2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Unified analysis of stochastic gradient projection methods for convex optimization with functional constraints

Nitesh Kumar Singh[1] and Ion Necoara[1,2]

Abstract— **This paper addresses a unified convergence analysis for a large family of stochastic gradient projection algorithms for dealing with constrained finite sum convex problems, with a smooth objective function satisfying a strong convexity condition and possibly nonsmooth functional constraints. At each iteration, the algorithm takes an optimal gradient step based on a stochastic unbiased estimate of the objective function aiming to minimize it and then a feasibility step reducing the infeasibility associated with randomly observed constraints. Our stochastic gradient estimate can take diverse forms (e.g., Stochastic Gradient Descent (SGD), Stochastic Average Gradient Acceleration (SAGA), Loopless-Stochastic Variance Reduced Gradient (L-SVRG)). We conduct a convergence analysis of the proposed unified stochastic gradient projection algorithm under a diminishing stepsize, resulting in sublinear convergence rates, which are optimal for stochastic gradient methods within this problem class. Numerical evidence supports the effectiveness of our approach.**

I. Introduction

Optimization problems with functional constraints have wide applications across various domains, ranging from machine learning and signal processing to operations research and optimal control. In this paper, we consider the problem:

$$\begin{array}{lll} f^* = \min_{x\in\mathcal{Y}\subseteq\mathbb{R}^n} & f(x)\left(:= \frac{1}{N}\sum_{i=1}^N f_i(x)\right) \\ \text{subject to} & h_j(x)\leq 0 \quad \forall j=1:m, \end{array} \tag{1}$$

where f is convex, smooth, and h_j's are assumed to be convex (possibly nonsmooth). This problem is highly versatile, encompassing a range of optimization applications, including optimal control [4], distributed control [18], machine learning and statistics [3],[23], signal processing [15],[22] and portfolio optimization [1]. The complexity of these problems is increasing with the growing number of variables and constraints. Consequently, (sub)gradient-based methods are commonly employed to tackle these optimization tasks. Notably, the projected gradient descent algorithm proves effective for solving (1) when the feasible set allows for straightforward projections [5]. However, this method's reliance on computing the full gradient of the objective and working with the full feasible set at each iteration can become computationally intensive, if not infeasible, in certain scenarios (see [18]). This motivates us to consider the settings where we can project only on a single functional constraint while having access to stochastic unbiased gradient estimates at each iteration [16], [8]. Such settings are prevalent in machine learning, where one has the minimization of expected objective functions including or excluding constraints [3], and in statistics, where the goal is to minimize a finite sum objective while adhering to functional constraints [22], [1]. Several previous works have explored similar optimization problems, such as [8],[17], and [16]. Particularly, [16] and [17] consider a similar problem as in (1) and propose (mini-batch) stochastic subgradient projection algorithms for solving it. For these stochastic methods, the authors derive sublinear convergence rates under basic properties of problem functions (convexity, bounded gradient type conditions) and access to only stochastic (sub)gradients. In [8] an (unconstrained) composite optimization problem is considered and the authors derive a unifying convergence framework for classical stochastic gradient descent (SGD) [15], quantized stochastic gradient methods [2], variance reduced methods [12], [9], and some randomized coordinate descent methods [14]. Nevertheless, it's important to note that the optimization problem, the algorithm, and consequently, the convergence analysis in [8], [17], and [16] significantly differ from the focus of this paper.

Contributions: In [8], under a unified assumption on the stochastic estimates of the gradients of the objective, it is shown that a large family of stochastic gradient methods converge (sub)linearly to a neighborhood of the minimizer for convex, smooth and quasi-strongly convex functions. Our contribution extends this line of inquiry to the settings where the objective function is satisfying a strong convexity condition instead of a quasi-strongly convex condition (see [8]), and, additionally, we consider explicit functional constraints, see (1). To accomplish this, we extend into an unified variant of the stochastic gradient projection algorithm from [16] for solving (1). At each iteration, the algorithm performs a gradient step based on a stochastic unbiased estimate, which can take the form of SGD [19], [21], or variance reduced type, e.g., SAGA [6] and L-SVRG [11], [13]. Then, we perform a step dedicated to reducing the feasibility violation associated with a randomly observed functional constraint. To enhance the adaptability of our method, we derive stepsize-switching rules that guide the transition from constant to decreasing stepsizes. As a result of these developments, our work provides a rigorous foundation for the sublinear convergence rates of the weighted averages of iterates in terms of expected distance to the minimizer as well as in terms of the distance to the feasible set. From our

[1] Automatic Control and Systems Engineering Department, National University of Science and Technology Politehnica Bucharest, Spl. Independentei 313, 060042 Bucharest, Romania.

[2] Gheorghe Mihoc-Caius Iacob Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, 050711 Bucharest, Romania. Emails: nitesh.nitesh@stud.acs.upb.ro, ion.necoara@upb.ro.

knowledge, this study is the first comprehensive convergence analysis of a unified stochastic gradient projection method for addressing the general problem (1).

Content: The structure of this paper is as follows: In Section II, we detail the assumptions and algorithm. Section III discusses the different forms of our algorithm, and Section IV presents our convergence results. Finally, Section V showcases numerical experiments to validate our findings.

II. Preliminaries

In this section, we will introduce the notations, several important inequalities, and the underlying assumptions. We also present a unified stochastic gradient projection algorithm, called (U-SGP). For given $x, y \in \mathbb{R}^n$, $\langle x, y\rangle$ denotes the standard Euclidean inner product, $\|x\|$ denotes its Euclidean norm, $(x)_+ = \max\{0, x\}$. We also denote $\Pi_{\mathcal{Y}}(x)$ as the projection of point x onto the set $\mathcal{Y}$ and $\text{dist}^2(x, \mathcal{Y}) = \|x - \Pi_{\mathcal{Y}}(x)\|^2$. Let dom$f$ represent the domain of a proper closed convex function f and $D_f(x, y)$ denote the Bregman divergence associated with f: $D_f(x, y) := f(x) - f(y) - \langle \nabla f(y), x - y\rangle$. The following inequalities are used throughout the paper in our analysis:

$$\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2. \tag{2}$$

$$\|\Pi_{\mathcal{X}}(x) - y\|^2 \leq \|x - y\|^2 \quad \forall x \in \mathbb{R}^n,\ y \in \mathcal{X}. \tag{3}$$

The feasible set of problem (1) is denoted by:

$$\mathcal{X} = \{x \in \mathcal{Y} :\ h_j(x) \leq 0 \ \ \forall j = 1 : m\}.$$

We assume that the optimal value $f^* > -\infty$. Let us also denote by $[N] = \{1, ..., N\}$ and $[m] = \{1, ..., m\}$.

A. Assumptions

For the given problem (1), the set $\mathcal{Y}$ is assumed to be convex and simple, i.e., it is easy to evaluate the projection onto $\mathcal{Y}$. Furthermore, we assume that the interior of $\mathcal{Y}$ is contained in the effective domains of the functions f_i and h_j. We make no assumptions on the differentiability of h_j and use the same expression for the gradient or the subgradient of h_j at x, that is $\nabla h_j(x) \in \partial h_j(x)$, where $\partial h_j(x)$ is the subdifferential, which has one element or is a nonempty set for any $j = 1 : m$. We consider additionally the following assumptions. First, we assume that the objective function f satisfies the convexity and smoothness condition.

Assumption 2.1: Each function f_i is L-smooth and convex, i.e., for any $x, y \in \mathbb{R}^n$, the following hold:

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \tag{4}$$

$$f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x\rangle. \tag{5}$$

Note that inequality (4) yields that the objective f is also L-smooth and (5) implies convexity of f. Next, we give a unifying assumption on the stochastic gradients $\nu(x)$, first introduced in [8], which allows us to simultaneously analyze the classical SGD and variance reduced methods such as SAGA and L-SVRG.

Assumption 2.2: The gradient estimates are unbiased:

$$\mathbb{E}[\nu(x)|x] = \nabla f(x). \tag{6}$$

We also assume that there exists a random sequence $\{\sigma_k^2\}_{k\geq 0}$ such that the following two relations hold for all $x \in \mathcal{Y}$:

$$\mathbb{E}[\|\nu(x) - \nabla f(x^*)\|^2|x] \leq 2AD_f(x, x^*) + B\sigma_k^2 + D_1, \tag{7}$$

$$\mathbb{E}[\sigma_{k+1}^2|x] \leq (1 - \rho)\sigma_k^2 + 2CD_f(x, x^*) + D_2, \tag{8}$$

where A, B, C, D_1, D_2, ρ are non-negative constants.
Note that when we the objective function is L-smooth, then $D_f(x, x^*) \leq \frac{L}{2}\|x - x^*\|^2$ and consequently the previous two inequalities implies:

$$\mathbb{E}[\|\nu(x) - \nabla f(x^*)\|^2|x] \leq AL\|x - x^*\|^2 + B\sigma_k^2 + D_1, \tag{9}$$

$$\mathbb{E}[\sigma_{k+1}^2|x] \leq (1 - \rho)\sigma_k^2 + CL\|x - x^*\|^2 + D_2. \tag{10}$$

Though we consider the relations (7), (8) as an assumption in our convergence analysis, in the next section we show that for particular algorithms (e.g., SGD, SAGA, L-SVRG) these inequalities hold with known constants. We also assume f to satisfy the following strong convexity condition:

Assumption 2.3: The function f satisfies a strong convexity condition, i.e., there exists non-negative constant $\mu \geq 0$ such that for any $x, y \in \mathbb{R}^n$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2. \tag{11}$$

Note that one relation equivalent to (11) is the following inequality (see Theorem 2.1.9 in[20]):

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2 \quad \forall x, y \in \mathbb{R}^n. \tag{12}$$

Also, we assume the following boundedness on the subgradient of the functional constraints:

Assumption 2.4: The functional constraints h_j have bounded subgradients on $\mathbb{R}^n$, i.e., there exists $B_h > 0$ such that for all $\nabla h_j(x) \in \partial h_j(x)$, we have :

$$\|\nabla h_j(x)\| \leq B_h \quad \forall x \in \mathbb{R}^n, \ \ j = 1 : m.$$

This assumption implies that the functional constraints h_j are Lipschitz continuous functions. Finally, we assume some regularity condition for the constraints.

Assumption 2.5: The functional constraints satisfy additionally the following Hölderian growth condition for some constants $c > 0$:

$$\text{dist}^2(x, \mathcal{X}) \leq c \cdot \mathbb{E}_j[(h_j(x))_+^2] \ \ \forall x \in \mathbb{R}^n,$$

where j is a random variable. Note that Assumption 2.5 has been frequently used in the context of stochastic optimization problems, see e.g., [16]. Particularly, it holds e.g., when the feasible set $\mathcal{X}$ has a nonempty interior or is polyhedral (see the discussion and references there in [16]). It also holds for more general sets, e.g., when the collection of functional constraints satisfies a strengthened Slater condition, such as the generalized Robinson condition [16].

B. Unified stochastic gradient projection algorithm

In this section, we introduce a unified stochastic gradient projection algorithm, called U-SGP. Given the iteration counter k, we usually choose two indices $i_k \in \{1, \cdots, N\}$ and $j_k \in \{1, \cdots, m\}$ uniformly at random. In this algorithm, at each iteration, we perform a stochastic gradient step by sampling an unbiased estimate $\nu(x_k)$ based on i_k aimed at minimizing the finite sum objective function and then a subsequent subgradient step minimizing the feasibility violation of the observed random j_k^{th} constraint (we use the convention 0/0 = 0).

Algorithm 1 U-SGP

```
Require: Choose $x_0 \in \mathcal{Y}$, tol, and stepsizes $\alpha_k > 0$, $\beta \in (0,2)$. Set $k \leftarrow 0$
  for $k \geq 0$ do
      Sample $\nu(x_k)$ according to alg. choice and compute:
      $v_k \leftarrow x_k - \alpha_k \nu(x_k)$
      Sample $j_k \in [m]$ uniformly at random and compute:
      $z_k \leftarrow v_k - \beta \frac{(h_{j_k}(v_k))_+}{\|\nabla h_{j_k}(v_k)\|^2} \nabla h_{j_k}(v_k)$
      $x_{k+1} \leftarrow \Pi_{\mathcal{Y}}(z_k)$.
  end for
```

In the algorithm, $\alpha_k > 0$ and $\beta > 0$ are deterministic stepsizes. Note that the unbiased estimate $\nu(x_k)$ can have many forms depending on the algorithm of interest. Our algorithm is different from [8] as we consider an additional feasibility step and also differs from [16] as we consider a general stochastic gradient step for optimality.

III. Many forms of U-SGP algorithm

As discussed in the previous sections, the estimate $\nu(x_k)$ can have many different forms. In this section, we give several examples for the unbiased estimate $\nu(x_k)$ and hence different algorithms to solve problem (1). We also provide the explicit expressions of the non-negative constants used in Assumption 2.2 under the smoothness condition of function f.

A. U-SGP *as SGD*

In this section the proposed algorithm U-SGP becomes of SGD type when the unbiased estimator $\nu(x_k)$ is sampled as in SGD. This leads to the following algorithm: Under the assumption that the functions f_i are differentiable and L_i-smooth, the constants A, B, ρ, C, D_1, D_2 from Assumption 2.2 have the following explicit expressions (the proof is given in Lemma A.1, equation (16), in [8]):

Algorithm 2 U-SGP as SGD

```
Require: Choose $x_0 \in \mathcal{Y}$, tol, and stepsizes $\alpha_k > 0$, $\beta \in (0,2)$. Set $k \leftarrow 0$
  for $k \geq 0$ do
      Sample $i_k \in [N]$ uniformly at random and compute:
      $\nu(x_k) = \nabla f_{i_k}(x_k)$
      $v_k \leftarrow x_k - \alpha_k \nu(x_k)$
      Sample $j_k \in [m]$ uniformly at random and compute:
      $z_k \leftarrow v_k - \beta \frac{(h_{j_k}(v_k))_+}{\|\nabla h_{j_k}(v_k)\|^2} \nabla h_{j_k}(v_k)$
      $x_{k+1} \leftarrow \Pi_{\mathcal{Y}}(z_k)$.
  end for
```

$$A = 2L, B = 0, \rho = 1, C = 0, D_1 = 2\sigma^2, D_2 = 0,$$

where $\sigma^2 = \mathbb{E}[\|\nabla f_i(x^*)\|^2]$ and $\sigma_k \equiv 0$ a.s.

B. U-SGP *as variance reduced type algorithms*

Now, we also show that the algorithm U-SGP can take the form of a variance reduced type algorithm. First, in this section we consider that the unbiased estimator $\nu(x_k)$ to be sampled as of SAGA type and hence we have the algorithm U-SGP as SAGA. When we consider the functions f_i to be L-smooth, then the constants A, B, ρ, C, D_1, D_2 from Assumption 2.2 have the following explicit expressions in this case (the proof is given in Lemma A.6 in [8]):

Algorithm 3 U-SGP as SAGA

```
Require: Choose $x_0 \in \mathcal{Y}$, tol, and stepsizes $\alpha_k > 0$, $\beta \in (0,2)$. Set $k \leftarrow 0$ and $\phi_0^i = x_0$ for each $i = 1 : N$
  for $k \geq 0$ do
      Sample $i_k \in [N]$ uniformly at random
      Set $\phi_{k+1}^{i_k} = x_k$ and $\phi_{k+1}^i = \phi_k^i$ for $i \neq i_k$
      $\nu(x_k) = \nabla f_{i_k}(\phi_{k+1}^{i_k}) - \nabla f_{i_k}(\phi_k^{i_k}) + \frac{1}{N}\sum_{i=1}^N \nabla f_i(\phi_k^i)$
      $v_k \leftarrow x_k - \alpha_k \nu(x_k)$
      Sample $j_k \in [m]$ uniformly at random
      $z_k \leftarrow v_k - \beta \frac{(h_{j_k}(v_k))_+}{\|\nabla h_{j_k}(v_k)\|^2} \nabla h_{j_k}(v_k)$
      $x_{k+1} \leftarrow \Pi_{\mathcal{Y}}(z_k)$.
  end for
```

$$A = 2L, B = 2, \rho = 1/N, C = L/N, D_1 = 0, D_2 = 0,$$

where we consider $\sigma_k^2 = \frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\phi_k^i) - \nabla f_i(x^*)\|^2$. Second, we consider the unbiased estimator $\nu(x_k)$ to be sampled as in L-SVRG and this will give us the algorithm U-SGP as L-SVRG. When the functions f_i are L-smooth, then the constants A, B, ρ, C, D_1, D_2 from Assumption 2.2 have the following explicit expressions in this case (the proof is given in Lemma A.11 in [8]):

Algorithm 4 U-SGP as L-SVRG

```
Require: Choose $x_0 \in \mathcal{Y}$, tol, $p \in (0,1]$ and stepsizes $\alpha_k > 0$, $\beta \in (0,2)$. Set $k \leftarrow 0$ and $\omega_0 = x_0$
  for $k \geq 0$ do
      Sample $i_k \in [N]$ uniformly at random and compute:
      $\nu(x_k) = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(\omega_k) + \nabla f(\omega_k)$
      $v_k \leftarrow x_k - \alpha_k \nu(x_k)$
      $\omega_{k+1} = \begin{cases} x_k, \text{with probability } p \\ \omega_k, \text{with probability } 1-p \end{cases}$
      Sample $j_k \in [m]$ uniformly at random and compute:
      $z_k \leftarrow v_k - \beta \frac{(h_{j_k}(v_k))_+}{\|\nabla h_{j_k}(v_k)\|^2} \nabla h_{j_k}(v_k)$
      $x_{k+1} \leftarrow \Pi_{\mathcal{Y}}(z_k)$.
  end for
```

$$A = 2L, B = 2, \rho = p, C = pL, D_1 = 0, D_2 = 0,$$

where we consider $\sigma_k^2 = \frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\omega_k) - \nabla f_i(x^*)\|^2$.

IV. Unified convergence analysis

In this section, we establish a unified convergence analysis for the U-SGP algorithm assuming a strongly convex objective function. First, we establish a relation between x_k and v_{k-1} (the proof is similar to Lemma 4.3 in [17]):

Lemma 4.1: Let h_j be convex functions. Additionally, Assumptions 2.4 and 2.5 hold. Then, we have the following relation for the iterates of U−SGP:

$$\mathbb{E}[\|x_k - x^*\|^2] \leq \mathbb{E}[\|v_{k-1} - x^*\|^2] - \frac{\beta(2-\beta)}{cB_h^2}\mathbb{E}\left[\text{dist}^2(x_k, \mathcal{X})\right]. \tag{13}$$

Next, we prove a relation for the sequences v_k and x_k which plays a key role in providing the rates later.

Lemma 4.2: Let f be L-smooth and Assumption 2.2, 2.3 be satisfied. Then, for all $k \geq 0$ and stepsize $\alpha_k > 0$, we have the following recursion for any $\eta > 0$:

$$\begin{aligned}&\mathbb{E}[\|v_k - x^*\|^2] + 2\frac{B}{\rho}\alpha_k^2\mathbb{E}[\sigma_{k+1}^2] \qquad (14)\\
&\leq (1-\mu\alpha_k)\,\mathbb{E}[\|x_k - x^*\|^2] + 2\frac{B}{\rho}\alpha_k^2\mathbb{E}[\sigma_k^2]\\
&-\alpha_k\left(\mu - 2\left(A + \frac{BC}{\rho}\right)L\alpha_k\right)\|x_k - x^*\|^2 + \eta\mathbb{E}\left[\text{dist}^2(x_k, \mathcal{X})\right]\\
&\quad + 2\alpha_k^2\left(D_1 + \frac{B}{\rho}D_2 + \left(1 + \frac{1}{2\eta}\right)\|\nabla f(x^*)\|^2\right).\end{aligned}$$

Proof: Using the definition of v_k from U−SGP and the inequality (3), we get:

$$\begin{aligned}\|v_k - x^*\|^2 &= \|x_k - x^* - \alpha_k\nu(x_k)\|^2\\
&\overset{(2)}{\leq} \|x_k - x^*\|^2 - 2\alpha_k\langle\nu(x_k), x_k - x^*\rangle\\
&\quad + 2\alpha_k^2\|\nu(x_k) - \nabla f(x^*)\|^2 + 2\alpha_k^2\|\nabla f(x^*)\|^2.\end{aligned}$$

Now, taking the expectation conditioned on x_k and using strong convexity of f, we have:

$$\begin{aligned}&\mathbb{E}[\|v_k - x^*\|^2|x_k] \leq \|x_k - x^*\|^2 - 2\alpha_k\langle\nabla f(x_k), x_k - x^*\rangle\\
&\quad + 2\alpha_k^2\mathbb{E}[\|\nu(x_k) - \nabla f(x^*)\|^2|x_k] + 2\alpha_k^2\|\nabla f(x^*)\|^2\\
&= \|x_k - x^*\|^2 - 2\alpha_k\langle\nabla f(x_k) - \nabla f(x^*), x_k - x^*\rangle\\
&\quad - 2\alpha_k(\langle\nabla f(x^*), x_k - \Pi_{\mathcal{X}}(x_k)\rangle + \langle\nabla f(x^*), \Pi_{\mathcal{X}}(x_k) - x^*\rangle)\\
&\quad + 2\alpha_k^2\mathbb{E}[\|\nu(x_k) - \nabla f(x^*)\|^2|x_k] + 2\alpha_k^2\|\nabla f(x^*)\|^2\\
&\overset{(12)}{\leq} \|x_k - x^*\|^2 - 2\mu\alpha_k\|x_k - x^*\|^2 + \eta\|x_k - \Pi_{\mathcal{X}}(x_k)\|^2\\
&\quad + 2\alpha_k^2\mathbb{E}[\|\nu(x_k) - \nabla f(x^*)\|^2|x_k] + 2\left(1 + \frac{1}{2\eta}\right)\alpha_k^2\|\nabla f(x^*)\|^2\\
&\overset{(9)}{\leq} (1-\mu\alpha_k)\|x_k - x^*\|^2 + \eta\|x_k - \Pi_{\mathcal{X}}(x_k)\|^2\\
&\quad - \alpha_k(\mu - 2AL\alpha_k)\|x_k - x^*\|^2\\
&\quad + 2\alpha_k^2\left(B\sigma_k^2 + D_1 + \left(1 + \frac{1}{2\eta}\right)\|\nabla f(x^*)\|^2\right),\end{aligned}$$

where the second inequality follows from the relation $2\langle a, b\rangle \leq \eta\|a\|^2 + \frac{1}{\eta}\|b\|^2$, for any $a, b \in \mathbb{R}^n, \eta > 0$ and the optimality condition $\langle\nabla f(x^*), \Pi_{\mathcal{X}}(x_k) - x^*\rangle \geq 0$. Now adding $2\frac{B}{\rho}\alpha_k^2\mathbb{E}[\sigma_{k+1}^2|x_k]$ on both sides of the inequality and using (10) on the right hand side of the inequality, we get:

$$\begin{aligned}&\mathbb{E}[\|v_k - x^*\|^2|x_k] + 2\frac{B}{\rho}\alpha_k^2\mathbb{E}[\sigma_{k+1}^2|x_k]\\
&\leq (1-\mu\alpha_k)\|x_k - x^*\|^2 + 2\frac{B}{\rho}\alpha_k^2((1-\rho)\sigma_k^2 + D_2)\\
&-\alpha_k\left(\mu - 2\left(A + \frac{BC}{\rho}\right)L\alpha_k\right)\|x_k - x^*\|^2 + \eta\|x_k - \Pi_{\mathcal{X}}(x_k)\|^2\\
&\quad + 2\alpha_k^2\left(B\sigma_k^2 + D_1 + \left(1 + \frac{1}{2\eta}\right)\|\nabla f(x^*)\|^2\right).\end{aligned}$$

After taking the full expectation and rearranging the terms, we get the desired result. ■

Before providing the next result we introduce a switching stepsize strategy. Define $k_0 = \lceil\frac{8L\left(A + \frac{BC}{\rho}\right)}{\mu^2}\rceil$ and $\alpha_k = \frac{\gamma_k}{\mu}$, where γ_k is defined as:

$$\gamma_k = \begin{cases}\frac{\mu^2}{4L\left(A + \frac{BC}{\rho}\right)}, & \text{if } k \leq k_0\\ \frac{2}{k+1}, & \text{if } k > k_0.\end{cases}$$

Equivalently, one can easily see that our stepsize can be written as $\alpha_k = \min\left(\frac{\mu}{4L\left(A + \frac{BC}{\rho}\right)}, \frac{2}{\mu(k+1)}\right)$. Note that since our stepsize $\alpha_k \leq \frac{\mu}{4L\left(A + \frac{BC}{\rho}\right)}$, this implies:

$$\mu - 2\left(A + \frac{BC}{\rho}\right)L\alpha_k \geq \frac{\mu}{2}. \tag{15}$$

For simplicity, we define $C_{\beta,c,B_h} = \frac{\beta(2-\beta)}{cB_h^2} > 0$ and $S = \frac{\mu}{4L\left(A + \frac{BC}{\rho}\right)}$. Let us also introduce the following Lyapunov function:

$$\hat{V}_k := \|v_k - x^*\|^2 + \frac{\mu S^2 B}{\rho}\sigma_{k+1}^2.$$

Next, we prove the following recurrence for U−SGP iterates.

Lemma 4.3: Let f and h_j be convex functions, for all $j = 1 : m$. Additionally, Assumptions 2.1–2.5 hold. Further, define the stepsizes $\alpha_k = \min\left(\frac{\mu}{4L\left(A + \frac{BC}{\rho}\right)}, \frac{2}{\mu(k+1)}\right)$, $\beta \in (0, 2)$. Then, the iterates of U−SGP satisfy the recurrence:

$$\begin{aligned}&\forall k \leq k_0,\ \ \mathbb{E}[\hat{V}_{k_0}] \leq \hat{V}_0 \qquad (16)\\
&\quad + 2S^2\left(D_1 + \frac{B}{\rho}D_2 + \left(1 + \frac{1}{(1-\mu S)\,C_{\beta,c,B_h}}\right)\|\nabla f(x^*)\|^2\right),\\
&\forall k > k_0,\ \ (k+1)^2\mathbb{E}[\|v_k - x^*\|^2] + \frac{8B}{\rho\mu^2}\mathbb{E}[\sigma_{k+1}^2]\\
&\quad + (k+1)\mathbb{E}[\|x_k - x^*\|^2] + \frac{(k+1)^2}{6}C_{\beta,c,B_h}\mathbb{E}\left[\text{dist}^2(x_k, \mathcal{X})\right]\\
&\leq k^2\mathbb{E}[\|v_{k-1} - x^*\|^2] + \frac{8B}{\rho\mu^2}\mathbb{E}[\sigma_k^2] \qquad (17)\\
&\quad + \frac{8}{\mu^2}\left(D_1 + \frac{B}{\rho}D_2 + \left(1 + \frac{3}{C_{\beta,c,B_h}}\right)\|\nabla f(x^*)\|^2\right).\end{aligned}$$

Proof: Combining the inequalities (14) and (13) and using the inequality (15) together with $\eta = (1 - \mu\alpha_k)C_{\beta,c,B_h}/2$, we get:

$$\begin{aligned}&\mathbb{E}[\|v_k - x^*\|^2] + 2\frac{B}{\rho}\alpha_k^2\mathbb{E}[\sigma_{k+1}^2] + \frac{\mu\alpha_k}{2}\mathbb{E}[\|x_k - x^*\|^2]\\
&\quad + \frac{(1-\mu\alpha_k)\,C_{\beta,c,B_h}}{2}\mathbb{E}\left[\text{dist}^2(x_k, \mathcal{X})\right] \qquad (18)\\
&\leq (1-\mu\alpha_k)\,\mathbb{E}[\|v_{k-1} - x^*\|^2] + 2\frac{B}{\rho}\alpha_k^2\mathbb{E}[\sigma_k^2]\\
&\quad + 2\alpha_k^2\left(D_1 + \frac{B}{\rho}D_2 + \left(1 + \frac{1}{(1-\mu\alpha_k)C_{\beta,c,B_h}}\right)\|\nabla f(x^*)\|^2\right).\end{aligned}$$

For $k \leq k_0$, we have $\alpha_k = \frac{\mu}{4L\left(A + \frac{BC}{\rho}\right)}(:= S)$. Using the fact that $(1 - \mu\alpha_k) \leq 1$, from (18), we obtain:

$$\begin{aligned}&\mathbb{E}[\hat{V}_{k_0}] \leq \hat{V}_0\\
&\quad + 2S^2\left(D_1 + \frac{B}{\rho}D_2 + \left(1 + \frac{1}{(1-\mu S)\,C_{\beta,c,B_h}}\right)\|\nabla f(x^*)\|^2\right).\end{aligned}$$

This proves the first statement (16). Further, for $k > k_0$ we have $\alpha_k = \frac{\gamma_k}{\mu} = \frac{2}{\mu(k+1)}$. Note that here $\gamma_k = \frac{2}{k+1}$ is a nonincreasing sequence and thus we get:

$$1 - \gamma_k = \frac{k-1}{k+1} \geq \frac{1}{3} \;\; \forall k \geq 2,$$

so by relation (18), we have:

$$\begin{aligned}
&\mathbb{E}[\|v_k - x^*\|^2] + \frac{8B}{\rho\mu^2(k+1)^2}\mathbb{E}[\sigma_{k+1}^2] \\
&\quad + \frac{1}{(k+1)}\mathbb{E}[\|x_k - x^*\|^2] + \frac{1}{6}C_{\beta,c,B_h}\mathbb{E}\left[\text{dist}^2(x_k, \mathcal{X})\right] \\
&\leq \frac{k-1}{k+1}\mathbb{E}[\|v_{k-1} - x^*\|^2] + \frac{8B}{\rho\mu^2(k+1)^2}\mathbb{E}[\sigma_k^2] \\
&\quad + \frac{8}{\mu^2(k+1)^2}\left(D_1 + \frac{B}{\rho}D_2 + \left(1 + \frac{3}{C_{\beta,c,B_h}}\right)\|\nabla f(x^*)\|^2\right).
\end{aligned}$$

Now, multiply the whole inequality by $(k+1)^2$, and using the fact $k^2 - 1 \leq k^2$, we also get (17). ■

Now, we are ready to prove the rates. Let us define for $k \geq k_0+1$ the sum $S_k = \sum_{j=k_0+1}^{k}(j+1)^2 \sim \mathcal{O}(k^3 + k_0^2 k + k^2 k_0)$, and the corresponding average sequences:

$$\hat{x}_k = \frac{1}{S_k}\sum_{j=k_0+1}^{k}(j+1)^2 x_j,$$

and

$$\hat{w}_k = \frac{1}{S_k}\sum_{j=k_0+1}^{k}(j+1)^2 \Pi_{\mathcal{X}}(x_j) \in \mathcal{X}.$$

Theorem 4.4: Let f and $h_j(\cdot)$ be convex functions, for all $j = 1 : m$. Additionally, Assumptions 2.1–2.5 hold. Further, consider the stepsizes-switching rule $\alpha_k = \min\left(\frac{\mu}{4L\left(A + \frac{BC}{\rho}\right)}, \frac{2}{\mu(k+1)}\right)$, $\beta \in (0, 2)$ and $k_0 = \lceil \frac{8L\left(A + \frac{BC}{\rho}\right)}{\mu^2} \rceil$. Then, for $k > k_0$ we have the following convergence rates for the average sequence $\hat{x}_k$ in terms of optimality and feasibility violation for problem (1) (keeping only the dominant terms):

$$\begin{aligned}
&\mathbb{E}[\|\hat{x}_k - x^*\|^2] \qquad\qquad (19) \\
&\quad \leq \mathcal{O}\left(\frac{(D_1 + MD_2)C_{\beta,c,B_h}^{-1}}{\mu^2(k^2 + kk_0 + k_0^2)} + \frac{(D_1 + MD_2)}{\mu^2(k - k_0)}\right),
\end{aligned}$$

$$\mathbb{E}[\text{dist}^2(\hat{x}_k, \mathcal{X})] \leq \mathcal{O}\left(\frac{(D_1 + MD_2)C_{\beta,c,B_h}^{-1}}{\mu^2(k^2 + kk_0 + k_0^2)}\right). \qquad (20)$$

Proof: For $k > k_0$, from Lemma 4.3, summing the inequality from $k_0 + 1$ to k, we get:

$$\begin{aligned}
&(k+1)^2\mathbb{E}[\|v_k - x^*\|^2] + \frac{8B}{\rho\mu^2}\mathbb{E}[\sigma_{k+1}^2] \\
&\quad + \sum_{j=k_0+1}^{k}(j+1)\mathbb{E}[\|x_k - x^*\|^2] \\
&\quad + \frac{C_{\beta,c,B_h}}{6}\sum_{j=k_0+1}^{k}(j+1)^2\mathbb{E}\left[\text{dist}^2(x_j, \mathcal{X})\right] \\
&\leq (k_0+1)^2\mathbb{E}[\|v_{k_0} - x^*\|^2] + \frac{8B}{\rho\mu^2}\mathbb{E}[\sigma_{k_0+1}^2] \\
&\quad + \frac{8}{\mu^2}\left(D_1 + \frac{B}{\rho}D_2 + \left(1 + \frac{3}{C_{\beta,c,B_h}}\right)\|\nabla f(x^*)\|^2\right)(k - k_0).
\end{aligned}$$

Now, using relation (3), the convexity of the norm and the linearity of the expectation operator, we obtain:

$$\begin{aligned}
&(k+1)^2\mathbb{E}[\|v_k - x^*\|^2] + \frac{8B}{\rho\mu^2}\mathbb{E}[\sigma_{k+1}^2] \\
&\quad + \frac{S_k}{(k+1)}\mathbb{E}[\|\hat{w}_k - x^*\|^2] + \frac{S_k C_{\beta,c,B_h}}{6}\mathbb{E}\left[\|\hat{w}_k - \hat{x}_k\|^2\right] \\
&\leq (k_0+1)^2\mathbb{E}[\|v_{k_0} - x^*\|^2] + \frac{8B}{\rho\mu^2}\mathbb{E}[\sigma_{k_0+1}^2] \\
&\quad + \frac{8}{\mu^2}\left(D_1 + \frac{B}{\rho}D_2 + \left(1 + \frac{3}{C_{\beta,c,B_h}}\right)\|\nabla f(x^*)\|^2\right)(k - k_0).
\end{aligned}$$

After simple calculations (keeping only the dominant terms):

$$\begin{aligned}
&\mathbb{E}[\|\hat{w}_k - x^*\|^2] \leq \mathcal{O}\left(\frac{\left(D_1 + \frac{B}{\rho}D_2 + \|\nabla f(x^*)\|^2\right)}{\mu^2(k - k_0)}\right), \\
&\mathbb{E}[\|\hat{w}_k - \hat{x}_k\|^2] \\
&\quad \leq \mathcal{O}\left(\frac{\left(D_1 + \frac{B}{\rho}D_2 + \|\nabla f(x^*)\|^2\right)C_{\beta,c,B_h}^{-1}}{\mu^2(k^2 + kk_0 + k_0^2)}\right).
\end{aligned}$$

Since $\hat{w}_k \in \mathcal{X}$, using the relation $\mathbb{E}[\text{dist}^2(\hat{x}_k, \mathcal{X})] \leq \mathbb{E}[\|\hat{w}_k - \hat{x}_k\|^2]$ we get the required result (20). Furthermore, since x^* is the minimizer of problem (1) and using the inequality (2), we further get convergence rate for the average sequence $\hat{x}_k$ in terms of optimality:

$$\mathbb{E}[\|\hat{x}_k - x^*\|^2] \leq 2\mathbb{E}[\|\hat{w}_k - x^*\|^2] + 2\mathbb{E}[\|\hat{w}_k - \hat{x}_k\|^2].$$

This gives the required result (19) ■

Our derived convergence rates are consistent with those in [17] and [16], assuming the smoothness and strong convexity of the objective function. To the best of our knowledge, this is the first unified convergence analysis for a broad spectrum of stochastic gradient projection algorithms, encompassing variance reduced techniques, designed to tackle problem (1). In the forthcoming section, we will explore the practical benefits of variance reduced iterations, such as SAGA or L-SVRG, over the standard SGD variant.

V. Numerical experiments

In this section, we apply our proposed algorithm, U-SGP (including its variants SGD, SAGA, L-SVRG), to address the problem of finding the minimum distance to specified points within the context of a finite number of half-space constraints and box constraints (as detailed in Section 4 of [7]). The problem can be framed as follows:

$$\begin{aligned}
&\min_{x \in \mathbb{R}^n} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\|x - c_i\|^2 \qquad (21) \\
&\text{s.t.} \quad \langle a_j, x\rangle \leq b_j \;\; \forall j \in [m], \;\; x \in [l, \; u],
\end{aligned}$$

where $c_i, a_j \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ and $l, u \in \mathbb{R}^n$ be the lower and upper limits of vector x. Note that problem (21) aligns with the assumptions outlined in our paper. In our experiments we generate all the data from a normal distribution, choose $\beta = 1.96$, $m = n = 10^2$, tol $= 10^{-2}$, and consider $N = 10^4$. The algorithms are stopped when $\|\max(0, h_j(x))\| \leq$ tol and $\|x - x^*\| \leq$ tol (we consider CVX solution [10] for computing x^*). The codes are written in Matlab R2023b and run on a PC with an i7 CPU at 2.1 GHz and 16 GB RAM memory. In Figure 1, we present the convergence behavior of all three variants of U−SGP algorithm, i.e., SGD, SAGA, and L-SVRG along epochs, with $N = 10^4, m = n = 100$ in terms of optimality (left) and feasibility (right) for solving the problem (21). One can easily see that the variance reduced methods, i.e., SAGA and L-SVRG variants, outperform the standard SGD variant in terms of the number of epochs required for convergence.

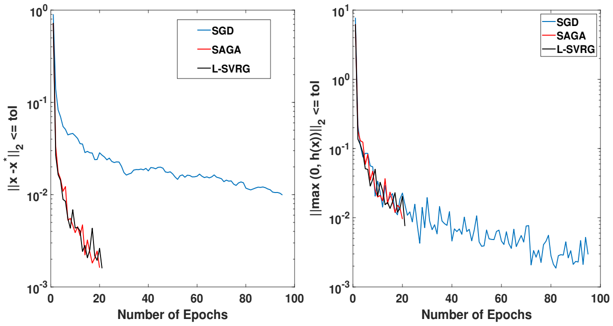

Fig. 1. Behaviour of U−SGP algorithms (SGD, SAGA and L-SVRG): optimality (left) and feasibility (right) for $N = 10^4, m = n = 10^2$.

In Table I, we compare the three variants of algorithm U−SGP (SGD, SAGA, and L-SVRG) with CVX in terms of epochs and cpu time for $N = 10^4$ and $m = n = 10^2$. From the table one can see that SAGA and L-SVRG methods perform better than SGD in both number of epochs and cpu time (seconds). Hence, variance reduced optimality steps in U−SGP have a beneficial effect on the overall convergence behavior of this algorithm. Moreover, all our stochastic gradient projection methods are much faster than CVX.

Characteristics Method	# Iter	Cpu Time (s)
SGD	95	2.78
SAGA	24	1.66
L-SVRG	21	0.72
CVX	**	179.38

TABLE I

PERFORMANCE OF U−SGP (SGD, SAGA, AND L-SVRG) AND CVX IN TERMS OF EPOCHS AND CPU TIME (SEC.) ($N = 10^4, m = n = 10^2$).

VI. CONCLUSIONS

In this work, we have focused on a convex finite sum problem with functional constraints. To solve this problem we have proposed a large family of stochastic gradient projection algorithms, called U−SGP, covering, in particular, SGD, but also variance reduced schemes. We provide a unified convergence analysis and derive sublinear convergence rates for the weighted average of the iterates in terms of expected distance to the constraint set, as well as for expected optimality of the distance to the optimal point. The numerical tests also prove the effectiveness of our algorithmic framework.

VII. ACKNOWLEDGEMENT

The research leading to these results has received funding from: UEFISCDI, Romania, PN-III-P4-PCE-2021-0720, under project L2O-MOC, nr. 70/2022.

REFERENCES

[1] A. Ali, J.Z. Kolter, S. Diamond and S.P. Boyd, ”Disciplined Convex Stochastic Programming: A New Framework for Stochastic Optimization”, *Conf. on Uncertainty in Artificial Intelligence*, 62-71, 2015.

[2] D. Alistarh, D. Grubic, J. Li, R. Tomioka and M. Vojnovic, ”QSGD: Communication-efficient SGD via gradient quantization and encoding”, *In Advances in Neural Information Processing Systems*, 1709–1720, 2017.

[3] C. Bhattacharyya, L.R. Grate, M.I. Jordan, L. El Ghaoui and S. Mian, ”Robust sparse hyperplane classifiers: Application to uncertain molecular profiling data”, *Journal of Computational Biology*, 11(6): 1073–1089, 2004.

[4] E. Berthier, J. Carpentier, A. Rudi and F. Bach, ”Infinite-Dimensional Sums-of-Squares for Optimal Control”, *Conf. on Decision and Control (CDC), Cancun, Mexico*, 577-582, 2022.

[5] P. L. Combettes and J.C. Pesquet, “Proximal splitting methods in signal processing”, *in Fixed-point algorithms for inverse problems in science and engineering, Springer*, 2011. ISBN : 978-1-4419-9568-1.

[6] A. Defazio, F. R. Bach, and S. Lacoste-Julien, ”SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives”, *In Advances in Neural Information Processing Systems 27*, 1646–1654, 2014.

[7] N. Ekkarntrong, T. Arunrat, and N. Nimana, ”Convergence of a distributed method for minimizing sum of convex functions with fixed point constraints”, *Journal of Inequalities and Applications*, 197: 2021.

[8] E. Gorbunov, F. Hanzely and P. Richtárik, ”A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent”, *Int. Conf. on Artificial Intelligence and Statistics*, 108: 2020.

[9] R. M. Gower, P. Richtárik and F. Bach, ”Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching”, *Mathematical Programming*, 188:135–192, 2021.

[10] M. Grant and S. Boyd, ”CVX: matlab software for disciplined convex programming, version 2.0 beta”, 2013.

[11] T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams, ”Variance reduced stochastic gradient descent with neighbors”, *In Advances in Neural Information Processing Systems 28*, 2305–2313, 2015.

[12] R. Johnson and T. Zhang, ”Accelerating stochastic gradient descent using predictive variance reduction”, *In Advances in Neural Information Processing Systems 26*, 315–323, 2013.

[13] D. Kovalev, S. Horvath, and P. Richtárik, ”Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop”, *Int. Conf. on Algo. Learning Theory*, 451–467, 2020.

[14] I. Necoara, D. Clipici, ”Parallel random coordinate descent methods for composite minimization: convergence analysis and error bounds”, *SIAM J. Optimization*, 26(1): 197-226, 2016.

[15] I. Necoara, ”General convergence analysis of stochastic first order methods for composite optimization”, *Journal of Optimization Theory and Applications*, 189: 66–95, 2021.

[16] I. Necoara and N. K. Singh, ”Stochastic subgradient for composite convex optimization with functional constraints”, *Journal of Machine Learning Research*, 23(265): 1–35, 2022.

[17] N. K. Singh, I. Necoara and V. Kungurtsev, ”Mini-batch stochastic subgradient for functional constrained optimization, *Optimization*, 1–27, 2023. doi: 10.1080/02331934.2023.2189015

[18] V. Nedelcu, I. Necoara and Q. Tran Dinh, ”Computational complexity of inexact gradient augmented Lagrangian methods: application to constrained MPC”, *SIAM J. Control Optim.*, 52(5): 3109–3134, 2014.

[19] A. Nemirovski and D.B. Yudin, ”Problem complexity and method efficiency in optimization”, *Wiley Interscience*, 1983.

[20] Yu. Nesterov, ”Lectures on Convex Optimization,” *Springer Optimization and Its Applications*, 137, 2018.

[21] H. Robbins and S. Monro, ”A Stochastic Approximation Method”, *The Annals of Mathematical Statistics*, 22(3): 400–407, 1951.

[22] R. Tibshirani, ”The solution path of the generalized lasso”, *Phd Thesis, Stanford University*, 2011.

[23] V. Vapnik, ”Statistical learning theory”, *John Wiley*, 1998.