Solving Stochastic Optimization with Expectation Constraints Efficiently by a Stochastic Augmented Lagrangian-Type Algorithm

Liwei Zhang
School of Mathematical Sciences, Dalian University of Technology, 116023 Dalian, China, lwzhang@dlut.edu.cn

Yule Zhang
School of Science, Dalian Maritime University, 116085 Dalian, China, ylzhang@dlmu.edu.cn

Jia Wu
School of Mathematical Sciences, Dalian University of Technology, 116023 Dalian, China, wujia@dlut.edu.cn

Xiantao Xiao
School of Mathematical Sciences, Dalian University of Technology, 116023 Dalian, China, xtxiao@dlut.edu.cn

This paper considers the problem of minimizing a convex expectation function with a set of inequality convex expectation constraints. We propose a stochastic augmented Lagrangian-type algorithm, namely the stochastic linearized proximal method of multipliers, to solve this convex stochastic optimization problem. This algorithm can be roughly viewed as a hybrid of stochastic approximation and the traditional proximal method of multipliers. Under mild conditions, we show that this algorithm exhibits $O(K^{-1/2})$ expected convergence rates for both objective reduction and constraint violation if parameters in the algorithm are properly chosen, where $K$ denotes the number of iterations. Moreover, we show that, with high probability, the algorithm has $O(\log(K)K^{-1/2})$ constraint violation bound and $O(\log^{3/2}(K)K^{-1/2})$ objective bound. Numerical results demonstrate that the proposed algorithm is efficient.

Key words: stochastic approximation; linearized proximal method of multipliers; expectation constrained stochastic program; expected convergence rate; high probability bound

1. Introduction

In this paper, we consider the following stochastic optimization problem

$$\min_{x \in \mathcal{C}} f(x) := \mathbb{E}[F(x, \xi)]$$

s.t. $g_i(x) := \mathbb{E}[G_i(x, \xi)] \leq 0$, $i = 1, \ldots, p.$

(1)
Here, $C \subset \mathbb{R}^n$ is a nonempty bounded closed convex set, $\xi$ is a random vector whose probability distribution is supported on $\Xi \subseteq \mathbb{R}^q$, $F : C \times \Xi \to \mathbb{R}$ and $G_i : C \times \Xi \to \mathbb{R}$, $i = 1, \ldots, p$. Let $\Phi$ be the feasible set of problem (1) as

$$\Phi := \{ x \in C : g_i(x) \leq 0, \ i = 1, \ldots, p \}. \quad (2)$$

We assume that

$$\mathbb{E}[F(x, \xi)] = \int_{\Xi} F(x, \xi) dP(\xi), \quad \mathbb{E}[G_i(x, \xi)] = \int_{\Xi} G_i(x, \xi) dP(\xi), \ i = 1, \ldots, p$$

are well defined and finite valued for every $x \in C$. Moreover, we assume that the functions $F(\cdot, \xi)$ and $G_i(\cdot, \xi)$ are continuous and convex on $C$ for almost every $\xi$. Hence, the expectation functions $f(\cdot)$ and $g_i(\cdot, \xi)$ are continuous and convex on $C$. Problems in the form of (1) are standard in stochastic programming (Ruszczynski and Shapiro 2003, Römisch 2003) and also arise frequently in many practical applications such as machine learning (Scott and Nowak 2005, Tong et al. 2016) and finance (Rockafellar and Uryasev 2000, Dentcheva and Ruszczyński 2003).

A computational difficulty of solving (1) is that expectation is a multidimensional integral and it cannot be computed with a high accuracy for large dimension $q$. In order to handle this issue, a popular approach is to use stochastic approximation (SA) technique which is based on the following general assumptions: (i) it is possible to generate i.i.d. sample $\xi^1, \xi^2, \ldots$, of realizations of random vector $\xi$; (ii) there is an oracle, which, for any point $(x, \xi) \in C \times \Xi$ returns stochastic subgradients $v_0(x, \xi)$, $v_1(x, \xi)$, $\ldots$, $v_p(x, \xi)$ of $F(x, \xi)$, $G_1(x, \xi)$, $\ldots$, $G_p(x, \xi)$ such that $v_i(x) = \mathbb{E}[v_i(x, \xi)]$, $i = 0, 1, \ldots, p$ are well defined and are subgradients of $f(\cdot)$, $g_1(\cdot)$, $\ldots$, $g_p(\cdot)$ at $x$, respectively, i.e., $v_0(x) \in \partial f(x), v_i(x) \in \partial g_i(x)$, $i = 1, \ldots, p$. 

Since the pioneering paper (Robbins and Monro 1951), due to low demand for computer memory and cheap computation cost at every iteration, SA type algorithms become widely used in stochastic optimization and machine learning, see, e.g. Pflug (1996), Bottou et al. (2018). If $f(\cdot)$ is twice continuously differentiable and strongly convex, in the classical analysis it is shown that the SA algorithm exhibits asymptotically optimal rate of convergence $\mathbb{E}[f(x^k) - f^*] = O(k^{-1})$, where $x^k$ is $k$th iterate and $f^*$ is the optimal value. An important improvement developed by Polyak (1990) and Polyak and Juditsky (1992) suggests that, larger stepsizes of SA algorithm can be adopted by consequently averaging the obtained
iterates. Moreover, Nemirovski et al. (2008) show that, without assuming smoothness and strong convexity, a properly modified SA method achieves the convergence rate $O(k^{-1/2})$ and remarkably outperforms the sample average approximation (SAA) approach for a certain class of convex stochastic problems. After the seminal work (Nemirovski et al. 2008), there are many significant results appeared, even for nonconvex stochastic optimization problems, see Bottou et al. (2018), Lan (2020) and references cited therein. Among all mentioned works, the feasible set is an abstract closed convex set and none of these SA algorithms are applicable to expectation constrained problems. The main reason is that the computation of projection $\Pi_\Phi$ is quite easy only when $\Phi$ is of a simple structure. However, when $\Phi$ is defined by (2), the computation is prohibitive.

As a first attempt for solving expectation constrained stochastic optimization problems with stochastic approximation technique, Lan and Zhou (2020) introduce a cooperative stochastic approximation (CSA) algorithm for solving (1) with single expectation constraint ($p = 1$), which is a stochastic counterpart of Polyak’s subgradient method (Polyak 1967). The authors show that CSA exhibits the optimal $O(1/\sqrt{K})$ rate of expected convergence, where $K$ is a fixed iteration number. In an online fashion, Yu et al. (2017) propose an algorithm (simply denoted by “YNW”) that can be easily extended to solve (1) with multiple expectation constraints. Under the Slater’s condition and the assumption that $\mathcal{C}$ is compact, they show that the algorithm can achieve $O(1/\sqrt{K})$ expected regret and $O(\log(K)/\sqrt{K})$ high probability regret. Xiao (2019) develops a penalized stochastic gradient (PSG) method and establishes its almost sure convergence and expected convergence rates. PSG can be roughly viewed as a hybrid of the classical penalty method for nonlinear programming and the stochastic quasi-gradient method (Wang et al. 2017) for stochastic composition problem. A stochastic level-set method (Lin et al. 2020), which ensures a feasible solution path with high probability, is proposed and analyzed. Akhtar et al. (2021) propose a conservative stochastic optimization algorithm (CSOA), which is in the similar primal-dual framework as PSG and YNW. In addition to CSOA, the authors also propose a projection-free algorithm named as FW-CSOA which can deal with the case that the projection $\Pi_\mathcal{C}$ is difficult to calculate. Yan and Xu (2022) study an adaptive primal-dual stochastic gradient method (APriD) for solving (1) and establish the convergence rate of $O(1/\sqrt{K})$ in terms of the objective error and the constraint violation.
All of the above mentioned methods for solving (1) can be cast into the family of stochastic first-order algorithms. Although the iteration in stochastic first-order algorithms is computationally cheap and these methods perform well for certain problems, there are plenty of practical experiences and evidences of their convergence difficulties and instability with respect to the choice of parameters. Recently, the success of augmented Lagrangian methods for various kinds of functional constrained optimization problems is witnessed. Parpas and Rustem (2007) study an augmented Lagrangian method for multistage stochastic problems. For solving semidefinite programming (SDP) problems, Zhao et al. (2010) consider an Newton-CG augmented Lagrangian method, which is shown to be very efficient even for large-scale SDP problems. Dentcheva et al. (2016) propose several methods based on augmented Lagrangian framework for optimization problems with stochastic-order constraints and analyze their convergence. Bai et al. (2021) study an augmented Lagrangian decomposition method for nonconvex chance-constrained problems, in which a convex subproblem and a 0-1 knapsack subproblem are solved at each iteration. The aim of this paper is to develop an efficient stochastic approximation-based augmented Lagrangian-type method for solving (1). To the best of our knowledge, this is still limited in the literature.

Zhang et al. (2020) propose a stochastic proximal method of multipliers (PMMSopt) for solving problem (1) and show that PMMSopt exhibits $O(K^{-1/2})$ convergence rates for both objective reduction and constraint violation. PMMSopt is partially inspired by the classic proximal method of multipliers (Rockafellar 1976), which is modeled through an augmented Lagrangian with an extra proximal term. However, the subproblem is difficult to solve, that makes PMMSopt an unimplementable algorithm, and hence no numerical results are given.

In this paper, based on PMMSopt, we propose a stochastic linearized proximal method of multipliers (SLPMM) for solving the stochastic convex optimization problem (1), and analyze its expected convergence rate as well as probability guarantee for both objective reduction and constraint violation. In specific, at the $k$th iteration in SLPMM, we consider the augmented Lagrangian function $L^k(x, \lambda)$ of a linearized problem with respect to the stochastic subgradients $v_i(x^k, \xi^k), i = 0, 1, \ldots, p$. Then, we obtain the next iterate $x^{k+1}$ by solving the problem $\min_{x \in C} L^k(x, \lambda) + \alpha \|x - x^k\|^2$ and update the Lagrange multiplier. The subproblem is the minimization of a strongly convex (approximately) quadratic function
and hence is relatively easy to solve. Assuming that the set \( C \) is compact, the subgradients are bounded and the Slater’s condition holds, if the parameters in SLPMM are chosen as \( \alpha = \sqrt{K} \) and \( \sigma = 1/\sqrt{K} \), we show that SLPMM attains \( O(1/\sqrt{K}) \) expected convergence rate with respect to both objective reduction and constraint violation. Under certain light-tail assumptions, we also establish the large-deviation properties of SLPMM. The numerical results on some practical applications such as Neyman-Pearson classification demonstrate that SLPMM performs efficiently and has certain advantages over the existing stochastic first-order methods.

The remaining parts of this paper are organized as follows. In Section 2, we develop some important properties of SLPMM. In Section 3, in the expectation sense we establish the convergence rate of SLPMM for problem (1). The high probability guarantees for objective reduction and constraint violation of SLPMM are investigated in Section 4. In Section 5, we report our numerical results. Finally, we draw a conclusion in Section 6.

2. Algorithmic framework, assumptions and auxiliary lemmas

In this section, we propose a stochastic linearized proximal method of multipliers (SLPMM) for solving problem (1) and establish some important auxiliary lemmas.

Let us define \([t]_+ := \max\{t, 0\}\) for any \( t \in \mathbb{R} \) and let \([y]_+ = \Pi_{R^p_+}[y]\) denote the projection of \( y \) onto \( \mathbb{R}^p_+ \) for any \( y \in \mathbb{R}^p \). We also define \([t]^2_+ := (\max\{t, 0\})^2\).

The detail of SLPMM is described in Algorithm 1. In specific, at each iteration, we first generate an i.i.d. sample \( \xi^k \) and choose the stochastic subgradients \( v_i(x^k, \xi^k), i = 0, 1, \ldots, p \) of \( F \) and \( G_i \), respectively. Then, in (3) we obtain \( x^{k+1} \) by computing the proximal point of \( \mathcal{L}_\sigma^k(x, \lambda) \), which is the augmented Lagrangian function of the linearized problem

\[
\min_{x \in C} F(x^k, \xi^k) + \langle v_0(x^k, \xi^k), x - x^k \rangle \\
\text{s.t. } G_i(x^k, \xi^k) + \langle v_i(x^k, \xi^k), x - x^k \rangle \leq 0, \quad i = 1, \ldots, p.
\]

Finally, in (5) we update the Lagrange multipliers.

Denote

\[
G(x, \xi) := (G_1(x, \xi), \ldots, G_p(x, \xi))^T, \quad g(x) := (g_1(x), \ldots, g_p(x))^T.
\]

Let

\[
V(x^k, \xi^k) := (v_1(x^k, \xi^k), \ldots, v_p(x^k, \xi^k))^T,
\]
Algorithm 1: A stochastic linearized proximal method of multipliers

1. Initialization: Choose an initial point $x^0 \in \mathcal{C}$ and select parameters $\sigma > 0, \alpha > 0$. Set $\lambda^0 = 0 \in \mathbb{R}^p$ and $k = 0$.

2. for $k = 0, 1, 2, \ldots$ do

3. Generate i.i.d. sample $\xi^k$ of $\xi$ and compute

   $$x^{k+1} = \arg\min_{x \in \mathcal{C}} \left\{ \mathcal{L}_u^k(x, \lambda^k) + \frac{\alpha}{2} \|x - x^k\|^2 \right\},$$  

   where

   $$\mathcal{L}_u^k(x, \lambda) := F(x^k, \xi^k) + \langle v_0(x^k, \xi^k), x - x^k \rangle$$

   $$+ \frac{1}{2\sigma} \left[ \sum_{i=1}^p \left[ \lambda_i + \sigma \left( G_i(x^k, \xi^k) + \langle v_i(x^k, \xi^k), x - x^k \rangle \right) \right]^2 - \|\lambda\|^2 \right],$$  

   and $v_i(x^k, \xi^k), i = 0, 1, \ldots, p$ are the corresponding stochastic subgradients.

4. Update the Lagrange multipliers by

   $$\lambda_i^{k+1} = \left[ \lambda_i^k + \sigma \left( G_i(x^k, \xi^k) + \langle v_i(x^k, \xi^k), x^{k+1} - x^k \rangle \right) \right]_+, \quad i = 1, \ldots, p.$$  

5. Set $k \leftarrow k + 1$.

then (6) can be rewritten as

$$\lambda^{k+1} = \left[ \lambda^k + \sigma \left( G(x^k, \xi^k) + V(x^k, \xi^k)(x^{k+1} - x^k) \right) \right]_+.$$  

In the following, we shall study the convergence of the stochastic process $\{x^k, \lambda^k\}$ generated by SLPMM with respect to the filtration $\mathcal{F}_k$ (sigma-algebra) which is generated by the random information $\{(\xi^0, \ldots, \xi^{k-1})\}$. Before that, we introduce the following assumptions.

**Assumption 1.** Let $R > 0$ be a positive parameter such that

$$\|x' - x''\| \leq R, \ \forall x', x'' \in \mathcal{C}.$$  

**Assumption 2.** There exists a constant $\nu_g > 0$ such that for each $\xi^k$,

$$\|G(x, \xi^k)\| \leq \nu_g, \ \forall x \in \mathcal{C}.$$
ASSUMPTION 3. There exist constants $\kappa_f > 0$ and $\kappa_g > 0$ such that for each $\xi^k$,

$$\|v_0(x, \xi^k)\| \leq \kappa_f, \quad \|v_i(x, \xi^k)\| \leq \kappa_g, \quad i = 1, \ldots, p, \quad \forall x \in C.$$

ASSUMPTION 4. The Slater’s condition holds, i.e., there exist $\varepsilon_0 > 0$ and $\hat{x} \in C$ such that

$$g_i(\hat{x}) \leq -\varepsilon_0, \quad i = 1, \ldots, p.$$

Assumption 3 shows that $C$ is a compact convex set with diameter $R$. Assumption 2 indicates that the constraint functions $G_i(\cdot, \xi^k)$ are bounded over $C$. This assumption is a bit restrictive, but it is required in the analysis of almost all existing stochastic methods for solving problem (1) (Lan and Zhou 2020, Yu et al. 2017, Lin et al. 2020, Xiao 2019). Assumption 3 requires that the stochastic subgradients $v_i(\cdot, \xi^k)$ are bounded over $C$. Assumption 4 is a standard Slater’s condition for optimization problem with functional constraints.

The following auxiliary lemma will be used several times in the sequel.

**Lemma 1.** For any $z \in C$, we have

$$\langle v_0(x^k, \xi^k), x^{k+1} - x^k \rangle + \frac{1}{2\sigma} \|\lambda^{k+1}\|^2 + \frac{\alpha}{2} \|x^{k+1} - x^k\|^2 \leq \langle v_0(x^k, \xi^k), z - x^k \rangle + \frac{1}{2\sigma} \left[ \sum_{i=1}^{p} \left[ \lambda^k_i + \sigma(G_i(x^k, \xi^k) + \langle v_i(x^k, \xi^k), z - x^k \rangle) \right]_+^2 \right] \tag{7}$$

$$+ \frac{\alpha}{2} (\|z - x^k\|^2 - \|z - x^{k+1}\|^2).$$

In particular, if we take $z = x^k$, it yields

$$\langle v_0(x^k, \xi^k), x^{k+1} - x^k \rangle + \frac{1}{2\sigma} \|\lambda^{k+1}\|^2 + \alpha \|x^{k+1} - x^k\|^2 \leq \frac{1}{2\sigma} \left[ \sum_{i=1}^{p} [\lambda^k_i + \sigma G_i(x^k, \xi^k)]_+^2 \right]. \tag{8}$$

**Proof.** By using the optimality conditions, we have from (3) that $x^{k+1}$ satisfies

$$0 \in \nabla_x L(x^{k+1}, x^k) + \alpha(x^{k+1} - x^k) + N_C(x^{k+1}), \tag{9}$$

where $N_C(x^{k+1})$ denotes the normal cone of $C$ at $x^{k+1}$ and

$$\nabla_x L(x^{k+1}, x^k) = v_0(x^k, \xi^k) + \sum_{i=1}^{p} v_i(x^k, \xi^k) \cdot [\lambda^k_i + \sigma(G_i(x^k, \xi^k) + \langle v_i(x^k, \xi^k), x^{k+1} - x^k \rangle)]_+.$$
Let us now consider the following auxiliary problem

\[
\min_{x \in \mathbb{C}} \langle v_0(x^k, \xi^k), x - x^k \rangle + \frac{1}{2\sigma} \left[ \sum_{i=1}^{p} [\lambda_i^k + \sigma (G_i(x^k, \xi^k) + \langle v_i(x^k, \xi^k), x - x^k \rangle)]^2 \right] + \frac{\sigma}{2} (\|x - x^k\|^2 - \|x - x^{k+1}\|^2).
\]

We can easily check that (10) is a convex optimization problem. Therefore, \( \hat{x} \) is an optimal solution to (10) if and only if

\[
0 \in v_0(x^k, \xi^k) + \sum_{i=1}^{p} v_i(x^k, \xi^k) \cdot [\lambda_i^k + \sigma (G_i(x^k, \xi^k) + \langle v_i(x^k, \xi^k), \hat{x} - x^k \rangle)] + \alpha (x^{k+1} - x^k) + \mathcal{N}_C(\hat{x}).
\]

Hence, it follows from (9) that \( x^{k+1} \) is an optimal solution to (10), which gives (7) and (8) obviously. □

In what follows, we estimate an upper bound of \( \|x^{k+1} - x^k\| \).

**Lemma 2.** Let Assumptions 1-3 be satisfied. Then, if the parameters satisfy \( 2\alpha - p\kappa_g^2\sigma > 0 \), we have

\[
\|x^{k+1} - x^k\| \leq \frac{1}{\alpha} (\kappa_f + \sqrt{p}\kappa_g\|\lambda^k\| + \sqrt{p}\nu_g\kappa_g\sigma).
\]

**Proof.** From (8) and Assumption 3 we have

\[
\alpha \|x^{k+1} - x^k\|^2 \leq \kappa_f \|x^{k+1} - x^k\| + \frac{1}{2\sigma} \sum_{i=1}^{p} ([a_i]^2 - [b_i]^2),
\]

in which, for simplicity, we use

\[
a_i := \lambda_i^k + \sigma G_i(x^k, \xi^k), \quad b_i := \lambda_i^k + \sigma (G_i(x^k, \xi^k) + \langle v_i(x^k, \xi^k), x^{k+1} - x^k \rangle).
\]

Noticing that

\[
[a_i]^2 - [b_i]^2 = ([a_i]^2 + [b_i]^2)([a_i]^2 - [b_i]^2) \\
\leq ([a_i] + [b_i]) \cdot |a_i - b_i| \\
\leq (2|a_i| + |b_i - a_i|) \cdot |a_i - b_i| \\
= 2|a_i| \cdot |a_i - b_i| + (a_i - b_i)^2 \\
\leq 2|\lambda_i^k + \sigma G_i(x^k, \xi^k)| \cdot \sigma \kappa_g \|x^{k+1} - x^k\| + \sigma^2 \kappa_g^2 \|x^{k+1} - x^k\|^2,
\]

we obtain

$$2\alpha \|x^{k+1} - x^k\| \leq 2\kappa_f + \sum_{i=1}^{p} (2\kappa_g |\lambda_i^k + \sigma G_i(x^k, \xi^k)| + \sigma \kappa_g^2 \|x^{k+1} - x^k\|).$$

If $2\alpha - p\kappa_g^2 \sigma > 0$, it yields

$$\|x^{k+1} - x^k\| \leq \frac{2}{2\alpha - p\kappa_g^2 \sigma} \left( \kappa_f + \sum_{i=1}^{p} (\kappa_g |\lambda_i^k + \sigma G_i(x^k, \xi^k)| \right).$$

Therefore, from the facts that $\sum_{i=1}^{p} |\lambda_i^k| \leq \sqrt{p} \|\lambda^k\|$ and

$$\sum_{i=1}^{p} |G_i(x^k, \xi^k)| \leq \sqrt{p} \|G(x^k, \xi^k)\| \leq \sqrt{p} \nu_g,$$

the claim is obtained. $\square$

Under the Slater’s condition, we derive the following conditional expected estimate of the multipliers.

**Lemma 3.** Let Assumption 4 be satisfied. Then, for any $t_2 \leq t_1 - 1$ where $t_1$ and $t_2$ are positive integers,

$$\mathbb{E} \left[ \langle \lambda_t^{t_1}, G(\hat{x}, \xi^{t_1}) \rangle | \mathcal{F}_{t_2} \right] \leq -\varepsilon_0 \mathbb{E} \left[ \|\lambda_t^{t_1}\| | \mathcal{F}_{t_2} \right].$$

**Proof.** For any $i \in \{1, \ldots, p\}$, noticing that $\lambda_i^{t_1} \in \mathcal{F}_{t_1}$ and $\mathcal{F}_{t_2} \subseteq \mathcal{F}_{t_1}$ for $t_2 \leq t_1 - 1$, we have

$$\mathbb{E} \left[ \lambda_i^{t_1} G_i(\hat{x}, \xi^{t_1}) | \mathcal{F}_{t_2} \right] = \mathbb{E} \left[ \mathbb{E} \left[ \lambda_i^{t_1} G_i(\hat{x}, \xi^{t_1}) | \mathcal{F}_{t_1} \right] | \mathcal{F}_{t_2} \right] = \mathbb{E} \left[ \lambda_i^{t_1} G_i(\hat{x}) | \mathcal{F}_{t_2} \right] \leq -\varepsilon_0 \mathbb{E} \left[ \lambda_i^{t_1} | \mathcal{F}_{t_2} \right].$$

Summing the above inequality over $i \in \{1, \ldots, p\}$ yields

$$\mathbb{E} \left[ \langle \lambda_t^{t_1}, G(\hat{x}, \xi^{t_1}) \rangle | \mathcal{F}_{t_2} \right] \leq -\varepsilon_0 \mathbb{E} \left[ \sum_{i=1}^{p} \lambda_i^{t_1} | \mathcal{F}_{t_2} \right] \leq -\varepsilon_0 \mathbb{E} \left[ \|\lambda_t^{t_1}\| | \mathcal{F}_{t_2} \right],$$

by using $\sum_{i=1}^{p} \lambda_i^{t_1} \geq \|\lambda_t^{t_1}\|$. $\square$

We next present some important relations of $\|\lambda^k\|$. 

**Lemma 4.** Let Assumptions 1–4 be satisfied and $s > 0$ be an arbitrary integer. Define $\beta_0 := \nu_g + \sqrt{p} \kappa_g R$ and

$$\vartheta(\sigma, \alpha, s) := \frac{\varepsilon_0 \sigma s}{2} + \sigma \beta_0 (s - 1) + \frac{\alpha R^2}{\varepsilon_0 s} + \frac{2\kappa_f R}{\varepsilon_0} + \frac{\sigma \nu_g^2}{\varepsilon_0}. \quad (11)$$
Then, the following holds:

\[ ||\lambda^{k+1}|| - ||\lambda^k|| \leq \sigma \beta_0 \]  

(12)

and

\[ \mathbb{E} [||\lambda^{k+s}|| - ||\lambda^k|| | \mathcal{F}_k] \leq \begin{cases} 
  s \sigma \beta_0, & \text{if } ||\lambda^k|| < \vartheta(\sigma, \alpha, s), \\
  -s \frac{\sigma \varepsilon_0}{2}, & \text{if } ||\lambda^k|| \geq \vartheta(\sigma, \alpha, s).
\end{cases} \]  

(13)

**Proof.** It follows from Assumptions 1–3, (6) and the nonexpansion property of the projection \( \Pi_{\mathbb{R}^p}(\cdot) \) that

\[
||\lambda^{k+1}|| - ||\lambda^k|| \\
\leq ||\lambda^{k+1} - \lambda^k|| = ||[\lambda^k + \sigma (G(x^k, \xi^k) + V(x^k, \xi^k)(x^{k+1} - x^k))]_+ - [\lambda^k]_+|| \\
\leq \sigma ||G(x^k, \xi^k) + V(x^k, \xi^k)(x^{k+1} - x^k)|| \\
\leq \sigma [\nu_g + \sqrt{p} \kappa_0 R],
\]

which implies (12). This also gives that \( ||\lambda^{k+s}|| - ||\lambda^k|| \leq s \sigma \beta_0 \). Hence, we only need to establish the second part in (13) under the case \( ||\lambda^k|| \geq \vartheta(\sigma, \alpha, s) \).

For a given positive integer \( s \), suppose that \( ||\lambda^k|| \geq \vartheta(\sigma, \alpha, s) \). For any \( l \in \{k, k+1, \ldots, k+s-1\} \), from (7) and the convexity of \( G_i(\cdot, \xi^l) \) one has

\[
\langle v_0(x^l, \xi^l), x^{l+1} - x^l \rangle + \frac{1}{2\sigma} ||\lambda^{l+1}||^2 + \frac{\alpha}{2} ||x^{l+1} - x^l||^2 \\
\leq \langle v_0(x^l, \xi^l), \hat{x} - x^l \rangle + \frac{1}{2\sigma} \left[ \sum_{i=1}^p [\lambda_i^l + \sigma (G_i(x^l, \xi^l) + \langle v_i(x^l, \xi^l), \hat{x} - x^l \rangle)]_+^2 \right] \\
+ \frac{\alpha}{2} \left( ||\hat{x} - x^l||^2 - ||\hat{x} - x^{l+1}||^2 \right) \\
\leq \langle v_0(x^l, \xi^l), \hat{x} - x^l \rangle + \frac{1}{2\sigma} ||\lambda^l + \sigma G(\hat{x}, \xi^l)||^2 + \frac{\alpha}{2} \left( ||\hat{x} - x^l||^2 - ||\hat{x} - x^{l+1}||^2 \right) \\
\leq \langle v_0(x^l, \xi^l), \hat{x} - x^l \rangle + \frac{1}{2\sigma} ||\lambda^l + \sigma G(\hat{x}, \xi^l)||^2 + \frac{\alpha}{2} \left( ||\hat{x} - x^l||^2 - ||\hat{x} - x^{l+1}||^2 \right). 
\]

Rearranging terms and using Assumption 2 we obtain

\[
\frac{1}{2\sigma} \left[ ||\lambda^{l+1}||^2 - ||\lambda^l||^2 \right] \\
\leq \langle v_0(x^l, \xi^l), \hat{x} - x^{l+1} \rangle + \langle \lambda^l, G(\hat{x}, \xi^l) \rangle + \frac{\sigma}{2} ||G(\hat{x}, \xi^l)||^2 \\
+ \frac{\alpha}{2} \left( ||\hat{x} - x^l||^2 - ||\hat{x} - x^{l+1}||^2 \right) \\
\leq \kappa_f R + \langle \lambda^l, G(\hat{x}, \xi^l) \rangle + \frac{\sigma}{2} \nu_g^2 + \frac{\alpha}{2} \left( ||\hat{x} - x^l||^2 - ||\hat{x} - x^{l+1}||^2 \right). 
\]
Making a summation over \( \{k, k+1, \ldots, k+s-1\} \) and taking the conditional expectation, we obtain from Lemma 3 that

\[
\frac{1}{2\sigma} \mathbb{E} \left[ \|\lambda^{k+s}\|^2 - \|\lambda^k\|^2 \mid \mathcal{F}_k \right] 
\leq (\kappa_f R + \frac{\sigma}{2} \nu_g^2) s + \sum_{l=k}^{k+s-1} \mathbb{E} \left[ \langle \lambda^l, G(\hat{x}, \xi^l) \rangle \mid \mathcal{F}_k \right] + \frac{\sigma}{2} \|\hat{x} - x^k\|^2 
\leq (\kappa_f R + \frac{\sigma}{2} \nu_g^2) s - \varepsilon_0 \sum_{l=0}^{s-1} \mathbb{E} \left[ \|\lambda^{k+l}\| \mid \mathcal{F}_k \right] + \frac{\sigma}{2} R^2 
\leq (\kappa_f R + \frac{\sigma}{2} \nu_g^2) s - \varepsilon_0 \sum_{l=0}^{s-1} \mathbb{E} \left[ \|\lambda^k\| - \sigma \beta_0 l \mid \mathcal{F}_k \right] + \frac{\sigma}{2} R^2 
\leq (\kappa_f R + \frac{\sigma}{2} \nu_g^2) s - \varepsilon_0 \frac{\sigma \beta_0 (s-1)}{2} - \varepsilon_0 s \|\lambda^k\| + \frac{\sigma}{2} R^2.
\]

Further, we get from Assumption 2 and (11) that

\[
\mathbb{E} \left[ \|\lambda^{k+s}\|^2 \mid \mathcal{F}_k \right] 
\leq \|\lambda^k\|^2 + 2\sigma (\kappa_f R + \frac{\sigma}{2} \nu_g^2) s + \varepsilon_0 \sigma^2 \beta_0 s (s-1) - 2\varepsilon_0 \sigma s \|\lambda^k\| + \sigma \alpha R^2 
\leq (\|\lambda^k\| - \frac{\varepsilon_0 \sigma s}{2})^2 + \varepsilon_0 \sigma^2 \beta_0 s (s-1) + 2\sigma (\kappa_f R + \frac{\sigma}{2} \nu_g^2) s + \sigma \alpha R^2 - \varepsilon_0 \sigma s \|\lambda^k\| 
\leq (\|\lambda^k\| - \frac{\varepsilon_0 \sigma s}{2})^2 + \varepsilon_0 \sigma s [\sigma \beta_0 (s-1) + \frac{2(\kappa_f R + \frac{\sigma}{2} \nu_g^2)}{\varepsilon_0} + \frac{\alpha R^2}{\varepsilon_0 s} - \vartheta(\sigma, \alpha, s)] 
\leq (\|\lambda^k\| - \frac{\varepsilon_0 \sigma s}{2})^2.
\]

This, together with Jensen’s inequality and the fact that \( \|\lambda^k\| \geq \vartheta(\sigma, \alpha, s) \geq \frac{\varepsilon_0 \sigma}{2} s \), implies that

\[
\mathbb{E} \left[ \|\lambda^{k+s}\| \mid \mathcal{F}_k \right] \leq \|\lambda^k\| - \frac{\varepsilon_0 \sigma}{2} s.
\]

The proof is completed. \( \square \)

Let us make some comments on inequality (13). This result may seem a bit confusing. From the proof, we actually show that: the inequality \( \mathbb{E}[\|\lambda^{k+s} - \lambda^k\| \mid \mathcal{F}_k] \leq s \sigma \beta_0 \) holds true under the conditions of Lemma 4; in addition, if \( \|\lambda^k\| \geq \vartheta(\sigma, \alpha, s) \), the bound can be improved to \( \mathbb{E}[\|\lambda^{k+s} - \lambda^k\| \mid \mathcal{F}_k] \leq -s \frac{\vartheta \sigma}{2} \). However, we state it in the form of (13) intentionally. Since this is only a middle result, our true purpose is to show that the conditions of the following lemma (Yu et al., 2017, Lemma 5) are satisfied for \( \|\lambda^k\| \).
**Lemma 5.** Let \( \{Z_t, t \geq 0\} \) be a discrete time stochastic process adapted to a filtration \( \{\mathcal{F}_t, t \geq 0\} \) with \( Z_0 = 0 \) and \( \mathcal{F}_0 = \{\emptyset, \Omega\} \). Suppose there exist an integer \( t_0 > 0 \), real constants \( \theta > 0 \), \( \delta_{\text{max}} > 0 \) and \( 0 < \zeta \leq \delta_{\text{max}} \) such that

\[
|Z_{t+1} - Z_t| \leq \delta_{\text{max}},
\]

\[
\mathbb{E}[Z_{t+t_0} - Z_t | \mathcal{F}_t] \leq \begin{cases} 
 t_0 \delta_{\text{max}}, & \text{if } Z_t < \theta, \\
 -t_0 \zeta, & \text{if } Z_t \geq \theta,
\end{cases}
\]

hold for all \( t \in \{1, 2, \ldots\} \). Then the following properties are satisfied.

(i) The following inequality holds,

\[
\mathbb{E}[Z_t] \leq \theta + t_0 \delta_{\text{max}} + t_0 \frac{4 \delta_{\text{max}}^2}{\zeta} \log \left( \frac{8 \delta_{\text{max}}^2}{\zeta^2} \right), \quad \forall t \in \{1, 2, \ldots\}. \tag{14}
\]

(ii) For any constant \( 0 < \mu < 1 \), we have

\[
\Pr [Z_t \geq z] \leq \mu, \quad \forall t \in \{1, 2, \ldots\},
\]

where

\[
z = \theta + t_0 \delta_{\text{max}} + t_0 \frac{4 \delta_{\text{max}}^2}{\zeta} \log \left( \frac{8 \delta_{\text{max}}^2}{\zeta^2} \right) + t_0 \frac{4 \delta_{\text{max}}^2}{\zeta} \log \left( \frac{1}{\mu} \right). \tag{15}
\]

It is not difficult to verify that, Lemma 4 implies that the conditions of Lemma 5 are satisfied with respect to \( \|\lambda^k\| \) if we take

\[
\theta = \vartheta(\sigma, \alpha, s), \quad \delta_{\text{max}} = \sigma \beta_0, \quad \zeta = \frac{\sigma}{2} \varepsilon_0, \quad t_0 = s.
\]

For simplicity, we define

\[
\psi(\sigma, \alpha, s) := \kappa_0 + \kappa_1 \frac{\alpha}{s} + \kappa_2 \sigma + \kappa_3 \sigma s, \quad \phi(\sigma, \alpha, s, \mu) := \psi(\sigma, \alpha, s) + \frac{8 \beta_0^2}{\varepsilon_0} \log \left( \frac{1}{\mu} \right) \sigma s,
\]

where \( \kappa_0, \kappa_1, \kappa_2, \kappa_3 \) are constants given by

\[
\kappa_0 = \frac{2 \kappa_f R}{\varepsilon_0}, \quad \kappa_1 = \frac{R^2}{\varepsilon_0}, \quad \kappa_2 = \frac{\nu^2}{\varepsilon_0} - \beta_0, \quad \kappa_3 = \left[ 2 \beta_0 + \frac{\varepsilon_0}{2} + \frac{8 \beta_0^2}{\varepsilon_0} \log \frac{32 \beta_0^2}{\varepsilon_0^2} \right]. \tag{16}
\]

We can also observe that \( \psi(\sigma, \alpha, s) \) and \( \phi(\sigma, \alpha, s, \mu) \) are exactly the same as the right-hand sides of (14) and (15), respectively. Therefore, in view of Lemma 4, the following lemma is a direct consequence of Lemma 5.

**Lemma 6.** Let Assumptions 1–4 be satisfied and \( s > 0 \) be an arbitrary integer. Then, it holds that

\[
\mathbb{E}[\|\lambda^k\|] \leq \psi(\sigma, \alpha, s). \tag{17}
\]

Moreover, for any constant \( 0 < \mu < 1 \), we have

\[
\Pr [\|\lambda^k\| \geq \phi(\sigma, \alpha, s, \mu)] \leq \mu. \tag{18}
\]
3. Expected convergence rates

In this section, we shall establish the expected convergence rates of SLPMM with respect to constraint violation and objective reduction.

In the following lemma, we derive a bound of the constraints in terms of the averaged iterate

\[ \hat{x}^K = \frac{1}{K} \sum_{k=0}^{K-1} x^k, \]

where \( K \) is a fixed iteration number.

**Lemma 7.** Let Assumptions 1-3 be satisfied. Then, if the parameters satisfy \( 2 \alpha - p \kappa_2^2 \sigma > 0 \), for each \( i = 1, \ldots, p \) we have

\[ \mathbb{E}[g_i(\hat{x}^K)] \leq \frac{1}{\sigma K} \mathbb{E}[\lambda_i^K] + \frac{K_2}{\alpha}(\kappa_f + \sqrt{p \nu_g \kappa_g} \sigma) + \frac{\sqrt{p \kappa_g^2}}{\alpha K} \sum_{k=0}^{K-1} \mathbb{E}[\|\lambda^k\|]. \]

**Proof.** From the definition \( \lambda_i^{k+1} = \left[ \lambda_i^k + \sigma(G_i(x^k, \xi^k) + \langle v_i(x^k, \xi^k), x^{k+1} - x^k \rangle) \right]_+ \), it follows that

\[ \lambda_i^{k+1} \geq \lambda_i^k + \sigma(G_i(x^k, \xi^k) - \kappa_g \|x^{k+1} - x^k\|). \]

Using Lemma 2, we have

\[ G_i(x^k, \xi^k) \leq \frac{1}{\sigma}(\lambda_i^{k+1} - \lambda_i^k) + \frac{K_2}{\alpha}(\kappa_f + \sqrt{p \kappa_g} \|\lambda^k\| + \sqrt{p \nu_g \kappa_g} \sigma). \] (19)

Taking conditional expectation with respect to \( \mathcal{F}_k \), it yields that

\[ g_i(x^k) \leq \frac{1}{\sigma}(\mathbb{E}[\lambda_i^{k+1}]|\mathcal{F}_k] - \lambda_i^k) + \frac{K_2}{\alpha}(\kappa_f + \sqrt{p \kappa_g} \mathbb{E}[\|\lambda^k\|] + \sqrt{p \nu_g \kappa_g} \sigma), \]

which further gives that

\[ \mathbb{E}[g_i(x^k)] \leq \frac{1}{\sigma}(\mathbb{E}[\lambda_i^{k+1}] - \mathbb{E}[\lambda_i^k]) + \frac{K_2}{\alpha}(\kappa_f + \sqrt{p \kappa_g} \mathbb{E}[\|\lambda^k\|] + \sqrt{p \nu_g \kappa_g} \sigma). \]

Summing over \( \{0, \ldots, K-1\} \) and noticing that \( \lambda^0 = 0 \), we obtain

\[ \sum_{k=0}^{K-1} \mathbb{E}[g_i(x^k)] \leq \frac{1}{\sigma} \mathbb{E}[\lambda^K] + \frac{K_2 K}{\alpha}(\kappa_f + \sqrt{p \nu_g \kappa_g} \sigma) + \frac{\sqrt{p \kappa_g^2}}{\alpha K} \sum_{k=0}^{K-1} \mathbb{E}[\|\lambda^k\|]. \]

Therefore, from the convexity of \( g_i \) and the definition of \( \hat{x}^K \) it follows

\[ \mathbb{E}[g_i(\hat{x}^K)] \leq \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[g_i(x^k)] \]

\[ \leq \frac{\mathbb{E}[\lambda^K]}{\sigma K} + \frac{K_2(K_f + \sqrt{p \nu_g \kappa_g} \sigma)}{\alpha} + \frac{\sqrt{p \kappa_g^2}}{\alpha K} \sum_{k=0}^{K-1} \mathbb{E}[\|\lambda^k\|]. \]

The proof is completed. □
In what follows, we present the bound of the objective reduction in terms of the averaged iterate.

**Lemma 8.** Let Assumptions 1-3 be satisfied. Then, for any \( z \in \Phi \),
\[
\mathbb{E}[f(\hat{x}^K)] - f(z) \leq \frac{\kappa_i^2}{2\alpha} + \frac{\sigma}{2} \nu^2 + \frac{\alpha}{2} R^2.
\]

**Proof.** For any \( z \in \Phi \), since \( v_0(x^k, \xi^k) \in \partial_x F(x^k, \xi^k) \), we have
\[
\langle v_0(x^k, \xi^k), z - x^k \rangle \leq F(z, \xi^k) - F(x^k, \xi^k).
\]
Then, in view of (7), one has
\[
F(x^k, \xi^k)
\leq F(z, \xi^k) + \left[ \langle v_0(x^k, \xi^k), x^k - x^{k+1} \rangle - \frac{\alpha}{2} \|x^{k+1} - x^k\|^2 \right]
+ \frac{1}{2\nu} \left[ \|\lambda^k + \sigma(G(x^k, \xi^k) + V(x^k, \xi^k)(z - x^k))\|_+^2 - \|\lambda^k\|^2 \right]
- \frac{1}{2\nu} \left[ \|\lambda^{k+1}\|^2 - \|\lambda^k\|^2 \right] + \frac{\alpha}{2} \|z - x^k\|^2 - \|z - x^{k+1}\|^2.
\]
From Assumption 3 and the fact that \( \langle x, y \rangle \leq \frac{\sigma}{2} \|x\|^2 + \frac{1}{2\alpha} \|y\|^2 \), we obtain that
\[
\langle v_0(x^k, \xi^k), x^k - x^{k+1} \rangle - \frac{\alpha}{2} \|x^{k+1} - x^k\|^2 \leq \frac{1}{2\alpha} \|v_0(x^k, \xi^k)\|^2 \leq \frac{\kappa_i^2}{2\alpha}.
\] (21)

For every \( i = 1, \ldots, p \), we have from \( v_i(x^k, \xi^k) \in \partial_x G_i(x^k, \xi^k) \) and \( [a]^2_+ \leq a^2 \) that
\[
[\lambda^k_i + \sigma(G_i(x^k, \xi^k) + \langle v_i(x^k, \xi^k), z - x^k \rangle)]^2_+ \leq [\lambda^k_i + \sigma G_i(z, \xi^k)]^2
\]
and hence
\[
\|\lambda^k + \sigma(G(x^k, \xi^k) + V(x^k, \xi^k)(z - x^k))\|_+^2 \leq \|\lambda^k + \sigma G(z, \xi^k)\|^2.
\]
Then, we obtain
\[
\|\lambda^k + \sigma(G(x^k, \xi^k) + V(x^k, \xi^k)(z - x^k))\|_+^2 - \|\lambda^k\|^2
\leq 2\sigma \langle \lambda^k, G(z, \xi^k) \rangle + \sigma^2 \|G(z, \xi^k)\|^2.
\] (22)
Substituting (21) and (22) into (20), we get
\[
F(x^k, \xi^k) \leq F(z, \xi^k) + \frac{\kappa_i^2}{2\alpha} - \frac{1}{2\nu} \left[ \|\lambda^{k+1}\|^2 - \|\lambda^k\|^2 \right] + \langle \lambda^k, G(z, \xi^k) \rangle
+ \frac{\alpha}{2} \|z - x^k\|^2 - \|z - x^{k+1}\|^2.
\] (23)
Taking conditional expectation with respect to $F_k$ and noticing that
\[ \mathbb{E}[\langle \lambda^k, G(z, \xi^k) \rangle | F_k] = \langle \lambda^k, g(z) \rangle \leq 0, \]
we have
\[ f(x^k) - f(z) \leq \frac{\kappa_j^2}{2\sigma} - \frac{1}{2\sigma} \left[ \mathbb{E}[\|\lambda^{k+1}\|^2 | F_k] - \|\lambda^k\|^2 \right] + \frac{\alpha \nu^2}{2} + \frac{\alpha}{2} \left[ \|z - x^k\|^2 - \mathbb{E}[\|z - x^{k+1}\|^2 | F_k] \right], \]
which further gives
\[ \mathbb{E}[f(x^k)] - f(z) \leq \frac{\kappa_j^2}{2\alpha} - \frac{1}{2\alpha} \left[ \mathbb{E}[\|\lambda^{k+1}\|^2] - \mathbb{E}[\|\lambda^k\|^2] \right] + \frac{\alpha \nu^2}{2} + \frac{\alpha}{2} \left[ \mathbb{E}[\|z - x^k\|^2] - \mathbb{E}[\|z - x^{k+1}\|^2] \right]. \]
Making a summation and noticing that $\lambda^0 = 0$, one has
\[ \sum_{k=0}^{K-1} \mathbb{E}[f(x^k)] \leq K \left[ f(z) + \frac{\kappa_j^2}{2\alpha} + \frac{\alpha \nu^2}{2} + \frac{\alpha}{2} \|z - x^0\|^2 \right]. \]
Therefore, from the convexity of $f$ and the definition of $\hat{x}^K$, it follows
\[ \mathbb{E}[f(\hat{x}^K)] \leq \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[f(x^k)] \leq f(z) + \frac{\kappa_j^2}{2\alpha} + \frac{\alpha \nu^2}{2} + \frac{\alpha}{2} R^2. \]
The proof is completed. \(\square\)

Based on Lemma 7 and Lemma 8, if we take $\alpha = \sqrt{K}$, $\sigma = 1/\sqrt{K}$ and $s = \lceil \sqrt{K} \rceil$, where $\lceil a \rceil$ denotes the ceiling function that returns the least integer greater than or equal to $a$, the expected convergence rates of SLPMM with respect to constraint violation and objective reduction are shown to be $O(1/\sqrt{K})$ in the following theorem.

**Theorem 1.** Let Assumptions 1-4 be satisfied. If we take $\alpha = \sqrt{K}$ and $\sigma = 1/\sqrt{K}$ in Algorithm 1, where $K$ is a fixed iteration number. Then, the following statements hold.

(i) If $K > \max\{1, p\kappa_j^2/2\}$, then we have
\[ \mathbb{E}[g_i(\hat{x}^K)] \leq \frac{1 + \sqrt{p\nu^2}}{\sqrt{K}} \kappa_i + \kappa_g \kappa_f + \frac{(1 + \sqrt{p\nu^2}) \kappa_2 + \sqrt{p\nu^2} \kappa_2^2}{K}, \quad i = 1, \ldots, p, \]
where $\tilde{\kappa} := \kappa_0 + \kappa_1 + 2\kappa_3$ and $\kappa_0, \kappa_1, \kappa_2, \kappa_3$ are defined in (16).

(ii) For all $K \geq 1$,
\[ \mathbb{E}[f(\hat{x}^K)] - f(x^*) \leq \frac{\kappa_f^2 + \nu^2 + R^2}{2\sqrt{K}}, \]
where $x^*$ is any optimal solution to (1).
Proof. Consider item (i). If \( K > p\kappa^2/2 \), we have \( \frac{2\alpha}{p\kappa^2} > 0 \), then it follows from Lemma 7 that

\[
\mathbb{E}[g_i(\hat{x}^K)] \leq \frac{1}{\sigma K} \mathbb{E}[\lambda_i^K] + \frac{\kappa_g}{\alpha} (\kappa_f + \sqrt{p\nu_g\kappa_g} \sigma) + \frac{\sqrt{p\kappa^2 \nu_g}}{\alpha K} \sum_{k=0}^{K-1} \mathbb{E}[\|\lambda^k\|]. \tag{24}
\]

If we take \( s = \lceil \sqrt{K} \rceil \), then from Lemma 6 one has

\[
\mathbb{E}[\|\lambda^k\|] \leq \psi(\sigma, \alpha, s) = \kappa_0 + \kappa_1 + \kappa_2 s + \kappa_3 \sigma s \leq \kappa_0 + \frac{\kappa_2}{\sqrt{K}} + 2\kappa_3 = \bar{\kappa} + \frac{\kappa_2}{\sqrt{K}}.
\]

Therefore, from \( \alpha = \sqrt{K}, \sigma = 1/\sqrt{K} \) and (24) we have

\[
\mathbb{E}[g_i(\hat{x}^K)] \leq \frac{1}{\sqrt{K}} \left( \bar{\kappa} + \frac{\kappa_2}{\sqrt{K}} \right) + \frac{\kappa_g \kappa_f}{\sqrt{K}} + \frac{\sqrt{p\nu_g\kappa_g^2}}{\sqrt{K}} \left( \bar{\kappa} + \frac{\kappa_2}{\sqrt{K}} \right),
\]

which verifies item (i).

By taking \( z = x^* \) in Lemma 8 we derive item (ii) since

\[
\mathbb{E}[f(\hat{x}^K)] - f(x^*) \leq \frac{\kappa_f^2}{2\alpha} + \frac{\sigma^2}{2} \nu^2 + \frac{R^2}{2K} = \frac{\kappa_f^2 + \nu^2 + R^2}{2\sqrt{K}}.
\]

The proof is completed. \(\square\)

Let us point out that all of the algorithms (Yu et al. 2017, Lan and Zhou 2020, Akhtar et al. 2021) have \( O(1/\sqrt{K}) \) expected convergence. However, the algorithm (Yu et al. 2017) is an extension of Zinkevich’s online algorithm (Zinkevich 2003), which is a variant of the projection gradient method, and the CSA method (Lan and Zhou 2020) is a stochastic counterpart of Polyak’s subgradient method (Polyak 1967). When problem (1) reduces to a deterministic problem, these algorithms have at most linear rate of convergence. In contrast, SLPMM becomes the (linearized) proximal method of multipliers, which has an asymptotic superlinear rate of convergence. Moreover, the iteration complexity analysis (Lan and Zhou 2020) is based on the selection of stepsizes, which are dependent on the parameters \( R, \kappa_f \) and \( \kappa_g \). However, these data are not known beforehand when problem (1) is put forward to solve. Note that, in SLPMM the stepsizes \( \sigma \) and \( \alpha \) are problem-independent.

4. High probability performance analysis

In this section, we shall establish the large-deviation properties of SLPMM. By Theorem 1 and Markov’s inequality, we have for all \( \rho_c > 0 \) and \( \rho_0 > 0 \) that

\[
\Pr \left[ g_i(\hat{x}^K) \leq \rho_c \left( \frac{1 + \sqrt{p\kappa^2 \nu}}{\sqrt{K}} \right) \right] \geq 1 - \frac{1}{\rho_c} \tag{25}
\]
However, these results are very weak. In the following, we will show that these high probability bounds can be significantly improved.

We introduce the following standard “light-tail” assumption, see (Lan 2016, Lan and Zhou 2020, Lin et al. 2020) for instance.

**Assumption 5.** There exists a constant \( \sigma_c > 0 \) such that, for any \( x \in C \),

\[
\mathbb{E}[\exp(\|G_i(x, \xi) - g_i(x)\|^2/\sigma_c^2)] \leq \exp(1), \quad i = 1, \ldots, p.
\]

From a well-known result (Lan 2020, Lemma 4.1), under Assumption 5, one has for any \( \rho \geq 0 \) and \( i = 1, \ldots, p \) that

\[
\Pr \left[ \frac{1}{K} \sum_{k=0}^{K-1} g_i(x^k) - \frac{1}{K} \sum_{k=0}^{K-1} G_i(x^k, \xi^k) \geq \rho \sigma_c \sqrt{\frac{K}{\log(K+1)}} \right] \leq \exp(-\rho^2/3). \quad (27)
\]

For the sake of readability, we define the following notations,

\[
\theta_1 := \sigma_c + (1 + \sqrt{p} \kappa_g^2) \frac{16 \beta_0}{\varepsilon_0}, \quad \theta_2 := \kappa_g \kappa_f + (1 + \sqrt{p} \kappa_g^2)(\kappa_0 + \kappa_1 + 2 \kappa_3)
\]

and

\[
\theta_3 := (1 + \sqrt{p} \kappa_g^2) \frac{16 \beta_0}{\varepsilon_0}, \quad \theta_4 := \sqrt{\nu_g \kappa_g^2} + (1 + \sqrt{p} \kappa_g^2) \kappa_2,
\]

in which \( \beta_0 \) is defined in Lemma 4, \( \kappa_0, \kappa_1, \kappa_2, \kappa_3 \) are defined in (16) and other parameters are defined in Assumptions 1-5.

We are now ready to state the main result on constraint violation.

**Theorem 2.** Let Assumptions 1-5 be satisfied. We take \( \alpha = \sqrt{K} \) and \( \sigma = 1/\sqrt{K} \) in Algorithm 1, where \( K \) is a fixed iteration number satisfying \( K > \max\{1, p \kappa_g^2/2\} \). Then, for any \( \rho \geq 0 \) and \( i = 1, \ldots, p \),

\[
\Pr \left[ g_i(\hat{x}^K) \leq \frac{\theta_1 \rho + \theta_2 + \theta_3 \log(K+1)}{\sqrt{K}} + \frac{\theta_4}{K} \right] \geq 1 - \exp(-\rho^2/3) - \exp(-\rho).
\]

**Proof.** Summing (19) over \( \{0, \ldots, K-1\} \), we have

\[
\frac{1}{K} \sum_{k=0}^{K-1} G_i(x^k, \xi^k) \leq \frac{\chi_i^K}{\sigma K} + \frac{\kappa_g (\kappa_f + \sqrt{\nu_g \kappa_g} \sigma)}{\alpha} + \frac{\sqrt{p} \kappa_g^2}{\alpha K} \sum_{k=0}^{K-1} \|\chi^k\|.
\]
Noticing that $\alpha = \sqrt{K}$, $\sigma = 1/\sqrt{K}$ and $g_i(\hat{x}_K) \leq \frac{1}{K} \sum_{k=0}^{K-1} g_i(x^k)$, one has

$$g_i(\hat{x}_K) \leq \frac{1}{K} \sum_{k=0}^{K-1} [g_i(x^k) - G_i(x^k, \xi^k)] + \frac{\lambda_i K}{\sqrt{K}} + \frac{\kappa g K_0^2}{K} + \frac{\sqrt{p_0^2 K_0^2}}{K^{3/2}} \sum_{k=0}^{K-1} \|\lambda^k\|. \quad (28)$$

We next consider the probability bound of $\lambda^k$. From (18), it follows that

$$\Pr[\|\lambda^k\| \geq \phi(\sigma, \alpha, s, \mu)] \leq \mu, \quad k = 0, 1, \ldots, K.$$  

If we take $s = \lceil \sqrt{K} \rceil$ and $\mu = \exp(-\rho)/(K+1)$, then

$$\phi(\sigma, \alpha, s, \mu) = \kappa_0 + \kappa_1 + \kappa_2 + \kappa_3 + \frac{8\beta_0^2}{\varepsilon_0} \log \left( \frac{1}{\mu} \right) \sigma s$$

$$\leq \kappa_0 + \frac{\kappa_2}{\sqrt{K}} + 2\kappa_3 + \frac{16\beta_0^2}{\varepsilon_0} (\rho + \log(K+1))$$

and hence for all $k = 0, 1, \ldots, K$,

$$\Pr[\|\lambda^k\| \geq \kappa_0 + \frac{\kappa_2}{\sqrt{K}} + 2\kappa_3 + \frac{16\beta_0^2}{\varepsilon_0} (\rho + \log(K+1))] \leq \frac{\exp(-\rho)}{K+1}. \quad (29)$$

Using (27) and (29) in (28), we conclude that

$$\Pr \left[ g_i(\hat{x}_K) \geq \frac{\rho(\sigma_c + (1+\sqrt{p_0^2}) (16\beta_0^2) \log(K+1))}{\sqrt{K}} + \frac{\kappa g K_0^2 + (1+\sqrt{p_0^2}) \kappa_0 + \kappa_2 \kappa_3}{\sqrt{K}} \right] \leq \exp(-\rho^2/3) + \exp(-\rho).$$

The proof is completed. \hfill \Box

In view of Theorem 2 if we take $\rho = \log(K)$, then we have

$$\Pr \left[ g_i(\hat{x}_K) \leq O \left( \frac{\log(K)}{\sqrt{K}} \right) \right] \geq 1 - \frac{1}{K^{2/3}} - \frac{1}{K}. $$

We next make the following “light-tail” assumption with respect to the objective function.

**Assumption 6.** There exists a constant $\sigma_0 > 0$ such that, for any $x \in \mathcal{C}$,

$$\mathbb{E} [\exp(\|F(x, \xi) - f(x)\|^2/\sigma_0^2)] \leq \exp(1).$$

Similar to (27), under Assumption 6 one has for any $\rho \geq 0$ that

$$\Pr \left[ \frac{1}{K} \sum_{k=0}^{K-1} f(x^k) - \frac{1}{K} \sum_{k=0}^{K-1} F(x^k, \xi^k) \geq \frac{\rho \sigma_0}{\sqrt{K}} \right] \leq \exp(-\rho^2/3) \quad (30)$$
and
\[
\Pr\left[ \frac{1}{K} \sum_{k=0}^{K-1} F(z^k, \xi^k) - \frac{1}{K} \sum_{k=0}^{K-1} f(z^k) \geq \frac{\rho \sigma_o}{\sqrt{K}} \right] \leq \exp \left( -\frac{\rho^2}{3} \right)
\]
for all \( z \in C \).

The following lemma is from (Yu et al. 2017, Lemma 9).

\textbf{Lemma 9.} Let \( \{ Z_t, t \geq 0 \} \) be a supermartingale adapted to a filtration \( \{ \mathcal{F}_t, t \geq 0 \} \) with \( Z_0 = 0 \) and \( \mathcal{F}_0 = \{ \emptyset, \Omega \} \), i.e. \( \mathbb{E}[Z_{t+1} | \mathcal{F}_t] \leq Z_t, \forall t \geq 0 \). Suppose there exists a constant \( c > 0 \) such that \( \{|Z_{t+1} - Z_t| > c\} \subseteq \{ Y_t > 0 \}, \forall t \geq 0 \), where each \( Y_t \) is adapted to \( \mathcal{F}_t \). Then, for all \( z > 0 \), we have
\[
\Pr[Z_t \geq z] \leq e^{-z^2/(2tc^2)} \sum_{j=0}^{t-1} \Pr[Y_j > 0], \forall t \geq 1.
\]

For any fixed \( z \in \Phi \), by taking \( Z_t := \sum_{k=0}^{t-1} \langle \lambda^k, G(z, \xi^k) \rangle \) in Lemma 9 we obtain the following lemma.

\textbf{Lemma 10.} For any fixed \( z \in \Phi \) and an arbitrary constant \( c > 0 \), let \( Z_0 := 0 \) and \( Z_t := \sum_{k=0}^{t-1} \langle \lambda^k, G(z, \xi^k) \rangle \) for \( t \geq 1 \). Let \( \mathcal{F}_0 = \{ \emptyset, \Omega \} \) and \( Y_t := \| \lambda^t \| - c/\nu_g \) for all \( t \geq 0 \). Then, for all \( \gamma > 0 \), we have
\[
\Pr[Z_t \geq \gamma] \leq e^{-\gamma^2/(2tc^2)} \sum_{j=0}^{t-1} \Pr[Y_j > 0], \forall t \geq 1.
\]

\textbf{Proof.} It is simple to check that \( \{ Z_t \} \) and \( \{ Y_t \} \) are both adapted to \( \{ \mathcal{F}_t, t \geq 0 \} \). Now we prove that \( \{ Z_t \} \) is a supermartingale. Since \( Z_{t+1} = Z_t + \langle \lambda^t, G(z, \xi^t) \rangle \), we have
\[
\mathbb{E}[Z_{t+1} | \mathcal{F}_t] = \mathbb{E}[Z_t + \langle \lambda^t, G(z, \xi^t) \rangle | \mathcal{F}_t]
\]
\[
= Z_t + \langle \lambda^t, \mathbb{E}[G(z, \xi^t) | \mathcal{F}_t] \rangle
\]
\[
= Z_t + \langle \lambda^t, g(z) \rangle
\]
\[
\leq Z_t,
\]
which follows from \( \lambda^t \in \mathcal{F}_t, \lambda^t \geq 0 \) and \( g(z) \leq 0 \). Thus, we obtain that \( \{ Z_t \} \) is a supermartingale.

From Assumption 2 we get
\[
\| Z_{t+1} - Z_t \| = \| \langle \lambda^t, G(z, \xi^t) \rangle \| \leq \nu_g \| \lambda^t \|.
\]
This implies that \( \|\lambda^t\| > c/\nu_g \) if \( |Z_{t+1} - Z_t| > c \) and hence

\[
\{ |Z_{t+1} - Z_t| > c \} \subseteq \{ Y_t > 0 \}.
\]

Therefore, we can observe that the conditions of Lemma \ref{Lemma9} are satisfied, and hence the claim is obtained. \( \square \)

Finally, we establish a high probability objective reduction bound in the following theorem.

**Theorem 3.** Let Assumptions \ref{Assump1-4} and \ref{Assump6} be satisfied. We take \( \alpha = \sqrt{K} \) and \( \sigma = 1/\sqrt{K} \) in Algorithm \ref{Algorithm1}, where \( K \geq 1 \) is a fixed iteration number. Then, for any \( \rho \geq 0 \),

\[
\Pr \left[ f(\hat{x}^K) - f(x^*) \leq \sqrt{2\rho \nu_g} \left( \frac{\kappa_0 + \kappa_1 + 2\kappa_3}{\sqrt{K}} + \frac{16\beta_0^2}{\kappa_0} (\rho + \log(K)) + \frac{\kappa_2}{K} \right) + \frac{2\sigma_0 \rho}{\sqrt{K}} + \frac{\theta_5}{\sqrt{K}} \right] \geq 1 - 2 \exp(-\rho^2/3) - 2 \exp(-\rho),
\]

where \( x^* \) is any fixed optimal solution to (1), \( \theta_5 := (\kappa_2^2 + \nu_g^2 + R^2)/2 \), \( \beta_0 \) is defined in Lemma \ref{Lemma4} and \( \kappa_0, \kappa_1, \kappa_2, \kappa_3 \) are defined in (16).

**Proof.** For any \( z \in \Phi \), summing (23) over \( \{0, \ldots, K-1\} \) and using the facts that \( \lambda^0 = 0 \), \( \|G(z, \xi^k)\| \leq \nu_g^2 \) and \( \|z - x^0\| \leq R \), we have

\[
\frac{1}{K} \sum_{k=0}^{K-1} F(x^k, \xi^k) \leq \frac{1}{K} \sum_{k=0}^{K-1} F(z, \xi^k) + \frac{1}{K} \sum_{k=0}^{K-1} \langle \lambda^k, G(z, \xi^k) \rangle + \frac{\kappa_2^2 + \nu_g^2 + R^2}{2\sqrt{K}}.
\]

Then, it follows from \( f(\hat{x}^K) \leq \frac{1}{K} \sum_{k=0}^{K-1} f(x^k) \) that

\[
f(\hat{x}^K) - f(z)
\leq \frac{1}{K} \sum_{k=0}^{K-1} [f(x^k) - F(x^k, \xi^k)] + \frac{1}{K} \sum_{k=0}^{K-1} [F(z, \xi^k) - f(z)]
\]

\[
+ \frac{1}{K} \sum_{k=0}^{K-1} \langle \lambda^k, G(z, \xi^k) \rangle + \frac{\kappa_2^2 + \nu_g^2 + R^2}{2\sqrt{K}}.
\]

(32)

By Lemma \ref{Lemma10} for any \( c > 0 \) and \( \gamma > 0 \) we have

\[
\Pr \left[ \frac{1}{K} \sum_{k=0}^{K-1} \langle \lambda^k, G(z, \xi^k) \rangle \geq \frac{\gamma}{K} \right] \leq \exp(-\gamma^2/(2Kc^2)) + \sum_{k=0}^{K-1} \Pr[\|\lambda^k\| \geq c/\nu_g].
\]
Let us take \( s = \lceil \sqrt{K} \rceil \) and \( \mu = \exp(-\rho)/K \), then
\[
\phi(\sigma, \alpha, s, \mu) \leq \kappa_0 + \kappa_1 + \frac{\kappa_2}{\sqrt{K}} + 2\kappa_3 + \frac{16\beta_0^2}{\varepsilon_0} (\rho + \log(K)).
\]
If we take \( c = \nu_g \phi(\sigma, \alpha, s, \mu) \), then from (18) we obtain
\[
\sum_{k=0}^{K-1} \Pr[\|\lambda^k\| \geq c/\nu_g] \leq K \mu = \exp(-\rho).
\]
Moreover, let us take \( \gamma = \sqrt{\frac{2\rho K}{c}} \), then
\[
\frac{\gamma}{K} = \sqrt{\frac{2\rho}{K}} \left( \frac{\kappa_0 + \kappa_1 + 2\kappa_3}{\sqrt{K}} + \frac{16\beta_0^2}{\varepsilon_0} (\rho + \log(K)) + \frac{\kappa_2}{K} \right)
\]
and hence
\[
\Pr \left[ \frac{1}{K} \sum_{k=0}^{K-1} \langle \lambda^k, G(z, \xi^k) \rangle \geq \sqrt{\frac{2\rho}{K}} \left( \frac{\kappa_0 + \kappa_1 + 2\kappa_3}{\sqrt{K}} + \frac{16\beta_0^2}{\varepsilon_0} (\rho + \log(K)) + \frac{\kappa_2}{K} \right) \right] \leq 2 \exp(-\rho).
\]
Using (30), (31) and (33) in (32), one has
\[
\Pr \left[ f(\hat{x}^K) - f(z) \geq \frac{2\sigma_0 \rho}{\sqrt{K}} + \sqrt{\frac{2\rho}{K}} \left( \frac{\kappa_0 + \kappa_1 + 2\kappa_3}{\sqrt{K}} + \frac{16\beta_0^2}{\varepsilon_0} (\rho + \log(K)) + \frac{\kappa_2}{K} \right) \right] \leq 2 \exp(-\rho^2/3) + 2 \exp(-\rho).
\]
The claim is derived by taking \( z = x^* \) in the above inequality. \( \square \)

In view of Theorem 3, if we take \( \rho = \log(K) \), then we have
\[
\Pr \left[ f(\hat{x}^K) - f(x^*) \leq O \left( \frac{\log^{3/2}(K)}{\sqrt{K}} \right) \right] \geq \frac{1 - 2}{K^{2/3}} - \frac{2}{K}.
\]
In contrast to (25) and (26), we can observe that the results in Theorem 2 and 3 are much finer.

5. Preliminary numerical experiments

In this section, we demonstrate the efficiency of the proposed stochastic linearized proximal method of multipliers on two preliminary numerical problems. All numerical experiments are carried out using MATLAB R2020a on a desktop computer with Intel(R) Xeon(R) E-2124G 3.40GHz and 32GB memory. The MATLAB code and test problems can be found on [https://bitbucket.org/Xiantao_Xiao/SLPMM](https://bitbucket.org/Xiantao_Xiao/SLPMM). All reported time is wall-clock time in seconds.
5.1. Solving subproblems

This subsection focuses on solving the subproblem (3) in SLPMM, that is

\[ x^{k+1} = \arg \min_{x \in \mathcal{C}} \left\{ \mathcal{L}_k^0(x, \lambda^k) + \frac{\sigma}{2} \| x - x^k \|^2 \right\}. \]

This problem is equivalent to

\[ \min_{x \in \mathcal{C}} \phi(x) := \frac{1}{2} \sum_{i=1}^{p} [a_i^T x + b_i]^2 + \frac{1}{2} \| x \|^2 + c^T x, \quad (34) \]

where

\[ a_i := \sqrt{\frac{\sigma}{\alpha}} v_i(x^k, \xi), \quad b_i := \frac{\lambda_i}{\sqrt{\sigma \alpha}} + \sqrt{\frac{\sigma}{\alpha}} G_i(x^k, \xi) - \left\langle \sqrt{\frac{\sigma}{\alpha}} v_i(x^k, \xi), x^k \right\rangle \]

and \( c := v_0(x^k, \xi)/\alpha - x^k \). Since \( \phi \) is obviously strongly convex, we could apply the following popular Nesterov’s accelerated gradient method to solve (34).

**APG**: Nesterov’s accelerated projected gradient method for (34).

**Step 0** Input \( x^0 \in \mathcal{C} \) and \( \eta > 1 \). Set \( y^0 = x^0, L_{-1} = 1 \) and \( t := 0 \).

**Step 1** Set

\[ x^{t+1} = T_{L_t}(y^t), \]

where \( T_L(y) := \Pi_{\mathcal{C}}[y - \frac{1}{L} \nabla \phi(y)] \), the stepsize \( L_t = L_{t-1} \eta^t \) and \( i_t \) is the smallest non-negative integer satisfies the following condition

\[ \phi(T_{L_{t-1} \eta^t}(y^t)) \leq \phi(y^t) + \langle \nabla \phi(y^t), T_{L_{t-1} \eta^t}(y^t) - y^t \rangle \]

\[ + \frac{L_{t-1} \eta^t}{2} \| T_{L_{t-1} \eta^t}(y^t) - y^t \|^2. \]

**Step 2** Compute

\[ y^{t+1} = x^{t+1} + \frac{t}{t+3} (x^{t+1} - x^t). \]

**Step 3** Set \( t := t + 1 \) and go to Step 1.

A well-known convergence result of the above method is that, if \( \phi \) is \( \mu \)-strongly convex and \( \nabla \phi \) is \( L \)-Lipschitz continuous, then \( \phi(x^t) - \phi(x^*) \leq O \left( (1 - \sqrt{\mu/L})^t \right) \). See (Beck 2017) for a detailed discussion on this topic. Here, we assume that the set \( \mathcal{C} \) is simple such that the projection \( \Pi_{\mathcal{C}} \) can be efficiently computed. For example, if

\[ \mathcal{C} := \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1, \ x \geq 0 \right\}, \]
the projection $\Pi_C$ can be computed by the method proposed in (Wang and Lu 2015).

When $\mathcal{C}$ is $\mathbb{R}^n$ or a polyhedron, the subproblem is equivalent to a convex quadratic programming (QP) problem as
\[
\min_{x,y} \frac{1}{2} \sum_{i=1}^{p} y_i^2 + \frac{1}{2} \|x\|^2 + c^T x
\]
\[\text{s.t. } a_i^T x + b_i - y_i \leq 0, \quad i = 1, 2, \ldots, p,
\]
\[x \in \mathcal{C}, \quad y \geq 0.
\]
In this case, the subproblem can also be solved by a QP solver. Let us also mention that, if $p = 1$, the closed form of the stationary point to the objective function in Problem (34) is given by
\[
\tilde{x} = \begin{cases} 
-c, & \text{if } -a_1^T c + b_1 \leq 0, \\
-(b_1 a_1 + c) + \frac{a_1^T (b_1 a_1 + c) a_1}{1 + a_1^T a_1}, & \text{otherwise.}
\end{cases}
\]
Then, $\tilde{x}$ is the unique optimal solution if it lies in the interior of $\mathcal{C}$.

5.2. Neyman-Pearson classification

For a classifier $h$ to predict 1 and $-1$, let us define the type I error (misclassifying class $-1$ as 1) and type II error (misclassifying class 1 as $-1$) respectively by

\[
\text{type I error} := \mathbb{E}[\varphi(-bh(a))|b = -1], \quad \text{type II error} := \mathbb{E}[\varphi(-bh(a))|b = 1],
\]

where $\varphi$ is some merit function. Unlike the conventional binary classification in machine learning, the Neyman-Pearson (NP) classification paradigm is developed to learn a classifier by minimizing type II error with type I error being below a user-specified level $\tau > 0$, see (Tong et al. 2016) and references therein. In specific, for a given class $\mathcal{H}$ of classifiers, the NP classification is to solve the following problem
\[
\min_{h \in \mathcal{H}} \mathbb{E}[\varphi(-bh(a))|b = 1]
\]
\[\text{s.t. } \mathbb{E}[\varphi(-bh(a))|b = -1] \leq \tau.
\]
In what follows, we consider its empirical risk minimization counterpart. Suppose that a labeled training dataset $\{a_i\}_{i=1}^{N}$ consists of the positive set $\{a_i^0\}_{i=1}^{N_0}$ and the negative set $\{a_i^1\}_{i=1}^{N_1}$. The associated empirical NP classification problem is
\[
\min_x f(x) := \frac{1}{N_0} \sum_{i=1}^{N_0} \ell(x^T a_i^0)
\]
\[\text{s.t. } g(x) := \frac{1}{N_1} \sum_{i=1}^{N_1} \ell(-x^T a_i^1) - \tau \leq 0,
\]

(35)
where $\ell(\cdot)$ is a loss function, e.g., logistic loss $\ell(y) := \log(1 + \exp(-y))$.

The datasets tested in our numerical comparison are summarized in Table 1. The datasets for multi-class classification have been manually divided into two types. For example, the MNIST dataset is used for classifying odd and even digits.

| Dataset | Data $N$ | Variable $n$ | Density | Reference               |
|---------|---------|--------------|---------|-------------------------|
| gisette | 6000    | 5000         | 12.97%  | (Guyon et al. 2004)     |
| CINA    | 16033   | 132          | 29.56%  | (workbench team 2008)   |
| MNIST   | 60000   | 784          | 19.12%  | (LeCun et al. 2010)     |

In the following experiment, we show the performance of SLPMM compared with CSA (Lan and Zhou 2020), PSG (Xiao 2019), YNW (Yu et al. 2017) and APriD (Yan and Xu 2022). For all five methods, we use an efficient mini-batch strategy, that is, at each iteration the stochastic gradients of the objective function and the constraint function are computed, respectively, by

$$v_0^k := \frac{1}{|N_0^k|} \sum_{i \in N_0^k} \nabla f_i(x^k), \quad v_1^k := \frac{1}{|N_1^k|} \sum_{i \in N_1^k} \nabla g_i(x^k),$$

where $f_i(x) := \ell(x^T a_i^0), i = 1, \ldots, N_0$ and $g_i(x) := \ell(-x^T a_i^1), i = 1, \ldots, N_1$. Here, the sets $N_0^k$ and $N_1^k$ are randomly chosen from the index sets $\{1, \ldots, N_0\}$ and $\{1, \ldots, N_1\}$, respectively. The batch sizes $|N_0^k|$ and $|N_1^k|$ are fixed to 1% of the data sizes $N_0$ and $N_1$, respectively. We choose $x^0 = 0$ as the initial point. The parameter $\tau$ is set to 1. The parameters in SLPMM is chosen as $\alpha = \sqrt{K}$ and $\sigma = 1/\sqrt{K}$. The maximum number of iterations is set to $K = 3000$.

In Figure 1, Figure 2, and Figure 3 we show the performance of all methods for solving the empirical NP classification problem with logistic loss. In each figure, the pictures (a) and (b) show the changes of the objective value and the constraint value with respect to epochs, and the pictures (c) and (d) represent the changes of the objective value and the constraint value with respect to cputime. Here, in (a) and (c) the horizontal dashed line represents a reference optimal objective value which is computed by the built-in MATLAB function fmincon. Moreover, one epoch denotes a full pass over a dataset. The results are averaged over 10 independent runs.
Generally, we can observe that the behaviors of CSA, PSG and YNW are similar since all of them are stochastic first-order methods. SLPMM obviously outperforms these three methods by combining the evaluations of both objective decreasing and constraint violation. In particular, the results demonstrate that SLPMM converges obviously faster than CSA and PSG both with respect to epochs and cputime. Our results also show that PSG usually generates solutions which are failed to satisfy the constraint. In contrast, CSA always gives feasible solutions, but the objective values are far from optimal. Finally, the performance of APriDi is very different from the others. The total performance of APriDi seems better than the others. However, the curves of APriD oscillate heavily even for the average of 10 runs, and the issue is much worse for each independent run.

![Figure 1 Comparison of algorithms on gisette for Neyman-Pearson classification.](image)

5.3. Stochastic quadratically constrained quadratical programming

In this subsection, we consider the following stochastic quadratically constrained quadratical programming

\[
\min_{x \in \mathbb{C}} f(x) := \mathbb{E} \left[ \frac{1}{2} x^T A^{(0)} x + (b^{(0)})^T x - c^{(0)} \right]
\]

s.t. \( g_i(x) := \mathbb{E} \left[ \frac{1}{2} x^T A^{(i)} x + (b^{(i)})^T x + c^{(i)} \right] \leq 0, \ i = 1, 2, \ldots, p, \)
Figure 2  Comparison of algorithms on CINA for Neyman-Pearson classification.

Figure 3  Comparison of algorithms on MNIST for Neyman-Pearson classification.
where \( A^{(i)} \in S_+^n \), \( b^{(i)} \in \mathbb{R}^n \), \( c^{(i)} \in \mathbb{R} \) for \( i = 0, 1, \ldots, p \). Here, \( S_+^n \) denotes the set of all \( n \times n \) positive semidefinite matrices. The expectations are taken with respect to the components of the parameters \( \{ A^{(i)}, b^{(i)}, c^{(i)} \}_{i=0}^p \), which are all random variables.

The following numerical example is partially motivated by Cao et al. (2021). The set \( C := \{ x \in \mathbb{R}^n : \| x \| \leq R \} \), where \( R > 0 \) is a constant. Let \( \hat{x} \in \mathbb{R}^n \) be a given point with its entry \( \hat{x}_i \) being uniformly generated from \( \left( -\frac{R}{\sqrt{n}}, \frac{R}{\sqrt{n}} \right) \). Let \( I_n \) be the identity matrix. For each \( i = 0, 1, \ldots, p \), the random matrix \( A^{(i)} = I_n + \Delta_i \), where \( \Delta_i \) is a symmetric matrix and its entry is uniformly distributed over \( [-0.1, 0.1] \). The random vector \( b^{(i)} \) is uniformly distributed from \( [-1, 1] \). The random variable \( c^{(i)} \) is constructed with a particular purpose. Let \( h^{(i)} \) be a random variable uniformly distributed over \( [0, 2i] \), then define \( c^{(i)} = -\left( \frac{1}{2} \hat{x}^T A^{(i)} \hat{x} + (b^{(i)})^T \hat{x} + h^{(i)} \right) \). In this setting, we can easily verify that \( g_i(\hat{x}) = -i < 0 \) for \( i = 1, \ldots, p \) and hence the Slater’s condition is satisfied. We can also get that the optimal solution is 0 and the optimal value is \( \frac{1}{2} \| \hat{x} \|^2 \).

In this experiment, we compare the performance of SLPMM with PSG, YNW and APriD. At each iteration of the algorithms, we generate the samples of \( \{ A^{(i)}, b^{(i)}, c^{(i)} \}_{i=0}^p \) based on the above distributions for function and gradient evaluation. We set \( n = 100 \), \( p = 5 \), \( R = 2 \). The maximum number of iterations is set to \( K = 1000 \). The initial point is set to \( x^0 = (\sqrt{R/n}, \sqrt{R/n}, \ldots, \sqrt{R/n})^T \).

The results in terms of time are shown in Figure 4. From picture (b) (plots the value of \( \max\{g_i(x^k)\} \)), we can see the iterations of all algorithms satisfy the constraints. From picture (a), we observe that SLPMM is comparable with PSG, and obviously outperforms over APriD and YNW.
5.4. Second-order stochastic dominance constrained portfolio optimization

In this subsection, we consider the following second-order stochastic dominance (SSD) constrained portfolio optimization problem

$$\min \mathbb{E}[-\xi^T x]$$

s.t. \( \mathbb{E}[(\eta - \xi^T x)_+] \leq \mathbb{E}[(\eta - Y)_+] \), \( \forall \eta \in \mathbb{R} \),

\( x \in C := \{ x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1, \bar{x} \geq x \geq 0 \} \),

where \( \bar{x} \) is the upper bound and \( Y \) stands for the random return of a benchmark portfolio dominated by the target portfolio in the SSD sense. Since it was first introduced by Dentcheva and Ruszczyński (2003), SSD has been widely used to control risk in financial portfolio (Kallio and Dehghan Hardoroudi 2018, Noyan 2018). Keçeci et al. (2016) showed that, if \( Y \) is discretely distributed with \( \{y_1, y_2, \ldots, y_p\} \), the SSD constrained portfolio optimization is reduced to

$$\min f(x) := \mathbb{E}[-\xi^T x]$$

s.t. \( g_i(x) := \mathbb{E}[(y_i - \xi^T x)_+] - \mathbb{E}[(y_i - Y)_+] \leq 0 \), \( i = 1, \ldots, p \),

\( x \in C := \{ x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1, \bar{x} \geq x \geq 0 \} \),

which is an instance of Problem (I).

Dentchevá et al. (2016) proposed several methods for solving SSD constrained optimization problems based on augmented Lagrangian framework and analyze their convergence. In particular, the proposed approximate augmented Lagrangian method with exact minimization (PALEM) has some similarities to SLPMM. At each iteration in PALEM, a minimization problem with respect to the augmented Lagrangian function of a reduced problem is solved to obtain \( x^k \), and the multiplier \( \mu^k \) is updated. They proved that the sequences \( \{x^k\} \) and \( \{\mu^k\} \) converge to the optimal solution of primal and dual problem, respectively. In contrast, although SLPMM is also constructed based on augmented Lagrangian framework as PALEM, they are quite different. The subproblem at each iteration in SLPMM is a minimization problem of a linearized augmented Lagrangian function together with a proximal term, which is easier to solve. The sampling strategy is different. In PALEM, the sample set is updated at each iteration based on the calculation of the expectation of constraint function. SLPMM only simply requires one sample at each iteration. Moreover,
since in our setting the expectation is assumed to be impossible to be calculated, we cannot obtain the convergence of the sequence to optimal solution.

In this experiment, we compare the performance of SLPMM with APriD, PSG, YNW and PALEM to solve Problem (36) on the following four datasets

\{“Dax_26_3046”, “DowJones_29_3020”, “SP100_90_3020”, “DowJones_76_30000”\}

from (Kececi et al. 2016). Take “DowJones_29_3020” for example, “DowJones” stands for Dow Jones Index, 29 is the number of stocks and 3020 is the number of scenarios, i.e., \(n = 29, p = 3020\). The initial point is set to 0. For PALEM, we use the MATLAB function \texttt{fmincon} to solve the subproblem. For SLPMM, we utilize the Nesterov’s accelerated projected gradient method (APG) to solve the subproblem \(\text{[34]}\), the stopping criterion of APG is set to \(\| y^t - T_{L_t}(y^t) \| \leq 10^{-6}\), and the projection \(\Pi_C\) is computed by the method proposed in (Wang and Lu 2015). In particular, since the number of the constraints of Problem \(\text{[36]}\) is large, we apply a sampling technique to reduce the computational cost. In specific, at each iteration, instead of using the whole constraint index set \(\{1, \ldots, p\}\) in the augmented Lagrangian function \(\text{[4]}\), we first randomly sample a subset \(I_k \subset \{1, \ldots, p\}\) and then replace \(\sum_{i=1}^{p} C_i y_i^t\) with \(\sum_{i \in I_k} C_i y_i^t\) in \(\text{[4]}\). This sampling strategy, which is also used in (Xiao 2019), is proven to be very efficient in practice. Let us also remark that, by taking an extra expectation with respect to \(I_k\), the expected convergence rates of SLPMM coupled with this sampling strategy can be established in a similar way as in Section 3. This is also pointed out in (Xiao 2019, Section 5).

The numerical results are presented in Figure 5. Since the maximum of \(p\) constraint values are always zero (which indicates that the constraints are satisfied), we omit the presentation of constraint violation. We only report the change of the objective value with respect to cputime. The horizontal dashed line in each picture represents a reference optimal objective value which is obtained from (Kececi et al. 2016). In general, we can observe that SLPMM has an obvious advantage compared with the other four algorithms. In view of dataset “DowJones_76_30000” which refers to a large scale optimization problem with 30,000 constraints, SLPMM converges to the optimal objective value less than 4 seconds. We can also observe that SLPMM is very robust and stable for all four datasets.
6. Conclusion

We present a hybrid method of stochastic approximation technique and proximal augmented Lagrangian method. It is shown that the expected convergence rates and the large-deviation properties are comparable with the existing related stochastic methods. On the other hand, the proposed method is parametric-independent. Numerical experiments also demonstrate the superiority in comparison with the stochastic first-order methods. Thus, both theoretical and numerical results suggest that the proposed algorithm is efficient for solving convex stochastic programming with expectation constraints.

However, there are still several valuable questions left to be answered. It is well-known that the deterministic augmented Lagrangian can achieve superlinear convergence. Therefore, the first question is whether the convergence rates can be improved to match the numerical performance and the rates in the deterministic setting. Secondly, it is worthwhile to consider the inexact method, that is, the subproblem is solved inexactly. Another interesting topic is how to use the techniques in this paper to deal with nonconvex stochastic optimization. The proposed algorithm in the current form is not applicable to solve...
nonconvex problems, such as chance constrained programs (Bai et al. 2021) and MIMO transmit signal design problem (Liu et al. 2019).

Finally, let us mention that the stochastic algorithms for stochastic optimization can be easily extended to solve online problems, and vice versa, see (Yu et al. 2017) for instance. Hence, the proposed method can be slightly revised to solve the corresponding online problems.

Acknowledgments
The authors would like to thank the anonymous reviewers and the associate editor for the valuable comments and suggestions that helped us to greatly improve the quality of the paper.

References
Akhtar Z, Bedi AS, Rajawat K (2021) Conservative stochastic optimization with expectation constraints. *IEEE Transactions on Signal Processing* 69:3190–3205.

Bai X, Sun J, Zheng X (2021) An augmented Lagrangian decomposition method for chance-constrained optimization problems. *INFORMS J. Comput.* 33(3):1056–1069.

Beck A (2017) *First-order methods in optimization*, volume 25 of *MOS-SIAM Series on Optimization* (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA).

Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *SIAM Rev.* 60(2):223–311.

Cao X, Zhang J, Poor HV (2021) Constrained online convex optimization with feedback delays. *IEEE Trans. Automat. Control* 66(11):5049–5064.

Dentcheva D, Martinez G, Wolfhagen E (2016) Augmented Lagrangian methods for solving optimization problems with stochastic-order constraints. *Oper. Res.* 64(6):1451–1465.

Dentcheva D, Ruszczyński A (2003) Optimization with stochastic dominance constraints. *SIAM J. Optim.* 14(2):548–566.

Guyon I, Gunn S, Ben-Hur A, Dror G (2004) Result analysis of the NIPS 2003 feature selection challenge. *Adv. in Neural Inf. Process. Syst.* 17, 545–552 (MIT Press).

Kallio M, Dehghan Hardoroudi N (2018) Second-order stochastic dominance constrained portfolio optimization: theory and computational tests. *European J. Oper. Res.* 264(2):675–685.

Keçeci NF, Kuzmenko V, Uryasev S (2016) Portfolios dominating indices: Optimization with second-order stochastic dominance constraints vs. minimum and mean variance portfolios. *Journal of Risk and Financial Management* 9(4):1–14.

Lan G (2016) Gradient sliding for composite optimization. *Math. Program.* 159(1-2, Ser. A):201–235.
Lan G (2020) First-order and Stochastic Optimization Methods for Machine Learning. Springer Series in the Data Sciences (Springer, Cham).

Lan G, Zhou Z (2020) Algorithms for stochastic optimization with function or expectation constraints. *Comput. Optim. Appl.* 76(2):461–498.

LeCun Y, Cortes C, Burges CJC (2010) The mnist database of handwritten digits. [http://yann.lecun.com/exdb/mnist/](http://yann.lecun.com/exdb/mnist/).

Lin Q, Nadarajah S, Soheili N, Yang T (2020) A data efficient and feasible level set method for stochastic convex optimization with expectation constraints. *Journal of Machine Learning Research* 21(143):1–45.

Liu A, Lau VKN, Kananian B (2019) Stochastic successive convex approximation for non-convex constrained stochastic optimization. *IEEE Trans. Signal Process.* 67(16):4189–4203.

Nemirovski A, Juditsky A, Lan G, Shapiro A (2008) Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* 19(4):1574–1609.

Noyan N (2018) Risk-averse stochastic modeling and optimization. *Recent Advances in Optimization and Modeling of Contemporary Problems*, chapter 10, 221–254 (INFORMS).

Parpas P, Rustem B (2007) Computational assessment of nested Benders and augmented Lagrangian decomposition for mean-variance multistage stochastic problems. *INFORMS J. Comput.* 19(2):239–247.

Pflug GC (1996) *Optimization of stochastic models*, volume 373 of *The Kluwer International Series in Engineering and Computer Science* (Kluwer Academic Publishers, Boston, MA), the interface between simulation and optimization.

Polyak BT (1967) A general method of solving extremum problems. *Doklady Akademii Nauk SSSR* 174(1):593–597.

Polyak BT (1990) New stochastic approximation type procedures. *Automat. i Telemekh.* 7:98–107.

Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* 30(4):838–855.

Robbins H, Monro S (1951) A stochastic approximation method. *Ann. Math. Statistics* 22:400–407.

Rockafellar RT (1976) Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.* 1(2):97–116.

Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. *Journal of Risk* 2:21–41.

Römisch W (2003) Stability of stochastic programming problems. *Stochastic programming*, volume 10 of *Handbooks Oper. Res. Management Sci.*, 483–554 (Elsevier Sci. B. V., Amsterdam).

Ruszczynski A, Shapiro A (2003) Stochastic programming models. *Stochastic programming*, volume 10 of *Handbooks Oper. Res. Management Sci.*, 1–64 (Elsevier Sci. B. V., Amsterdam).

Scott C, Nowak R (2005) A Neyman-Pearson approach to statistical learning. *IEEE Trans. Inform. Theory* 51(11):3806–3819.
Tong X, Feng Y, Zhao A (2016) A survey on Neyman-Pearson classification and suggestions for future research. Wiley Interdiscip. Rev. Comput. Stat. 8(2):64–81.

Wang M, Fang EX, Liu H (2017) Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. Math. Program. 161(1-2, Ser. A):419–449.

Wang W, Lu C (2015) Projection onto the capped simplex, https://arxiv.org/abs/1503.01002.

workbench team C (2008) A marketing dataset. http://www.causality.inf.ethz.ch/data/CINA.html

Xiao X (2019) Penalized stochastic gradient methods for stochastic convex optimization with expectation constraints, optimization-online.

Yan Y, Xu Y (2022) Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs. Math. Program. Comput. 14(2):319–363.

Yu H, Neely MJ, Wei X (2017) Online convex optimization with stochastic constraints. Advances in Neural Information Processing Systems, 1428–1438.

Zhang L, Zhang Y, Wu J (2020) Stochastic approximation proximal method of multipliers for convex stochastic programming, https://arxiv.org/abs/1907.12226.

Zhao XY, Sun D, Toh KC (2010) A Newton-CG augmented Lagrangian method for semidefinite programming. SIAM J. Optim. 20(4):1737–1765.

Zinkevich M (2003) Online convex programming and generalized infinitesimal gradient ascent. Proceedings of the 20th International Conference on Machine Learning, 928–935.