Journal of Machine Learning Research 23 (2022) 1-35 Submitted 9/21; Revised 4/22; Published 9/22

Stochastic subgradient projection methods for composite optimization with functional constraints

Ion Necoara ION.NECOARA@UPB.RO
Automatic Control and Systems Engineering Department, University Politehnica Bucharest, Spl. Independentei 313, 060042 Bucharest, Romania.
Gheorghe Mihoc-Caius Iacob Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, 050711 Bucharest, Romania.

Nitesh Kumar Singh NITESH.NITESH@STUD.ACS.UPB.RO
Automatic Control and Systems Engineering Department, University Politehnica Bucharest, Spl. Independentei 313, 060042 Bucharest, Romania.

Editor: Silvia Villa

Abstract

In this paper we consider optimization problems with stochastic composite objective function subject to (possibly) infinite intersection of constraints. The objective function is expressed in terms of expectation operator over a sum of two terms satisfying a stochastic bounded gradient condition, with or without strong convexity type properties. In contrast to the classical approach, where the constraints are usually represented as intersection of simple sets, in this paper we consider that each constraint set is given as the level set of a convex but not necessarily differentiable function. Based on the flexibility offered by our general optimization model we consider a stochastic subgradient method with random feasibility updates. At each iteration, our algorithm takes a stochastic proximal (sub)gradient step aimed at minimizing the objective function and then a subsequent subgradient step minimizing the feasibility violation of the observed random constraint. We analyze the convergence behavior of the proposed algorithm for diminishing stepsizes and for the case when the objective function is convex or has a quadratic functional growth, unifying the nonsmooth and smooth cases. We prove sublinear convergence rates for this stochastic subgradient algorithm, which are known to be optimal for subgradient methods on this class of problems. When the objective function has a linear least-square form and the constraints are polyhedral, it is shown that the algorithm converges linearly. Numerical evidence supports the effectiveness of our method in real problems.

Keywords: Stochastic optimization, convex functional constraints, stochastic subgradient, rate of convergence, constrained least-squares, robust/sparse svm.

1. Introduction

The large sum of functions in the objective function and/or the large number of constraints in most of the practical optimization applications, including machine learning and statistics (Vapnik, 1998; Bhattacharyya et al., 2004), signal processing (Necoara, 2021; Tibshirani, 2011), computer science (Kundu et al., 2018), distributed control (Nedelcu et al., 2014), operations research and finance (Rockafellar and Uryasev, 2000), create several theoretical and computational challenges. For example, these problems are becoming increasingly large

in terms of both the number of variables and the size of training data and they are usually nonsmooth due to the use of regularizers, penalty terms and the presence of constraints. Due to these challenges, (sub)gradient-based methods are widely applied. In particular, proximal gradient algorithms (Necoara, 2021; Rosasco et al., 2019) are natural in applications where the function to be minimized is in composite form, i.e. the sum of a smooth term and a nonsmooth regularizer (e.g., the indicator function of some simple constraints), as e.g. the empirical risk in machine learning. In these algorithms, at each iteration the proximal operator defined by the nonsmooth component is applied to the gradient descent step for the smooth data fitting component. In practice, it is important to consider situations where these operations cannot be performed exactly. For example, the case where the gradient, the proximal operator or the projection can be computed only up to an error have been considered in (Devolder et al., 2014; Nedelcu et al., 2014; Necoara and Patrascu, 2018; Rasch and Chambolle, 2020), while the situation when only stochastic estimates of these operators are available have been studied in (Rosasco et al., 2019; Hardt et al., 2016; Patrascu and Necoara, 2018; Hermer et al., 2020). This latter setting, which is also of interest here, is very important in machine learning, where we have to minimize an expected objective function with or without constraints from random samples (Bhattacharyya et al., 2004), or in statistics, where we need to minimize a finite sum objective subject to functional constraints (Tibshirani, 2011).

Previous work. A very popular approach for minimizing an expected or finite sum objective function is the stochastic gradient descent (SGD) (Robbins and Monro, 1951; Nemirovski and Yudin, 1983; Hardt et al., 2016) or the stochastic proximal point (SPP) algorithms (Moulines and Bach, 2011; Nemirovski et al., 2009; Necoara, 2021; Patrascu and Necoara, 2018; Rosasco et al., 2019). In these studies sublinear convergence is derived for SGD or SPP with decreasing stepsizes under the assumptions that the objective function is smooth and (strongly) convex. For nonsmooth stochastic convex optimization one can recognize two main approaches. The first one uses stochastic variants of the subgradient method combined with different averaging techniques, see e.g. (Nemirovski et al., 2009; Yang and Lin, 2016). The second line of research is based on stochastic variants of the proximal gradient method under the assumption that the expected objective function is in composite form (Duchi and Singer, 2009; Necoara, 2021; Rosasco et al., 2019). For both approaches, using decreasing stepsizes, sublinear convergence rates are derived under the assumptions that the objective function is (strongly) convex. Even though SGD and SPP are well-developed methodologies, they only apply to problems with simple constraints, requiring the whole feasible set to be projectable.

In spite of its wide applicability, the study on efficient solution methods for optimization problems with many constraints is still limited. The most prominent line of research in this area is the alternating projections, which focus on applying random projections for solving problems that involve intersection of a (infinite) number of sets. The case when the objective function is not present in the formulation, which corresponds to the convex feasibility problem, stochastic alternating projection algorithms were analyzed e.g., in (Bauschke and Borwein, 1996), with linear convergence rates, provided that the sets satisfy some linear regularity condition. Stochastic forward-backward algorithms have been also applied to solve optimization problems with many constraints. However, the papers introducing those

general algorithms focus on proving only asymptotic convergence results without rates, or they assume the number of constraints is finite, which is more restricted than our settings, see e.g. (Bianchi et al., 2019; Xu, 2020; Wang et al., 2015). In the case where the number of constraints is finite and the objective function is deterministic, Nesterov's smoothing framework is studied in (Tran-Dinh et al., 2018) under the setting of accelerated proximal gradient methods. Incremental subgradient methods or primal-dual approaches were also proposed for solving problems with finite intersection of simple sets through an exact penalty reformulation in (Bertsekas, 2011; Kundu et al., 2018).

The papers most related to our work are (Nedich, 2011; Nedich and Necoara, 2019), where subgradient methods with random feasibility steps are proposed for solving convex problems with *deterministic* objective and many functional constraints. However, the optimization problem, the algorithm and consequently the convergence analysis are different from the present paper. In particular, our algorithm is a *stochastic proximal gradient* extension of the algorithms proposed in (Nedich, 2011; Nedich and Necoara, 2019). Additionally, the stepsizes in (Nedich, 2011; Nedich and Necoara, 2019) are chosen decreasing, while in the present work for strongly like objective functions we derive insightful stepsize-switching rules which describe when one should switch from a constant to a decreasing stepsize regime. Furthermore, in (Nedich, 2011) and (Nedich and Necoara, 2019) sublinear convergence rates are established either for convex or strongly convex deterministic objective functions, respectively, while in this paper we prove (sub)linear rates under an expected composite objective function which is either convex or satisfies relaxed strong convexity conditions. Moreover, (Nedich, 2011; Nedich and Necoara, 2019) present separately the convergence analysis for smooth and nonsmooth objective, while in this paper we present a unified convergence analysis covering both cases through the so-called stochastic bounded gradient condition. Hence, since we deal with stochastic composite objective functions, smooth or nonsmooth, and relaxed strong convexity assumptions, and since we consider a stochastic proximal gradient with new stepsize rules, our convergence analysis requires additional insights that differ from that of (Nedich, 2011; Nedich and Necoara, 2019).
In (Patrascu and Necoara, 2018) a stochastic optimization problem with infinite intersection of sets is considered and stochastic proximal point steps are combined with alternating projections for solving it. However, in order to prove sublinear convergence rates, (Patrascu and Necoara, 2018) requires strongly convex and smooth objective functions, while our results are valid for a more relaxed strong convexity condition and possibly non-smooth functions. Lastly, (Patrascu and Necoara, 2018) assumes the projectability of individual sets, whereas in our case, the constraints might not be projectable. Finally, in all these studies the nonsmooth component is assumed to be proximal, i.e. one can easily compute its proximal operator. This assumption is restrictive, since in many applications the nonsmooth term is also expressed as expectation or finite sum of nonsmooth functions, which individually are proximal, but it is hard to compute the proximal operator for the whole nonsmooth component, as e.g., in support vector machine or generalized lasso problems (Villa et al., 2014; Rosasco et al., 2019). Moreover, all the convergence results from the previous papers are derived separately for the smooth and the nonsmooth stochastic optimization problems.

Contributions. In this paper we remove the previous drawbacks. We propose a stochastic subgradient algorithm for solving general optimization problems having the objective

function expressed in terms of expectation operator over a sum of two terms subject to (possibly infinite number of) functional constraints. The only assumption we require is to have access to an unbiased estimate of the (sub)gradient and of the proximal operator of each of these two terms and to the subgradients of the constraints. To deal with such problems, we propose a stochastic subgradient method with random feasibility updates. At each iteration, the algorithm takes a stochastic subgradient step aimed at only minimizing the expected objective function, followed by a feasibility step for minimizing the feasibility violation of the observed random constraint achieved through Polyak's subgradient iteration, see (Polyak, 1969). In doing so, we can avoid the need for projections onto the whole set of constraints, which may be expensive computationally. The proposed algorithm is applicable to the situation where the whole objective function and/or constraint set are not known in advance, but they are rather learned in time through observations.

We present a general framework for the convergence analysis of this stochastic subgradient algorithm which is based on the assumptions that the objective function satisfies a stochastic bounded gradient condition, with or without strong convexity type properties and the subgradients of the functional constraints are bounded. These assumptions include the most well-known classes of objective functions and of constraints analyzed in the literature: composition of a nonsmooth function and a smooth function, with or without strong convexity, and nonsmooth Lipschitz functions, respectively. Moreover, when the objective function satisfies some relaxed strong convexity conditions, we prove insightful stepsize-switching rules which describe when one should switch from a constant to a decreasing stepsize regime. Then, we prove sublinear convergence rates for the weighted averages of the iterates in terms of expected distance to the constraint set, as well as for the expected optimality of the function values/distance to the optimal set. Under some special conditions we also derive linear rates. Our convergence estimates are known to be optimal for this class of stochastic subgradient schemes for solving nonsmooth (convex) problems with functional constraints. Besides providing a general framework for the design and analysis of stochastic subgradient methods, in special cases, where complexity bounds are known for some particular problems, our convergence results recover the existing bounds. In particular, for problems without functional constraints we recover the complexity bounds from (Necoara, 2021) and for linearly constrained least-squares problems we get similar convergence bounds as in (Leventhal and Lewis, 2010).

Content. In Section 2 we present our optimization problem and the main assumptions. In Section 3 we design a stochastic subgradient projection algorithm and analyze its convergence properties, while is Section 4 we adapt this algorithm to constrained least-squares problems and derive linear convergence. Finally, in Section 5 detailed numerical simulations are provided that support the effectiveness of our method in real problems.

2. Problem formulation and assumptions

We consider the general composite optimization problem with expected objective function and functional constraints:

$$\begin{array}{rl} F^* = & \min\limits_{x \in \mathcal{Y} \subseteq \mathbb{R}^n} F(x) \quad (:= \mathbb{E}[f(x,\zeta) + g(x,\zeta)]) \\ & \text{subject to } \; h(x,\xi) \leq 0 \;\; \forall \xi \in \Omega_2, \end{array} \tag{1}$$

where the composite objective function F has a stochastic representation in the form of expectation w.r.t. a random variable $\zeta \in \Omega_1$, i.e., $\mathbb{E}[f(x,\zeta) + g(x,\zeta)]$, Ω_2 is an arbitrary collection of indices and $\mathcal{Y}$ is a closed convex set. Here $f(\cdot,\zeta), g(\cdot,\zeta)$ and $h(\cdot,\xi)$ are proper lower-semicontinuous functions containing $\mathcal{Y}$ in their domains. Moreover, $f(\cdot,\zeta)$ are possibly nonconvex, while $g(\cdot,\zeta), h(\cdot,\xi)$ are assumed convex. One can notice that more commonly, one sees a single g representing the regularizer on the parameters in the formulation (1). However, there are also applications where one encounters terms of the form $\mathbb{E}[g(x,\zeta)]$ or $\sum_{\xi\in\Omega_2} g(x,\xi)$, as e.g., in Lasso problems with mixed $\ell_1 - \ell_2$ regularizers or regularizers with overlapping groups, see e.g., (Villa et al., 2014). Multiple functional constraints, onto which is difficult to project, can arise from robust classification in machine learning, chance constrained problems, min-max games and control, see (Bhattacharyya et al., 2004; Rockafellar and Uryasev, 2000; Patrascu and Necoara, 2018). For further use, we define the following functions: $F(x,\zeta) = f(x,\zeta) + g(x,\zeta)$, $f(x) = \mathbb{E}[f(x,\zeta)]$ and $g(x) = \mathbb{E}[g(x,\zeta)]$. Moreover, $\mathcal{Y}$ is assumed to be simple, i.e. one can easily compute the projection of a point onto this set. Let us define the individual sets $\mathcal{X}_\xi$ as $\mathcal{X}_\xi = \{x \in \mathbb{R}^n : \ h(x,\xi) \leq 0\}$ for all $\xi \in \Omega_2$. Denote the feasible set of (1) by:

$$\mathcal{X} = \{x \in \mathcal{Y} : \ h(x,\xi) \leq 0 \ \ \forall \xi \in \Omega_2\} = \mathcal{Y} \cap (\cap_{\xi\in\Omega_2} \mathcal{X}_\xi)\,.$$

We assume $\mathcal{X}$ to be nonempty. We also assume that the optimization problem (1) has finite optimum and we let F^* and $\mathcal{X}^*$ denote the optimal value and the optimal set, respectively:

$$F^* = \min_{x\in\mathcal{X}} F(x) := \mathbb{E}[F(x,\zeta)], \quad \mathcal{X}^* = \{x \in \mathcal{X} \mid F(x) = F^*\} \neq \emptyset.$$

Further, for any $x \in \mathbb{R}^n$ we denote its projection onto the optimal set $\mathcal{X}^*$ by $\bar{x}$, that is:

$$\bar{x} = \Pi_{\mathcal{X}^*}(x).$$

For the objective function we assume that the first term $f(\cdot,\zeta)$ is either differentiable or nondifferentiable function and we use, with some abuse of notation, the same notation for the gradient or the subgradient of $f(\cdot,\zeta)$ at x, that is $\nabla f(x,\zeta) \in \partial f(x,\zeta)$, where the subdifferential $\partial f(x,\zeta)$ is either a singleton or a nonempty set for any $\zeta \in \Omega_1$. The other term $g(\cdot,\zeta)$ is assumed to have an easy proximal operator for any $\zeta \in \Omega_1$:

$$\text{prox}_{\gamma g(\cdot,\zeta)}(x) = \arg\min_{y\in\mathbb{R}^m} g(y,\zeta) + \frac{1}{2\gamma}\|y - x\|^2.$$

Recall that the proximal operator of the indicator function of a closed convex set becomes the projection. We consider additionally the following assumptions on the objective function and constraints:

Assumption 1 *The (sub)gradients of F satisfy a stochastic bounded gradient condition, that is there exist non-negative constants $L \geq 0$ and $B \geq 0$ such that:*

$$B^2 + L(F(x) - F^*) \geq \mathbb{E}[\|\nabla F(x,\zeta)\|^2] \quad \forall x \in \mathcal{Y}. \tag{2}$$

We also assume F to satisfy some regularity condition:

Assumption 2 *The function F satisfies a quadratic functional growth condition, i.e. there exists non-negative constant $\mu \geq 0$ such that:*

$$F(x) - F^* \geq \frac{\mu}{2}\|x - \bar{x}\|^2 \quad \left(:= \frac{\mu}{2}\text{dist}^2(x, \mathcal{X}^*)\right) \quad \forall x \in \mathcal{Y}. \tag{3}$$

Note that when $\mu = 0$ relation (3) holds automatically. Finally, we assume boundedness on the subgradients of the functional constraints:

Assumption 3 *The functional constraints $h(\cdot, \xi)$ have bounded subgradients on $\mathcal{Y}$, i.e., there exists non-negative constant $B_h > 0$ such that for all $\nabla h(x, \xi) \in \partial h(x, \xi)$, we have:*

$$\|\nabla h(x, \xi)\| \leq B_h \quad \forall x \in \text{dom}\, g \;\; \textit{and} \;\; \xi \in \Omega_2.$$

Note that our assumptions are quite general and cover the most well-known classes of functions analyzed in the literature. In particular, Assumptions 1 and 2, related to the objective function, covers the class of non-smooth Lipschitz functions and composition of a (potentially) non-smooth function and a smooth function, with or without strong convexity, as the following examples show.

Example 1 [Non-smooth (Lipschitz) functions satisfy Assumption 1]: Assume that the functions f and g have bounded (sub)gradients:

$$\|\nabla f(x, \zeta)\| \leq B_f \quad \text{and} \quad \|\nabla g(x, \zeta)\| \leq B_g \quad \forall x \in \mathcal{Y}.$$

Then, obviously Assumption 1 holds with $L = 0 \quad \text{and} \quad B^2 = 2B_f^2 + 2B_g^2$.

Example 2 [Smooth (Lipschitz gradient) functions satisfy Assumption 1]: Condition (2) includes the class of functions formed as a sum of two terms, $f(\cdot, \zeta)$ having Lipschitz continuous gradient and $g(\cdot, \zeta)$ having bounded subgradients, and $\mathcal{Y}$ bounded. Indeed, let us assume that $f(\cdot, \zeta)$ has Lipschitz continuous gradient, i.e. there exists $L_f(\zeta) > 0$ such that:

$$\|\nabla f(x, \zeta) - \nabla f(\bar{x}, \zeta)\| \leq L_f(\zeta)\|x - \bar{x}\| \quad \forall x \in \mathcal{Y}.$$

Using standard arguments (see Theorem 2.1.5 in (Nesterov, 2018)), we have:

$$f(x, \zeta) - f(\bar{x}, \zeta) \geq \langle \nabla f(\bar{x}, \zeta), x - \bar{x} \rangle + \frac{1}{2L_f(\zeta)}\|\nabla f(x, \zeta) - \nabla f(\bar{x}, \zeta)\|^2.$$

Assuming $g(\cdot, \zeta)$ convex, then adding $g(x, \zeta) - g(\bar{x}, \zeta) \geq \langle \nabla g(\bar{x}, \zeta), x - \bar{x} \rangle$ in the previous inequality, where $\nabla g(\bar{x}, \zeta) \in \partial g(\bar{x}, \zeta)$, we get:

$$F(x, \zeta) - F(\bar{x}, \zeta) \geq \langle \nabla F(\bar{x}, \zeta), x - \bar{x} \rangle + \frac{1}{2L_f(\zeta)}\|\nabla f(x, \zeta) - \nabla f(\bar{x}, \zeta)\|^2,$$

where we used that $\nabla F(\bar{x}, \zeta) = \nabla f(\bar{x}, \zeta) + \nabla g(\bar{x}, \zeta) \in \partial F(\bar{x}, \zeta)$. Taking expectation w.r.t. ζ and assuming that the set $\mathcal{Y}$ is bounded, with the diameter D, then after using Cauchy-Schwartz inequality, we get (we also assume $L_f(\zeta) \leq L_f$ for all ζ):

$$F(x) - F^* \geq \frac{1}{2L_f}\mathbb{E}[\|\nabla f(x, \zeta) - \nabla f(\bar{x}, \zeta)\|^2] - D\|\nabla F(\bar{x})\| \quad \forall x \in \mathcal{Y}.$$

Therefore, for any $\nabla g(x,\zeta) \in \partial g(x,\zeta)$, we have:

$$\begin{aligned}\mathbb{E}[\|\nabla F(x,\zeta)\|^2] &= \mathbb{E}[\|\nabla f(x,\zeta) - \nabla f(\bar{x},\zeta) + \nabla g(x,\zeta) + \nabla f(\bar{x},\zeta)\|^2] \\ &\leq 2\mathbb{E}[\|\nabla f(x,\zeta) - \nabla f(\bar{x},\zeta)\|^2] + 2\mathbb{E}[\|\nabla g(x,\zeta) + \nabla f(\bar{x},\zeta)\|^2] \\ &\leq 4L_f(F(x) - F^*) + 4L_f D\|\nabla F(\bar{x})\| + 2\mathbb{E}[\|\nabla g(x,\zeta) + \nabla f(\bar{x},\zeta)\|^2].\end{aligned}$$

Assuming now that the regularization functions g have bounded subgradients on $\mathcal{Y}$, i.e., $\|\nabla g(x,\zeta)\| \leq B_g$, then the bounded gradient condition (2) holds on on $\mathcal{Y}$ with:

$$L = 4L_f \quad \text{and} \quad B^2 = 4\left(B_g^2 + \max_{\bar{x}\in\mathcal{X}^*}\left(\mathbb{E}[\|\nabla f(\bar{x},\zeta)\|^2] + DL_f\|\nabla F(\bar{x})\|\right)\right).$$

Note that B^2 is finite, since the optimal set $\mathcal{X}^*$ is compact (recall that in this example $\mathcal{Y}$ is assumed bounded). Further, many practical problems satisfy the quadratic functional growth condition (3), the most relevant one is given next.

Example 3 [Composition between a strongly convex function and a linear map satisfy Assumption 2]: Assume $F(x) = \hat{f}(A^T x) + g(x)$, where $\hat{f}$ is a strongly convex function with constant $\sigma_f > 0$, A is a general nonzero matrix of appropriate dimension and g is a polyhedral function (i.e., the epigraph of g is a polyhedral set). Assume also that index set Ω_1 is finite and $h(\cdot,\zeta)$ are also polyhedral functions. Then, the quadratic functional growth condition (3) holds on any sublevel set, that is for any $M > 0$ there exists $\mu(M) > 0$ such that (Necoara et al., 2019):

$$F(x) - F^* \geq \frac{\mu(M)}{2}\|x - \bar{x}\|^2 \quad \forall x \in \mathcal{X} :\ F(x) - F^* \leq M.$$

Quadratic functional growth (3) is a relaxation of the well-known strong convexity notion, see (Necoara et al., 2019) for a detailed discussion. Clearly, any strongly convex function F satisfies (3). Moreover, Assumption 3, related to the functional constraints, covers the class of nonsmooth Lipschitz functions. Finally, note that Assumption 2 may look contradictory with Example 1 when F is convex, since the following relations need to hold:

$$\frac{\mu}{2}\|x - \bar{x}\|^2 \leq F(x) - F(\bar{x}) \leq \langle \nabla F(x), x - \bar{x}\rangle \leq B_F\|x - \bar{x}\| \quad \forall x \in \mathcal{Y},$$

where the second inequality follows from the convexity of F and the third one from the Cauchy-Schwartz inequality and boundedness of the subgradients of F (according to Example 1). This implies that $\|x - \bar{x}\| \leq 2B_F/\mu$ for any $x \in \mathcal{Y}$. Note that this inequality is always valid provided that the set $\mathcal{Y}$ is compact and our optimization model (1) allows us to impose such an assumption on the set $\mathcal{Y}$.

Finally, we also impose some regularity condition for the constraints.

Assumption 4 *The functional constraints satisfy a regularity condition, i.e., there exists non-negative constant $c > 0$ such that:*

$$\operatorname{dist}^2(y,\mathcal{X}) \leq c \cdot \mathbb{E}\left[(h(y,\xi))_+^2\right] \quad \forall y \in \operatorname{dom} g. \tag{4}$$

Note that this assumption holds e.g., when the index set Ω_2 is arbitrary and the feasible set $\mathcal{X}$ has an interior point, see (Polyak, 2001), or when the feasible set is polyhedral, see relation (22) below. However, Assumption (4) holds for more general sets, e.g., when a strengthened Slater condition holds for the collection of convex functional constraints, such as the generalized Robinson condition, as detailed in Corollary 3 of (Lewis and Pang, 1998).

3. Stochastic subgradient projection algorithm

Given the iteration counter k, we consider independent random variables ζ_k and ξ_k sampled from Ω_1 and Ω_2 according to probability distributions $\mathbf{P}_1$ and $\mathbf{P}_2$, respectively. Then, we define the following stochastic subgradient projection algorithm, where at each iteration we perform a stochastic proximal (sub)gradient step aimed at minimizing the expected composite objective function and then a subsequent subgradient step minimizing the feasibility violation of the observed random constraint (we use the convention $0/0 = 0$):

Algorithm 1 (SSP):
Choose $x_0 \in \mathcal{Y}$ and stepsizes $\alpha_k > 0$ and $\beta \in (0, 2)$
For $k \geq 0$ repeat:

Sample independently $\zeta_k \sim \mathbf{P}_1$ and $\xi_k \sim \mathbf{P}_2$ and update :

$$v_k = \text{prox}_{\alpha_k g(\cdot,\zeta_k)}\left(x_k - \alpha_k \nabla f(x_k, \zeta_k)\right) \tag{5}$$

$$z_k = v_k - \beta \frac{(h(v_k, \xi_k))_+}{\|\nabla h(v_k, \xi_k)\|^2} \nabla h(v_k, \xi_k) \tag{6}$$

$$x_{k+1} = \Pi_{\mathcal{Y}}(z_k). \tag{7}$$

Here, $\alpha_k > 0$ and $\beta > 0$ are deterministic stepsizes and $(x)_+ = \max\{0, x\}$. Note that v_k represents a stochastic proximal (sub)gradient step (or a stochastic forward-backward step) at x_k for the expected composite objective function $F(x) = \mathbb{E}[f(x, \zeta) + g(x, \zeta)]$. Note that the optimality step selects a random pair of functions $(f(x_k, \zeta_k), g(x_k, \zeta_k))$ from the composite objective function F according to the probability distribution $\mathbf{P}_1$, i.e., the index variable ζ_k with values in the set Ω_1. Also, we note that the random feasibility step selects a random constraint $h(\cdot, \xi_k) \leq 0$ from the collection of constraints set according to the probability distribution $\mathbf{P}_2$, independently from ζ_k, i.e. the index variable ξ_k with values in the set Ω_2. The vector $\nabla h(v_k, \xi_k)$ is chosen as:

$$\nabla h(v_k, \xi_k) = \begin{cases} \nabla h(v_k, \xi_k) \in \partial((h(v_k, \xi_k))_+) & \text{if } (h(v_k, \xi_k))_+ > 0 \\ s_h \neq 0 & \text{if } (h(v_k, \xi_k))_+ = 0, \end{cases}$$

where $s_h \in \mathbb{R}^n$ is any nonzero vector. If $(h(v_k, \xi_k))_+ = 0$, then for any choice of nonzero s_h, we have $z_k = v_k$. Note that the feasibility step (6) has the special form of the Polyak's subgradient iteration, see e.g., (Polyak, 1969). Moreover, when $\beta = 1$, z_k is the projection of v_k onto the hyperplane:

$$\mathcal{H}_{v_k,\xi_k} = \{z : \ h(v_k, \xi_k) + \nabla h(v_k, \xi_k)^T (z - v_k) \leq 0\},$$

that is $z_k = \Pi_{\mathcal{H}_{v_k,\xi_k}}(v_k)$ for $\beta = 1$. Indeed, if $v_k \in \mathcal{H}_{v_k,\xi_k}$, then $(h(v_k,\xi_k))_+ = 0$ and the projection is the point itself, i.e., $z_k = v_k$. On the other hand, if $v_k \notin \mathcal{H}_{v_k,\xi_k}$, then the projection of v_k onto $\mathcal{H}_{v_k,\xi_k}$ reduces to the projection onto the corresponding hyperplane:

$$z_k = v_k - \frac{h(v_k,\xi_k)}{\|\nabla h(v_k,\xi_k)\|^2}\nabla h(v_k,\xi_k).$$

Combining these two cases, we finally get our update (6) . Note that when the feasible set of optimization problem (1) is described by (infinite) intersection of simple convex sets, see e.g. (Patrascu and Necoara, 2018; Bianchi et al., 2019; Xu, 2020; Wang et al., 2015):

$$\mathcal{X} = \mathcal{Y} \cap \left(\cap_{\xi\in\Omega_2}\mathcal{X}_\xi\right),$$

where each set X_ξ admits an easy projection, then one can choose the following functional representation in problem (1):

$$h(x,\xi) = (h(x,\xi))_+ = \operatorname{dist}(x,\mathcal{X}_\xi) \quad \forall \xi \in \Omega_2.$$

One can easily notice that this function is convex, nonsmooth and with bounded subgradients, since we have (Mordukhovich and Nam, 2005):

$$\frac{x - \Pi_{\mathcal{X}_\xi}(x)}{\|x - \Pi_{\mathcal{X}_\xi}(x)\|} \in \partial h(x,\xi).$$

In this case, step (6) in SSP algorithm becomes a usual random projection step:

$$z_k = v_k - \beta(v_k - \Pi_{\mathcal{X}_{\xi_k}}(v_k)).$$

Hence, our formulation is more general than (Patrascu and Necoara, 2018; Bianchi et al., 2019; Xu, 2020; Wang et al., 2015), as it allows to also deal with constraints that might not be projectable, but one can easily compute a subgradient of h. We mention also the possibility of performing iterations in parallel in steps (5) and (6) of SSP algorithm. However, here we do not consider this case. A thorough convergence analysis of minibatch iterations when the objective function is deterministic and strongly convex can be found in (Nedich and Necoara, 2019). For the analysis of SSP algorithm, let us define the stochastic (sub)gradient mapping (for simplicity we omit its dependence on stepsize α):

$$\mathcal{S}(x,\zeta) = \alpha^{-1}\left(x - \operatorname{prox}_{\alpha g(\cdot,\zeta)}(x - \alpha\nabla f(x,\zeta))\right).$$

Then, it follows immediately that the stochastic proximal (sub)gradient step (5) can be written as:

$$v_k = x_k - \alpha_k \mathcal{S}(x_k,\zeta_k).$$

Moreover, from the optimality condition of the prox operator it follows that there exists $\nabla g(v_k,\zeta_k) \in \partial g(v_k,\zeta_k)$ such that:

$$\mathcal{S}(x_k,\zeta_k) = \nabla f(x_k,\zeta_k) + \nabla g(v_k,\zeta_k).$$

Let us also recall a basic property of the projection onto a closed convex set $\mathcal{X} \subseteq \mathbb{R}^n$, see e.g., (Nedich and Necoara, 2019):

$$\|\Pi_{\mathcal{X}}(v) - y\|^2 \leq \|v - y\|^2 - \|\Pi_{\mathcal{X}}(v) - v\|^2 \qquad \forall v \in \mathbb{R}^n \text{ and } y \in \mathcal{X}. \tag{8}$$

Define also the filtration:

$$\mathcal{F}_{[k]} = \{\zeta_0, \cdots, \zeta_k, \;\; \xi_0, \cdots, \xi_k\}.$$

The next lemma provides a key descent property for the sequence v_k and for the proof we use as main tool the stochastic (sub)gradient mapping $\mathcal{S}(\cdot)$.

Lemma 5 *Let $f(\cdot, \zeta)$ and $g(\cdot, \zeta)$ be convex functions. Additionally, assume that the bounded gradient condition from Assumption 1 holds. Then, for any $k \geq 0$ and stepsize $\alpha_k > 0$, we have the following recursion:*

$$\mathbb{E}[\|v_k - \bar{v}_k\|^2] \leq \mathbb{E}[\|x_k - \bar{x}_k\|^2] - \alpha_k(2 - \alpha_k L)\, \mathbb{E}[F(x_k) - F(\bar{x}_k)] + \alpha_k^2 B^2. \tag{9}$$

Proof Recalling that $\bar{v}_k = \Pi_{\mathcal{X}^*}(v_k)$ and $\bar{x}_k = \Pi_{\mathcal{X}^*}(x_k)$, from the definition of v_k and using (8) for $y = \bar{x}_k \in \mathcal{X}^*$ and $v = v_k$, we get:

$$\begin{aligned} \|v_k - \bar{v}_k\|^2 &\overset{(8)}{\leq} \|v_k - \bar{x}_k\|^2 = \|x_k - \bar{x}_k - \alpha_k \mathcal{S}(x_k, \zeta_k)\|^2 \\ &= \|x_k - \bar{x}_k\|^2 - 2\alpha_k \langle \mathcal{S}(x_k, \zeta_k), x_k - \bar{x}_k \rangle + \alpha_k^2 \|\mathcal{S}(x_k, \zeta_k)\|^2 \\ &= \|x_k - \bar{x}_k\|^2 - 2\alpha_k \langle \nabla f(x_k, \zeta_k) + \nabla g(v_k, \zeta_k), x_k - \bar{x}_k \rangle + \alpha_k^2 \|\mathcal{S}(x_k, \zeta_k)\|^2. \end{aligned} \tag{10}$$

Now, we refine the second term. First, from convexity of f we have:

$$\langle \nabla f(x_k, \zeta_k), x_k - \bar{x}_k \rangle \geq f(x_k, \zeta_k) - f(\bar{x}_k, \zeta_k).$$

Then, from convexity of $g(\cdot, \zeta)$ and the definition of the gradient mapping $\mathcal{S}(\cdot, \zeta)$, we have:

$$\begin{aligned} &\langle \nabla g(v_k, \zeta_k), x_k - \bar{x}_k \rangle = \langle \nabla g(v_k, \zeta_k), x_k - v_k \rangle + \langle \nabla g(v_k, \zeta_k), v_k - \bar{x}_k \rangle \\ &\geq \alpha_k \|\mathcal{S}(x_k, \zeta_k)\|^2 - \alpha_k \langle \nabla f(x_k, \zeta_k), \mathcal{S}(x_k, \zeta_k) \rangle + g(v_k, \zeta_k) - g(\bar{x}_k, \zeta_k) \\ &\geq \alpha_k \|\mathcal{S}(x_k, \zeta_k)\|^2 - \alpha_k \langle \nabla f(x_k, \zeta_k) + \nabla g(x_k, \zeta_k), \mathcal{S}(x_k, \zeta_k) \rangle + g(x_k, \zeta_k) - g(\bar{x}_k, \zeta_k). \end{aligned}$$

Replacing the previous two inequalities in (10), we obtain:

$$\begin{aligned} \|v_k - \bar{v}_k\|^2 \leq \|x_k - \bar{x}_k\|^2 &- 2\alpha_k \left(f(x_k, \zeta_k) + g(x_k, \zeta_k) - f(\bar{x}_k, \zeta_k) - g(\bar{x}_k, \zeta_k) \right) \\ &+ 2\alpha_k^2 \langle \nabla f(x_k, \zeta_k) + \nabla g(x_k, \zeta_k), \mathcal{S}(x_k, \zeta_k) \rangle - \alpha_k^2 \|\mathcal{S}(x_k, \zeta_k)\|^2. \end{aligned}$$

Using that $2\langle u, v \rangle - \|v\|^2 \leq \|u\|^2$ for all $v \in \mathbb{R}^n$, that $F(x_k, \zeta_k) = f(x_k, \zeta_k) + g(x_k, \zeta_k)$ and $\nabla F(x_k, \zeta_k) = \nabla f(x_k, \zeta_k) + \nabla g(x_k, \zeta_k) \in \partial F(x_k, \zeta_k)$, we further get:

$$\|v_k - \bar{v}_k\|^2 \leq \|x_k - \bar{x}_k\|^2 - 2\alpha_k \left(F(x_k, \zeta_k) - F(\bar{x}_k, \zeta_k) \right) + \alpha_k^2 \|\nabla F(x_k, \zeta_k)\|^2.$$

Since v_k depends on $\mathcal{F}_{[k-1]} \cup \{\zeta_k\}$, not on ξ_k, and the stepsize α_k does not depend on (ζ_k, ξ_k), then from basic properties of conditional expectation, we have:

$$\begin{aligned}
&\mathbb{E}_{\zeta_k}[\|v_k - \bar{v}_k\|^2 | \mathcal{F}_{[k-1]}] \\
&\leq \|x_k - \bar{x}_k\|^2 - 2\alpha_k \mathbb{E}_{\zeta_k}[F(x_k, \zeta_k) - F(\bar{x}_k, \zeta_k) | \mathcal{F}_{[k-1]}] + \alpha_k^2 \mathbb{E}_{\zeta_k}[\|\nabla F(x_k, \zeta_k)\|^2 | \mathcal{F}_{[k-1]}] \\
&= \|x_k - \bar{x}_k\|^2 - 2\alpha_k \left(F(x_k) - F(\bar{x}_k)\right) + \alpha_k^2 \mathbb{E}_{\zeta_k}[\|\nabla F(x_k, \zeta_k)\|^2 | \mathcal{F}_{[k-1]}] \\
&\leq \|x_k - \bar{x}_k\|^2 - 2\alpha_k \left(F(x_k) - F(\bar{x}_k)\right) + \alpha_k^2 \left(B^2 + L\left(F(x_k) - F(\bar{x}_k)\right)\right) \\
&= \|x_k - \bar{x}_k\|^2 - \alpha_k(2 - \alpha_k L)\left(F(x_k) - F(\bar{x}_k)\right) + \alpha_k^2 B^2,
\end{aligned}$$

where in the last inequality we used $x_k \in \mathcal{Y}$ and the stochastic bounded gradient inequality from Assumption 1. Now, taking expectation w.r.t. $\mathcal{F}_{[k-1]}$ we get:

$$\mathbb{E}[\|v_k - \bar{v}_k\|^2] \leq \mathbb{E}[\|x_k - \bar{x}_k\|^2] - \alpha_k(2 - \alpha_k L)\mathbb{E}[(F(x_k) - F(\bar{x}_k))] + \alpha_k^2 B^2,$$

which concludes our statement. ■

Next lemma gives a relation between x_k and v_{k-1} (see also (Nedich and Necoara, 2019)).

Lemma 6 *Let $h(\cdot, \xi)$ be convex functions. Additionally, Assumption 3 holds. Then, for any $y \in \mathcal{Y}$ such that $(h(y, \xi_{k-1}))_+ = 0$ the following relation holds:*

$$\|x_k - y\|^2 \leq \|v_{k-1} - y\|^2 - \beta(2 - \beta)\left[\frac{(h(v_{k-1}, \xi_{k-1}))_+^2}{B_h^2}\right].$$

Proof Consider any $y \in \mathcal{Y}$ such that $(h(y, \xi_{k-1}))_+ = 0$. Then, using the nonexpansive property of the projection and the definition of z_{k-1}, we have:

$$\begin{aligned}
\|x_k - y\|^2 &= \|\Pi_{\mathcal{Y}}(z_{k-1}) - y\|^2 \leq \|z_{k-1} - y\|^2 \\
&= \|v_{k-1} - y - \beta \frac{(h(v_{k-1}, \xi_{k-1}))_+}{\|\nabla h(v_{k-1}, \xi_{k-1})\|^2} \nabla h(v_{k-1}, \xi_{k-1})\|^2 \\
&= \|v_{k-1} - y\|^2 + \beta^2 \frac{(h(v_{k-1}, \xi_{k-1}))_+^2}{\|\nabla h(v_{k-1}, \xi_{k-1})\|^2} \\
&\quad - 2\beta \frac{(h(v_{k-1}, \xi_{k-1}))_+}{\|\nabla h(v_{k-1}, \xi_{k-1})\|^2} \langle v_{k-1} - y, \nabla h(v_{k-1}, \xi_{k-1})\rangle \\
&\leq \|v_{k-1} - y\|^2 + \beta^2 \frac{(h(v_{k-1}, \xi_{k-1}))_+^2}{\|\nabla h(v_{k-1}, \xi_{k-1})\|^2} - 2\beta \frac{(h(v_{k-1}, \xi_{k-1}))_+}{\|\nabla h(v_{k-1}, \xi_{k-1})\|^2} (h(v_{k-1}, \xi_{k-1}))_+,
\end{aligned}$$

where the last inequality follows from convexity of $(h(\cdot, \xi))_+$ and our assumption that $(h(y, \xi_{k-1}))_+ = 0$. After rearranging the terms and using that $v_{k-1} \in \operatorname{dom} g$, we get:

$$\begin{aligned}
\|x_k - y\|^2 &\leq \|v_{k-1} - y\|^2 - \beta(2 - \beta)\left[\frac{(h(v_{k-1}, \xi_{k-1}))_+^2}{\|\nabla h(v_{k-1}, \xi_{k-1})\|^2}\right] \\
&\overset{\text{Assumption 3}}{\leq} \|v_{k-1} - y\|^2 - \beta(2 - \beta)\left[\frac{(h(v_{k-1}, \xi_{k-1}))_+^2}{B_h^2}\right],
\end{aligned}$$

which concludes our statement. ■

In the next sections, based on the previous two lemmas, we derive convergence rates for SSP algorithm depending on the (convexity) properties of $f(\cdot, \zeta)$.

3.1 Convergence analysis: convex objective function

In this section we consider that the functions $f(\cdot,\zeta), g(\cdot,\zeta)$ and $h(\cdot,\xi)$ are convex. First, it is easy to prove that $cB_h^2 > 1$ (we can always choose c sufficiently large such that this relation holds), see also (Nedich and Necoara, 2019). For simplicity of the exposition let us introduce the following notation:

$$C_{\beta,c,B_h} := \frac{\beta(2-\beta)}{cB_h^2}\left(1 - \frac{\beta(2-\beta)}{cB_h^2}\right)^{-1} > 0.$$

Let us also define the average sequence generated by the algorithm SSP:

$$\hat{x}_k = \frac{\sum_{j=1}^k \alpha_j x_j}{S_k}, \quad \text{where } S_k = \sum_{j=1}^k \alpha_j.$$

The next theorem derives sublinear convergence rates for the average sequence $\hat{x}_k$.

Theorem 7 *Let $f(\cdot,\zeta), g(\cdot,\zeta)$ and $h(\cdot,\xi)$ be convex functions. Additionally, Assumptions 1, 3 and 4 hold. Further, assume a nonincreasing positive stepsize sequence α_k, with $\alpha_0 \in \left(0, \frac{1}{L}\right)$, satisfying $\sum_{k\geq 0}\alpha_k = \infty$ and $\sum_{k\geq 0}\alpha_k^2 < \infty$, and stepsize $\beta \in (0,2)$. Then, we have the following convergence rates for the average sequence $\hat{x}_k$ in terms of optimality and feasibility violation for problem* (1)*:*

$$\mathbb{E}\left[F(\hat{x}_k) - F^*\right] \leq \frac{\|v_0 - \overline{v}_0\|^2}{S_k} + \frac{B^2\sum_{j=1}^k \alpha_j^2}{S_k},$$
$$\mathbb{E}\left[\text{dist}^2(\hat{x}_k, \mathcal{X})\right] \leq \frac{\alpha_0\|v_0 - \overline{v}_0\|^2}{C_{\beta,c,B_h}\cdot S_k} + \frac{\alpha_0 B^2\sum_{j=1}^k \alpha_j^2}{C_{\beta,c,B_h}\cdot S_k}.$$

Proof Recall that from Lemma 5, we have:

$$\mathbb{E}\left[\|v_k - \bar{v}_k\|^2\right] \leq \mathbb{E}\left[\|x_k - \bar{x}_k\|^2\right] - \alpha_k(2-\alpha_k L)\mathbb{E}\left[F(x_k) - F(\bar{x}_k)\right] + \alpha_k^2 B^2. \tag{11}$$

Now, for $y = \bar{v}_{k-1} \in \mathcal{X}^* \subseteq \mathcal{X} \subseteq \mathcal{Y}$ we have that $(h(\bar{v}_{k-1}, \xi_{k-1}))_+ = 0$, and thus using Lemma 6, we get:

$$\|x_k - \bar{x}_k\|^2 \overset{(8)}{\leq} \|x_k - \bar{v}_{k-1}\|^2 \leq \|v_{k-1} - \bar{v}_{k-1}\|^2 - \beta(2-\beta)\left[\frac{(h(v_{k-1},\xi_{k-1}))_+^2}{B_h^2}\right].$$

Taking conditional expectation on ξ_{k-1} given $\mathcal{F}_{[k-1]}$, we get:

$$\begin{aligned}\mathbb{E}_{\xi_{k-1}}[\|x_k - \bar{x}_k\|^2|\mathcal{F}_{[k-1]}] &\leq \|v_{k-1} - \bar{v}_{k-1}\|^2 - \beta(2-\beta)\mathbb{E}_{\xi_{k-1}}\left[\frac{(h(v_{k-1},\xi_{k-1}))_+^2}{B_h^2}|\mathcal{F}_{[k-1]}\right] \\ &\overset{(4)}{\leq} \|v_{k-1} - \bar{v}_{k-1}\|^2 - \frac{\beta(2-\beta)}{cB_h^2}\text{dist}^2(v_{k-1},\mathcal{X}).\end{aligned}$$

Taking now full expectation, we obtain:

$$\mathbb{E}[\|x_k - \bar{x}_k\|^2] \leq \mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] - \frac{\beta(2-\beta)}{cB_h^2}\mathbb{E}[\text{dist}^2(v_{k-1},\mathcal{X})], \tag{12}$$

and using this relation in (11), we get:

$$\mathbb{E}\left[\|v_k-\bar{v}_k\|^2\right]+\frac{\beta(2-\beta)}{cB_h^2}\mathbb{E}\left[\mathrm{dist}^2(v_{k-1},\mathcal{X})\right]+\alpha_k(2-\alpha_kL)\mathbb{E}\left[F(x_k)-F(\bar{x}_k)\right]$$
$$\leq \mathbb{E}[\|v_{k-1}-\bar{v}_{k-1}\|^2]+\alpha_k^2B^2. \tag{13}$$

Similarly, for $y=\Pi_{\mathcal{X}}(v_{k-1})\subseteq\mathcal{X}\subseteq\mathcal{Y}$ we have that $(h(\Pi_{\mathcal{X}}(v_{k-1}),\xi_{k-1}))_+=0$, and thus using again Lemma 6, we obtain:

$$\mathrm{dist}^2(x_k,\mathcal{X})=\|x_k-\Pi_{\mathcal{X}}(x_k)\|^2\leq\|x_k-\Pi_{\mathcal{X}}(v_{k-1})\|^2$$
$$\leq \mathrm{dist}^2(v_{k-1},\mathcal{X})-\beta(2-\beta)\frac{(h(v_{k-1},\xi_{k-1}))_+^2}{B_h^2}.$$

Taking conditional expectation on ξ_{k-1} given $\mathcal{F}_{[k-1]}$, we get:

$$\mathbb{E}_{\xi_{k-1}}\left[\mathrm{dist}^2(x_k,\mathcal{X})|\mathcal{F}_{[k-1]}\right]$$
$$\leq \mathrm{dist}^2(v_{k-1},\mathcal{X})-\frac{\beta(2-\beta)}{B_h^2}\mathbb{E}_{\xi_{k-1}}\left[(h(v_{k-1},\xi_{k-1}))_+^2|\mathcal{F}_{[k-1]}\right]$$
$$\overset{(4)}{\leq}\left(1-\frac{\beta(2-\beta)}{cB_h^2}\right)\mathrm{dist}^2(v_{k-1},\mathcal{X}).$$

After taking full expectation, we get:

$$\mathbb{E}\left[\mathrm{dist}^2(x_k,\mathcal{X})\right]\leq\left(1-\frac{\beta(2-\beta)}{cB_h^2}\right)\mathbb{E}\left[\mathrm{dist}^2(v_{k-1},\mathcal{X})\right]. \tag{14}$$

Using (14) in (13), we obtain:

$$\mathbb{E}\left[\|v_k-\bar{v}_k\|^2\right]+\frac{\beta(2-\beta)}{cB_h^2}\left(1-\frac{\beta(2-\beta)}{cB_h^2}\right)^{-1}\mathbb{E}\left[\mathrm{dist}^2(x_k,\mathcal{X})\right] \tag{15}$$
$$+\alpha_k(2-\alpha_kL)\mathbb{E}\left[F(x_k)-F(\bar{x}_k)\right]\leq\mathbb{E}\left[\|v_{k-1}-\bar{v}_{k-1}\|^2\right]+\alpha_k^2B^2.$$

Summing (15) from 1 to k , we get:

$$\mathbb{E}\left[\|v_k-\bar{v}_k\|^2\right]+C_{\beta,c,B_h}\sum_{j=1}^k\mathbb{E}\left[\mathrm{dist}^2(x_j,\mathcal{X})\right]$$
$$+\sum_{j=1}^k\alpha_j(2-\alpha_jL)\mathbb{E}\left[F(x_j)-F^*\right]\leq\|v_0-\bar{v}_0\|^2+B^2\sum_{j=1}^k\alpha_j^2.$$

Since $\alpha_0\in(0,1/L)$ and $\alpha_j\leq\alpha_0$, then $\alpha_j\in(0,1/L)$ and $2-\alpha_jL\geq1$ for all $j\geq0$. Moreover, since α_j is a nonincreasing sequence, we have $\alpha_j/\alpha_0\leq1$. Thus, we obtain:

$$\mathbb{E}\left[\|v_k-\bar{v}_k\|^2\right]+\frac{C_{\beta,c,B_h}}{\alpha_0}\sum_{j=1}^k\alpha_j\mathbb{E}\left[\mathrm{dist}^2(x_j,\mathcal{X})\right]+\sum_{j=1}^k\alpha_j\mathbb{E}\left[F(x_j)-F^*\right]$$
$$\leq\|v_0-\bar{v}_0\|^2+B^2\sum_{j=1}^k\alpha_j^2.$$

Using the definition of the average sequence $\hat{x}_k$ and the convexity of F and of $\text{dist}^2(\cdot, \mathcal{X})$, we further get sublinear rates in expectation for the average sequence in terms of optimality:

$$\mathbb{E}\left[F(\hat{x}_k) - F^*\right] \leq \sum_{j=1}^{k} \frac{\alpha_j}{S_k} \mathbb{E}\left[F(x_j) - F^*\right] \leq \frac{\|v_0 - \bar{v}_0\|^2}{S_k} + B^2 \frac{\sum_{j=1}^{k} \alpha_j^2}{S_k},$$

and feasibility violation:

$$\begin{aligned} \frac{C_{\beta,c,B_h}}{\alpha_0} \mathbb{E}\left[\text{dist}^2(\hat{x}_k, \mathcal{X})\right] &\leq \frac{C_{\beta,c,B_h}}{\alpha_0} \sum_{j=1}^{k} \frac{\alpha_j}{S_k} \mathbb{E}\left[\text{dist}^2(x_j, \mathcal{X})\right] \\ &\leq \frac{\|v_0 - \bar{v}_0\|^2}{S_k} + B^2 \frac{\sum_{j=1}^{k} \alpha_j^2}{S_k}. \end{aligned}$$

These conclude our statements. ■

Note that for stepsize $\alpha_k = \frac{\alpha_0}{(k+1)^\gamma}$, with $\gamma \in [1/2, 1)$, we have:

$$S_k = \sum_{j=1}^{k} \alpha_j \geq \mathcal{O}(k^{1-\gamma}) \quad \text{and} \quad \sum_{j=1}^{k} \alpha_j^2 \leq \begin{cases} \mathcal{O}(1) & \text{if } \gamma > 1/2 \\ \mathcal{O}(\ln(k)) & \text{if } \gamma = 1/2. \end{cases}$$

Consequently for $\gamma \in (1/2, 1)$ we obtain from Theorem 7 the following sublinear convergence rates:

$$\begin{aligned} \mathbb{E}\left[(F(\hat{x}_k) - F^*)\right] &\leq \frac{\|v_0 - \bar{v}_0\|^2}{k^{1-\gamma}} + \frac{B^2}{k^{1-\gamma}}, \\ \mathbb{E}\left[\text{dist}^2(\hat{x}_k, \mathcal{X})\right] &\leq \frac{\alpha_0 \|v_0 - \bar{v}_0\|^2}{C_{\beta,c,B_h} \cdot k^{1-\gamma}} + \frac{\alpha_0 B^2}{C_{\beta,c,B_h} \cdot k^{1-\gamma}}. \end{aligned}$$

For the particular choice $\gamma = 1/2$, if we neglect the logarithmic terms, we get from Theorem 7 sublinear convergence rates of order:

$$\mathbb{E}\left[(F(\hat{x}_k) - F^*)\right] \leq \mathcal{O}\left(\frac{1}{k^{1/2}}\right) \quad \text{and} \quad \mathbb{E}\left[\text{dist}^2(\hat{x}_k, \mathcal{X})\right] \leq \mathcal{O}\left(\frac{1}{k^{1/2}}\right).$$

It is important to note that when $B = 0$, from Theorem 7 improved rates can be derived for SSP algorithm in the convex case. More precisely, for stepsize $\alpha_k = \frac{\alpha_0}{(k+1)^\gamma}$, with $\gamma \in [0, 1)$, we obtain convergence rates for $\hat{x}_k$ in optimality and feasibility violation of order $\mathcal{O}\left(\frac{1}{k^{1-\gamma}}\right)$. In particular, for $\gamma = 0$ (i.e. constant stepsize $\alpha_k = \alpha_0 \in (0, 1/L)$ for all $k \geq 0$) the previous convergence estimates yield rates of order $\mathcal{O}\left(\frac{1}{k}\right)$. From our best knowledge these rates are new for stochastic subgradient methods applied on the class of optimization problems (1).

3.2 Convergence analysis: quadratic growth convex objective function

In this section we additionally assume the quadratic growth inequality from Assumption 2. The next lemma derives an improved recurrence relation for the sequence v_k under the

quadratic growth inequality. Our proof below is different from the one in (Nedich and Necoara, 2019), since that paper makes use heavily on the strong convexity condition, the uniqueness of the optimum point, and the boundedness of the subgradients of the objective function, while here all these conditions do not hold anymore.

Lemma 8 *Let $f(\cdot,\zeta), g(\cdot,\zeta)$ and $h(\cdot,\xi)$ be convex functions. Additionally, Assumptions 1–4 hold. Define $k_0 = \lceil \frac{8L}{\mu} \rceil$, $\beta \in (0,2)$, $\theta_{L,\mu} = 1-\mu/(2L)$ and $\alpha_k = \frac{4}{\mu}\gamma_k$, where γ_k is given by:*

$$\gamma_k = \begin{cases} \frac{\mu}{4L} & \text{if } k \le k_0 \\ \frac{2}{k+1} & \text{if } k > k_0. \end{cases}$$

Then, the iterates of SSP algorithm satisfy the following recurrence:

$$\mathbb{E}[\|v_{k_0} - \bar{v}_{k_0}\|^2] \le \begin{cases} \frac{B^2}{L^2} & \text{if } \theta_{L,\mu} \le 0 \\ \theta_{L,\mu}^{k_0}\|v_0 - \bar{v}_0\|^2 + \frac{1-\theta_{L,\mu}^{k_0}}{1-\theta_{L,\mu}}\frac{B^2}{L^2} & \text{if } \theta_{L,\mu} > 0, \end{cases}$$

$$\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \gamma_k \mathbb{E}[\|\Pi_{\mathcal{X}}(x_k) - \bar{x}_k\|^2] + \frac{1}{3}C_{\beta,c,B_h}\mathbb{E}[\mathrm{dist}^2(x_k,\mathcal{X})]$$
$$\le (1-\gamma_k)\,\mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] + \frac{16}{\mu^2}\gamma_k^2 B^2 \qquad \forall k > k_0.$$

Proof One can easily see that our stepsize can be written equivalently as $\alpha_k = \min\left(\frac{1}{L}, \frac{8}{\mu(k+1)}\right)$. Since for this choice of the stepsize we have $\alpha_k \le \frac{1}{L}$, then $(2-\alpha_k L) \ge 1$ and using this in Lemma 5 combined with Assumption 2, we get:

$$\begin{aligned} \mathbb{E}[\|v_k - \bar{v}_k\|^2] &\le \mathbb{E}[\|x_k - \bar{x}_k\|^2] - \alpha_k \mathbb{E}[(F(x_k) - F(\bar{x}_k))] + \alpha_k^2 B^2 \\ &\overset{\text{Assumption 2}}{\le} \mathbb{E}[\|x_k - \bar{x}_k\|^2] - \frac{\mu\alpha_k}{2}\mathbb{E}[\|x_k - \bar{x}_k\|^2] + \alpha_k^2 B^2. \end{aligned} \tag{16}$$

From (12), we also have:

$$\begin{aligned} \mathbb{E}[\|x_k - \bar{x}_k\|^2] &\le \mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] - \beta(2-\beta)\frac{\mathbb{E}[\mathrm{dist}^2(v_{k-1},\mathcal{X})]}{cB_h^2} \\ &\overset{(14)}{\le} \mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] - \frac{\beta(2-\beta)}{cB_h^2}\left(1 - \frac{\beta(2-\beta)}{cB_h^2}\right)^{-1}\mathbb{E}[\mathrm{dist}^2(x_k,\mathcal{X})] \\ &= \mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] - C_{\beta,c,B_h}\mathbb{E}[\mathrm{dist}^2(x_k,\mathcal{X})]. \end{aligned} \tag{17}$$

For $k \le k_0$, we have $\alpha_k = \frac{1}{L}$ and combining (16) with (17), we obtain:

$$\begin{aligned} \mathbb{E}[\|v_k - \bar{v}_k\|^2] &\le \left(1 - \frac{\mu}{2L}\right)\mathbb{E}[\|x_k - \bar{x}_k\|^2] + \frac{B^2}{L^2} \\ &\le \max\left(\left(1 - \frac{\mu}{2L}\right)\mathbb{E}[\|x_k - \bar{x}_k\|^2] + \frac{B^2}{L^2}, \frac{B^2}{L^2}\right) \\ &\le \max\left(\left(1 - \frac{\mu}{2L}\right)\mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] + \frac{B^2}{L^2}, \frac{B^2}{L^2}\right). \end{aligned}$$

Using the geometric sum formula and recalling that $\theta_{L,\mu} = 1 - \mu/(2L)$, we obtain the first statement. Further, for $k > k_0$, from relation (16), we have:

$$\begin{aligned}\mathbb{E}[\|v_k - \bar{v}_k\|^2] &= \mathbb{E}[\|x_k - \bar{x}_k\|^2] - \left(\frac{\mu\alpha_k}{4} + \frac{\mu\alpha_k}{4}\right)\mathbb{E}[\|x_k - \bar{x}_k\|^2] + \alpha_k^2 B^2 \\ &= \left(1 - \frac{\mu\alpha_k}{4}\right)\mathbb{E}[\|x_k - \bar{x}_k\|^2] - \left(\frac{\mu\alpha_k}{4}\right)\mathbb{E}[\|x_k - \bar{x}_k\|^2] + \alpha_k^2 B^2.\end{aligned}$$

Further, if we take $v = x_k \in \mathbb{R}^n$ and $y = \bar{x}_k \in \mathcal{X}^* \subset \mathcal{X}$ in (8), we have:

$$\|\Pi_{\mathcal{X}}(x_k) - \bar{x}_k\|^2 \leq \|x_k - \bar{x}_k\|^2$$

and using this in the previous recurrence, we get:

$$\mathbb{E}[\|v_k - \bar{v}_k\|^2] \leq \left(1 - \frac{\mu\alpha_k}{4}\right)\mathbb{E}[\|x_k - \bar{x}_k\|^2] - \frac{\mu\alpha_k}{4}\mathbb{E}[\|\Pi_{\mathcal{X}}(x_k) - \bar{x}_k\|^2] + \alpha_k^2 B^2.$$

Combining this recurrence with (17) and using that $0 \leq 1 - \mu\alpha_k/4 \leq 1$ for all $k > k_0$, we further obtain:

$$\begin{aligned}&\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \frac{\mu\alpha_k}{4}\mathbb{E}[\|\Pi_{\mathcal{X}}(x_k) - \bar{x}_k\|^2] \\ &\leq \left(1 - \frac{\mu\alpha_k}{4}\right)\mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] - \left(1 - \frac{\mu\alpha_k}{4}\right)C_{\beta,c,B_h}\mathbb{E}[\text{dist}^2(x_k, \mathcal{X})] + \alpha_k^2 B^2.\end{aligned}$$

After rearranging the terms, we have:

$$\begin{aligned}\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \frac{\mu\alpha_k}{4}\mathbb{E}[\|\Pi_{\mathcal{X}}(x_k) - \bar{x}_k\|^2] + \left(1 - \frac{\mu\alpha_k}{4}\right)C_{\beta,c,B_h}\mathbb{E}[\text{dist}^2(x_k, \mathcal{X})] \\ \leq \left(1 - \frac{\mu\alpha_k}{4}\right)\mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] + \alpha_k^2 B^2.\end{aligned}$$

Since $k > k_0 = \lceil \frac{8L}{\mu} \rceil$ and $\alpha_k = \frac{4}{\mu}\gamma_k$, we get:

$$\begin{aligned}\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \gamma_k\mathbb{E}[\|\Pi_{\mathcal{X}}(x_k) - \bar{x}_k\|^2] + (1 - \gamma_k)\, C_{\beta,c,B_h}\mathbb{E}[\text{dist}^2(x_k, \mathcal{X})] \\ \leq (1 - \gamma_k)\,\mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] + \frac{16}{\mu^2}\gamma_k^2 B^2 \quad \forall k > k_0.\end{aligned}$$

Note that in this case $\gamma_k = 2/(k+1)$ is a decreasing sequence, an thus we have:

$$1 - \gamma_k = \frac{k-1}{k+1} \geq \frac{1}{3} \quad \forall k \geq 2.$$

Using this bound in the previous recurrence, we also get the second statement. ■

Now, we are ready to derive sublinear rates when we assume additionally a quadratic growth on the objective function. Let us define for $k \geq k_0 + 1$ the sum:

$$\bar{S}_k = \sum_{j=k_0+1}^{k} (j+1)^2 \sim \mathcal{O}(k^3 - k_0^3),$$

and the corresponding average sequences:

$$\hat{x}_k = \frac{\sum_{j=k_0+1}^{k}(j+1)^2 x_j}{\bar{S}_k}, \quad \hat{x}_k^* = \bar{S}_k^{-1}\sum_{j=k_0+1}^{k}(j+1)^2\bar{x}_j \in \mathcal{X}^*$$
$$\text{and} \quad \hat{w}_k = \frac{\sum_{j=k_0+1}^{k}(j+1)^2\Pi_{\mathcal{X}}(x_j)}{\bar{S}_k} \in \mathcal{X}.$$

Theorem 9 *Let $f(\cdot,\zeta), g(\cdot,\zeta)$ and $h(\cdot,\xi)$ be convex functions. Additionally, Assumptions 1–4 hold. Further, consider stepsize $\alpha_k = \min\left(\frac{1}{L}, \frac{8}{\mu(k+1)}\right)$ and $\beta \in (0,2)$. Then, for $k > k_0$, where $k_0 = \lceil \frac{8L}{\mu} \rceil$, we have the following sublinear convergence rates for the average sequence $\hat{x}_k$ in terms of optimality and feasibility violation for problem (1) (we keep only the dominant terms):*

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

Proof From Lemma 8, the following recurrence is valid for any $k > k_0$:

$$\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \gamma_k\mathbb{E}[\|\Pi_{\mathcal{X}}(x_k) - \bar{x}_k\|^2] + \frac{1}{3}C_{\beta,c,B_h}\mathbb{E}[\text{dist}^2(x_k,\mathcal{X})]$$
$$\leq (1-\gamma_k)\,\mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] + \frac{16}{\mu^2}\gamma_k^2 B^2.$$

From definition of $\gamma_k = \frac{2}{k+1}$ and multiplying the whole inequality with $(k+1)^2$, we get:

$$(k+1)^2\mathbb{E}[\|v_k - \bar{v}_k\|^2] + 2(k+1)\mathbb{E}[\|\Pi_{\mathcal{X}}(x_k) - \bar{x}_k\|^2] + \frac{C_{\beta,c,B_h}}{3}(k+1)^2\mathbb{E}[\text{dist}^2(x_k,\mathcal{X})]$$
$$\leq k^2\mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] + \frac{64}{\mu^2}B^2.$$

Summing this inequality from k_0+1 to k, we get:

$$(k+1)^2\mathbb{E}[\|v_k - \bar{v}_k\|^2] + 2\sum_{j=k_0+1}^{k}(j+1)\mathbb{E}[\|\Pi_{\mathcal{X}}(x_j) - \bar{x}_j\|^2]$$
$$+ \frac{C_{\beta,c,B_h}}{3}\sum_{j=k_0+1}^{k}(j+1)^2\mathbb{E}[\text{dist}^2(x_j,\mathcal{X})]$$
$$\leq (k_0+1)^2\mathbb{E}[\|v_{k_0} - \bar{v}_{k_0}\|^2] + \frac{64}{\mu^2}B^2(k-k_0).$$

By linearity of the expectation operator and using convexity of the norm, we get:

$$(k+1)^2\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \frac{2\bar{S}_k}{(k+1)}\mathbb{E}[\|\hat{w}_k - \hat{x}_k^*\|^2] + \frac{\bar{S}_k C_{\beta,c,B_h}}{3}\mathbb{E}[\|\hat{w}_k - \hat{x}_k\|^2]$$
$$\leq (k_0+1)^2\mathbb{E}[\|v_{k_0} - \bar{v}_{k_0}\|^2] + \frac{64}{\mu^2}B^2(k-k_0),$$

After some simple calculations and keeping only the dominant terms, we get the following:

$$\mathbb{E}[\|\hat{w}_k - \hat{x}_k^*\|^2] \leq \mathcal{O}\left(\frac{B^2}{\mu^2(k+1)}\right),$$
$$\mathbb{E}[\|\hat{w}_k - \hat{x}_k\|^2] \leq \mathcal{O}\left(\frac{B^2}{\mu^2 C_{\beta,c,B_h}(k^2+kk_0+k_0^2)}\right).$$

Since $\hat{w}_k \in \mathcal{X}$, we get the following convergence rate for the average sequence $\hat{x}_k$ in terms of feasibility violation:

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

Furthermore, since $\hat{x}_k^* \in \mathcal{X}^*$ and using the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we also get convergence rate for the average sequence $\hat{x}_k$ in terms of optimality:

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

This proves our statements. ■

Recall the expression of C_{β,c,B_h}:

$$C_{\beta,c,B_h} = \left(\frac{\beta(2-\beta)}{cB_h^2}\right)\left(1 - \frac{\beta(2-\beta)}{cB_h^2}\right)^{-1} = \left(\frac{\beta(2-\beta)}{cB_h^2 - \beta(2-\beta)}\right).$$

For the particular choice of the stepsize $\beta = 1$, we have:

$$C_{1,c,B_h} = \left(\frac{1}{cB_h^2 - 1}\right) > 0,$$

since we always have $cB_h^2 > 1$. Using this expression in the convergence rates of Theorem 9, we obtain:

$$\mathbb{E}\left[\text{dist}^2(\hat{x}_k, \mathcal{X})\right] \leq \mathcal{O}\left(\frac{B^2(cB_h^2-1)}{\mu^2(k^2+kk_0+k_0^2)}\right),$$
$$\mathbb{E}\left[\text{dist}^2(\hat{x}_k, \mathcal{X}^*)\right] \leq \mathcal{O}\left(\frac{B^2(cB_h^2-1)}{\mu^2(k^2+kk_0+k_0^2)} + \frac{B^2}{\mu^2(k+1)}\right).$$

We can easily notice from Lemma 8 that for $B = 0$ we can get better convergence rates. More specifically, for this particular case, taking constant stepsize, we get linear rates for the last iterate x_k in terms of optimality and feasibility violation. We state this result in the next corollary.

Corollary 10 *Under the assumptions of Theorem 9, with $B = 0$, the last iterate x_k generated by SSP algorithm with constant stepsize $\alpha_k \equiv \alpha < \min(1/L, 2/\mu)$ converges linearly in terms of optimality and feasibility violation.*

Proof When $B = 0$ and the stepsize satisfies $\alpha_k = \alpha < \min(1/L, 2/\mu)$, if we combine (16) with (17), we obtain:

$$\mathbb{E}[\|x_{k+1} - \bar{x}_{k+1}\|^2] + C_{\beta,c,B_h}\mathbb{E}[\text{dist}^2(x_{k+1}, \mathcal{X})] \leq \mathbb{E}[\|v_k - \bar{v}_k\|^2] \leq \left(1 - \frac{\mu\alpha}{2}\right)\mathbb{E}[\|x_k - \bar{x}_k\|^2].$$

Then, since $1 - \mu\alpha/2 \in (0, 1)$, we get immediately that:

$$\mathbb{E}[\|x_k - \bar{x}_k\|^2] \leq \left(1 - \frac{\mu\alpha}{2}\right)^k \|x_0 - \bar{x}_0\|^2,$$
$$C_{\beta,c,B_h}\mathbb{E}[\text{dist}^2(x_k, \mathcal{X})] \leq \left(1 - \frac{\mu\alpha}{2}\right)^k \|x_0 - \bar{x}_0\|^2,$$

which proves our statements. ■

Note that in (Necoara, 2021) it has been proved that stochastic first order methods are converging linearly on optimization problems of the form (1) without functional constraints and satisfying Assumption 1 with $B = 0$. This paper extends this result to a stochastic subgradient projection method on optimization problems with functional constraints (1) satisfying Assumption 1 with $B = 0$. To the best of our knowledge, these convergence rates are new for stochastic subgradient projection methods applied on the general class of optimization problems (1). In Section 4 we provide an example of an optimization problem with functional constraints, that is the constrained linear least-squares, which satisfies Assumption 1 with $B = 0$.

3.3 Convergence analysis: strongly quasi-convex objective function

In this section, we consider for our problem (1) that $g = 0$ and $f(\cdot, \zeta)$ are nonconvex functions. Hence, we have, $f(x) = \mathbb{E}[f(x, \zeta)]$, i.e. $F(x) = f(x)$. In these settings the update for v_k takes the form $v_k = x_k - \alpha_k \nabla f(x_k, \zeta_k)$. We also replace Assumption 2 with the following strongly quasi-convex condition (see Necoara et al. (2019)):

Assumption 2' *The function f satisfies a strongly quasi-convex condition, i.e. there exists non-negative constant $\mu > 0$ such that:*

$$f(\bar{x}) \geq f(x) + \langle \nabla f(x), \bar{x} - x\rangle + \frac{\mu}{2}\|\bar{x} - x\|^2 \quad \forall x \in \mathcal{Y}. \tag{18}$$

In (Necoara et al., 2019) it has been proved that (18) is a stronger condition than the one given in Assumption 2. In this case we can still derive a recurrence similar to Lemma 5, without requiring $f(\cdot, \zeta)$ to be convex functions.

Lemma 11 *Let Assumptions 1 and 2' hold and $g = 0$. Then, for any $k \geq 0$ and stepsize $\alpha_k > 0$, we have the following recursion:*

$$\mathbb{E}[\|v_k - \bar{v}_k\|^2] \leq (1 - \mu\alpha_k)\mathbb{E}[\|x_k - \bar{x}_k\|^2] - \alpha_k(2 - \alpha_k L)\,\mathbb{E}[f(x_k) - f(\bar{x}_k)] + \alpha_k^2 B^2. \tag{19}$$

Proof From the definition of v_k iteration, we have:

$$\begin{aligned}\|v_k - \bar{v}_k\|^2 &\leq \|v_k - \bar{x}_k\|^2 = \|x_k - \bar{x}_k - \alpha_k \nabla f(x_k, \zeta_k)\|^2 \\ &= \|x_k - \bar{x}_k\|^2 - 2\alpha_k\langle \nabla f(x_k, \zeta_k), x_k - \bar{x}_k\rangle + \alpha_k^2\|\nabla f(x_k, \zeta_k)\|^2.\end{aligned}$$

Since v_k depends on $\mathcal{F}_{[k-1]} \cup \{\zeta_k\}$, not on ξ_k, and the stepsize α_k does not depend on (ζ_k, ξ_k), then from basic properties of conditional expectation, we have:

$$\begin{aligned}
&\mathbb{E}_{\zeta_k}[\|v_k - \bar{v}_k\|^2 | \mathcal{F}_{[k-1]}] \\
&\leq \|x_k - \bar{x}_k\|^2 - 2\alpha_k \mathbb{E}_{\zeta_k}[\langle \nabla f(x_k, \zeta_k), x_k - \bar{x}_k \rangle | \mathcal{F}_{[k-1]}] + \alpha_k^2 \mathbb{E}_{\zeta_k}[\|\nabla f(x_k, \zeta_k)\|^2 | \mathcal{F}_{[k-1]}] \\
&= \|x_k - \bar{x}_k\|^2 - 2\alpha_k \langle \nabla f(x_k), x_k - \bar{x}_k \rangle + \alpha_k^2 \mathbb{E}_{\zeta_k}[\|\nabla f(x_k, \zeta_k)\|^2 | \mathcal{F}_{[k-1]}] \\
&\overset{(18)}{\leq} (1 - \mu\alpha_k)\|x_k - \bar{x}_k\|^2 - 2\alpha_k (f(x_k) - f(\bar{x}_k)) + \alpha_k^2 \mathbb{E}_{\zeta_k}[\|\nabla f(x_k, \zeta_k)\|^2 | \mathcal{F}_{[k-1]}] \\
&\overset{(2)}{\leq} (1 - \mu\alpha_k)\|x_k - \bar{x}_k\|^2 - 2\alpha_k \left(f(x_k) - f(\bar{x}_k)\right) + \alpha_k^2 \left(B^2 + L\left(f(x_k) - f(\bar{x}_k)\right)\right) \\
&= (1 - \mu\alpha_k)\|x_k - \bar{x}_k\|^2 - \alpha_k(2 - \alpha_k L)\left(f(x_k) - f(\bar{x}_k)\right) + \alpha_k^2 B^2.
\end{aligned}$$

Now, taking expectation w.r.t. $\mathcal{F}_{[k-1]}$, we get:

$$\mathbb{E}[\|v_k - \bar{v}_k\|^2] \leq (1 - \mu\alpha_k)\mathbb{E}[\|x_k - \bar{x}_k\|^2] - \alpha_k(2 - \alpha_k L)\mathbb{E}[f(x_k) - f(\bar{x}_k)] + \alpha_k^2 B^2,$$

which concludes our statement. ■

Lemma 12 *Let Assumptions 1, 2', 3 and 4 hold and $g = 0$. Define $k_0 = \lceil \frac{2L}{\mu} \rceil$, $\beta \in (0, 2)$, $\theta_{L,\mu} = 1 - \mu/L$ and $\alpha_k = \frac{\gamma_k}{\mu}$, where γ_k is given by:*

$$\gamma_k = \begin{cases} \frac{\mu}{L} & \text{if } k \leq k_0 \\ \frac{2}{k+1} & \text{if } k > k_0. \end{cases}$$

Then, the iterates of SSP algorithm satisfy the following recurrence:

$$\begin{aligned}
&\mathbb{E}[\|v_{k_0} - \bar{v}_{k_0}\|^2] \leq \begin{cases} \frac{B^2}{L^2} & \text{if } \theta_{L,\mu} \leq 0 \\ \theta_{L,\mu}^{k_0}\|v_0 - \bar{v}_0\|^2 + \frac{1-\theta_{L,\mu}^{k_0}}{1-\theta_{L,\mu}} \frac{B^2}{L^2} & \text{if } \theta_{L,\mu} > 0, \end{cases} \\
&\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \frac{\gamma_k}{\mu}\mathbb{E}[f(x_k) - f(\bar{x}_k)] + \frac{1}{3}C_{\beta,c,B_h}\mathbb{E}[\mathrm{dist}^2(x_k, \mathcal{X})] \\
&\leq (1 - \gamma_k)\,\mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] + \frac{\gamma_k^2}{\mu^2}B^2 \qquad \forall k > k_0.
\end{aligned}$$

Proof One can easily see that our stepsize can be written equivalently as $\alpha_k = \min\left(\frac{1}{L}, \frac{2}{\mu(k+1)}\right)$. Since for this choice of the stepsize we have $\alpha_k \leq \frac{1}{L}$, then $(2 - \alpha_k L) \geq 1$ and using this in Lemma 11, we get:

$$\mathbb{E}[\|v_k - \bar{v}_k\|^2] \leq (1 - \mu\alpha_k)\mathbb{E}[\|x_k - \bar{x}_k\|^2] - \alpha_k\,\mathbb{E}[f(x_k) - f(\bar{x}_k)] + \alpha_k^2 B^2.$$

For $k \leq k_0$, proceeding in the same manner as in Lemma 8 we get the statement with $\theta_{L,\mu} = 1 - \mu/L$. For $k > k_0$, combining the last inequality with (17) and using $0 \leq 1 - \mu\alpha_k \leq 1$ for all $k > k_0$, we get:

$$\begin{aligned}
&\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \alpha_k\mathbb{E}[f(x_k) - f(\bar{x}_k)] + (1 - \mu\alpha_k)\,C_{\beta,c,B_h}\mathbb{E}[\mathrm{dist}^2(x_k, \mathcal{X})] \\
&\qquad \leq (1 - \mu\alpha_k)\,\mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] + \alpha_k^2 B^2.
\end{aligned}$$

Finally, using the same reasoning as in Lemma 8, we also get the second statement. ■

Now, we are ready to derive convergence rates also for the case when the individual functions $f(\cdot,\zeta)$ are nonconvex, but f is still convex. Recall that $\bar{S}_k$, $\hat{x}_k$, $\hat{x}_k^*$ and $\hat{w}_k$ have been already introduced in Section 3.2.

Theorem 13 *Let Assumptions 1, 2', 3 and 4 hold, $g = 0$ and f be convex. Further, consider stepsizes $\alpha_k = \min\left(\frac{1}{L}, \frac{2}{\mu(k+1)}\right)$ and $\beta \in (0,2)$. Then, for $k > k_0$, where $k_0 = \lceil \frac{2L}{\mu} \rceil$, we have the following sublinear convergence rates for the average sequence $\hat{x}_k$ in terms of optimality and feasibility violation (keeping only the dominant terms):*

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

Proof From Lemma 12, the following recurrence is valid for any $k > k_0$:

$$\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \frac{\gamma_k}{\mu}\mathbb{E}[f(x_k) - f(\bar{x}_k)] + \frac{1}{3}C_{\beta,c,B_h}\mathbb{E}[\mathrm{dist}^2(x_k,\mathcal{X})]$$
$$\leq (1-\gamma_k)\,\mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] + \frac{\gamma_k^2}{\mu^2}B^2 \qquad \forall k > k_0.$$

From definition of $\gamma_k = \frac{2}{k+1}$, we further get:

$$\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \frac{2}{\mu(k+1)}\mathbb{E}[f(x_k) - f(\bar{x}_k)] + \frac{1}{3}C_{\beta,c,B_h}\mathbb{E}[\mathrm{dist}^2(x_k,\mathcal{X})]$$
$$\leq \left(\frac{k-1}{k+1}\right)\mathbb{E}[\|v_{k-1} - \bar{v}_{k-1}\|^2] + \frac{4}{\mu^2}\frac{1}{(k+1)^2}B^2.$$

Multiplying the whole inequality with $(k+1)^2$, and summing the inequality from k_0+1 to k, we get:

$$(k+1)^2\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \frac{2}{\mu}\sum_{j=k_0+1}^{k}(j+1)\mathbb{E}[f(x_k) - f(\bar{x}_k)]$$
$$+ \frac{C_{\beta,c,B_h}}{3}\sum_{j=k_0+1}^{k}(j+1)^2\mathbb{E}[\mathrm{dist}^2(x_j,\mathcal{X})]$$
$$\leq (k_0+1)^2\mathbb{E}[\|v_{k_0} - \bar{v}_{k_0}\|^2] + \frac{4}{\mu^2}B^2(k-k_0).$$

By linearity of the expectation operator and using convexity of f and of the norm, we get:

$$(k+1)^2\mathbb{E}[\|v_k - \bar{v}_k\|^2] + \frac{2\bar{S}_k}{\mu(k+1)}\mathbb{E}[f(\hat{x}_k) - f(\hat{x}_k^*)] + \frac{\bar{S}_k C_{\beta,c,B_h}}{3}\mathbb{E}[\|\hat{w}_k - \hat{x}_k\|^2]$$
$$\leq (k_0+1)^2\mathbb{E}[\|v_{k_0} - \bar{v}_{k_0}\|^2] + \frac{4}{\mu^2}B^2(k-k_0).$$

After some simple calculations and keeping only the dominant terms, we get the bounds:

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

Since $\hat{w}_k \in \mathcal{X}$, we get the following convergence rate for the average sequence $\hat{x}_k$ in terms of feasibility violation:

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

This proves our statements. ■

Note that the convergence rate in optimality from Theorem 13 is in function values, while in Theorem 9 is in terms of distance to the optimal solution set. This is due to the fact that we replace Assumption 2 with Assumption 2' and we do not impose convexity on the individual functions $f(\cdot, \zeta)$.

4. Stochastic subgradient for constrained least-squares

In this section we consider the problem of finding a solution to a system of linear equalities and inequalities, see also equation (11) in (Leventhal and Lewis, 2010):

$$\text{find } x \in \mathcal{Y} : \ Ax = b, \ Cx \leq d, \tag{20}$$

where $A \in \mathbb{R}^{m \times n}$, $C \in \mathbb{R}^{p \times n}$ and $\mathcal{Y}$ is a simple polyhedral set. We assume that this system is consistent, i.e. it has at least one solution. This problem can be reformulated equivalently as a particular case of the optimization problem with functional constraints (1):

$$\begin{aligned} \min_{x \in \mathcal{Y}} \ & f(x) \quad \left(:= \frac{1}{2}\mathbb{E}\left[\|A_\zeta^T x - b_\zeta\|^2\right]\right) \\ \text{subject to } & C_\xi^T x - d_\xi \leq 0 \quad \forall \xi \in \Omega_2, \end{aligned} \tag{21}$$

where A_ζ^T and C_ξ^T are (block) rows partitions of matrices A and C, respectively. Clearly, problem (21) is a particular case of problem (1), with $f(x, \zeta) = \frac{1}{2}\|A_\zeta^T x - b_\zeta\|^2$, $g(x, \zeta) = 0$, and $h(x, \xi) = C_\xi^T x - d_\xi$ (provided that C_ξ is a row of C). Let us define the polyhedral subset partitions $\mathcal{C}_\xi = \{x \in \mathbb{R}^n : C_\xi^T x - d_\xi \leq 0\}$ and $\mathcal{A}_\zeta = \{x \in \mathbb{R}^n : A_\zeta^T x - b_\zeta = 0\}$. In this case the feasible set is the polyhedron $\mathcal{X} = \{x \in \mathcal{Y} : \ Cx \leq d\}$ and the optimal set is the polyhedron:

$$\mathcal{X}^* = \{x \in \mathcal{Y} : \ Ax = b, \ Cx \leq d\} = \mathcal{Y} \cap_{\zeta \in \Omega_1} \mathcal{A}_\zeta \cap_{\xi \in \Omega_2} \mathcal{C}_\xi.$$

Note that for the particular problem (21) Assumption 1 holds with $B = 0$ and e.g. $L = 2\max_\zeta \|A_\zeta\|^2$, since $f^* = 0$ and we have:

$$\mathbb{E}[\|\nabla f(x, \zeta)\|^2] = \mathbb{E}[\|A_\zeta(A_\zeta^T x - b_\zeta)\|^2] \leq (2\max_\zeta \|A_\zeta\|^2)\left(\frac{1}{2}\mathbb{E}\left[\|A_\zeta^T x - b_\zeta\|^2\right]\right) = L\, f(x).$$

It is also obvious that Assumption 3 holds, since the functional constraints are linear. Moreover, for the constrained least-squares problem, we replace Assumptions 2 and 4 with the well-known Hoffman property of a polyhedral set, see Example 3 from Section 2 and also (Pena et al., 2021; Leventhal and Lewis, 2010; Necoara et al., 2019):

$$\operatorname{dist}^2(u, \mathcal{X}^*) \leq c \cdot \mathbb{E}\left[\operatorname{dist}^2(u, \mathcal{A}_\zeta) + \operatorname{dist}^2(u, \mathcal{C}_\xi)\right] \quad \forall u \in \mathcal{Y}, \tag{22}$$

for some $c \in (0, \infty)$. Recall that the Hoffman condition (22) always holds for *nonempty* polyhedral sets (Pena et al., 2021). For the constrained least-squares problem the SSP algorithm becomes:

Algorithm 2 (SSP-LS):
Choose $x_0 \in \mathcal{Y}$, stepsizes $\alpha_k > 0$ and $\beta \in (0, 2)$
For $k \geq 0$ repeat:

Sample independently $\zeta_k \sim \mathbf{P}_1$ and $\xi_k \sim \mathbf{P}_2$ and update:
$v_k = x_k - \alpha_k A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})$
$z_k = (1-\beta)v_k + \beta \Pi_{\mathcal{C}_{\xi_k}}(v_k)$
$x_{k+1} = \Pi_{\mathcal{Y}}(z_k)$.

Note that the update for v_k can be written as step (5) in SSP for $f(x, \zeta) = \frac{1}{2}\|A_\zeta^T x - b_\zeta\|^2$ and $g = 0$. In contrast to the previous section however, here we consider an adaptive stepsize:

$$\alpha_k = \delta \frac{\|A_{\zeta_k}^T x_k - b_{\zeta_k}\|^2}{\|A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2}, \text{ where } \delta \in (0, 2).$$

Note that when C_ξ is a row of C, then z_k has the explicit expression:

$$z_k = v_k - \beta \frac{(C_{\xi_k}^T v_k - d_{\xi_k})_+}{\|C_{\xi_k}\|^2} C_{\xi_k},$$

which coincides with step (6) in SSP for $h(x, \xi) = C_\xi^T x - d_\xi$. Note that we can use e.g., probability distributions dependent on the (block) rows of matrices A and C:

$$\mathbf{P}_1(\zeta = \zeta_k) = \frac{\|A_{\zeta_k}\|_F^2}{\|A\|_F^2} \text{ and } \mathbf{P}_2(\xi = \xi_k) = \frac{\|C_{\xi_k}\|_F^2}{\|C\|_F^2},$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Note that our algorithm SSP-LS is different from Algorithm 4.6 in (Leventhal and Lewis, 2010) through the choice of the stepsize α_k, of the sampling rules and of the update law for x_k and it is more general as it allows to work with block of rows of the matrices A and C. Moreover, SSP-LS includes the classical Kaczmarz's method when solving linear systems of equalities. In the next section we derive linear convergence rates for SSP-LS algorithm, provided that the system of equalities and inequalities is consistent, i.e. $\mathcal{X}^*$ is nonempty.

4.1 Linear convergence

In this section we prove linear convergence for the sequence generated by the SSP-LS algorithm for solving the constrained least-squares problem (21). Let us define *maximum block condition number* over all the submatrices A_ζ:

$$\kappa_{\text{block}} = \max_{\zeta \sim \mathbf{P}_1} \|A_\zeta^T\| \cdot \|(A_\zeta^T)^\dagger\|,$$

where $(A_\zeta^T)^\dagger$ denotes the pseudoinverse of A_ζ^T. Note that if A_ζ^T has full rank, then $(A_\zeta^T)^\dagger = A_\zeta (A_\zeta^T A_\zeta)^{-1}$. Then, we have the following result.

Theorem 14 *Assume that the polyhedral set $\mathcal{X}^* = \{x \in \mathcal{Y} : \ Ax = b, \ Cx \leq d\}$ is nonempty. Then, we have the following linear rate of convergence for the sequence x_k generated by the SSP-LS algorithm:*

$$\mathbb{E}\left[\text{dist}^2(x_k, \mathcal{X}^*)\right] \leq \left(1 - \frac{1}{c} \min\left(\frac{\delta(2-\delta)}{2\kappa_{block}^2}, \frac{2-\delta}{4\delta}, \frac{\beta(2-\beta)}{2}\right)\right)^k \text{dist}^2(x_0, \mathcal{X}^*).$$

Proof From the updates of the sequences x_{k+1}, z_k and v_k in SSP-LS algorithm, we have:

$$\begin{aligned}
&\|x_{k+1} - \bar{x}_{k+1}\|^2 \leq \|x_{k+1} - \bar{x}_k\|^2 = \|\Pi_{\mathcal{Y}}(z_k) - \Pi_{\mathcal{Y}}(\bar{x}_k)\|^2 \leq \|z_k - \bar{x}_k\|^2 \\
&= \|v_k - \bar{x}_k + \beta(\Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k)\|^2 \\
&= \|v_k - \bar{x}_k\|^2 + \beta^2 \|\Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k\|^2 + 2\beta \langle v_k - \bar{x}_k, \Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k \rangle \\
&= \|x_k - \bar{x}_k\|^2 + \alpha_k^2 \|A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2 - 2\alpha_k \langle x_k - \bar{x}_k, A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k}) \rangle \\
&\quad + \beta^2 \|\Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k\|^2 + 2\beta \langle v_k - \bar{x}_k, \Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k \rangle.
\end{aligned}$$

Using the definition of α_k and that $A_{\zeta_k}^T(x_k - \bar{x}_k) = A_{\zeta_k}^T x_k - b_{\zeta_k}$, we further get:

$$\begin{aligned}
\|x_{k+1} - \bar{x}_{k+1}\|^2 &\leq \|x_k - \bar{x}_k\|^2 - \delta(2-\delta) \frac{\|A_{\zeta_k}^T x_k - b_{\zeta_k}\|^4}{\|A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2} + \beta^2 \|\Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k\|^2 \\
&\quad + 2\beta \langle v_k - \bar{x}_k, \Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k \rangle \\
&= \|x_k - \bar{x}_k\|^2 - \delta(2-\delta) \frac{\|A_{\zeta_k}^T x_k - b_{\zeta_k}\|^4}{\|A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2} + \beta^2 \|\Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k\|^2 \\
&\quad - 2\beta \langle \Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k, \Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k \rangle + 2\beta \langle \Pi_{\mathcal{C}_{\xi_k}}(v_k) - \bar{x}_k, \Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k \rangle.
\end{aligned}$$

From the optimality condition of the projection we always have $\langle \Pi_{\mathcal{C}_{\xi_k}}(v_k) - z, \Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k \rangle \leq 0$ for all $z \in \mathcal{C}_{\xi_k}$. Taking $z = \bar{x}_k \in \mathcal{X}^* \subseteq \mathcal{C}_{\xi_k}$ in the previous relation, we finally get:

$$\begin{aligned}
&\|x_{k+1} - \bar{x}_{k+1}\|^2 \\
&\leq \|x_k - \bar{x}_k\|^2 - \delta(2-\delta) \frac{\|A_{\zeta_k}^T x_k - b_{\zeta_k}\|^4}{\|A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2} - \beta(2-\beta) \|\Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k\|^2.
\end{aligned} \tag{23}$$

From the definition of v_k and α_k, we have:

$$\begin{aligned} v_k = x_k - \alpha_k A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k}) &\iff \alpha_k^2 \|A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2 = \|v_k - x_k\|^2 \\ &\iff \frac{\|A_{\zeta_k}^T x_k - b_{\zeta_k}\|^4}{\|A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2} = \frac{1}{\delta^2}\|v_k - x_k\|^2. \end{aligned}$$

Also, from the definition of z_k, we have:

$$\|\Pi_{\mathcal{C}_{\xi_k}}(v_k) - v_k\|^2 = \frac{1}{\beta^2}\|z_k - v_k\|^2. \tag{24}$$

Now, replacing these two relations in (23), we get:

$$\begin{aligned} &\|x_{k+1} - \bar{x}_{k+1}\|^2 \\ &\quad \le \|x_k - \bar{x}_k\|^2 - \frac{\delta(2-\delta)}{\delta^2}\|v_k - x_k\|^2 - \frac{\beta(2-\beta)}{\beta^2}\|z_k - v_k\|^2 \\ &\quad \le \|x_k - \bar{x}_k\|^2 - \frac{\delta(2-\delta)}{2\kappa_{\text{block}}^2}\frac{\kappa_{\text{block}}^2}{\delta^2}\|v_k - x_k\|^2 \\ &\qquad - \min\left(\frac{\delta(2-\delta)}{4\delta^2}, \frac{\beta(2-\beta)}{2}\right)\left(2\|v_k - x_k\|^2 + \frac{2}{\beta^2}\|z_k - v_k\|^2\right). \end{aligned} \tag{25}$$

First, let us consider the subset $\mathcal{C}_{\xi_k}$. Then, we have:

$$\begin{aligned} \text{dist}^2(x_k, \mathcal{C}_{\xi_k}) = \|x_k - \Pi_{\mathcal{C}_{\xi_k}}(x_k)\|^2 &\le \|x_k - \Pi_{\mathcal{C}_{\xi_k}}(v_k)\|^2 \le 2\|x_k - v_k\|^2 + 2\|v_k - \Pi_{\mathcal{C}_{\xi_k}}(v_k)\|^2 \\ &\overset{(24)}{\le} 2\|x_k - v_k\|^2 + \frac{2}{\beta^2}\|v_k - z_k\|^2. \end{aligned}$$

Second, let us consider the subset $\mathcal{A}_{\zeta_k}$. Since the corresponding $A_{\zeta_k}^T$ represents a block of rows of matrix A, the update for v_k in SSP-LS can be written as:

$$v_k = (1-\delta)x_k + \delta T_{\zeta_k}(x_k),$$

where the operator T_{ζ_k} is given by the following expression

$$T_{\zeta_k}(x_k) = x_k - \frac{\|A_{\zeta_k}^T x_k - b_{\zeta_k}\|^2}{\|A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2} A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k}).$$

Further, the projection of x_k onto the subset $\mathcal{A}_{\zeta_k}$ is, see e.g., (Horn and Johnson, 2012):

$$\Pi_{\mathcal{A}_{\zeta_k}}(x_k) = x_k - (A_\zeta^T)^\dagger (A_{\zeta_k}^T x_k - b_{\zeta_k}).$$

Hence, we have:

$$\begin{aligned}\text{dist}^2(x_k, \mathcal{A}_{\zeta_k}) &= \|x_k - \Pi_{\mathcal{A}_{\zeta_k}}(x_k)\|^2 = \|(A_\zeta^T)^\dagger (A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2 \\ &\leq \|(A_\zeta^T)^\dagger\|^2 \|A_{\zeta_k}^T x_k - b_{\zeta_k}\|^2 \\ &= \|(A_\zeta^T)^\dagger\|^2 \frac{\|A_{\zeta_k}^T x_k - b_{\zeta_k}\|^2}{\|A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2} \|A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2 \\ &\leq \|(A_{\zeta_k}^T)^\dagger\|^2 \|A_{\zeta_k}^T\|^2 \frac{\|A_{\zeta_k}^T x_k - b_{\zeta_k}\|^4}{\|A_{\zeta_k}(A_{\zeta_k}^T x_k - b_{\zeta_k})\|^2} \\ &= \kappa_{\text{block}}^2 \|T_{\zeta_k}(x_k) - x_k\|^2 \\ &= \frac{\kappa_{\text{block}}^2}{\delta^2} \|x_k - v_k\|^2.\end{aligned}$$

Using these two relations in (25), we finally get the following recurrence:

$$\begin{aligned}&\|x_{k+1} - \bar{x}_{k+1}\|^2 \\ &\leq \|x_k - \bar{x}_k\|^2 - \min\left(\frac{\delta(2-\delta)}{2\kappa_{\text{block}}^2}, \frac{2-\delta}{4\delta}, \frac{\beta(2-\beta)}{2}\right)\left(\text{dist}^2(x_k, \mathcal{A}_{\zeta_k}) + \text{dist}^2(x_k, \mathcal{C}_{\xi_k})\right).\end{aligned} \tag{26}$$

Now, taking conditional expectation w.r.t. $\mathcal{F}_{[k-1]}$ in (26) and using Hoffman inequality (22), we obtain:

$$\begin{aligned}&\mathbb{E}_{\zeta_k, \xi_k}[\|x_{k+1} - \bar{x}_{k+1}\|^2 | \mathcal{F}_{[k-1]}] \\ &\leq \|x_k - \bar{x}_k\|^2 - \frac{1}{c}\min\left(\frac{\delta(2-\delta)}{2\kappa_{\text{block}}^2}, \frac{2-\delta}{4\delta}, \frac{\beta(2-\beta)}{2}\right) \text{dist}^2(x_k, \mathcal{X}^*) \\ &= \left(1 - \frac{1}{c}\min\left(\frac{\delta(2-\delta)}{2\kappa_{\text{block}}^2}, \frac{2-\delta}{4\delta}, \frac{\beta(2-\beta)}{2}\right)\right) \|x_k - \bar{x}_k\|^2.\end{aligned}$$

Finally, taking full expectation, recursively we get the statement of the theorem. ■

Note that for $\delta = \beta = 1$ and A_ζ^T a single row of matrix A, we have $\kappa_{\text{block}} = 1$ and we get a simplified estimate for linear convergence $\mathbb{E}\left[\text{dist}^2(x_k, \mathcal{X}^*)\right] \leq (1 - 1/(4c))^k \text{dist}^2(x_0, \mathcal{X}^*)$, which is similar to convergence estimate for the algorithm in (Leventhal and Lewis, 2010). In the block case, for $\delta = \beta = 1$ and assuming that $\kappa_{\text{block}} \geq 2$, we get the linear convergence $\mathbb{E}\left[\text{dist}^2(x_k, \mathcal{X}^*)\right] \leq \left(1 - 1/(2c\kappa_{\text{block}}^2)\right)^k \text{dist}^2(x_0, \mathcal{X}^*)$, i.e., our rate depends explicitly on the geometric properties of the submatrices A_ζ and C_ζ (recall that both constants c and κ_{block} are defined in terms of these submatrices). From our best knowledge, this is the first time when such convergence bounds are obtained for a stochastic subgradient type algorithm solving constrained least-squares.

5. Illustrative examples and numerical tests

In this section, we present several applications where our algorithm can be applied, such as the robust sparse SVM classification problem (Bhattacharyya et al., 2004), sparse SVM classification problem (Weston et al., 2003), constrained least-squares and linear programs (Tibshirani, 2011), accompanied by detailed numerical simulations. The codes were written in Matlab and run on a PC with i7 CPU at 2.1 GHz and 16 GB memory.

5.1 Robust sparse SVM classifier

We consider a two class dataset $\{(z_i, y_i)\}_{i=1}^N$, where z_i is the vector of features and $y_i \in \{-1, 1\}$ is the corresponding label. A robust classifier is a hyperplane parameterized by a weight vector w and an offset from the origin d, in which the decision boundary and set of relevant features are resilient to uncertainty in the data, see equation (2) in (Bhattacharyya et al., 2004) for more details. Then, the robust sparse classification problem can be formulated as:

$$\begin{aligned} &\min_{w,d,u} \ \lambda \sum_{i=1}^N u_i + \|w\|_1 \\ &\text{subject to :} \ \ y_i(w^T \bar{z}_i + d) \geq 1 - u_i \quad \forall \bar{z}_i \in \mathcal{Z}_i, \ \ u_i \geq 0 \quad \forall i = 1:N, \end{aligned}$$

where $\mathcal{Z}_i$ is the uncertainty set in the data z_i, the parameter $\lambda > 0$ and 1-norm is added in the objective to induce sparsity in w. To find a hyperplane that is robust and generalized well, each $\mathcal{Z}_i$ would need to be specified by a large corpus of pseudopoints. In particular, finding a robust hyperplane can be simplified by considering a data uncertainty model in the form of ellipsoids. In this case we can convert infinite number of linear constraints into a single nonlinear constraint and thus recasting the above set of robust linear inequalities as second order cone constraints. Specifically, if the uncertainty set for ith data is defined by an ellipsoid with the center z_i and the shape given by the positive semidefinite matrix $Q_i \succeq 0$, i.e. $\mathcal{Z}_i = \{\bar{z}_i : \langle Q_i(\bar{z}_i - z_i), \bar{z}_i - z_i\rangle \leq 1\}$, then a solution to the robust hyperplane classification problem is one in which the hyperplane in (w, d) does not intersect any ellipsoidal data uncertainty model (see Appendix 1 for a proof):

$$y_i(w^T z_i + d) \geq \|Q_i^{-1/2} w\| + 1 - u_i. \tag{27}$$

Hence, the robust classification problem can be recast as a convex optimization problem with many functional constraints that are either linear or second order cone constraints:

$$\begin{aligned} &\min_{w,d,u} \ \lambda \sum_{i=1}^N u_i + \|w\|_1 \\ &\text{subject to :} \ \ y_i(w^T z_i + d) \geq 1 - u_i, \quad u_i \geq 0 \quad \forall i = 1:N \\ &\qquad\qquad y_i(w^T z_i + d) \geq \|Q_i^{-1/2} w\| + 1 - u_i \ \ \forall i = 1:N. \end{aligned}$$

This is a particular form of problem (1) and thus we can solve it using our algorithm SSP. Since every one of the N data points has its own covariance matrix Q_i, this formulation results in a large optimization problem so it is necessary to impose some restrictions on the shape of these matrices. Hence, we consider two scenarios: (i) class-dependent covariance matrices, i.e., $Q_i = Q_+$ if $y_i = +1$ or $Q_i = Q_-$ if $y_i = -1$; (ii) class-independent covariance matrix, i.e., $Q_i = Q_\pm$ for all y_i. For more details on the choice of covariance matrices see (Bhattacharyya et al., 2004). Here, each covariance matrix is assumed to be diagonal. In a class-dependent diagonal covariance matrix, the diagonal elements of Q_+ or Q_- are unique, while in class-independent covariance matrix, all diagonal elements of $Q_\pm$ are identical. Computational experiments designed to evaluate the performance of SSP require datasets

in which the level of variability associated with the data can be quantified. Here, a noise level parameter, $0 \leq \rho \leq 1$, is introduced to scale each diagonal element of the covariance matrix, i.e., ρQ_+ or ρQ_- or $\rho Q_\pm$. When $\rho = 0$, data points are associated with no noise (the nominal case). The ρ value acts as a proxy for data variability. For classifying a point we consider the following rules. The "ordinary rule" for classifying a data point z is as follows: if $w_*^T z + d_* > 0$, then z is assigned to the $+1$ class; if $w_*^T z + d_* < 0$, then z is identified with the -1 class. An ordinary error occurs when the class predicted by the hyperplane differs from the known class of the data point. The "worst case rule" determines whether an ellipsoid with center z intersects the hyperplane. Hence, some allowable values of z will be classified incorrectly if $|w_*^T z + d_*| < \|Q_i^{-1/2} w_*\|^2$ (worst case error).

Table 1: Comparison between nominal and robust classifiers on training data: class-dependent covariance matrix (first half), class-independent covariance matrix (second half).

λ	ρ	**Robust**			**Nominal**	
		$w_* \neq 0$	worst case error	ordinary error	$w_* \neq 0$	ordinary error
0.1	0.01	1699	3	332	546	198
0.2		406	25	116	767	113
0.3		698	14	111	774	67
0.1	0.3	1581	63	331	546	198
0.2		1935	86	326	767	113
0.3		1963	77	311	774	67
0.1	0.01	1734	0	331	546	198
0.2		1822	1	268	767	113
0.3		1937	0	266	774	67
0.1	0.3	1629	19	316	546	198
0.2		1899	19	296	767	113
0.3		2050	20	282	774	67

Tables 1 and 2 give the results of our algorithm SSP for robust ($\rho > 0$) and nominal ($\rho = 0$) classification formulations. We choose the parameters $\lambda = 0.1, 0.2, 0.3$, $\beta = 1.96$, and stopping criterion 10^{-2}. We consider a dataset of CT scan images having two classes, covid and non-covid, available at `https://www.kaggle.com/plameneduardo/sarscov2-ctscan-dataset`. This dataset contains CT scan images of dimension ranging from 190×190 to 410×386 pixels. To implement our algorithm we have taken 1488 data in which 751 are of Covid patients and 737 of Non-Covid patients. Then, we divide them into training data and testing data. For training data we have taken 1240 images in which 626 are of Covid and 614 are of Non-Covid. For testing data we have taken 248 images in which 125 are of Covid and 123 of Non-Covid. We also resize all images into 190×190 pixels. First half of the Tables 1 and 2 correspond to feature-dependent covariance matrix and the second half to feature-independent covariance matrix. Table 1 shows the results for the training data and Table 2 for the testing data. As one can see from Tables 1 and 2, the robust classifier yields better accuracies on both training and testing datasets.

Table 2: Comparison between nominal and robust classifiers on testing data: class-dependent covariance matrix (first half), class-independent covariance matrix (second half).

λ	ρ	**Robust**	**Nominal**
		accuracy (ordinary rule)	accuracy (ordinary rule)
0.1	0.01	180, 72.5%	166, 66.9%
0.2		200, 80.6%	198, 79.8%
0.3		198, 79.8%	193, 77.9%
0.1	0.3	180, 72.5%	168, 67.8%
0.2		200, 80.6%	174, 70.2%
0.3		198, 79.8%	172, 69.4%
0.1	0.01	180, 72.5%	167, 67.4%
0.2		200, 80.6%	196, 79.1%
0.3		198, 79.8%	193, 77.9%
0.1	0.3	180, 72.5%	165, 66.6%
0.2		200, 80.6%	183, 73.8%
0.3		198, 79.8%	159, 64.1%

5.2 Constrained least-squares

Next, we consider constrained least-squares problem (21). We compare the performance of our algorithm SSP-LS and the algorithm in (Leventhal and Lewis, 2010) on synthetic data matrices A and C generated from a normal distribution. Both algorithms were stopped when $\max(\|Ax-b\|, \|(Cx-d)_+\|) \leq 10^{-3}$. The results for different sizes of matrices A and C are given in Table 3. One can easily see from Table 3 the superior performance of our algorithm in both, number of full iterations (epochs, i.e. number of passes through data) and cpu time (in seconds).

5.3 Linear programs

Next, we consider solving linear programs (LP) of the form:

$$\min_{\mathbf{z}\geq 0} \mathbf{c}^T\mathbf{z} \quad \text{subject to} \quad \mathbf{C}\mathbf{z} \leq \mathbf{d}.$$

Using the primal-dual formulation this problem is equivalent to (20):

$$\text{find } \mathbf{z} \in [0,\infty)^n, \nu \in [0,\infty)^p : \mathbf{c}^T\mathbf{z} + \mathbf{d}^T\nu = 0,\ \mathbf{C}\mathbf{z} \leq \mathbf{d},\ \mathbf{C}^T\nu + \mathbf{c} \geq 0.$$

Therefore, we can easily identify in (20):

$$x = \begin{bmatrix}\mathbf{z}\\ \nu\end{bmatrix} \in \mathcal{Y} = [0,\infty)^{n+p},\ A = [\mathbf{c}^T\ \mathbf{d}^T] \in \mathbb{R}^{n+p},\ C = \begin{bmatrix}\mathbf{C} & 0_{p\times p}\\ 0_{n\times n} & -\mathbf{C}^T\end{bmatrix}.$$

Hence, we can use our algorithm SSP-LS to solve LPs. In Table 4 we compare the performance of our algorithm, the algorithm in (Leventhal and Lewis, 2010) and Matlab solver *lsqlin* for solving the least-squares formulation of LPs taken from the Netlib library available on `https://www.netlib.org/lp/data/index.html`, and Matlab format LP library available on `https://users.clas.ufl.edu/hager/coap/Pages/matlabpage.html`. The first

Table 3: Comparison between SSP-LS and algorithm in (Leventhal and Lewis, 2010) in terms of epochs and cpu time (sec) on random least-squares problems.

$\delta = \beta$	m	p	n	SSP-LS		(Leventhal and Lewis, 2010)	
				epochs	cpu time (s)	epochs	cpu time (s)
0.96	900	900	10^3	**755**	**26.0**	817	29.9
1.96	900	900	10^3	**591**	**20.1**	787	26.2
0.96	900	1100	10^3	**624**	**23.2**	721	23.9
1.96	900	1100	10^3	**424**	**16.7**	778	27.3
0.96	9000	9000	10^4	**1688**	**5272.0**	1700	5778.1
1.96	9000	9000	10^4	**1028**	**3469.2**	1662	5763.7
0.96	9000	11000	10^4	**1224**	**5716.0**	1437	5984.8
1.96	9000	11000	10^4	**685**	**2693.7**	1461	5575.3
0.96	900	10^5	10^3	**5**	**105.0**	9	163.9
1.96	900	10^5	10^3	**4**	**77.7**	9	163.9
0.96	9000	10^5	10^4	**64**	**2054.5**	214	5698.5
1.96	9000	10^5	10^4	**42**	**1306.5**	213	5156.7
0.96	9000	9000	10^5	**23**	**1820.3**	23	1829.4
0.96	9000	11000	10^5	**21**	**2158.7**	23	2193.1
1.96	9000	11000	10^5	**19**	**1939.7**	23	2216.9
0.96	900	900	10^5	**14**	**65.7**	17	86.8
1.96	900	1100	10^5	**13**	**74.3**	15	75.6

two algorithms were stopped when $\max(\|Ax - b\|, \|(Cx - d)_+\|) \leq 10^{-3}$ and we choose $\delta = \beta = 1.96$. In Table 4, in the first column after the name of the LP we provide the dimension of the matrix $\mathbf{C}$. From Table 4 we observe that SSP-LS is always better than the algorithm in (Leventhal and Lewis, 2010) and for large dimensions it also better than *lsqlin*, a Matlab solver specially dedicated for solving constrained least-squares problems. Moreover, for "qap15" dataset *lsqlin* yields out of memory.

5.4 Sparse linear SVM

Finally, we consider the sparse linear SVM classification problem:

$$\begin{aligned} &\min_{w,d,u} \ \lambda \sum_{i=1}^{n} u_i + \|w\|_1 \\ &\text{subject to}: \quad y_i(w^T z_i + d) \geq 1 - u_i, \quad u_i \geq 0 \quad \forall i = 1:N. \end{aligned}$$

This can be easily recast as an LP and consequently as a least-squares problem. In Table 5 we report the results provided by our algorithms SSP-LS for solving sparse linear SVM problems. Here, the ordinary error has the same meaning as in Section 5.1. We use several datasets: covid dataset from `www.kaggle.com/plameneduardo/sarscov2-ctscan-dataset`; PMU-UD, sobar-72 and divorce datasets available on `https://archive-beta.ics.uci.edu/ml/datasets`; and the rest from LIBSVM library `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`. In the first column, the first argument represents the

Table 4: Comparison between SSP-LS, algorithm in (Leventhal and Lewis, 2010) and Matlab solver lsqlin in terms of epochs and cpu time (sec) on real data LPs.

LP	SSP-LS		(Leventhal and Lewis, 2010)		lsqlin
	epochs	time (s)	epochs	time (s)	time (s)
afiro (21×51)	**1163**	**1.9**	5943	2.7	0.09
beaconfd (173×295)	**1234**	**9.9**	9213	63.5	1.0
kb2 (43×68)	**10**	**0.02**	17	0.03	0.14
sc50a (50×78)	**9**	**0.04**	879	1.9	0.14
sc50b (50×78)	**25**	**0.1**	411	0.8	0.2
share2b (96×162)	**332**	**1.8**	1691	84.9	0.2
degen2 (444×757)	**4702**	**380.6**	5872	440.6	9.8
fffff800 (524×1028)	**44**	**5.5**	80	9.3	3.4
israel (174×316)	**526**	**5.4**	3729	312.9	0.3
lpi bgdbg1 (348×649)	**476**	**15.4**	9717	263.1	0.5
osa 07 (1118×25067)	**148**	**3169.8**	631	7169.7	3437.1
qap15 (6330×22275)	**70**	**2700.7**	373	7794.6	*
fit2p (3000×13525)	**7**	**65.7**	458	2188.3	824.9
maros r7 (3137×9408)	**635**	**2346.8**	1671	3816.2	3868.8
qap12 (3193×8856)	**54**	**374.1**	339	1081.8	3998.4

number of features and the second argument represents the number of data. We divided each dataset into 80% for training and 20% for testing. For the LP formulation we use the simple observation that any scalar u can be written as $u = u_+ - u_-$, with $u_+, u_- \geq 0$. We use the same stopping criterion as in Section 5.3. In Table 5 we provide the number of relevant features, i.e., the number of nonzero elements of the optimal w_*, and the number of misclassified data (ordinary error) on real training and testing datasets.

Appendix

1. *Proof of inequality* (27). One can easily see that the robust linear inequality

$$y_i(w^T\bar{z}_i + d) \geq 1 - u_i \quad \forall \bar{z}_i \in \mathcal{Z}_i$$

over the ellipsoid with the center in z_i

$$\mathcal{Z}_i = \{\bar{z}_i : \langle Q_i(\bar{z}_i - z_i), \bar{z}_i - z_i\rangle \leq 1\},$$

can be written as optimization problem whose minimum value must satisfy:

$$\begin{aligned} 1 - u_i \leq \min_{\bar{z}_i} \; & y_i(w^T\bar{z}_i + d) \\ & \text{subject to: } (\bar{z}_i - z_i)^T Q_i(\bar{z}_i - z_i) - 1 \leq 0. \end{aligned}$$

The corresponding dual problem is as follows:

$$\max_{\lambda \geq 0} \min_{\bar{z}_i} y_i(w^T\bar{z}_i + d) + \lambda((\bar{z}_i - z_i)^T Q_i(\bar{z}_i - z_i) - 1). \tag{28}$$

Table 5: Performance of sparse linear SVM classifier: ordinary error and sparsity of w_* on real training and testing datasets.

Dataset	λ	$w_* \neq 0$	ordinary error (train)	ordinary error (test)
sobar-72 (38×72)	0.1	7/38	9/58	3/14
	0.5	12/38	1/58	2/14
breastcancer (18×683)	0.1	9/18	12/547	13/136
	0.5	9/18	17/547	7/136
divorce (108×170)	0.1	23/108	3/136	0/34
	0.5	13/108	0/136	1/34
caesarian (10×80)	0.1	0/10(NA)	*	*
	0.5	5/10	14/64	9/16
cryotherapy (12×90)	0.1	3/12	8/72	5/18
	0.5	6/12	6/72	1/18
PMU-UD (19200×1051)	0.1	13/19200	0/841	70/210
	0.5	37/19200	0/841	47/210
Covid (20000×2481)	0.1	428/20000	146/1985	101/496
	0.5	752/20000	142/1985	97/496
Nomao (16×34465)	0.1	12/14	3462/27572	856/6893
	0.5	11/14	3564/27572	885/6893
Training (60×11055)	0.1	31/60	630/8844	152/2211
	0.5	30/60	611/8844	159/2211
leukemia (14258×38)	0.1	182/14258	9/31	2/7
	0.5	271/14258	9/31	2/7
mushrooms (42×8124)	0.1	30/42	426/6500	121/1624
	0.5	31/42	415/6500	105/1624
ijcnn1 (28×49990)	0.1	11/28	3831/39992	1021/9998
	0.5	10/28	3860/39992	992/9998
phishing (60×11055)	0.1	56/60	2953/8844	728/2211
	0.5	51/60	2929/8844	725/2211

Minimizing the above problem with respect to $\bar{z}_i$, we obtain:

$$y_i w + 2\lambda Q_i(\bar{z}_i - z_i) = 0 \iff \bar{z}_i = z_i - \frac{y_i}{2\lambda} Q_i^{-1} w. \tag{29}$$

By replacing this value of $\bar{z}_i$ into the dual problem (28), we get:

$$\max_{\lambda \geq 0} \left[-\frac{y_i^2}{4\lambda} w^T Q_i^{-1} w + y_i(w^T z_i + d) - \lambda \right].$$

For the dual optimal solution

$$\lambda^* = \frac{1}{2}\sqrt{y_i^2 (w^T Q_i^{-1} w)} = \frac{1}{2}\sqrt{(w^T Q_i^{-1} w)},$$

we get the primal optimal solution

$$\bar{z}_i^* = z_i - \frac{y_i Q_i^{-1} w}{\sqrt{w^T Q_i^{-1} w}},$$

and consequently the second order cone condition

$$y_i \left(w^T z_i + d \right) \geq 1 + \| Q_i^{-1/2} w \| - u_i.$$

Acknowledgments

The research leading to these results has received funding from: NO Grants 2014–2021, under project ELO-Hyp, contract no. 24/2020; UEFISCDI PN-III-P4-PCE-2021-0720, under project L2O-MOC, nr. 70/2022.

References

H. Bauschke and J. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Review, 38(3): 367–376, 1996.

D.P. Bertsekas, *Incremental proximal methods for large scale convex optimization*, Mathematical Programming, 129(2): 163–195, 2011.

P. Bianchi, W. Hachem and A. Salim, *A constant step forward-backward algorithm involving random maximal monotone operators*, Journal of Convex Analysis, 26(2): 397–436, 2019.

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.

O. Devolder, F. Glineur and Yu. Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Mathematical Programming, 146: 37–75, 2014.

J. Duchi and Y. Singer, *Efficient online and batch learning using forward backward splitting*, Journal of Machine Learning Research, 10: 2899–2934, 2009.

M. Hardt, B. Recht and Y. Singer, *Train faster, generalize better: stability of stochastic gradient descent*, International Conference on Machine Learning, 2016.

N. Hermer, D.R. Luke and A. Sturm, *Random function iterations for stochastic fixed point problems*, arXiv:2007.06479, 2020.

R. A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 2012.

A. Kundu, F. Bach and C. Bhattacharya, *Convex optimization over inter-section of simple sets: improved convergence rate guarantees via an exact penalty approach*, International Conference on Artificial Intelligence and Statistics, 2018.

A. Lewis and J. Pang, *Error bounds for convex inequality systems*, Generalized Convexity, Generalized Monotonicity (J. Crouzeix, J. Martinez-Legaz and M. Volle eds.), Cambridge University Press, 75–110, 1998.

D. Leventhal and A.S. Lewis. *Randomized Methods for linear constraints: convergence rates and conditioning*, Mathematics of Operations Research, 35(3): 641–654, 2010.

E. Moulines and F. Bach, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, Advances in Neural Information Processing Systems Conf., 2011.

B.S. Mordukhovich and N.M. Nam, *Subgradient of distance functions with applications to Lipschitzian stability*, Mathematical Programming, 104: 635–668, 2005.

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

I. Necoara, Yu. Nesterov and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Mathematical Programming, 175(1): 69–107, 2019.

V. Nedelcu, I. Necoara and Q. Tran Dinh, *Computational complexity of inexact gradient augmented Lagrangian methods: application to constrained MPC*, SIAM Journal on Control and Optimization, 52(5): 3109–3134, 2014.

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

A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal Optimization, 19(4): 1574–1609, 2009.

Yu. Nesterov, *Lectures on Convex Optimization*, Springer Optimization and Its Applications, 137, 2018.

A. Nedich, *Random algorithms for convex minimization problems*, Mathematical Programming, 129(2): 225–273, 2011.

A. Nedich and I. Necoara, *Random minibatch subgradient algorithms for convex problems with functional constraints*, Applied Mathematics and Optimization, 8(3): 801–833, 2019.

A. Patrascu and I. Necoara, *On the convergence of inexact projection first order methods for convex minimization*, IEEE Transactions Automatic Control, 63(10): 3317–3329, 2018.

A. Patrascu and I. Necoara, *Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization*, Journal of Machine Learning Research, 18(198): 1–42, 2018.

J. Pena, J. Vera and L. Zuluaga, *New characterizations of Hoffman constants for systems of linear constraints*, Mathematical Programming, 187: 79–109, 2021.

B.T. Polyak, *Minimization of unsmooth functionals* , USSR Computational Mathematics and Mathematical Physics, 9 (3), 14-29, 1969.

B.T. Polyak, *Random algorithms for solving convex inequalities*, Studies in Computational Mathematics, 8: 409–422, 2001.

H. Robbins and S. Monro, *A Stochastic approximation method*, The Annals of Mathematical Statistics, 22(3): 400–407, 1951.

R.T. Rockafellar and S.P. Uryasev, *Optimization of conditional value-at-risk*, Journal of Risk, 2: 21–41, 2000.

L. Rosasco, S. Villa and B.C. Vu, *Convergence of stochastic proximal gradient algorithm*, Applied Mathematics and Optimization, 82: 891–917 , 2019.

J. Rasch and A. Chambolle, *Inexact first-order primal–dual algorithms*, Computational Optimization and Applications, 76: 381—430, 2020.

R. Tibshirani, *The solution path of the generalized lasso*, Phd Thesis, Stanford Univ., 2011.

Q. Tran-Dinh, O. Fercoq, and V. Cevher, *A smooth primal-dual optimization framework for nonsmooth composite convex minimization*, SIAM Journal on Optimization, 28(1): 96–134, 2018.

S. Villa, L. Rosasco, S. Mosci and A. Verri, *Proximal methods for the latent group lasso penalty*, Computational Optimization and Applications, 58: 381–407, 2014.

V. Vapnik, *Statistical learning theory*, John Wiley, 1998.

J. Weston, A. Elisseeff and B. Scholkopf, *Use of the zero norm with linear models and kernel methods*, Journal of Machine Learning Research, 3: 1439–1461, 2003.

M. Wang, Y. Chen, J. Liu and Y. Gu, *Random multiconstraint projection: stochastic gradient methods for convex optimization with many constraints*, arXiv: 1511.03760, 2015.

Y. Xu, *Primal-dual stochastic gradient method for convex programs with many functional constraints*, SIAM Journal on Optimization, 30(2): 1664–1692, 2020.

T. Yang and Q. Lin, *RSG: Beating subgradient method without smoothness and strong convexity*, Journal of Machine Learning Research, 19(6): 1–33, 2018.