	$P(f(x^N) - f^* > \varepsilon)$	$O(C_\varepsilon e^{-\beta_\varepsilon N})$
SAA[28]	convex	$P(x^N \notin S^*)$	$O(C_0 e^{-\beta_0 N})$
robust SA[17]	convex	$\mathbf{E}\ \lvert f(x^N) - f^*\rvert$	$O(N^{-1/2})$
SA[17]	strongly convex	$\mathbf{E}\ \lVert x^N - x^*\rVert$	$O(N^{-1/2})$
	Lipschitz gradient	$\mathbf{E}\ \lvert f(x^N) - f^*\rvert$	$O(N^{-1})$
SD	two-stage SQLP/SQQP strongly convex	$\mathbf{E}\ \lVert x^N - x^*\rVert$	$O(N^{-1})$

The rest of this paper is organized as follows. In section 2 we start with the setup of two-stage SQLP and SQQP problems as well as some standing assumptions for these models. In section 3, motivated by a standard SAA process, we introduce the SD algorithm for SQLP and SQQP together with its statistical convergence result. In section 4 we present the convergence rate of SD in two-stage SQLP/SQQP problems. In section 5, we design the "in-sample" stopping rule by constructing consistent bootstrap estimators for the optimality gap. Finally a discussion of our contributions and possible extensions are presented in section 6.

2 Problem Setup

We clarify the notations which are used in the paper. Because we are in the convex optimization world, there is no loss of generality assuming that all subdifferentials coincide. For any convex function $g(x)$ we thus write $\partial g(x)$ to denote the subdifferential of g at x. Moreover, let $[a]_+ \triangleq \max\{a, 0\}$ for any $a \in \mathbb{R}$. For a positive integer m, let $[m]$ denote the set $\{1, \ldots, m\}$. For any vector $d \in \mathbb{R}^m$, let $d_{(i)}$ denote the ith component of d for $i \in [m]$. Let $\|\cdot\|$ denote the ℓ_2 norm of a vector without the subscript since we only consider the ℓ_2 norm in this paper. For a square matrix B, $\theta_{min}(B)$ and $\theta_{max}(B)$ denote the smallest and the largest eigenvalues of B respectively.
In addition, the projection operator onto a set $X \subseteq \mathbb{R}^n$ is defined by

$$P_X(x) \triangleq \underset{z \in X}{\operatorname{argmin}} \ \|z - x\|^2, \tag{1}$$

and the proximal mapping point (see Rockafellar [23]) with respect to a convex function g is defined by

$$T_{\alpha g}^X(x) \triangleq \underset{z \in X}{\operatorname{argmin}} \left\{ \alpha g(z) + \frac{1}{2} \|z - x\|^2 \right\}. \tag{2}$$

In this paper, we may drop the superscript X of the proximal mapping for simplicity in some cases without causing confusion.

2.1 Two-stage stochastic quadratic programming

In the mathematical formulations of two-stage SQLP problems and two-stage SQQP problems, we use the subscript QL and QQ to identify SQLP and SQQP problems respectively. In two classes of problems, let x and y respectively denote the first-stage and the second-stage decision variables with x belonging to the set $X \subseteq \mathbb{R}^{n_1}$ and y belonging to a polyhedron in $\mathbb{R}^{n_2}$.
The mathematical formulation of a two-stage SQLP is given below.

$$\begin{array}{ll} \text{minimize} & f_{QL}(x) \triangleq \dfrac{1}{2} x^\top Q x + c^\top x + \mathbf{E}\left[h_{QL}(x, \widetilde{\omega})\right] \\ \text{subject to} & x \in X = \{x : Ax \leq b\} \subseteq \mathbb{R}^{n_1}, \end{array} \tag{3}$$

where the recourse function h_{QL} is defined as,

$$\begin{array}{rll} h_{QL}(x, \omega) \triangleq \text{minimum} & & d^\top y \\ \text{subject to} & & Dy = \xi(\omega) - C(\omega)x \\ & & y \geq 0, \ y \in \mathbb{R}^{n_2}. \end{array} \tag{4}$$

Here $A \in \mathbb{R}^{m_1 \times n_1}$ is a deterministic matrix with row vectors denoted by $\{a_i\}_{i=1}^{m_1}$ and $D \in \mathbb{R}^{m_2 \times n_2}$ is a deterministic matrix. In addition, $\widetilde{\omega}$ denotes a (vector) random variable in a probability space $(\Omega, \mathcal{F}, P)$ with Ω being the sample space, $\mathcal{F}$ being the σ-algebra generated by subsets of Ω and P being a probability measure defined on $\mathcal{F}$. Then, $\xi(\widetilde{\omega})$ denotes a random vector, $C(\widetilde{\omega})$ denotes a random matrix, and $\mathbf{E}\left[\cdot\right]$ denotes the expectation with respect to the probability measure of $\widetilde{\omega}$. Moreover, we use ω to denote an observation of the random variable $\widetilde{\omega}$. To be consistent with previous stochastic decomposition algorithms, we assume that the second-stage cost vector d is fixed. If $Q = \mathbf{0}$, then (3) becomes the general two-stage stochastic linear programming (SLP) problem. However, here we assume Q to be a positive definite matrix in SQLP/SQQP problems.
The definition of the recourse function in (4) shows that once the first-stage decision x is determined and an outcome of the random variable ω is observed, h_{QL} is the optimal value of a linear optimization problem. This optimal value reflects the cost associated with adapting to the information revealed through an outcome ω. Nevertheless, the first-stage decision x must be chosen before

the randomness $\widetilde{\omega}$ is realized. So the "cost-to-go" function is evaluated by its expectation in the first-stage objective. It is also interesting to note that ordinary support vector machines (SVM) in ML are simple two-stage SQLP problems. Here, the first stage decision (x) gives the weights of the SVM, the quadratic term is derived from the regularizer of the SVM, and the second stage denotes the penalty term for being on the "wrong" side of the separating hyperplane. In the terminology of SP, this penalty formulation is the so-called "simple recourse" model of SP.
As for a two-stage SQQP, we introduce a quadratic objective function in the recourse program and the SQQP problem is defined below.

$$\begin{array}{ll} \text{minimize} & f_{QQ}(x) \triangleq \dfrac{1}{2}x^\top Qx + c^\top x + \mathbf{E}\left[h_{QQ}(x,\widetilde{\omega})\right] \\ \text{subject to} & x \in X = \{x : Ax \leq b\} \subseteq \mathbb{R}^{n_1}, \end{array} \tag{5}$$

where h_{QQ} is the value function of a quadratic program defined as follows,

$$\begin{array}{rll} h_{QQ}(x,\omega) \triangleq \text{minimum} & \dfrac{1}{2}y^\top Py + d^\top y \\ \text{subject to} & Dy = \xi(\omega) - C(\omega)x \\ & y \geq 0,\ y \in \mathbb{R}^{n_2}. \end{array} \tag{6}$$

The notations are the same as in the two-stage SQLP problem except that the matrix $P \in \mathbb{R}^{n_2 \times n_2}$ is a positive definite matrix. We make the assumptions for two-stage SQLP and SQQP problems below.

(A1) Q and P are symmetric and positive definite matrices.

(A2) The set X is convex and compact. The outcome set Ω is compact.

(A3) The second-stage problem satisfies the relatively complete recourse property, i.e. the recourse function $h_{QL}(x,\omega)$ and $h_{QQ}(x,\omega)$ are finite for all $x \in X$ and almost every $\omega \in \Omega$.

It is appropriate to comment on the nature of these assumptions. In $(A1)$, since the square matrix Q is assumed to be positive definite, the SQLP/SQQP (3) and (5) are convex problems. In addition, the matrix P is assumed to be positive definite because of the dual constructions in the SD algorithm. However, this assumption could be relaxed to positive semi-definiteness using a hybrid scheme between the SD algorithms of SQLP and SQQP. The assumption $(A2)$ follows the previous work since we focus on the convex problems here. Moreover, the assumption $(A3)$ means that the recourse function achieves the optimum at one of extreme points or faces of the dual LP/QP in the second stage for any (x,ω) pair almost surely.

2.2 Approximating the second-stage problem

One of the more demanding aspects of an SP model is the need to convey the impact of uncertainty in the recourse function on decisions of the first stage. In order to do so, it is the best to take advantage of the structure of the recourse function, i.e. the value function of a linear or quadratic program. Based on the dual form of the recourse function, SD approximates the sample average of recourse functions by using a collection of piecewise linear or quadratic functions. These families of functions are updated sequentially so that SD is able to discover the structure of the recourse function as the sequential Monte Carlo sampling scheme proceeds. This particular way of using dual approximations of the recourse functions allows SD to create more accurate approximations in areas of the feasible set where the algorithm tends to visit.

We first present the recourse function $h_{QL}(x,\omega)$ as the optimal value of a dual problem.

$$\begin{aligned} h_{QL}(x,\omega) = \text{maximize} \quad & \pi^\top(\xi(\omega) - C(\omega)x) \\ \text{subject to} \quad & \pi \in \Pi = \{\pi : D^\top \pi \leq d\}. \end{aligned} \tag{7}$$

Let $\{\pi^1, \pi^2, \ldots, \pi^l\}$ denote the extreme points in the polyhedron Π. Since h_{QL} is finite almost surely for any solution x and realization ω, the optimal value of the dual problem could be achieved at one of the extreme points almost surely. Therefore

$$h_{QL}(x,\omega) = \max_{i=1,\ldots,l} \quad (\pi^i)^\top(\xi(\omega) - C(\omega)x). \tag{8}$$

Consequently given any observation ω, $h_{QL}(\cdot,\omega)$ is the maximum of l affine functions, thus a piecewise linear function with respect to x.
The recourse function h_{QQ} is a Type III quadratic program according to [7]. Since P is positive definite, by linear transformations, we derive the dual representation ([7]) as follows.

$$\begin{aligned} h_{QQ}(x,\omega) = \text{maximize} \quad & g_{QQ}(t,s;x,\omega) := -\frac{1}{2}(-d + D^\top t + s)^\top P^{-1}(-d + D^\top t + s) \\ & \qquad + (\xi(\omega) - C(\omega)x)^\top t \\ \text{subject to} \quad & s \in \mathbb{R}^{n_2},\ s \geq 0 \\ & t \in \mathbb{R}^{m_2},\ t \text{ is free} \end{aligned} \tag{9}$$

When D has full row rank, $DP^{-1}D^\top$ is invertible. Then given t is free while maximizing g_{QQ}, by eliminating t via unconstrained optimality, we derive the non-negative quadratic programming (NNQP) in (10).

$$\begin{aligned} h_{QQ}(x,\omega) = \quad & \text{maximize} \quad & -\tfrac{1}{2}s^\top H s + e(x,\omega)^\top s \\ & \text{subject to} \quad & s \in \mathbb{R}^{n_2}, s \geq 0 \end{aligned} \tag{10}$$

where

$$\begin{aligned} & M = DP^{-1/2}, \quad H = P^{-1/2}(I - M^\top(MM^\top)^{-1}M)P^{-1/2}, \\ & e(x,\omega) = Hd - P^{-1/2}(MM^\top)^{-1}(\xi(\omega) - C(\omega)x). \end{aligned}$$

Both two dual formulations (9) and (10) are convex quadratic programs with the cost vector linearly parameterized in x. Since the constraint sets are polyhedra with finitely many faces, the value function $h_{QQ}(\cdot,\omega)$ are convex piecewise quadratic function in x.
In the SD approach, approximations are built from the dual of the second-stage linear or quadratic optimization problems. As a result, asymptotic convergence analysis of SD relies on approximating piecewise linear or quadratic structures of the recourse functions. Because of the polyhedral nature of the structure that randomness only appears on the right hand side of constraints in the second stage, pertinent faces of dual problems remain fixed and finite both in SQLP and SQQP. As these dual polyhedra are shared by all scenarios, dual approximation schemes could be used in SD in a manner that exploits such commonality across scenarios and maintains a list of dual faces visited by the algorithm.

3 Stochastic Decomposition (SD) Algorithm

If we have access to the distribution of random vectors, we could solve two-stage SQLP and SQQP as deterministic quadratic programming problems. However, in most practical cases when the scenario space is potentially very large, or the random variable has a continuous but unknown distribution, it is impossible to find the exact optimum. Thus, sample-based methods are common ways to estimate the true objective by sampled-based approximations. In this section, we first present the Sample Average Approximation (SAA) approach with its convergence result. Then motivated by SAA, we design the Stochastic Decomposition (SD) algorithms for two-stage SQLP and SQQP problems.

3.1 Motivation underlying SD

An SAA instance with N samples $\{\omega^i\}_{i=1}^N$ in a two-stage SQLP/SQQP problem can be formulated as follows.

$$\begin{aligned} \text{minimize} \quad & F_N(x) \triangleq \frac{1}{2}x^\top Qx + c^\top x + \frac{1}{N}\sum_{i=1}^{N} h(x,\omega^i) \\ \text{subject to} \quad & x \in X = \{x : Ax \leq b\} \subseteq \mathbb{R}^{n_1}. \end{aligned} \tag{11}$$

In (11), an SAA function denoted by $F_N(x)$, is defined to be the sum of the quadratic term and the sample average of the recourse functions. The recourse function should be $h_{QL}(x,\omega^i)$ or $h_{QQ}(x,\omega^i)$ respectively in two-stage SQLP and SQQP problems. We present the computational SAA process in table 2 following the setup in [27] by Sen and Liu.

Table 2: Computational SAA Process

For a fixed number of replications M, do the following.

1. **Optimization**: For each replication $m \in [M]$ with the sample set of size N, create an SAA function $F_N^m(x)$ where the superscript represents the replication number. Solve the SAA instance (11) by a numerical optimization algorithm with an optimal solution $\hat{x}_N^m$ and the corresponding optimal value $\hat{v}_N^m$.

2. **Statistical Validation**: Estimate the lower bound confidence interval using the optimal values $\{\hat{v}_N^m\}_{m=1}^M$ and the upper bound confidence interval using the potential solutions $\{\hat{x}_N^m\}_{m=1}^M$ at a specified level of accuracy.

3. **Pessimistic Gap**: If the upper end of the estimated upper bound confidence interval is close to the lower end of the estimated lower bound confidence interval at an acceptable level of accuracy, then stop. Else, increase the sample size N and repeat from Step 1.

In step 2, the estimated confidence in the statistical validation step follows the work of Mak et al. in [15]. Following that, in step 3 a Pessimistic Gap, the worst case gap of estimated confidence intervals, evaluates the quality of sampling-based approximation approach. Since the solution $\hat{x}^N$ of the SAA instance (11) is a random variable with respect to the samples $\{\omega^i\}_{i=1}^N$, from a theoretical viewpoint, we consider the probability of $\hat{x}^N$ being in the ε-optimal solution set in terms of the number of samples N. Under some assumptions, here is an extended result of the convergence rate of SAA by A. Shapiro et al. in [28] and [29].

Proposition 1 (Convergence rate of SAA)**.** *In SQLP and SQQP problems* (3) *and* (5)*, besides the assumptions* $(A1)-(A3)$*, we assume the moment conditions* $(M4)$ *and* $(M5)$ *in [28] and that the optimal solution set* S *is nonempty. Let* S_ε *denote the set of* ε*-optimal solutions of the problem considered. Then in each problem (SQLP/SQQP), for any* $\varepsilon > 0$ *there exist constants* $C_\varepsilon > 0$ *and* $\beta_\varepsilon > 0$ *such that the inequality*

$$1 - P(\hat{x}^N \in S_\varepsilon) \leq C_\varepsilon e^{-\beta_\varepsilon N} \tag{12}$$

holds for all $N \in \mathbb{N}$*. Moreover, if the optimal solution set* S *is sharp, then we have*

$$1 - P(\hat{x}^N \in S) \leq C_0 e^{-\beta_0 N}. \tag{13}$$

The probability here is with respect to N *iid samples and the sharpness of the solution set is defined in Definition 5.22 [28].*

The above proposition shows that the solution provided by solving the SAA instance is in the suboptimal set with exponentially increasing probability as the sample size increases. Besides, for smaller value of ε the exponential constant β_ε is smaller.

3.2 The SD algorithm for two-stage SQLP and SQQP

Although the SAA method exhibits a linear convergence rate to the suboptimal solution set, there are some algorithmic issues when actually implementing SAA for large scale instances: a) The computational effort in solving the SAA instance using a numerical QP solver might be very expensive in a large-scale problem, b) the sample size recommended by the theory is often too conservative for practical purposes. Moreover, many of the parameters required to estimate the sample size (e.g., Lipschitz constants) are unknown before the iterations begin. As a result, it is customary to use multiple trials of sample sizes which leads to a large increase of computational time, and c) there is a lack of the integration of statistical stopping and numerical optimization, so that approximations created during the course of certain algorithms (e.g., Benders' decomposition) cannot be used in subsequent runs. These issues are mainly because SAA does not explicitly specify an algorithmic mechanism to exploit the structure of recourse functions. Hence, the SD algorithm is designed to accommodate these issues, while maintaining approximations in the spirit of the SAA approach.
For two-stage stochastic linear programming (SLP) problems, Higle and Sen in [8] designed the Stochastic Decomposition (SD) algorithm to accommodate the structure of SLP. In SD, the objective function is approximated by a bundle of stochastic minorants using approximate subgradients of the sample average of recourse functions. In order to make use of the minorants during the process for subsequent runs, a unified lower bound of recourse functions is required for all $x \in X$ and almost every $\omega \in \Omega$. Without loss of generality, we borrow the non-negativity assumption for the recourse functions which was originally made in [8].

$(A4)$ The recourse function $h(x, \omega)$ is non-negative for all $x \in X$ and almost every $\omega \in \Omega$.

Because the context (SQLP or SQQP) should be clear from the problem statement, we use one notation h for the recourse function in the assumption $(A4)$ without subscripts. Incidentally, it is interesting to note that the assumption of non-negative recourse functions holds for most stochastic optimization models in machine learning since loss functions are usually non-negative.
We present the SD algorithm for two-stage SQLP problems in Algorithm 1 with minor modifications of its predecessor, the SD algorithm for two-stage SLP problems in [27]. The idea is that we successively create a sequence of value function approximations $\{f_k\}$, each of which is a piecewise lower bound of the sample average function. The earliest version of the SD algorithm in [8]

obtains the solution sequence by minimizing the value function approximations $\{f_k\}$ successively. Subsequently, proximity control was added with the step size as a regularization term by Higle and Sen in [9]. Its interpretation as a stochastic proximal mapping, similar to its deterministic version (Boyd and Parikh [20]) leads to a more succinct algorithmic statement. Therefore, at iteration k, the updated solution x^k is computed as a proximal mapping point $T_{\alpha_{k-1}f_{k-1}}(\hat{x}^{k-1})$ where α_{k-1} is the step size, f_{k-1} is a value function approximation and $\hat{x}^{k-1}$ is an incumbent solution, all of which are constructed in the previous iteration. It is worth noticing that one might view SA as a stochastic forward Euler method, whereas such the stochastic proximal algorithm can be viewed as a stochastic backward Euler method. However, it is different from the stochastic proximal iteration (SPI) algorithm proposed by Ryu and Boyd in [25] because the value function approximation is created using the sample information of the entire history.

With the newly updated solution x^k, we now explain the construction of the value function approximation f_k in detail. We first generate a new sample ω^k independently of samples $\{\omega^i\}_{i=1}^{k-1}$ in the history. Because of the fixed recourse assumption, all realizations $\omega \in \Omega$ share the same collection of dual extreme points. We solve (7) at (x^k, ω^k) and include any new dual optimal vector supporting $h(\cdot, \omega^k)$ at x^k in the set $V_k \subseteq \Pi$, so the extreme points will be discovered during the course of SD. We then create two minorants of the SAA function which are defined in line 8 in Algorithm 1. In the notation of these two minorants, $h_k^k(x)$ and $h_{-k}^k(x)$, the superscripts refer to the iteration number and the subscripts refer to the index number of the minorants. Notice that these two minorants are constructed using the true subgradient for the sample ω^k and the subgradient approximations for the samples $\{\omega^i\}_{i=1}^{k-1}$ by restricting the dual variables in the second-stage program in the set V_k. Moreover, we keep those previously generated minorants which have nonzero Lagrangian multipliers at $T_{\alpha_{k-1}f_{k-1}}(\hat{x}^{k-1})$ and reweigh them such that they are the lower bounds of the SAA function F_k as well. From the optimality condition of $T_{\alpha_{k-1}f_{k-1}}(\hat{x}^{k-1})$, Higle and Sen in [9] have shown that at each iteration there exist Lagrangian multipliers such that only $n_1 + 1$ of them are positive where n_1 is the dimension of the first-stage decision variable. Therefore, with two newly constructed minorants, the total number of minorants kept in the index set $\mathcal{J}_k$ is no greater than n_1+3. These minorants are called sample average subgradient approximation (SASA) functions in SQLP problems. Consequently, a value function approximation f_k is created to be the sum of the quadratic function in the first stage and the maximum of the SASA functions recorded in $\mathcal{J}_k$ which is formally defined at line 9 in Algorithm 1.

Besides the construction of value function approximations in the SD algorithm, there are three more improvements on efficiency compared to the SAA method. First, to achieve the estimated improvements in objective, we introduce a competition between the current incumbent solution $\hat{x}^{k-1}$ and the newly updated solution x^k, which is called a candidate solution. The ratio between the reductions $f_k(x^k) - f_k(\hat{x}^k)$ and $f_{k-1}(x^k) - f_{k-1}(\hat{x}^k)$ decides whether we should accept the candidate solution, i.e., $\hat{x}^k = x^k$ or proceed to the next iteration with the unchanged incumbent solution, i.e. $\hat{x}^k = \hat{x}^{k-1}$. Note that the estimates $f_k(x^k)$ and $f_k(\hat{x}^k)$ are positively correlated, so as in the simulation-optimization literature, this correlation reduces variance in estimation of the difference $F_k(x^k) - F_k(\hat{x}^k)$. Moreover, this update rule is also closely related to the traditional update rule in trust region methods (see Conn et al. [6]) except that the functions used in measuring reductions are value function approximations instead of the true function values. Such interplay between optimization and statistical features makes SD unique in its design.

Another improvement is that instead of having a fixed sample size N as in SAA, an in-sample stopping rule with bootstrap estimators is presented in section 5 so that the sampling process stops automatically during the optimization procedure. Finally, we mention that SD also facilitates an additional step of variance reduction via replications. This replication mechanism, which was presented in Sen and Liu in [27], recommends optimizing a "grand mean" value function based on

the terminal value functions from each replication. Since the results of this paper are restricted to only one replication, we refer the readers to [27] for further details.

Algorithm 1 SD-QL

Require:

$\tau > 0,\ \alpha_0 = \tau,\ r \in (0,1),\ k = 0,\ V_0 = \emptyset$ and $\mathcal{J}_0 = \{0\}$.

a feasible solution $\hat{x}^0 \in X$.

$h_{j,0}(x) = 0, \forall x \in X$ and $\forall j \in \mathcal{J}_0$.

$f_0(x) = \frac{1}{2}x^\top Qx + c^\top x + \max\{h_{j,0}(x), j \in \mathcal{J}_0\}, \forall x \in X$.

Ensure:

while the stopping rule (see section 5) is not satisfied **do**
  $k \leftarrow k + 1$
  compute $x^k = T^X_{\alpha_{k-1} f_{k-1}}(\hat{x}^{k-1})$ following the definition in (2) and record the Lagrangian multipliers $\{\mu_j^{k-1}\}_{j \in \mathcal{J}_{k-1}}$ of $\{h_j^{k-1}(x)\}_{j \in \mathcal{J}_{k-1}}$ respectively
  compute $\widetilde{\pi}_k = \text{argmax}_{\pi \in \Pi}\ \pi^\top(\xi(\omega^k) - C(\omega^k)x^k)$ with a new sample ω^k
  update $V_k = V_{k-1} \bigcup \{\widetilde{\pi}_k\}$ and $\mathcal{J}_k = \mathcal{J}_{k-1} \bigcup \{-k, k\} \backslash \{j \in \mathcal{J}_{k-1} : \mu_j^{k-1} = 0\}$
  compute $\pi_{k,i} \in \text{argmax}_{\pi \in V_k}\ \pi^\top(\xi(\omega^i) - C(\omega^i)x^k)$ for $i \in [k]$
  compute $\hat{\pi}_{k,i} \in \text{argmax}_{\pi \in V_k}\ \pi^\top(\xi(\omega^i) - C(\omega^i)\hat{x}^{k-1})$ for $i \in [k]$
  construct $h_k^k(x) = \frac{1}{k}\sum_{i=1}^k \pi_{k,i}^\top(\xi(\omega^i) - C(\omega^i)x)$,
    $h_{-k}^k(x) = \frac{1}{k}\sum_{i=1}^k \hat{\pi}_{k,i}^\top(\xi(\omega^i) - C(\omega^i)x)$,
    $h_j^k(x) = \frac{|j|}{k}\ h_j^{|j|}(x) \quad \text{for } j \in \mathcal{J}_k \backslash \{-k, k\}$
  construct $f_k(x) = \frac{1}{2}x^T Qx + c^T x + \max\{h_j^k(x), j \in \mathcal{J}_k\}$
  if $f_k(x^k) - f_k(\hat{x}^{k-1}) \le r(f_{k-1}(x^k) - f_{k-1}(\hat{x}^{k-1}))$ **then**
    $\hat{x}^k = x^k, \quad \alpha_k = \frac{\tau}{k+1}$
  else
    $\hat{x}^k = \hat{x}^{k-1}, \quad \alpha_k = \alpha_{k-1}$
  end if
end while

As for two-stage SQQP problems, if the matrix P is positive semidefinite, the SD algorithm for SQLP remains valid since SASA functions can be constructed by solving the second-stage quadratic programming (6) with the efficient solvers. However, if the matrix P is positive definite, such curvature of the second-stage problem can be incorporated into the SD algorithm for better performance. When P is positive definite, the dual form (9) given x and ω is a quadratic program with only non-negativity constraints. Methods to solve (9) include active-set methods, iterative algorithms and also some state-of-art algorithms with faster convergence. We refer the interested readers to [2], [3] and [14] for details. Given the wide variety of efficient algorithms currently available, we assume that such a method can be used for solving the dual problem (9). Therefore, we design a method of tracking dual faces of the second-stage problem in SQQP, instead of dual extreme points in SQLP. Specifically, in the SD algorithm for SQQP, we construct a set $U_k = \{\tilde{u}_t\}_{t=1}^k$ where each $\tilde{u}_t$ records indexes of zero components of the dual variable s^t discovered at t-th iteration in (9). Hence the sample average quadratic approximation (SAQA) functions are constructed to approximate the sample average of the recourse functions. Then the value function approximation is the sum of the quadratic function in the first stage and the maximum of the SAQA functions recorded in $\mathcal{J}_k$. With such modifications, we present the SD algorithm for two-stage SQQP problem in Algorithm 2. When D has full row rank, we have the dual formulation (10) of the recourse function h_{QQ} as

a NNQP problem. Then the SD algorithm can be modified to reduce the computational effort. Specifically, in step 4 we compute dual variables by solving a NNQP problem (10) instead of (9). The optimization problem to compute $(t_{k,i}, s_{k,i})$ and $(\hat{t}_{k,i}, \hat{s}_{k,i})$ can be modified accordingly.

Algorithm 2 SD-QQ

Require:

$\tau > 0,\ \alpha_0 = \tau,\ r \in (0,1),\ k = 0,\ U_0 = \emptyset$ and $\mathcal{J}_0 = \{0\}$.

a feasible solution $\hat{x}^0 \in X$.

$h_{j,0}(x) = 0,\ \forall x \in X$ and $\forall j \in \mathcal{J}_0$.

$f_0(x) = \frac{1}{2}x^\top Qx + c^\top x + \max\{h_{j,0}(x), j \in \mathcal{J}_0\},\ \forall x \in X$.

Ensure:

1: **while** the stopping rule (see section 5) are not satisfied **do**

2: $k \leftarrow k + 1$

3: **compute** $x^k = T_{\alpha_{k-1} f_{k-1}}(\hat{x}^{k-1})$ following the definition in (2) and record the Lagrangian multipliers $\{\mu_j^{k-1}\}_{j \in \mathcal{J}_{k-1}}$ of $\{h_j^{k-1}(x)\}_{j \in \mathcal{J}_{k-1}}$ respectively

4: **compute** dual variables (s^k, t^k) in the second stage by solving (9) at (x^k, ω^k).

5: **update** $\tilde{u}_k = \{i : s^k_{(l)} = 0, 1 \le l \le n_2\}$, $U_k = U_{k-1} \bigcup \{\tilde{u}_k\}$,

$$\mathcal{J}_k = \mathcal{J}_{k-1} \bigcup \{-k, k\} \backslash \{j \in \mathcal{J}_{k-1} : \mu_j^{k-1} = 0\}$$

6: **compute** $(t_{k,i}, s_{k,i}) \in \underset{\{(t,s,u):\, u \in U_k,\, s \ge 0,\, s_{(l)} = 0,\, \forall l \in u\}}{\operatorname{argmax}} g_{QQ}(t, s; x^k, \omega^i)$ for $i \in [k]$

7: **compute** $(\hat{t}_{k,i}, \hat{s}_{k,i}) \in \underset{\{(t,s,u):\, u \in U_k,\, s \ge 0,\, s_{(l)} = 0,\, \forall l \in u\}}{\operatorname{argmax}} g_{QQ}(t, s; \hat{x}^{k-1}, \omega^i)$ for $i \in [k]$

8: **construct** $h_k^k(x) = \frac{1}{k}\sum_{i=1}^k g_{QQ}(t_{k,i}, s_{k,i}; x, \omega^i)$,

$h_{-k}^k(x) = \frac{1}{k}\sum_{i=1}^k g_{QQ}(\hat{t}_{k,i}, \hat{s}_{k,i}; x, \omega^i)$,

$h_j^k(x) = \frac{|j|}{k}\ h_j^{|j|}(x) \quad$ for $j \in \mathcal{J}_k \backslash \{-k, k\}$

9: **construct** $f_k(x) = \frac{1}{2}x^T Qx + c^T x + \max\{h_j^k(x), j \in \mathcal{J}_k\}$

10: **if** $f_k(x^k) - f_k(\hat{x}^{k-1}) \le r(f_{k-1}(x^k) - f_{k-1}(\hat{x}^{k-1}))$ **then**

11: $\hat{x}^k = x^k, \quad \alpha_k = \frac{\tau}{k+1}$

12: **else**

13: $\hat{x}^k = \hat{x}^{k-1},\ \alpha_k = \alpha_{k-1}$

14: **end if**

15: **end while**

Because of the tremendous recent growth in the application of non-smooth optimization in the area of SP as well as ML, we should mention connections between SD and several methods in SP and ML. We begin by observing that SD can be seen as one of the inexact bundle methods (see Oliveira et al. in [19]) in which SD at each iteration utilizes an exact subgradient of a random sample function and inexact subgradient approximations of all other sample functions in the construction of value function approximations. Similarly, the proximal Stochastic Variance Reduction Gradient (SVRG) algorithm developed by Xiao and Zhang in [32] is also one of the inexact gradient methods for solving the ERM problems. Thus, both SD and SVRG share a similar idea which is the inclusion of exact as well as inexact gradients/subgradients for all samples in the history in order to reduce the variance along the stochastic path with small computational cost. However, because of the structural differences between SQLP/SQQP problems and ERM problems, we are able to develop subgradient approximations in the SD algorithm, which are much different from the inexact oracles in proximal SVRG. Moreover, in ERM problems the performance is measured by empirical risk, while in SQLP/SQQP problems one is measured by the expectation for the generalizability.

3.3 The convergence result of the SD algorithm

Higle and Sen in [8] and [9] studied the convergence properties of SD in solving two-stage SLP problems, which can be viewed as a special version of two-stage SQLP when $Q = 0$. Thus the SD algorithm for the SLP problem is almost the same as Algorithm 1 except that the value function approximations do not include the quadratic term. For the sequence of solutions generated by the SD algorithm of two-stage SLP problems, we summarize the convergence result (see Higle & Sen [8]) in the following lemma.

Lemma 2. *Suppose the assumptions* $(A2) - (A4)$ *hold for SQLP problem* (3) *with* $Q = 0$. *Let* $\{f_k(x)\}, \{x^k\}, \{\hat{x}^k\}$ *respectively be the sequence of value function approximations, candidate solutions and incumbent solutions, generated by the SD algorithm with the diminishing step size in two-stage SLP. Let* K *denote the set of iteration numbers where the incumbent solution updates.*
(a) If there exists $x^* \in X$ *such that* $\{\hat{x}^k\}_{k\in K} \to x^*$, *then* $\{f_k(\hat{x}^k)\}_{k\in K} \to f(x^*)$ *with probability 1.*
(b) With probability 1, there exists a subsequence of iterations, indexed by a set K_*, *such that* $\lim_{k\in K_*} f_{k-1}(x^k) - f_{k-1}(\hat{x}^{k-1}) + \frac{1}{2}\left\|x^k - \hat{x}^{k-1}\right\|^2 = 0$. *Moreover, every accumulation point of* $\{\hat{x}^k\}_{k\in K_*}$ *is an optimal solution of the two-stage SLP problem.*

Proof.
(a) See Theorem 2 by Higle and Sen in [8] .
(b) See Theorem 5 by Higle and Sen in [9]. □

Since in two-stage SQLP and SQQP, there is a finite number of extreme points or faces in the second-stage problem, following the proofs in [8] and [9], theorem 2 can be easily shown to hold for two-stage SQLP and SQQP with the assumption $(A1)$. Using Lemma 3.2, the convergence of the entire sequence, shown by Sen and Liu in [27] based on theorem 2 can be extended to the two-stage SQLP and SQQP as well. We present the convergence result in the following proposition showing that the entire sequence of incumbent solutions converges to an optimal solution with probability 1.

Proposition 3 (Convergence result of SD)**.** *Suppose the assumptions* $(A1) - (A4)$ *hold. Then in two-stage SQLP/SQQP, the sequence of incumbent solutions* $\{\hat{x}^k\}$ *generated by the SD algorithm with the diminishing step size converges to the unique optimal solution* x^* *with probability one.*

Proof. See Theorem 1 by Sen and Liu in [27]. □

4 Convergence Rate of SD

From theorem 3, the SD algorithm is able to produce a sequence of incumbent solutions converging to the unique optimal solution with probability one. Thereupon the convergence rate is the next issue of focus. In SD, the vital step dominating its convergence is the proximal mapping of the value function approximation. In this section, by analyzing a contraction property of the stochastic proximal mapping, we derive a sublinear convergence rate of SD for two-stage SQLP and SQQP problems.
It will become clear subsequently that two classes of problems share the same convergence analysis, therefore in the following analysis we unify the notation (x^*, λ^*) to be the optimal solution pair of SQLP or SQQP, F_k to be the SAA function defined in (11) with the sample set of size k without identifying its specific structures, i.e. the recourse functions h_{QL} or h_{QQ}. Similarly, let f_k represent the value function approximation created at iteration k with the definition at line 9 in Algorithm 1

or line 9 in Algorithm 2 without recognizing its specific structures, i.e., SASA or SAQA functions. Moreover, since we consider the convergence result of the sequence of incumbent solutions, without loss of generality we filter out all candidate solutions which are rejected as incumbent solutions. For the sake of propositions and theorems in this section, we simplify the notation by denoting the sequence of incumbent solutions using $\{x^k\}$.

In studying the convergence rate of a randomized algorithm such as SD, it is customary to treat x^{k+1} as a random variable which is governed by the randomness of Monte Carlo sampling $\{\omega^i\}_{i=1}^k$ generated in the entire history of the algorithm. Let $\{\mathcal{F}_k\}_{k\in\mathbb{N}_+}$ denote the filtration with $\mathcal{F}_k = \sigma\{\widetilde{\omega}^1, \ldots, \widetilde{\omega}^k\}$. For any $k \in \mathbb{N}_+$, let $\hat{\mathbf{E}}_k$ denotes the expectation taken with respect to the product of probability measures of $\{\widetilde{\omega}^i\}_{i=1}^k$ in the sense that $\hat{\mathbf{E}}_k[\cdot] = \mathbf{E}[\ \cdot\ |\mathcal{F}_k]$. Therefore, the idea is to analyze $\hat{\mathbf{E}}_k\ \left\|x^{k+1} - x^*\right\|$ in terms of the number of iterations. Besides assumptions $(A1)-(A4)$, we make following assumptions for the convergence rate analysis of the SD algorithm in this section.

- $(B1)$ Linear independence constraint qualification (LICQ) holds at the optimum x^* for the feasible solution set $X = \{x : Ax \leq b\}$.
- $(B2)$ There exists a nonnegative-valued measurable function $L(\omega) : \Omega \to [0, \infty)$ with $\mathbf{E}[L(\omega)] < \infty$, such that $|h(x,\omega) - h(x',\omega)| \leq L(\omega)\,\|x - x'\|$ for all $x, x' \in X$ and almost every $\omega \in \Omega$.
- $(B3)$ There exists a neighborhood $B(x^*, \delta)$ with $\delta > 0$ such that the recourse function $h(x,\omega)$ is differentiable for all $x \in X \cap B(x^*, \delta)$ and almost every $\omega \in \Omega$.
- $(B4)$ The unique optimum x^* is sharp, i.e., there exists a positive constant ρ such that $f(x) \geq f(x^*) + \rho\,\|x - x^*\|$ for any $x \in X$.
- $(B5)$ Strict complementarity holds at x^*, which means $b_r - a_r^\top x^* + \lambda_r^* > 0$ for all $r \in [m]$ where $a_r^\top$ is rth row vector of the matrix A and b_r is rth component of the vector b.

At this point it is appropriate to comment about the above assumptions. The LICQ assumption in $(B1)$ is defined such that active constraints at x^* are linearly independent. This local condition at x^* will be used for analyzing the convergence rate, whereas in nonlinear programming, LICQ is intended to manage difficulties which might arise due to the linearization of nonlinear constraints. Assumption $(B2)$ actually can be directly derived from the assumption $(A3)$ that the recourse matrix D and d is fixed and the recourse function is relatively complete. However, due to the usage of the Lipschitz constant in the convergence rate analysis, we list it as an assumption here. Moreover, it is worth noticing that a sufficient condition for the local Lipschitz continuity of the optimal value function was presented in Lemma 5.1 in [21]. Assumption $(B3)$ can be derived from the almost sure differentiability of $h(x,\omega)$ at x^* with an additional condition that the radius of the neighborhoods centered at x^* within which $h(x,\omega)$ is differentiable have a uniform positive lower bound for almost every $\omega \in \Omega$. A stronger version of differentiability is assumed in Assumption 3 in [24] in which the recourse function is differentiable for all $x \in X$ and almost every $\omega \in \Omega$. Assumption $(B4)$ is a consequence of our use of the finite exponential convergence rate of SAA, which will be discussed in the context of the proof technique, i.e., equation (14) and theorem 1. With local differentiability in $(B3)$ and strong convexity, the unique minimizer satisfies the quadratic growth condition (see [26]). The linear growth condition holds naturally for two-stage stochastic linear program with finitely many scenarios (see [28]). Moreover, it is equivalent to $f'(x^*, q) \geq r\,\|q\|$ for any $q \in \mathcal{T}_X(x^*)$, where $\mathcal{T}_X(x^*)$ denotes the tangent cone of set X at point x^* (see (5.135) in [28]). Hence as a local condition, it suffices to impose the assumption $(B4)$ in a neighborhood of x^*. Finally, the assumption $(B5)$ of strict complementarity is needed for showing the stability of active constraints of stochastic proximal mappings even for the case of linearly constrained problems.

Recall that $\hat{\mathbf{E}}_k$ denotes the expectation taken with respect to the product of probability measures of $\{\omega^i\}_{i=1}^k$. By using the triangle inequality, we are able to separate the expected distance between the current incumbent solution and the unique optimal solution into three terms.

$$\begin{aligned}\hat{\mathbf{E}}_k \left\|x^{k+1} - x^*\right\| \leq\ & \hat{\mathbf{E}}_k \left\|T^X_{\alpha_k f_k}(x^k) - T^X_{\alpha_k F_k}(x^k)\right\| \\ & + \hat{\mathbf{E}}_k \left\|T^X_{\alpha_k F_k}(x^k) - T^X_{\alpha_k F_k}(x^*)\right\| + \hat{\mathbf{E}}_k \left\|T^X_{\alpha_k F_k}(x^*) - x^*\right\|.\end{aligned} \tag{14}$$

On the right hand side of (14), the first term is the distance of the proximal mapping point of the value function approximation f_k and the SAA function F_k. We will analyze two cases in theorem 5 showing that this distance is bounded in $O(\alpha_k)$ and $O(\alpha_k^2)$ respectively. For the second term, we will show the contraction property in theorem 9 for the stochastic proximal mapping of the SAA function with constraints. With quadratic programming in the first stage and the sequential sampling process of SD, this contraction property can be seen as an improvement of the well-known non-expansive property of deterministic proximal methods. Moreover, the third term will be shown to converge exponentially to zero in theorem 10 from the exponential convergence rate of the SAA method under the assumption $(B4)$. By combining these three propositions together with a lemma on limiting properties, in theorem 12 we derive a recursion showing that the distance between the incumbent solution and the optimal solution diminishes at a sublinear convergence rate.

Lemma 4. *Suppose assumptions $(A1)-(A4)$ and $(B3)$ hold for the two-stage SQLP (3) or SQQP (5) considered. There exists a finite number $\hat{k}_1$, such that for any $k \geq \hat{k}_1$, points $T^X_{\alpha_k f_k}(x^k)$ and x^k are on the same piece of f_k with probability 1.*

Proof. According to the algorithm, for any $k \in \mathbb{Z}_+$, $V_k \subseteq V_{k+1} \subseteq V$ with probability 1. Since the set V has finitely many elements, by the Monotone Convergence Theorem, we have $\lim_{k\to\infty} V_k = \bar{V} \subseteq V$ with probability 1 (see Lemma 1 in [8]). Similarly, for the two-stage SQQP, we have $\lim_{k\to\infty} U_k = \bar{U} \subseteq U$ with probability 1 where U is the set of the indexes of all the dual faces in the second-stage program of SQQP. Then there exist finite numbers k_1' and k_2' such that when $k \geq k_a = \max\{k_1', k_2'\}$, we have $V_k = \bar{V}$ and $U_k = \bar{U}$. It means that after finitely many iterations, V_k are stable to include all the necessary extreme points ($\bar{V}$) with probability 1. Similarly, after finitely many iterations, U_k are stable to include all the necessary indexes of zero dual variables ($\bar{U}$) with probability 1. Therefore, the sample average of subgradient approximation $h_k^k(x^k)$ is equal to the sample average of recourse function at x^k, i.e., $h_k^k(x^k) = \frac{1}{k}\sum_{i=1}^k h(x^k, \omega^i)$ with probability 1. Following that, we have $f_k^k(x^k) = F_k(x^k)$ for both SQLP and SQQP problems for large enough k with probability 1.

From theorem 3 and assumption $(B3)$, there exists a finite number k_1, such that for any $k \geq k_1$, $h(x, \omega)$ is differentiable with respect to x at x^k for almost every $\omega \in \Omega$. Hence when $k \geq \max\{k_1, k_1', k_2'\}$, both $F_k(x)$ and $f_k(x)$ are differentiable at x^k with the same gradient, i.e., $\nabla|_{x=x^k} f_k(x) = \nabla|_{x=x^k} F_k(x)$. Moreover, $\frac{1}{2}\left\|T^X_{\alpha_k f_k}(x^k) - x^k\right\|^2 \leq \alpha_k \left(f_k(x^k) - f_k(T^X_{\alpha_k f_k}(x^k))\right) \leq C\alpha_k$ for some constant $C \in \mathbb{R}_+$, so we have $\lim_{k\to\infty}\left\|T^X_{\alpha_k f_k}(x^k) - x^k\right\| = 0$ with probability 1. Hence, there exists a finite number $\hat{k}_1$ such that for any $k \geq \hat{k}_1$, the points $T^X_{\alpha_k f_k}(x^k)$ and x^k are on the same linear or quadratic piece of f_k with probability 1 and the function f_k is differentiable at x^k and $T^X_{\alpha_k f_k}(x^k)$ with probability 1. □

Proposition 5. *Suppose assumptions $(A1)-(A4)$, $(B2)$ and $(B3)$ hold for the two-stage SQLP (3) or SQQP (5) considered. There exist finite numbers $k_a \leq k_b$, such that at iteration k the proximal mapping over a feasible solution set X satisfies:*

(a) $\hat{\mathbf{E}}_k \left\| T^X_{\alpha_k f_k}(x^k) - T^X_{\alpha_k F_k}(x^k) \right\| \le 4M_0\alpha_k, \quad$ for any $k \ge k_a$.

(b) $\hat{\mathbf{E}}_k \left\| T^X_{\alpha_k f_k}(x^k) - T^X_{\alpha_k F_k}(x^k) \right\| \le 4M_0M_1\alpha_k^2, \quad$ for any $k \ge k_b$,

where $M_0 \triangleq \max_{x\in X} \|c + Qx\| + \mathbf{E}[L(\widetilde{\omega})]$ *and* M_1 *is a positive constant independent of* k.

Proof. First we notice that the proximal mapping can be represented as an implicit projection step, i.e.,

$$T^X_{\alpha_k f_k}(x^k) = P_X(x^k - \alpha_k \xi_{f_k}(T^X_{\alpha_k f_k} x^k)),$$

where $P_X(\cdot)$ is a projection mapping defined in (1), $\xi_{f_k}(T^X_{\alpha_k f_k} x^k)$ denotes the subgradient of f_k under optimality conditions at the proximal mapping point $T^X_{\alpha_k f_k} x^k$. Similarly, let $\xi_{F_k}(T^X_{\alpha_k F_k} x^k)$ denote the subgradient of F_k under optimality conditions at the proximal mapping point $T^X_{\alpha_k F_k} x^k$. According to the reasoning in theorem 4, there exists a finite number k_a such that when $k \ge k_a$, the sets V_k and U_k include all the dual extreme points and faces respectively in the second-stage dual problems of SQLP and SQQP. Therefore, F_k and f_k have the same value at x^k with at least one common subgradient when $k \ge k_a$. Let $\xi_{F_k}(x^k) = \xi_{f_k}(x^k)$ be one common subgradient of F_k and f_k at x^k. With representations of implicit projection steps, we derive the following inequality.

$$\begin{aligned}
\left\| T^X_{\alpha_k f_k}(x^k) - T^X_{\alpha_k F_k}(x^k) \right\| &\le \left\| P_X(x^k - \alpha_k \xi_{f_k}(T^X_{\alpha_k f_k}(x^k)) - P_X(x^k - \alpha_k \xi_{f_k}(x^k)) \right\| \\
&\quad + \left\| P_X(x^k - \alpha_k \xi_{F_k}(x^k)) - P_X(x^k - \alpha_k \xi_{F_k}(T^X_{\alpha_k F_k}(x^k)) \right\| \\
&\le \alpha_k \left\| \xi_{f_k}(T^X_{\alpha_k f_k}(x^k)) - \xi_{f_k}(x^k) \right\| \\
&\quad + \alpha_k \left\| \xi_{F_k}(x^k) - \xi_{F_k}(T^X_{\alpha_k F_k}(x^k)) \right\|. \qquad (15)
\end{aligned}$$

By assumptions $(A2)$ and $(B2)$, subgradients of the value function approximation can be bounded uniformly for all $x \in X$, i.e., $\|\xi_{f_k}(x)\| \le \max_{x\in X} \|c + Qx\| + \frac{1}{k}\sum_{i=1}^k L(\omega^i)$. Hence $\hat{\mathbf{E}}_k \|\xi_{f_k}(x)\| \le M_0 \triangleq \max_{x\in X} \|c + Qx\| + \mathbf{E}[L(\widetilde{\omega})]$ for any $x \in X$. Such bound also holds for the sample average function F_k. From (15), for any $k \ge k_a$,

$$\hat{\mathbf{E}}_k \left\| T^X_{\alpha_k f_k}(x^k) - T^X_{\alpha_k F_k}(x^k) \right\| \le 4M_0\alpha_k.$$

It proves the statement (a).
Moreover, from theorem 4 there exists a finite number k_b such that for any $k \ge k_b$, $T^X_{\alpha_k f_k}(x^k)$ and x^k are on the same linear/quadratic piece of a piecewise function f_k with probability 1. Therefore, there exists a constant M_1 such that

$$\hat{\mathbf{E}}_k \left\| \xi_{f_k}(T^X_{\alpha_k f_k}(x^k)) - \xi_{f_k}(x^k) \right\| \le M_1 \hat{\mathbf{E}}_k \left\| T^X_{\alpha_k f_k}(x^k) - x^k \right\|. \qquad (16)$$

The argument of theorem 4 also holds for F_k following almost the same analysis. Without loss of generality, with the same constant M_1 we have

$$\hat{\mathbf{E}}_k \left\| \xi_{F_k}(T^X_{\alpha_k F_k}(x^k)) - \xi_{F_k}(x^k) \right\| \le M_1 \hat{\mathbf{E}}_k \left\| T^X_{\alpha_k F_k}(x^k) - x^k \right\|. \qquad (17)$$

From the optimality condition of $T^X_{\alpha_k f_k}(x^k)$ and the boundness of the subgradient of f_k, we derive

$$\begin{aligned}\frac{1}{2}\left\|T^X_{\alpha_k f_k}(x^k)-x^k\right\|^2 &\le \alpha_k(f_k(x^k)-f_k(T^X_{\alpha_k f_k}(x^k)))\\ &\le \alpha_k\left(\max_{x\in X}\|c+Qx\|+\frac{1}{k}\sum_{i=1}^{k}L(\omega^i)\right)\left\|T^X_{\alpha_k f_k}(x^k)-x^k\right\|.\end{aligned}$$

Since $\left\|T^X_{\alpha_k f_k}(x^k)-x^k\right\|\neq 0$ with probability 1, dividing it and taking expectations on both sides, we derive

$$\hat{\mathbf{E}}_k\ \left\|T^X_{\alpha_k f_k}(x^k)-x^k\right\|\le 2M_0\alpha_k.$$

Similarly for the SAA function F_k, we have

$$\hat{\mathbf{E}}_k\ \left\|T^X_{\alpha_k F_k}(x^k)-x^k\right\|\le 2M_0\alpha_k.$$

Combining the above two inequalities with inequalities (15), (16) and (17), we prove the statement (b), i.e., $\hat{\mathbf{E}}_k\ \left\|T^X_{\alpha_k f_k}(x^k)-T^X_{\alpha_k F_k}(x^k)\right\|\le 4M_1M_0\alpha_k^2$. □

It is known that the proximal mapping obeys only a non-expansive property (see [20] and [23]). However as for the second term in the triangle inequality (14), we will improve the result showing that the proximal mapping of strongly convex SAA functions with constraints has a *contraction* property. A similar contraction property of stochastic proximal iteration has been analyzed by Ryu and Boyd [25] under an assumption of M-Restricted strong convexity without constraints. In order to do so, we start by showing that under some assumptions, active constraints of the proximal mapping $T^X_{\alpha F_k}(x^k)$ should stabilize after finitely many iterations with probability 1.

Lemma 6. *Suppose the assumptions* $(A1)-(A4)$, $(B1)$, $(B3)$ *and* $(B5)$ *hold for the two-stage SQLP* (3) *or SQQP* (5). *Let* $I^k\triangleq\{r: a_r^\top T^X_{\alpha F_k}(x^k)=b_r\}$ *and* $I^*\triangleq\{r: a_r^\top x^*=b_r\}$. *Then there exists a finite number* $\hat{k}_2$, *such that for any* $k\ge\hat{k}_2$, $I^k=I^*$ *with probability 1.*

Proof. See Appendix A. □

Let $\bar{X}\triangleq\{x: a_r^\top x=b_r, r\in I^*\}$, as the feasible solution set constructed by active constraints at x^*. From theorem 6, $T^X_{\alpha F_k}(x^k)=T^{\bar{X}}_{\alpha F_k}(x^k)$ when $k\ge\hat{k}_2$. Therefore, in order to bound the second term in (14) for the asymptotic convergence rate, we only need to consider $T^{\bar{X}}_{\alpha F_k}(x^k)$ by restricting the feasible solution set to be $\bar{X}$. In doing so, we first prove that the proximal mapping of a quadratic function with the positive definiteness obeys the contraction property. Following that, we prove that the proximal mapping of the sum of a quadratic function and a convex random function obeys the contraction property as well. Then by interpreting the sample average function as the sum of a quadratic function and a sample average of convex functions, we thus have the contraction property for the proximal mapping. We should mention that for simplicity we omit the superscript $\bar{X}$ in the notation of proximal mappings in the following two lemmas. Instead we claim that the proximal mapping is taken over a feasible solution set $\bar{X}$ in the statements.

Lemma 7. *Suppose* $\check{g}(z)=\frac{1}{2}z^\top Qz$ *where* Q *is a positive definite matrix and* $\bar{A}$ *is a matrix consisting of linearly independent row vectors. Then the proximal mapping* $T_{\alpha\check{g}}$ *over a feasible solution set* $\bar{X}=\{x:\bar{A}x=\bar{b}\}$ *satisfies the following,*

$$T_{\alpha\check{g}}x-T_{\alpha\check{g}}x'=K_\alpha(x-x'),\quad\forall x,x'\in\bar{X},\tag{18}$$

where $G_\alpha = (I + \alpha Q)^{-1}$ *and* $K_\alpha = G_\alpha - G_\alpha \bar{A}^\top (\bar{A} G_\alpha \bar{A}^\top)^{-1} \bar{A} G_\alpha$. *Moreover, we have*

$$\| T_{\check{g}} x - T_{\check{g}} x' \| \leq \theta_{max}(G_\alpha) \cdot \| x - x' \|, \tag{19}$$

where $\theta_{max}(\cdot)$ *denotes the largest eigenvalue of the matrix.*

Proof. See Appendix A. □

With theorem 7 we then prove that the proximal mapping of the sum of a quadratic function and a random function also obeys a contraction property with the same contraction factor.

Lemma 8. *Suppose* $\check{g}(z) = \frac{1}{2} z^\top Q z$ *where* Q *is a positive definite matrix,* $r(z)$ *is a random convex function, and* $\bar{A}$ *is a matrix consisting of linearly independent row vectors. Let* $g(z) = \check{g}(z) + r(z)$. *Then with probability 1 the proximal mapping* $T_{\alpha g}$ *over a feasible solution set* $\bar{X} = \{x : \bar{A}x = \bar{b}\}$ *satisfies the following,*

$$\left\| T_{\alpha g} x - T_{\alpha g} x' \right\| \leq \theta_{\max}(G_\alpha) \cdot \left\| x - x' \right\|, \quad \forall x, x' \in \bar{X}, \tag{20}$$

where $G_\alpha = (I + \alpha Q)^{-1}$ *and* $\theta_{max}(\cdot)$ *denotes the largest eigenvalue of the matrix.*

Proof. Assume $x \neq x'$ and $T_{\alpha g} x \neq T_{\alpha g} x'$, otherwise there is nothing to show. Let $\mu_x \in \partial r(T_{\alpha g}(x))$ with which optimality conditions hold for the solution pair $(T_{\alpha g} x, \lambda_x)$ as follows.

$$\begin{pmatrix} I + \alpha Q & \bar{A}^\top \\ \bar{A} & 0 \end{pmatrix} \begin{pmatrix} T_{\alpha g} x \\ \lambda_x \end{pmatrix} = \begin{pmatrix} x - \alpha \mu_x \\ \bar{b} \end{pmatrix}. \tag{21}$$

Similarly, let $\mu_{x'} \in \partial r(T_{\alpha g}(x'))$ with which the optimal conditions hold at $T_{\alpha g} x'$.
Then let $\tilde{r}_\epsilon$ be a quadratic function for some $\epsilon > 0$ defined below,

$$\tilde{r}_\epsilon(z) \triangleq \mu_x^\top z + \frac{1}{2}(z - T_{\alpha g} x)^\top \frac{v_\epsilon v_\epsilon^\top}{a_\epsilon}(z - T_{\alpha g} x),$$

where $v_\epsilon \triangleq -\mu_x + \mu_{x'} + \epsilon(T_{\alpha g} x' - T_{\alpha g} x)$, and $a_\epsilon \triangleq v_\epsilon^T (T_{\alpha g} x' - T_{\alpha g} x)$. In this construction, $v_\epsilon \in -\partial r(T_{\alpha g} x) + \partial r(T_{\alpha g} x') + \epsilon(T_{\alpha g} x' - T_{\alpha g} x)$. Moreover, since ∂r is a monotone operator, we derive

$$a_\epsilon = v_\epsilon^\top (T_{\alpha g} x' - T_{\alpha g} x) \geq \epsilon \| T_{\alpha g} x' - T_{\alpha g} x \|^2 \geq 0.$$

Thus by design $\tilde{r}_\epsilon$ is a convex quadratic function determined by x and x'. In addition, simple calculations show that

$$\nabla \tilde{r}_\epsilon(T_{\alpha g} x) = \mu_x, \tag{22}$$

$$\nabla \tilde{r}_\epsilon(T_{\alpha g} x') = \mu_{x'} + \epsilon(T_{\alpha g} x' - T_{\alpha g} x), \tag{23}$$

$$\nabla \tilde{r}_\epsilon(z) = \mu_x + \frac{v_\epsilon v_\epsilon^\top}{a_\epsilon}(z - T_{\alpha g} x). \tag{24}$$

We take (22) into the optimality conditions (21) and derive

$$\begin{pmatrix} I + \alpha Q & \bar{A}^\top \\ \bar{A} & 0 \end{pmatrix} \begin{pmatrix} T_{\alpha g} x \\ \lambda_x \end{pmatrix} = \begin{pmatrix} x - \alpha \nabla \tilde{r}_\epsilon(T_{\alpha g} x) \\ \bar{b} \end{pmatrix},$$

which is the optimality condition of the proximal mapping $T_{\alpha(\check{g} + \tilde{r}_\epsilon)}(x)$. Therefore

$$T_{\alpha(\check{g} + \tilde{r}_\epsilon)} x = T_{\alpha g} x. \tag{25}$$

Similarly we take (23) into the optimality conditions at $(T_{\alpha g}x', \lambda_{x'})$ and derive,

$$\begin{aligned}\begin{pmatrix} I+\alpha Q & \bar{A}^\top \\ \bar{A} & 0 \end{pmatrix}\begin{pmatrix} T_{\alpha g}x' \\ \lambda_{x'} \end{pmatrix} &= \begin{pmatrix} x' - \alpha(\nabla\tilde{r}_\epsilon(T_{\alpha g}x') - \epsilon(T_{\alpha g}x' - T_{\alpha g}x)) \\ \bar{b} \end{pmatrix} \\ &= \begin{pmatrix} x' - \alpha\nabla\tilde{r}_\epsilon(T_{\alpha g}x') \\ \bar{b} \end{pmatrix} + \begin{pmatrix} \alpha\epsilon(T_{\alpha g}x' - T_{\alpha g}x) \\ 0 \end{pmatrix}. \end{aligned} \tag{26}$$

Moreover, optimality conditions hold at the solution pair $(T_{\alpha(\check{g}+\tilde{r}_\epsilon)}x', \lambda_{\epsilon,x'})$,

$$\begin{pmatrix} I+\alpha Q & \bar{A}^\top \\ \bar{A} & 0 \end{pmatrix}\begin{pmatrix} T_{\alpha(\check{g}+\tilde{r}_\epsilon)}x' \\ \lambda_{\epsilon,x'} \end{pmatrix} = \begin{pmatrix} x' - \alpha\nabla\tilde{r}_\epsilon(T_{\alpha(\check{g}+\tilde{r}_\epsilon)}x') \\ \bar{b} \end{pmatrix}. \tag{27}$$

From (24), we have $\nabla\tilde{r}_\epsilon(T_{\alpha g}x') - \nabla\tilde{r}_\epsilon(T_{\alpha(\check{g}+\tilde{r}_\epsilon)}x') = \frac{v_\epsilon v_\epsilon^\top}{a_\epsilon}(T_{\alpha g}x' - T_{\alpha(\check{g}+\tilde{r}_\epsilon)}x')$. Thus by subtracting two optimality conditions (27) and (26), we derive

$$\begin{pmatrix} I+\alpha(Q+\frac{v_\epsilon v_\epsilon^\top}{a_\epsilon}) & \bar{A}^\top \\ \bar{A} & 0 \end{pmatrix}\begin{pmatrix} T_{\alpha g}x' - T_{\alpha(\check{g}+\tilde{r}_\epsilon)}x' \\ \lambda_{x'} - \lambda_{\epsilon,x'} \end{pmatrix} = \alpha\epsilon(T_{\alpha g}x' - T_{\alpha g}x)\begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Since $\bar{A}$ has linearly independent rows and Q is positive definite, we have

$$\lim_{\epsilon\to 0} T_{\alpha(\check{g}+\tilde{r}_\epsilon)}x' = T_{\alpha g}x' = T_{\alpha(\check{g}+r)}x'. \tag{28}$$

Let $G_{\alpha,\epsilon} \triangleq (I+\alpha(Q+\frac{v_\epsilon v_\epsilon^\top}{a_\epsilon}))^{-1}$. Because $\frac{v_\epsilon v_\epsilon^\top}{a_\epsilon}$ is positive semidefinite, from the positive semidefinite ordering we have $G_{\alpha,\epsilon} \preceq G_\alpha$. Then from theorem 7 we derive

$$\frac{\parallel T_{\alpha(\check{g}+\tilde{r}_\epsilon)}x - T_{\alpha(\check{g}+\tilde{r}_\epsilon)}x' \parallel}{\parallel x - x' \parallel} \le \theta_{max}(G_{\alpha,\epsilon}) \le \theta_{\max}(G_\alpha) < 1. \tag{29}$$

With the limiting equation in (25) and (28), let $\epsilon \to 0$ in (29) and we have

$$\left\|T_{\alpha g}x - T_{\alpha g}x'\right\| \le \theta_{\max}(G_\alpha)\ \left\|x - x'\right\| \quad \text{with probability 1.}$$

□

Now equipped with theorem 6 and theorem 8, the following proposition shows a contraction property of the proximal mapping of the SAA function in expectation.

Proposition 9. *In two-stage SQLP/SQQP problems with assumptions* $(A1)-(A4)$, $(B1)$, $(B3)$ *and* $(B5)$, *when* $k \ge \hat{k}_2$ *with* $\hat{k}_2$ *defined in theorem 6, the proximal mapping* $T^X_{\alpha_k F_k}(\cdot)$ *satisfies*

$$\hat{\boldsymbol{E}}_k\ \|T^X_{\alpha_k F_k}x^k - T^X_{\alpha_k F_k}x^*\| \le \gamma(\alpha_k)\ \hat{\boldsymbol{E}}_{k-1}\ \|\ x^k - x^*\ \|, \tag{30}$$

where $G_\alpha = (I+\alpha Q)^{-1}$ *and* $\gamma(\alpha) = \theta_{\max}(G_\alpha)$.

Proof. From theorem 6, when $k \ge \hat{k}_2$, we have $T^X_{\alpha_k F_k}x^k = T^{\bar{X}}_{\alpha_k F_k}x^k$. With almost the same analysis, the argument in theorem 6 should also hold for $T^X_{\alpha_k F_k}x^*$ such that $T^X_{\alpha_k F_k}x^* = T^{\bar{X}}_{\alpha_k F_k}x^*$ when $k \ge \hat{k}_2$. Therefore, to apply the result in theorem 8, we take $\check{g}(x)$ as the quadratic function $\frac{1}{2}x^\top Qx$ and $r(x)$ as a random function in the form of $c^\top x + \frac{1}{k}\sum_{i=1}^k h(x,\omega^i)$ where $\{\omega^i\}_{i=1}^k$ are iid random variables. Then $\check{g}+r$ is the sample average function F_k. From the assumption $(B1)$, $\bar{A}$ has independent row vectors. By applying theorem 8 for the SAA function F_k, with probability 1 we have

$$\|T_{\alpha_k F_k}x^k - T_{\alpha_k F_k}x^*\| \le \theta_{\max}(G_{\alpha_k})\ \|\ x^k - x^*\ \|. \tag{31}$$

Then by the law of iterated expectations, we have the contraction property in expectation where $\mathbf{E}$ is taken with respect to the probability measure of ω^k and $\hat{\mathbf{E}}_k$ is taken with respect to the product of probability measures of $\{\omega^i\}_{i=1}^k$.

$$\begin{aligned}\hat{\mathbf{E}}_k \ \|T_{\alpha_k F_k}x^k - T_{\alpha_k F_k}x^*\| &= \hat{\mathbf{E}}_{k-1}\ \left[\mathbf{E}\ \left[\|T_{\alpha_k F_k}x^k - T_{\alpha_k F_k}x^*\|\Big|\{\omega^i\}_{i=1}^{k-1}\right]\right]\\ &\le \hat{\mathbf{E}}_{k-1}\ \left[\mathbf{E}\ \left[\theta_{\max}(G_{\alpha_k})\ \ \| \ x^k - x^* \ \| \ \Big|\{\omega^i\}_{i=1}^{k-1}\right]\right]\\ &= \theta_{\max}(G_{\alpha_k})\ \hat{\mathbf{E}}_{k-1}\ \| \ x^k - x^* \ \| \ .\end{aligned}$$

□

theorem 9 presents a valuable contraction factor $\gamma(\alpha)$ for the proximal mapping with respect to the sample average function $F_k(x)$. However, the stochastic proximal mapping still may not have a fixed point. Here in the case of stochastic proximal mapping of the sample average approximation functions, we show that the expectation of the proximal mapping gap at x^* converges to zero exponentially.

Proposition 10 (Convergence of the fixed-point gap)**.** *Let F_k be the SAA function defined in* (11)*. Suppose assumptions* $(A1)-(A3)$ *and* $(B4)$ *hold for the two-stage SQLP/SQQP problems considered. Then there exist constants $C>0$ and $\beta>0$ such that for the optimal solution x^*, we have*

$$\hat{\mathbf{E}}_k \ \ \| \ T^X_{\alpha F_k}x^* - x^* \ \| < Ce^{-\beta k}. \tag{32}$$

Proof. Let $\tilde{x}^k \triangleq \underset{x\in X}{\operatorname{argmin}}\ F_k(x)$ and x^* be the unique and sharp optimal solution of a two-stage SQLP/SQQP. From theorem 1, $P\left(\| \ \tilde{x}^k - x^* \ \| > 0\right) \le C_0 e^{-\beta_0 k}$. Moreover, from the optimality conditions of $T^X_{\alpha F_k}(x^*)$, we have

$$F_k(T^X_{\alpha F_k}x^*) + \frac{1}{2\alpha}\left\|T^X_{\alpha F_k}x^* - x^*\right\|^2 < F_k(\tilde{x}^k) + \frac{1}{2\alpha}\left\|\tilde{x}^k - x^*\right\|^2. \tag{33}$$

Since $F_k(T^X_{\alpha F_k}x^*) > F_k(\tilde{x}^k)$, from (33) we have $\|T_{\alpha F_k}x^* - x^*\| < \left\|\tilde{x}^k - x^*\right\|$.
From the assumption $(A2)$, there exists a constant C_1 such that for any x and x' in the set X, we have $\| \ x - x' \ \| \le C_1$. Therefore, we have

$$\hat{\mathbf{E}}_k \ \left\|T^X_{\alpha F_k}x^* - x^*\right\| < \hat{\mathbf{E}}_k \ \left\|\tilde{x}^k - x^*\right\| \le C_1 P\left(\left\|\tilde{x}^k - x^*\right\| > 0\right) \le C_1 C_0 e^{-\beta_0 k},$$

which proves the statement. □

The following lemma will be used in deriving the convergence rate in theorem 12.

Lemma 11. *For $\Lambda>0$ and any two large integers k and k' satisfying $k' \ll k$, let $Z_\Lambda(k',k) \triangleq \Pi_{i=k'+1}^k(1-\frac{\Lambda}{i})$. We have the following limiting properties.*

1. $Z_\Lambda(k',k) \lesssim \left(\frac{k'}{k}\right)^\Lambda$.
2. $\sum_{i=\hat{k}+1}^k e^{-\beta i} Z_\Lambda(i,k) \lesssim \begin{cases}\frac{1}{e^{\beta k}} & \Lambda \ge 1\\ \frac{k^{1-\Lambda}}{e^{\beta k}} & \Lambda < 1.\end{cases}$

3. $\sum_{i=\hat{k}+1}^{k} \frac{1}{i} Z_\Lambda(i,k) \lesssim \frac{1}{\Lambda}$

4. $\sum_{i=\hat{k}+1}^{k} \frac{1}{i^2} Z_\Lambda(i,k) \lesssim \begin{cases} \frac{log(k)}{k} & \Lambda = 1 \\ \frac{1}{(\Lambda-1)k} & \Lambda > 1 \\ \frac{1}{(1-\Lambda)k^\Lambda} & \Lambda < 1 \end{cases}$

Proof. See Appendix A. □

Notice that the properties in theorem 4 and theorem 6 hold under the existence of some finite numbers, such as k_a, k_b and $\hat{k}_2$. In the following theorem, we inherit these finite numbers in deriving the sublinear convergence rate of the SD algorithm.

Theorem 12 (Convergence rate of SD in SQLP)**.** *Suppose the assumptions* $(A1)-(A4)$ *and* $(B1)-(B5)$ *hold for the two-stage SQLP* (3) *or SQQP* (5) *considered. Let* x^* *denote the optimal solution and* $\{x^k\}$ *denote the sequence of incumbent solutions provided by SD with the step size* $\{\frac{\tau}{k+1}\}$ *for a given constant* $\tau > 0$*. Let* $\Lambda \triangleq \tau \cdot \theta_{min}(Q)$ *where* $\theta_{min}(\cdot)$ *denotes the smallest eigenvalue of the matrix. Then,*
(a) for any $k \gg \hat{k}_a \triangleq \max\{k_a, \hat{k}_2\}$,

$$\hat{\mathbf{E}}_k \parallel x^{k+1} - x^* \parallel \lesssim \hat{\mathbf{E}}_{\hat{k}_a - 1} \parallel x^{\hat{k}_a} - x^* \parallel \left(\frac{\hat{k}_a}{k}\right)^\Lambda + C\frac{k^{[1-\Lambda]_+}}{e^{\beta k}} + \frac{4M_0}{\Lambda},$$

(b) for any $k \gg \hat{k}_b \triangleq \max\{k_b, \hat{k}_2\}$,

$$\hat{\mathbf{E}}_k \parallel x^{k+1} - x^* \parallel \lesssim \hat{\mathbf{E}}_{\hat{k}_b - 1} \parallel x^{\hat{k}_b} - x^* \parallel \left(\frac{\hat{k}_b}{k}\right)^\Lambda + C\frac{k^{[1-\Lambda]_+}}{e^{\beta k}} + 4M_0M_1\tau^2 \begin{cases} \frac{log(k)}{k} & \Lambda = 1 \\ \frac{1}{(\Lambda-1)k} & \Lambda > 1 \\ \frac{1}{(1-\Lambda)k^\Lambda} & \Lambda < 1, \end{cases}$$

where C *and* β *are constants defined in theorem 10, and* M_0 *and* M_1 *are positive constants defined in theorem 5.*

Proof. Since we only consider the sequence of incumbent solutions in SD, every x^k is a proximal mapping point of the value function approximation. The idea of deriving the convergence rate is to obtain the recurrence by bounding the expectation of the distance between x^k and x^* using the triangle inequality below,

$$\begin{aligned} \hat{\mathbf{E}}_k \left\| x^{k+1} - x^* \right\| \leq\ & \hat{\mathbf{E}}_k \left\| T^X_{\alpha_k f_k}(x^k) - T^X_{\alpha_k F_k}(x^k) \right\| \\ & + \hat{\mathbf{E}}_k \left\| T^X_{\alpha_k F_k}(x^k) - T^X_{\alpha_k F_k}(x^*) \right\| + \hat{\mathbf{E}}_k \left\| T^X_{\alpha_k F_k}(x^*) - x^* \right\|. \end{aligned}$$

From theorem 5, the first term in the above triangle inequality is bounded by $O(\alpha_k)$ and $O(\alpha_k^2)$ respectively when $k \geq k_a$ and when $k \geq k_b$. Therefore, for the convergence analysis, these two cases are considered separately here.

Case (a): when $k \geq \max\{k_a, \hat{k}_2\}$, we combine the results in theorem 5, theorem 9, and theorem 10 with the triangle inequality (14) to derive the following recurrence,

$$\hat{\mathbf{E}}_k \parallel x^{k+1} - x^* \parallel \leq \gamma(\alpha_k)\ \hat{\mathbf{E}}_{k-1} \parallel x^k - x^* \parallel + Ce^{-\beta k} + 4M_0\alpha_k. \tag{34}$$

From assumption $(A1)$, $\theta_{min}(Q) > 0$. Moreover, $\alpha_k = \frac{\tau}{k+1}$ then we have

$$\gamma(\alpha_k) = \theta_{max}(G_{\alpha_k}) = \theta_{max}((I + \alpha_k Q)^{-1}) = 1 - \frac{\tau}{k+1}\theta_{min}(Q) + o(\frac{\tau}{k}). \tag{35}$$

Suppose in the notation of the product $\prod_i^k$, it equals to 1 when $i > k$. By recursively applying (34) till the $\hat{k}_a$th iteration, we derive

$$\begin{aligned}
\hat{\mathbf{E}}_k||x^{k+1} - x^*|| &\leq \hat{\mathbf{E}}_{\hat{k}_a-1}||x^{\hat{k}_a} - x^*|| \prod_{i=\hat{k}_a+1}^{k} \gamma(\alpha_i) \\
&\quad + \sum_{i=\hat{k}_a+1}^{k} \left[\left(Ce^{-\beta i} + 4M_0\alpha_i \right) \prod_{j=i+1}^{k} \gamma(\alpha_j) \right] \\
&\lesssim \hat{\mathbf{E}}_{\hat{k}_a-1}||x^{\hat{k}_a} - x^*|| \left[\prod_{i=\hat{k}_a+1}^{k} \left(1 - \frac{\tau}{i+1}\theta_{min}(Q) \right) \right] \\
&\quad + \sum_{i=\hat{k}_a+1}^{k} \left[\left(Ce^{-\beta i} + \frac{4M_0\tau}{i+1} \right) \prod_{j=i+1}^{k} \left(1 - \frac{\tau}{j+1}\theta_{min}(Q) \right) \right].
\end{aligned}$$

Let $\Lambda \triangleq \tau \cdot \theta_{min}(Q)$ and $Z_\Lambda(k', k) \triangleq \prod_{i=k'}^{k}(1 - \frac{\Lambda}{i+1})$ for any positive integers $k' < k$. Then the above inequality can be rewritten below.

$$\hat{\mathbf{E}}_k||x^{k+1} - x^*|| \lesssim \hat{\mathbf{E}}_{\hat{k}_a-1}||x^{\hat{k}_a} - x^*|| Z_\Lambda(\hat{k}_a + 1, k) + \sum_{i=\hat{k}_a+1}^{k} (Ce^{-\beta i} + \frac{4M_0\tau}{i+1}) Z_\Lambda(i+1, k). \tag{36}$$

When $k \gg \hat{k}_a = \max\{k_a, \hat{k}_2\}$, by applying theorem 11 to inequality (36) we have,

$$\hat{\mathbf{E}}_k \parallel x^{k+1} - x^* \parallel \lesssim \hat{\mathbf{E}}_{\hat{k}_a} \parallel x^{\hat{k}_a} - x^* \parallel \left(\frac{\hat{k}_a - 1}{k} \right)^{\Lambda} + C\ \frac{k^{|1-\Lambda|_+}}{e^{\beta k}} + \frac{4M_0\tau}{\Lambda}.$$

Case (b): when $k \geq \hat{k}_b = \max\{k_b, \hat{k}_1\}$, we combine the results in theorem 5, theorem 9, and theorem 10 with the triangle inequality (14) and derive,

$$\hat{\mathbf{E}}_k \parallel x^{k+1} - x^* \parallel \leq \gamma(\alpha_k)\ \hat{\mathbf{E}}_{k-1} \parallel x^k - x^* \parallel + Ce^{-\beta k} + 4M_0M_1\alpha_k^2.$$

With the first-order approximation of $\gamma(\alpha_k)$, we follow the similar derivation of the recursion as in case (a) and derive

$$\begin{aligned}
\hat{\mathbf{E}}_k \parallel x^{k+1} - x^* \parallel \ &\lesssim\ \hat{\mathbf{E}}_{\hat{k}_b-1} \parallel x^{\hat{k}_b} - x^* \parallel Z_\Lambda(\hat{k}_b + 1, k) \\
&\quad + \sum_{i=\hat{k}_b+1}^{k} \left(Ce^{-\beta i} + \frac{4M_0M_1\tau^2}{(i+1)^2} \right) Z_\Lambda(i+1, k).
\end{aligned}$$

With some limiting properties in theorem 11, we have the asymptotical result stated in (b). □

In theorem 12, statement (a) shows the non-asymptotic convergence result with a constant error when k is small, while statement (b) proves that when k larger than a finite number $\hat{k}_b$, the distance between the incumbent solution and the optimum in expectation is controlled by three diminishing terms. We also observe that when $\Lambda > 1$ the overall rate of convergence is $O(\frac{1}{k})$. Therefore, we conclude that by appropriately choosing an initial step size and a sequence of diminishing step sizes, the sequence of incumbent solutions obtained in SD converges to the optimum with a sublinear convergence rate. Moreover, theorem 12 indicates that when $k \geq \hat{k}_b$, the larger Λ is, the faster the solution sequence converges. In fact $\Lambda = \tau \cdot \theta_{min}(Q)$ implies that Λ can be large if we choose a large initial step size τ. However, if the initial step size is too large, the number of iterations that are required to generate a stable set of extreme points or faces may become very large. This kind of stability has been discussed experimentally by Sen and Liu in [27] which was interpreted as the shadow price stability and taken into the consideration in the design of the in-sample stopping rule. Therefore, our analysis indicates a trade-off between the convergence rate within a neighborhood of the optimal solution and the rate of entering into that neighborhood. However, the effect of initial step size on this trade-off needs more study to be fully understood.

5 SD Stopping Rule: Consistent Bootstrap Estimators

For the SD algorithm in two-stage SQLP and SQQP problems, in practice the goal is to end with a good incumbent solution at some finite iteration. In other words, we need a stopping rule to recognize whether the solution produced by the SD algorithm can be accepted with a predetermined accuracy. In this section we design a stopping rule using bootstrap estimators to test whether the optimality gap between the optimal value and the objective value at $\hat{x}^k$ is statistically acceptable or not.

Higle and Sen in [8] first applied bootstrap directly to linear programs of primal and dual in a stopping rule. The dual multiplier in the solution pair might be infeasible in resampled linear programs, therefore it requires the solution of the dual problem at each resampled instance. Then Higle and Sen in [10] proposed proximal mapping updates in SLP for better convergence and showed that the dual multiplier can remain feasible in all resampled problems. Accordingly, by using bootstrap they took the duality gap at the incumbent solution in the proximal mapping of the value function approximation as the measurement of optimality. However, such measurement does not fully explain the optimality gap at the incumbent solution because the duality gap is associated with the value function approximation, not the true objective function. In this section, for two-stage SQLP and SQQP, we propose an approximate optimality gap and design an "in-sample" stopping rule using consistent bootstrap estimators for the approximate optimality gap. This scheme of an "in-sample" stopping rule should be applicable in two-stage SLP problems as well.

At iteration k, suppose the SD algorithm has generated a set of k samples $\{\omega^i\}_{i=1}^k$ and a sequence of incumbent solutions $\{\hat{x}^i\}_{i=1}^k$. To test the optimality of the incumbent solution $\hat{x}^k$, we define the optimality gap d_k as the difference between the current objective value $f(\hat{x}^k)$ and the optimal value $f(x^*)$, i.e., $d_k \triangleq f(\hat{x}^k) - f(x^*)$. Since we do not have any knowledge about the mathematical formulation of the distribution of random variables ω and nor have the optimal solution x^*, we need to construct statistical estimations of $f(\hat{x}^k)$ and $f(x^*)$ in order to estimate the optimality gap d_k. We first notice that $f(\hat{x}^k)$ has an unbiased estimator, the sample average value $F_k(\hat{x}^k)$ with a sample set $\{\overline{\omega}^i\}_{i=1}^k$ of size k independent of $\{\omega^i\}_{i=1}^k$. However, as for $f(x^*)$, the sample average minimal value $\min_{x \in X} F_k(x)$ is not only biased but also computationally expensive when

k is large. Therefore, we consider constructing a statistical lower bound of $f(x^*)$ using the value function approximation satisfying $\hat{\mathbf{E}}_k\ [\min_{x\in X} f_k(x)] \leq f(x^*)$. This is because under assumption $(A4)$, f_k is a lower bound of the sample average functions F_k, i.e. $f_k(x) \leq F_k(x)$ for all $x \in X$. By taking the expectation and minimization, with Jensen's inequality we have

$$\hat{\mathbf{E}}_k\ [\min_{x\in X}\ f_k(x)] \leq \min_{x\in X}\ \hat{\mathbf{E}}_k\ [f_k(x)] \leq f(x^*).$$

Let $\mathbf{E}'_k$ denote the expectation with respect to the product of probability measures of the sample set $\{\overline{\omega}^i\}_{i=1}^k$ used in the construction of the unbiased estimator of $f(\hat{x}^k)$. Then the optimality gap can be bounded as below,

$$d_k = f(\hat{x}^k) - f(x^*) \leq\ \mathbf{E}'_k\ [F_k(\hat{x}^k)] - \hat{\mathbf{E}}_k\ [\min_{x\in X} f_k(x)]. \tag{37}$$

Motivated by (37), we aim to obtain a conservative confidence interval of the optimality gap by using the consistent bootstrap estimators of statistics $F_k(\hat{x}^k)$ and $\min_{x\in X} f_k(x)$ with multiple replications. The idea of bootstrap is to estimate the distribution of a statistic of independent observations by the distribution of the same statistic of the resample set. This comes from the intuition that the sample set should be a good representation of the population of the true data. Under mild conditions, the bootstrap estimator yields a consistent estimator of a statistic's distribution, which means the distribution of bootstrap estimator is uniformly close to the asymptotic distribution of original statistic when the sample size is large. A definition of consistency is provided here following Definition 2.1 in [13].

Definition 13. *Let $\{X_i\}_{i=1}^n$ be a random sample generated from a probability distribution F_0 and $\{Z_i\}_{i=1}^n$ be a random sample generated from the empirical distribution F_n of $\{X_i\}_{i=1}^n$. We consider the statistic $T_n = T(X_1, X_2, \ldots, X_n)$ and its estimator $\widetilde{T}_n = T(Z_1, Z_2, \ldots, Z_n)$. We define*

$$G_n(\tau, F_n) \triangleq \mathbb{P}(\widetilde{T}_n \leq \tau), \quad G_\infty(\tau, F_0) \triangleq \lim_{n\to\infty} \mathbb{P}(T_n \leq \tau).$$

Let $\mathbb{P}_n$ be the joint probability measure of sample $\{Z_i\}_{i=1}^n$. Then the bootstrap estimator $\widetilde{T}_n$ is a consistent estimator of T_n if for each $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}_n \left(\sup_{\tau\in\mathbb{R}} |G_n(\tau, F_n) - G_\infty(\tau, F_0)| > \epsilon \right) = 0. \tag{38}$$

In the literature, Singh [31] proved that pivotal statistic and asymptotically pivotal statistic have consistent bootstrap estimators when the sample size increases to infinity. Moreover, Mammen [16] shows that the sample average of the functions has consistent bootstrap estimators if and only if the statistic satisfies the asymptotic normality condition. According to the literature study on the consistency properties, we construct the bootstrap estimators of $F_k(\hat{x}^k)$ and $\min_{x\in X} f_k(x)$ respectively. First we independently generate two resample sets $\{\overline{\omega}^i\}_{i=1}^k$ for the bootstrap estimator of $F_k(\hat{x}^k)$ and $\{\underline{\omega}^i\}_{i=1}^k$ for the bootstrap estimator of $\min_x f_k(x)$ from the empirical distribution of the sample set $\{\omega^i\}_{i=1}^k$. Then the bootstrap estimator of $F_k(\hat{x}^k)$ is constructed as below.

$$\bar{F}_k(\hat{x}^k) \triangleq \frac{1}{2}(\hat{x}^k)^\top Q\hat{x}^k + c^\top \hat{x}^k + \frac{1}{k}\sum_{i=1}^k h(\hat{x}^k, \overline{\omega}^i). \tag{39}$$

Since $F_k(\hat{x}^k)$ is the sample average of objective values, it is asymptotically pivotal and $\bar{F}_k(\hat{x}^k)$ is a consistent bootstrap estimator.

Then we consider the bootstrap estimator of $\min_{x\in X} f_k(x)$. Recall that SASA functions or SAQA functions $\{h_j^k\}_{j\in\mathcal{J}_k}$ are the sample average of functions defined in line 8 in Algorithm 1 and line 8 in Algorithm 2. Thus from Mammen in [16], SASA functions and SAQA functions have consistent bootstrap estimators. Formally, a resampled SASA function in two-stage SQLP is constructed such that for each $j \in \mathcal{J}_k$,

$$\tilde{h}_j^k(x) \triangleq \begin{cases} \frac{1}{k}\sum_{i=1}^{|j|}(\pi'_{j,i})^\top[\xi(\underline{\omega}^i) - C(\underline{\omega}^i)x] & \text{if } j \neq 0, \\ 0 & \text{if } j = 0, \end{cases} \tag{40}$$

$\pi'_{j,i} = \begin{cases} \pi_{j,l} & \text{if } j > 0, \\ \hat{\pi}_{j,l} & \text{if } j < 0, \end{cases}$ where l is a constant such that $\underline{\omega}^i = \omega^l$, $\pi_{j,l}$ and $\hat{\pi}_{j,l}$ are computed respectively following line 6 and line 7 in Algorithm 1.
In addition, a resampled SAQA function in two-stage SQQP is constructed as

$$\tilde{h}_j^k(x) \triangleq \begin{cases} \frac{1}{k}\sum_{i=1}^{|j|} g_{QQ}(t'_{j,i}, s'_{j,i}; x, \underline{\omega}^i) & \text{if } j \neq 0, \\ 0 & \text{if } j = 0, \end{cases} \tag{41}$$

Let l be the constant such that $\underline{\omega}^i = \omega^l$. Then in (41), $t'_{j,i} = \begin{cases} t_{j,l} & \text{if } j > 0, \\ \hat{t}_{j,l} & \text{if } j < 0, \end{cases}$, $s'_{j,i} = \begin{cases} s_{j,l} & \text{if } j > 0, \\ \hat{s}_{j,l} & \text{if } j < 0, \end{cases}$, and $t_{j,l}$, $\hat{t}_{j,l}$, $s_{j,l}$ and $\hat{s}_{j,l}$ are computed respectively following line 6 and line 7 in Algorithm 2.
Accordingly, the resampled lower bound estimation is

$$\min_{x\in X} \bar{f}_k(x) \triangleq \frac{1}{2}x^\top Qx + c^\top x + \max\{\tilde{h}_j^k(x), j \in \mathcal{J}_k\}. \tag{42}$$

Hence, based on the bound of the optimality gap d_k in (37) and bootstrap estimators constructed in (39) and (42), we design the "In-sample" stopping rule in table 3 to test the statistical performance of the optimality gap.

6 Conclusion and Further Directions

This paper studies the convergence rate of SD algorithms for two-stage SQLP and SQQP problems in which we have the quadratic program in the first stage and the linear/quadratic program in the second stage. With the assumption of the positive definiteness of the quadratic matrices, we present the contraction property of stochastic proximal mapping with constraints. We then prove a sublinear convergence rate of SD in SQLP and SQQP problems. The effect of the curvature of the second-stage problem in SQQP has not been incorporated into the convergence rate analysis. However, it has the potential to improve the rate with modifications in the convergence analysis. A deeper look at the convergence analysis reveals an interesting trade-off. By appropriately choosing a large initial step size τ such that $\Lambda = \tau \cdot \theta_{min}(Q)$ is greater than one, we could increase the asymptotic convergence rate. However, it should not be too large because of the trade-off between the convergence rate within a neighborhood of the optimum and the rate to enter into that neighborhood.
There are previous works which gave the convergence rate of algorithms for the SAA scenario problem. For example, Chen et al. in [4] gave the convergence rate of Newton's method for two-stage SQLP to the optimal point in the SAA problem. However, to the best of our knowledge, there is no prior theoretical analysis giving a convergence rate of SD for solving the two-stage

Table 3: "In-sample" Stopping Rule

At iteration k, with the sample set $\{\omega^i\}_{i=1}^k$ and the incumbent solution $\hat{x}^k$,

1. **Initialization:** let ϵ be the predetermined parameter and $\widetilde{F}^k$ be the empirical distribution of the sample data $\{\omega^i\}_{i=1}^k$.

For each replication $m = 1, \ldots, M (M \geq 30)$, do the following.

2. **Bootstrap estimators:**
   - generate a resample $\{\overline{\omega}^i\}_{i=1}^k$ of size k from $\widetilde{F}^k$. Compute the bootstrap estimator of sample average value $\bar{F}_k^m(\hat{x}^k)$ according to (39).
   - generate a new resample $\{\underline{\omega}^i\}_{i=1}^k$ of size k from $\widetilde{F}^k$. Compute the bootstrap estimator $\min_{x \in X} \bar{f}_k^m(x)$ according to (42).
   - compute the optimality gap estimation $\hat{d}_k^m \triangleq \bar{F}_k^m(\hat{x}^k) - \min_{x \in X} \bar{f}_k^m(x)$.
3. **Optimality gap:** compute the mean of the optimality gap estimations $\bar{d}_k \triangleq \sum_{m=1}^M \hat{d}_k^m / M$ and the sample variance $Var_k \triangleq \frac{1}{M-1} \sum_{m=1}^M (\hat{d}_k^m - \bar{d}_k)^2$. If $\bar{d}_k \leq t_{0.01}^{M-1} \sqrt{Var_k / M} + \epsilon$, then the incumbent solution x^k is accepted as the approximation of the optimum. Otherwise, we move forward to the next iteration of the SD algorithm.

stochastic programming in the decision space. Our work thus provides the theoretical support of the convergence of SD algorithms in two-stage stochastic quadratic programming problems. Specifically, our work proves a convergence rate $O(N^{-1})$ of SD in the distance $\left\|x^{N+1} - x^*\right\|$ for two-stage SQLP and SQQP under assumptions $(B1)$, $(B3)$, $(B4)$, and $(B5)$ on the optimal solution besides the common assumptions $(A1)$, $(A2)$, $(A3)$ and $(B2)$. Compared to the SA method, SD uses subgradients for iterative updates differently. It is possible, though still open, that SD could obtain a similar non-asymptotic result as SA by relaxing those assumptions on the optimal solution. However, since the SA method does not highlight the contraction result which appears in the proximal iteration, it is not clear that whether SA can be extended to obtain the same higher rate of convergence under those conditions at the optimal solution.

Acknowledgement

This work was supported, in part, by NSF Grant CMMI-1822327, CMMI 1538605, ECCS 1548847 and AFOSR Grant FA9550-15-1-0267. The authors would like to thank Jong-Shi Pang for his interest in this line of work, and his timely comments on a previous version of this paper.

References

[1] Dirk Banholzer, Jörg Fliege, and Ralf Werner. On almost sure rates of convergence for sample average approximations. *SIAM Journal on Optimization*, 2017.

[2] Rasmus Bro and Sijmen De Jong. A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 11(5):393–401, 1997.

[3] Donghui Chen and Robert J Plemmons. Nonnegativity constraints in numerical analysis. *The Birth of Numerical Analysis*, 10, 2009.

[4] Xiaojun Chen, Liqun Qi, and Robert S Womersley. Newton's method for quadratic stochastic programs with recourse. *Journal of Computational and Applied Mathematics*, 60(1):29–46, 1995.

[5] Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954.

[6] Andrew R Conn, Nicholas IM Gould, and Ph L Toint. *Trust region methods*, volume 1. SIAM, 2000.

[7] William S Dorn. Duality in quadratic programming. *Quarterly of Applied Mathematics*, 18(2):155–162, 1960.

[8] Julia L Higle and Suvrajeet Sen. Stochastic decomposition: An algorithm for two-stage linear programs with recourse. *Mathematics of Operations Research*, 16(3):650–669, 1991.

[9] Julia L Higle and Suvrajeet Sen. Finite master programs in regularized stochastic decomposition. *Mathematical Programming*, 67(1-3):143–168, 1994.

[10] Julia L Higle and Suvrajeet Sen. Statistical approximations for stochastic linear programming problems. *Annals of Operations Research*, 85:173–193, 1999.

[11] David C Hoaglin and Roy E Welsch. The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22, 1978.

[12] Tito Homem-de Mello and Güzin Bayraksan. Monte carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85, 2014.

[13] Joel L Horowitz. The bootstrap. In *Handbook of Econometrics*, volume 5, pages 3159–3228. Elsevier, 2001.

[14] Charles L Lawson and Richard J Hanson. *Solving least squares problems*, volume 15. SIAM, 1995.

[15] Wai-Kei Mak, David P Morton, and R Kevin Wood. Monte carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24(1):47–56, 1999.

[16] Enno Mammen. *When does bootstrap work?: asymptotic results and simulations*, volume 77. Springer Science & Business Media, 2012.

[17] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[18] Arkadii Nemirovskii, David Borisovich Yudin, and ER Dawson. Problem complexity and method efficiency in optimization. 1983.

[19] Welington Oliveira, Claudia Sagastizábal, and Susana Scheimberg. Inexact bundle methods for two-stage stochastic programming. *SIAM Journal on Optimization*, 21(2):517–544, 2011.

[20] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[21] Daniel Ralph and Huifu Xu. Convergence of stationary points of sample average two-stage stochastic programs: A generalized equation approach. *Mathematics of Operations Research*, 36(3):568–592, 2011.

[22] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[23] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[24] Johannes O Royset. Optimality functions in stochastic programming. *Mathematical Programming*, 135(1-2):293–321, 2012.

[25] Ernest K Ryu and Stephen Boyd. Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. *Author website, early draft*, 2014.

[26] Rüdiger Schultz. Strong convexity in stochastic programs with complete recourse. *Journal of Computational and Applied Mathematics*, 56(1-2):3–22, 1994.

[27] Suvrajeet Sen and Yifan Liu. Mitigating uncertainty via compromise decisions in two-stage stochastic linear programming: Variance reduction. *Operations Research*, 64(6):1422–1437, 2016.

[28] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[29] Alexander Shapiro and Tito Homem-de Mello. On the rate of convergence of optimal solutions of monte carlo approximations of stochastic programs. *SIAM Journal on Optimization*, 11(1):70–86, 2000.

[30] Alexander Shapiro and Arkadi Nemirovski. On complexity of stochastic programming problems. In *Continuous Optimization*, pages 111–146. 2005.

[31] Kesar Singh. On the asymptotic accuracy of efron's bootstrap. *The Annals of Statistics*, pages 1187–1195, 1981.

[32] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[33] Huifu Xu. Uniform exponential convergence of sample average random functions under general sampling with applications in stochastic programming. *Journal of Mathematical Analysis and Applications*, 368(2), 2010.

A Proofs of Lemmas

theorem 6 *Suppose the assumptions* $(A1)-(A4)$, $(B1)$, $(B3)$ *and* $(B5)$ *hold for the two-stage SQLP* (3) *or SQQP* (5) *considered. Let* $\{x^k\}$ *be the sequence of incumbent solutions generated by SD,* $I^k \triangleq \{r : a_r^\top T^X_{\alpha F_k}(x^k) = b_r\}$ *and* $I^* \triangleq \{r : a_r^\top x^* = b_r\}$. *Then there exists a finite number* $\hat{k}_2$, *such that for any* $k \geq \hat{k}_2$, $I^k = I^*$ *with probability 1.*

Proof. Under the assumptions $(A1), (A2)$ and $(A3)$, as k increases to infinity, the sample average function $F_k(x)$ uniformly converges to $f(x)$ on X (see Theorem 7.48 in [28]). Moreover, x^k converges to x^* with probability 1 from theorem 3. Thus, $T^X_{\alpha F_k}(x^k)$ converges to x^* with probability 1 as k increases to infinity.
With the assumption $(B3)$, there exists a constant k_2 such that for any $k \geq k_2$, $h(x, \omega)$ is differentiable at $T^X_{\alpha F_k}(x^k)$ for almost every $\omega \in \Omega$. Then we present the optimality conditions of the original optimization problem at the solution pair (x^*, λ^*) and of the proximal map at the solution pair $(T_{\alpha F_k}(x^k), \lambda^k)$ respectively as follows.

$$Q\ x^* + \mathbf{E}\ \nabla[h(x^*, \widetilde{\omega})] = -A^\top \lambda^*$$

$$Q\ T^X_{\alpha F_k}(x^k) + \frac{1}{k}\sum_{i=1}^{k} \nabla h(T^X_{\alpha F_k}(x^k), \omega^i) + \frac{1}{\alpha}(T^X_{\alpha F_k}(x^k) - x^k) = -A^\top \lambda^k.$$

As k increases to infinity, from the law of large numbers the difference of the right hand sides of the above two optimality conditions converges to zero. Therefore,

$$\lim_{k\to\infty} A^\top(\lambda^k - \lambda^*) = 0. \tag{43}$$

Recall that $\lim_{k\to\infty} T^X_{\alpha F_k}(x^k) = x^*$. Then there exists a finite number k_2' such that for any $k \geq k_2'$, if $r \in (I^*)^c$, then $a_r^\top T^X_{\alpha F_k}(x^k) < b_r$. This yields that $(I^*)^c \subseteq (I^k)^c$ which is equivalent to

$$I^k \subseteq I^* \text{ for any } k \geq k_2'. \tag{44}$$

We denote the index set $R^k \triangleq \{r : \lambda^k_{(r)} > 0\}$. Because of the slackness complementarity at x^k, we have $R^k \subseteq I^k \subseteq I^*$ for any $k \geq k_2'$. In other words,

$$\lambda^*_{(r)} = \lambda^k_{(r)} = 0 \quad \text{for any } r \in (I^*)^c, \text{ and } k \geq k_2'. \tag{45}$$

Then we show that for any $r \in I^*$, $\lim_{k\to\infty} \lambda^k_{(r)} = \lambda^*_{(r)}$. First the limit (43) with (45) can be rewritten as follows.

$$\lim_{k\to\infty} \sum_{r=1}^{m} (\lambda^k_{(r)} - \lambda^*_{(r)}) a_r = \lim_{k\to\infty} \sum_{r\in I^*} (\lambda^k_{(r)} - \lambda^*_{(r)}) a_r = 0$$

From the assumption $(B1)$ that the active constraint are linearly independent at the optimal solution x^*, we must have

$$\lim_{k\to\infty} \lambda^k_{(r)} = \lambda^*_{(r)} \quad \text{for any } r \in I^*.$$

Because of the strict complementarity in the assumption $(B5)$ at x^*, $\lambda_{(r)} > 0$ for any $r \in I^*$. Therefore, there exists a finite number k_2'' such that for any $k > k_2''$, $\lambda^k_{(r)} > 0$ for any $r \in I^*$. It follows that for any $k \geq k_2''$, $I^* \subseteq I^k$. Combining with (44), we thus have $I^k = I^*$ with probability 1 for any $k \geq \hat{k}_2 = \max\{k_2, k_2', k_2''\}$. □

theorem 7 *Suppose $\check{g}(z) = \frac{1}{2} z^\top Q z$ where Q is a positive definite matrix and $\bar{A}$ is a matrix consisting of the linearly independent row vectors. Then the proximal mapping $T_{\alpha\check{g}}$ over a feasible solution set $\bar{X} = \{x : \bar{A}x = \bar{b}\}$ satisfies the following,*

$$T_{\alpha\check{g}} x - T_{\alpha\check{g}} x' = K_\alpha (x - x'), \quad \forall x, x' \in \bar{X},$$

where $G_\alpha = (I + \alpha Q)^{-1}$ *and* $K_\alpha = G_\alpha - G_\alpha \bar{A}^\top (\bar{A} G_\alpha \bar{A}^\top)^{-1} \bar{A} G_\alpha$. *Moreover, we have*

$$\| T_{\check{g}} x - T_{\check{g}} x' \| \leq \theta_{max}(G_\alpha) \cdot \| x - x' \|,$$

where $\theta_{max}(\cdot)$ *denotes the largest eigenvalue of the matrix.*

Proof. Assume $x \neq x'$ and $T_{\alpha g} x \neq T_{\alpha g} x'$, otherwise there is nothing to show. For the proximal mapping $T_{\alpha \check{g}} x$, the optimal solution pair $(T_{\alpha \check{g}} x,\ \lambda)$ satisfies the following KKT conditions.

$$\begin{pmatrix} I + \alpha Q & \bar{A}^\top \\ \bar{A} & 0 \end{pmatrix} \begin{pmatrix} T_{\alpha \check{g}} x \\ \lambda \end{pmatrix} = \begin{pmatrix} x \\ \bar{b} \end{pmatrix}. \tag{46}$$

Since the matrix $\bar{A}$ has linearly independent rows, from the formula of the inverse of a block matrix, the unique global optimal solution can be represented as,

$$T_{\alpha \check{g}} x = K_\alpha x + J_\alpha \bar{b}, \tag{47}$$

where $K_\alpha = G_\alpha - G_\alpha \bar{A}^\top (\bar{A} G_\alpha \bar{A}^\top)^{-1} \bar{A} G_\alpha$ and $G_\alpha = (I + \alpha Q)^{-1}$. The same representation also holds for x' such that $T_{\alpha \check{g}} x' = K_\alpha x' + J_\alpha \bar{b}$. Therefore

$$T_{\alpha \check{g}} x - T_{\alpha \check{g}} x' = K_\alpha (x - x').$$

Moreover,

$$\frac{\| T_{\alpha \check{g}} x - T_{\alpha \check{g}} x' \|}{\| x - x' \|} = \Big(\frac{(x - x')^T K_\alpha^2 (x - x')}{\| x - x' \|^2} \Big)^{1/2} \leq \big(\theta_{max}(K_\alpha^2) \big)^{1/2} = \theta_{max}(K_\alpha).$$

From the definition, $K_\alpha = (G_\alpha^{1/2})\ (I - G_\alpha^{1/2} \bar{A}^\top (\bar{A} G_\alpha \bar{A}^\top)^{-1} \bar{A} G_\alpha^{1/2})\ (G_\alpha^{1/2})$. Notice that $G_\alpha^{1/2} \bar{A}^\top (\bar{A} G_\alpha \bar{A}^\top)^{-1} \bar{A} G_\alpha^{1/2}$ is a hat matrix which is defined formally in [11] with eigenvalues 0 and 1. We then derive

$$\begin{aligned} \theta_{max}(K_\alpha) &\leq \theta_{max}(G_\alpha^{1/2})\ \theta_{max}(I - G_\alpha^{1/2} \bar{A}^\top (\bar{A} G_\alpha \bar{A}^\top)^{-1} \bar{A} G_\alpha^{1/2})\ \theta_{max}(G_\alpha^{1/2}) \\ &\leq \theta_{max}(G_\alpha) < 1. \end{aligned}$$

Hence, the proximal mapping $T_{\check{g}} x$ has the contraction property satisfying that

$$\| T_{\check{g}} x - T_{\check{g}} x' \| \leq \theta_{max}(G_\alpha) \cdot \| x - x' \|, \qquad \forall x, x' \in \bar{X}.$$

□

theorem 11 *Let* Λ *be a positive constant. For any two integers* k *and* k' *satisfying* $k' \ll k$, *let*

$$Z_\Lambda(k', k) \triangleq \Pi_{i=k'}^{k} (1 - \frac{\Lambda}{i}).$$

We have the following three limiting properties regarding to $Z_\Lambda(k', k)$ *when* k' *is a large number.*

1. $Z_\Lambda(k', k) \lesssim \left(\frac{k'}{k} \right)^\Lambda$.
2. $\sum_{i=\hat{k}+1}^{k} e^{-\beta i} Z_\Lambda(i, k) \lesssim \begin{cases} \frac{1}{e^{\beta k}} & \Lambda \geq 1 \\ \frac{k^{1-\Lambda}}{e^{\beta k}} & \Lambda < 1. \end{cases}$

3. $\sum_{i=\hat{k}+1}^{k} \frac{1}{i^2} Z_\Lambda(i,k) \lesssim \begin{cases} \frac{log(k)}{k} & \Lambda = 1 \\ \frac{1}{(\Lambda-1)k} & \Lambda > 1 \\ \frac{1}{(1-\Lambda)k^\Lambda} & \Lambda < 1 \end{cases}$

4. $\sum_{i=\hat{k}+1}^{k} \frac{1}{i} Z_\Lambda(i,k) \lesssim \frac{1}{\Lambda}$

Proof. Since k' is a large integer, we take the approximation that $log(1-\frac{\Lambda}{i}) \sim \frac{\Lambda}{i}$ for $i \geq k'$. Then

$$Z_\Lambda(k',k) = exp\{\sum_{i=k'}^{k} log(1-\frac{\Lambda}{i})\} \sim exp\{-\sum_{i=k'}^{k} \frac{\Lambda}{i}\}.$$

Besides,

$$\sum_{i=k'}^{k} \frac{\Lambda}{i} \leq \int_{k'}^{k} \frac{\Lambda}{u} \mathrm{d}u = \Lambda\ log(\frac{k}{k'}).$$

Therefore, claim 1 holds.

Next consider the summation $\sum_{i=\hat{k}+1}^{k} e^{-\beta i} Z_\Lambda(i,k)$. Since $Z_\Lambda(k',k) \lesssim \left(\frac{k'}{k}\right)^\Lambda$, we have

$$\sum_{i=\hat{k}+1}^{k} e^{-\beta i} Z_\Lambda(i,k) \lesssim \sum_{i=\hat{k}+1}^{k} e^{-\beta i} \left(\frac{i}{k}\right)^\Lambda \leq \frac{1}{k^\Lambda} \int_{\hat{k}+1}^{k} e^{-\beta u} \cdot u^\Lambda \mathrm{d}u.$$

Through some computation, the integral can be bounded such that

$$\int_{\hat{k}+1}^{k} e^{-\beta u} \cdot u^\Lambda \mathrm{d}u \ \lesssim \ \begin{cases} \frac{k^\Lambda}{e^{\beta k}} & \Lambda \geq 1 \\ \frac{k}{e^{\beta k}} & \Lambda < 1 \end{cases}$$

Therefore, the summation is bounded as follows

$$\sum_{i=\hat{k}+1}^{k} e^{-\beta i} Z_\Lambda(i,k) \lesssim \ \begin{cases} \frac{1}{e^{\beta k}} & \Lambda \geq 1 \\ \frac{k^{1-\Lambda}}{e^{\beta k}} & \Lambda < 1 \end{cases}$$

We then consider the summation $\sum_{i=\hat{k}+1}^{k} \frac{1}{i^2} Z_\Lambda(i,k)$.

$$\sum_{i=\hat{k}+1}^{k} \frac{1}{i^2} Z_\Lambda(i,k) \lesssim \sum_{i=\hat{k}+1}^{k} \frac{1}{i^2} \left(\frac{i}{k}\right)^\Lambda \leq \frac{1}{k^\Lambda} \int_{\hat{k}+1}^{k} u^{\Lambda-2} \mathrm{d}u$$

Therefore,

$$\sum_{i=\hat{k}+1}^{k} \frac{1}{i^2} Z_\Lambda(i,k) \lesssim \begin{cases} \frac{log(k)}{k} & \Lambda = 1 \\ \frac{1}{(\Lambda-1)k} & \Lambda > 1 \\ \frac{1}{(1-\Lambda)k^\Lambda} & \Lambda < 1 \end{cases}$$

We finally consider the summation $\sum_{i=\hat{k}+1}^{k} \frac{1}{i} Z_\Lambda(i,k)$.

$$\sum_{i=\hat{k}+1}^{k} \frac{1}{i} Z_\Lambda(i,k) \lesssim \sum_{i=\hat{k}+1}^{k} \frac{1}{i} \left(\frac{i}{k}\right)^\Lambda \leq \frac{1}{k^\Lambda} \int_{\hat{k}+1}^{k} u^{\Lambda-1} \mathrm{d}u \approx \frac{1}{\Lambda}.$$

□