Journal of Machine Learning Research 18 (2018) 1-42 Submitted 6/17; Revised 11/17; Published 4/18

Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization

Andrei Patrascu ANDREI.PATRASCU@FMI.UNIBUC.RO
Department of Computer Science
University of Bucharest
Str. Academiei 14, 010014 Bucharest

Ion Necoara ION.NECOARA@ACSE.PUB.RO
Automatic Control and Systems Engineering Department
University Politehnica of Bucharest
Spl. Independentei 313, 060042 Bucharest

Editor: Mark Schmidt

Abstract

A popular approach for solving stochastic optimization problems is the stochastic gradient descent (SGD) method. Although the SGD iteration is computationally cheap and its practical performance may be satisfactory under certain circumstances, there is recent evidence of its convergence difficulties and instability for unappropriate choice of parameters. To avoid some of the drawbacks of SGD, stochastic proximal point (SPP) algorithms have been recently considered. We introduce a new variant of the SPP method for solving stochastic convex problems subject to (in)finite intersection of constraints satisfying a linear regularity condition. For the newly introduced SPP scheme we prove new nonasymptotic convergence results. In particular, for convex Lipschitz continuous objective functions, we prove nonasymptotic convergence rates in terms of the expected value function gap of order $\mathcal{O}\left(\frac{1}{k^{1/2}}\right)$, where k is the iteration counter. We also derive better nonasymptotic convergence rates in terms of expected quadratic distance from the iterates to the optimal solution for smooth strongly convex objective functions, which in the best case is of order $\mathcal{O}\left(\frac{1}{k}\right)$. Since these convergence rates can be attained by our SPP algorithm only under some natural restrictions on the stepsize, we also introduce a restarting variant of SPP that overcomes these difficulties and derive the corresponding nonasymptotic convergence rates. Numerical evidence supports the effectiveness of our methods in real problems.

Keywords: Stochastic convex optimization, intersection of convex constraints, stochastic proximal point, nonasymptotic convergence analysis, rates of convergence.

1. Introduction

The randomness in most of the practical optimization applications led the stochastic optimization field to become an essential tool for many applied mathematics areas, such as machine learning (Polyak and Juditsky, 1992), distributed optimization (Necoara et al., 2011), control (Karimi and Kammer, 2017), and sensor networks problems (Blatt and Hero, 2006). Since the randomness usually enters the problem through the cost function and/or the constraints set, in this paper we approach both randomness sources and consider stochastic objective functions subject to stochastic constraints. Usually, in the literature, the following

unconstrained stochastic model has been considered:

$$\min_{x\in\mathbb{R}^n} F(x) = \left(\mathbb{E}[f(x;S)]\right), \tag{1}$$

where the expectation is taken w.r.t. the random variable S. In the following subsections, we recall some popular numerical optimization algorithms for solving the previous unconstrained stochastic optimization problem and set the context for our contributions.

1.1 Previous work

A very popular approach for solving the unconstrained stochastic problem (1) is the stochastic gradient method (SGD) (Nemirovski et al., 2009; Moulines and Bach, 2011; Rosasco et al., 2014; Polyak and Juditsky, 1992). At each iteration k, the SGD algorithm randomly samples S and takes a step along the gradient of the chosen individual function:

$$x^{k+1} = x^k - \mu_k \nabla f(x^k; S_k),$$

where μ_k is a positive stepsize. Convergence behavior of SGD for the last iterate sequence has been analyzed in (Nemirovski et al., 2009) and for the average of the iterates sequence has been given in (Polyak and Juditsky, 1992). However, there is a recent nonasymptotic convergence analysis of SGD provided in (Moulines and Bach, 2011), under various differentiability assumptions on the objective function. While the SGD scheme is the method of choice in practice for many machine learning applications due to its superior empirical performance, the theoretical estimates obtained in (Moulines and Bach, 2011) highlights several difficulties regarding its practical limitations and robustness. For example, the stepsize is highly constrained to small values by an exponential term from the convergence rate which could be catastrophically increased by uncontrolled variations of the stepsize. More precisely, the convergence rates of SGD with decreasing stepsize $\mu_k = \frac{\mu_0}{k}$, given for the quadratic mean $\{\mathbb{E}[\|x^k - x^*\|^2]\}_{k\geq 0}$, where x^* is the optimal solution of (1), contains certain exponential terms (depending on the initial stepsize) of the following form (Moulines and Bach, 2011):

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \frac{C_1 e^{C_2\mu_0^2}}{k^{\alpha\mu_0}} + \mathcal{O}\left(\frac{1}{k}\right), \tag{2}$$

for $\mu_0 > 2/\alpha$ and for appropriate positive constants C_1, C_2 and α. Note that this convergence rate holds under strong convexity and gradient Lipschitz assumptions on the objective function F. From (2) we observe that $\{\mathbb{E}[\|x^k - x^*\|^2]\}_{k\geq 0}$ can grow exponentially until the stepsizes becomes sufficiently small, a behavior which can be also observed in practical simulations.

Since these drawbacks are naturally introduced by the SGD iteration, other essential modifications of this scheme have been applied for avoiding the issues. One resulted method is the stochastic proximal point (SPP) algorithm for solving the unconstrained stochastic problem (1) having the following iteration (Ryu and Boyd, 2016; Toulis et al., 2016; Bianchi, 2016):

$$x^{k+1} = \arg\min_{z\in\mathbb{R}^n} \left[f(z;S_k) + \frac{1}{2\mu_k}\|z - x^k\|^2 \right].$$

Note that SGD represents a particular SPP iteration applied to the linearization of $f(z;S_k)$ in x^k, that is to the linear function $l_f(z;x^k,S_k) = f(x^k;S_k) + \langle \nabla f(x^k;S_k), z - x^k\rangle$. Of course, when f has an easily computable proximal operator, it is natural to use f instead of its linearization l_f. In (Ryu and Boyd, 2016), the SPP algorithm has been applied to problems with the objective function having Lipschitz continuous gradient and the following *restricted strong convexity* property:

$$f(x;S) \geq f(y;S) + \langle \nabla f(y;S), x-y\rangle + \frac{1}{2}\langle M_S(x-y), x-y\rangle \quad \forall x,y \in \mathbb{R}^n, \tag{3}$$

for some matrix $M_S \succeq 0$, satisfying $\lambda = \lambda_{\min}(\mathbb{E}[M_S]) > 0$. In (Ryu and Boyd, 2016) the asymptotic global convergence of SPP with decreasing stepsize $\mu_k = \frac{\mu_0}{k}$ is derived, followed by a nonasymptotic analysis for the SPP with constant stepsize. In particular, it has been proven that SPP converges linearly to a noise-dominated region around the optimal solution. Moreover, the following *asymptotic* (i.e. for a *sufficiently large* k) convergence rate in the quadratic mean have been given:

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \left(\frac{1}{e}\right)^{\mu_0\lambda \ln(k+1)} C_1 + \begin{cases} \frac{C_2}{(\mu_0\lambda - 1)k} & \text{if} \quad \mu_0\lambda > 1 \\ \frac{C_2 \ln(k)}{k} & \text{if} \quad \mu_0\lambda = 1 \\ \frac{C_2}{(1-\mu_0\lambda)k^{\mu_0\lambda}} & \text{if} \quad \mu_0\lambda < 1, \end{cases}$$

where C_1 and C_2 are some positive constants. With the essential difference that no exponential terms depending on μ_0 are encountered, these rates of convergence have similar orders with those for the variable stepsize SGD method. Although in this paper we make similar assumptions on the objective function, we additionally assume the presence of *convex constraints* and provide a *nonasymptotic* convergence analysis of the SPP for a more general stepsize $\mu_k = \frac{\mu_0}{k^\gamma}$, with $\gamma > 0$. Moreover, the Moreau smoothing framework used in our paper leads to more elegant and intuitive proofs. Another paper related to the SPP algorithm is (Toulis et al., 2016), where the considered stochastic model involves minimization of the expectation of random particular components $f(x;S)$ defined by the composition of a smooth function and a linear operator, i.e.:

$$f(x;S) = f(a_S^T x),$$

where $a_S \in \mathbb{R}^n$. Moreover, the objective function $F(x) = \mathbb{E}[f(a_S^T x)]$ needs to satisfy $\lambda_{\min}\left(\nabla^2 F(x)\right) \geq \lambda > 0$ for all $x \in \mathbb{R}^n$. The nonasymptotic convergence of the SPP with decreasing stepsize $\mu_k = \frac{\mu_0}{k^\gamma}$, with $\gamma \in (1/2, 1]$, has been analyzed in the quadratic mean and the following convergence rate has been derived in (Toulis et al., 2016):

$$\mathbb{E}[\|x^k - x^*\|^2] \leq C\left(\frac{1}{1+\lambda\mu_0\alpha}\right)^{k^{1-\gamma}} + \mathcal{O}\left(\frac{1}{k^\gamma}\right),$$

where C and α are some positive constants. However, the analysis used in (Toulis et al., 2016) cannot be extended to general convex objective functions and complicated constraints, since it is essential in the proofs that each component of the objective function has the form $f(a_S^T x)$, where $a_S \in \mathbb{R}^n$. In our paper we consider general convex objective functions, which lack the previously discussed structure, with (in)finite number of convex constraints.

Further, in (Bianchi, 2016) a general asymptotic convergence analysis of several variants of SPP scheme within operator theory settings has been provided, under mild convexity assumptions. A particular optimization model instance analyzed in (Bianchi, 2016), related to our paper, is:

$$\min_x \ f(x) \quad \text{s.t. } \ x \in \cap_{i=1}^m X_i,$$

for which has been derived the following SPP type algorithm:

$$x^{k+1} = \begin{cases} \arg\min_{z\in\mathbb{R}^n} \left[f(z) + \frac{1}{2\mu_k}\|z - x^k\|^2 \right] & \text{if } S_k = 0 \\ \Pi_{X_{S_k}}(x^k) & \text{otherwise,} \end{cases}$$

where S_k is randomly chosen in $\Omega = \{0, 1, \cdots, m\}$ according to a probability distribution $\mathbb{P}$. Although this scheme is very similar to the SPP algorithm, only the almost sure asymptotic convergence has been provided in (Bianchi, 2016). Convergence results of order $\mathcal{O}\left(\frac{1}{k}\right)$ in the strongly convex case, as well as almost sure convergence results under weaker assumptions, are also provided in (Rosasco et al., 2017) for the stochastic proximal gradient algorithm on convex composite optimization problems. In (Combettes and Pesquet, 2016) the asymptotic behavior of a stochastic forward-backward splitting algorithm for finding a zero of the sum of a maximally monotone set-valued operator and a co-coercive operator in Hilbert spaces is investigated. Weak and strong almost sure convergence properties of the iterates are established under mild conditions on the underlying stochastic processes.

A particular case of the stochastic optimization problem (1) is the discrete stochastic model, where the random variable S is discrete and thus, usually the objective function is given as a finite sum of functional components. There exists a large amount of work in the literature on deterministic and randomized algorithms for the finite sum optimization problem. Linear convergence of SGD for solving convex feasibility problems is proven recently in (Necoara, 2017). Convergence analysis of SGD for minimizing an objective function subject to a finite number of convex constraints is provided e.g. in (Necoara, 2017; Nedic, 2011). Linear convergence results on a restarted variant of SGD for finite-sum problems is given in (Yang and Lin, 2016). On the deterministic side, the cyclic incremental gradient methods were extensively analyzed e.g. in (Bertsekas, 2011). Recently, highly efficient algorithms with improved convergence estimates (compared to SGD) for finite sums have been developed using aggregated (averaged) or variance reduction techniques. The first category is based on the common idea of updating the current iterate along the aggregated (averaged) gradient step: e.g. incremental aggregated gradient (IAG) (Vanli et al., 2017), stochastic averaged gradient (SAG) (Roux et al., 2012) and its generalization SAGA (Defazio et al., 2014). Regarding the second category, there are simpler schemes, but memory intensive, such as stochastic variance reduced gradient (SVRG) method introduced in (Johnson and Zhang, 2013). It has been proved that all these schemes can achieve linear convergence under strong convexity and gradient Lipschitz assumptions on the finite sum objective function. Similar optimal performances on finite sum minimization, as for the previous two classes of algorithms, are obtained also for the stochastic dual coordinate ascent (SDCA) method, which has been analyzed in (Shalev-Shwartz and Zhang, 2013).

Other stochastic proximal (gradient) schemes together with their theoretical guarantees are studied in several recent papers as we further exemplify. In (Atchade et al., 2014) a perturbed proximal gradient method is considered for solving composite optimization

problems, where the gradient is intractable and approximated by Monte Carlo methods. Conditions on the stepsize and the Monte Carlo batch size are derived under which the convergence is guaranteed. Two classes of stochastic approximation strategies (stochastic iterative Tikhonov regularization and the stochastic iterative proximal point) are analyzed in (Koshal et al., 2013) for monotone stochastic variational inequalities and almost sure convergence results are presented. A new stochastic optimization method is analyzed in (Yurtsever et al., 2016) for the minimization of the sum of three convex functions, one of which has Lipschitz continuous gradient and satisfies a restricted strong convexity condition. In (Xu, 2011) a finite sample analysis for the averaged SGD is provided, which shows that it usually takes a huge number of samples for averaged SGD to reach its asymptotic region, for improperly chosen learning rate (stepsize). Moreover, simple strategies to properly set the learning rate are derived in the same paper so that it takes a reasonable amount of data for averaged SGD to reach its asymptotic region. In (Niu et al., 2011) it is shown through a novel theoretical analysis that SGD can be implemented in a parallel fashion without any locking. Moreover, for sparse optimization problems (meaning that the most gradient updates only modify small parts of the decision variable) the developed scheme achieves a nearly optimal rate of convergence. A regularized stochastic version of the BFGS method is proposed in (Mokhtari and Ribeiro, 2014) to solve convex optimization problems. Convergence analysis shows that lower and upper bounds on the Hessian eigenvalues of the sample functions are sufficient to guarantee convergence of order $\mathcal{O}\left(\frac{1}{k}\right)$. A comprehensive survey on modern optimization algorithms for machine learning problems is given recently in (Bottou et al., 2016). Based on experience, theoretical results are presented on a straightforward, yet versatile SGD algorithm, its practical behavior is discussed, and opportunities are highlighted for designing new algorithms with improved performance.

1.2 Contributions

In this paper we consider both randomness sources (i.e. objective function and constraints) and thus our problem of interest involves stochastic objective functions subject to (in)finite intersection of constraints. Given the clear superior features of SPP algorithm over the classical SGD scheme, we consider the SPP scheme for solving our problem of interest. The main contributions of this paper are:

(i) More general stochastic optimization model and a new stochastic proximal point algorithm: While most of the existing papers from the stochastic optimization literature consider convex models without constraints or simple (easy projection onto) constraints, in this paper we consider stochastic convex optimization problems subject to (in)finite intersection of constraints satisfying a linear regularity type condition. It turns out that many practical applications, including those from machine learning, fits into this framework: e.g. classification, regression, finite sum minimization, portfolio optimization, convex feasibility, optimal control problems. For this general stochastic optimization model we introduce a new stochastic proximal point (SPP) algorithm. It is worth to mention that although the analysis of an SPP method for stochastic models with complicated constraints is nontrivial and does not follow from the analysis corresponding to the unconstrained setting, our framework allows us to deal with even an infinite number of constraints. To the best of

our knowledge, our SPP method is the first stochastic proximal point algorithm that can tackle optimization problems with complicated constraints.

(ii) New nonasymptotic convergence results for the SPP method: For the newly introduced SPP scheme we prove new nonasymptotic convergence results. In particular, for convex and Lipschitz continuous objective functions, we prove nonasymptotic estimates for the rate of convergence of the SPP scheme in terms of the expected value function gap and feasibility violation of order $\mathcal{O}\left(\frac{1}{k^{1/2}}\right)$, where k is the iteration counter. We also derive better nonasymptotic bounds for the rate of convergence of SPP scheme with decreasing stepsize $\mu_k = \frac{\mu_0}{k^\gamma}$, with $\gamma \in (0, 1]$, for smooth strongly convex objective functions. For this case the convergence rates are given in terms of expected quadratic distance from the iterates to the optimal solution and are of order:

$$\mathbb{E}[\|x^k - x^*\|^2] \leq C \left(\mathbb{E}\left[\frac{1}{1 + \bar{\alpha}_S \mu_0} \right] \right)^{k^{1-\gamma}} + \mathcal{O}\left(\frac{1}{k^\gamma} \right),$$

where C and $\bar{\alpha}_S$ are appropriate nonnegative constants. Note that the derived rates of convergence do not contain any exponential term in μ_0, as it is the case for the SGD scheme, which makes SPP more robust than SGD even in the constrained case. This can be also observed in numerical simulations, see Section 7 below.

(iii) Restarted variant of SPP algorithm and the corresponding convergence analysis: Since the best complexity of our basic SPP scheme can be attained only under some natural restrictions on the initial stepsize μ_0, we also introduce a restarting stochastic proximal point algorithm that overcomes these difficulties. The main advantage of this restarted variant of SPP algorithm is that it is parameter-free and thus it is easily implementable in practice. Under strong convexity and smoothness assumptions on the objective function, for $\gamma > 0$ and epoch counter t, the restarting SPP scheme with the constant stepsize (per epoch) $\frac{1}{t^\gamma}$ provides a nonasymptotic complexity of order $\mathcal{O}\left(\frac{1}{\epsilon^{1+\frac{1}{\gamma}}}\right)$.

Paper outline. The paper is organized as follows. In Section 2 the problem of interest is formulated and analyzed. Further in Section 3, a new stochastic proximal point algorithm is introduced and its relations with the previous work are highlighted. We provide in Section 4 the first main result of this paper regarding the nonasymptotic convergence of SPP in the convex case. Further, stronger convergence results are presented in Section 5 for smooth strongly convex objective functions. In order to improve the convergence of the simple SPP scheme, in Section 6 we introduce a restarted variant of SPP algorithm. In Section 7 we provide some preliminary numerical simulations to highlight the empirical performance of our schemes. Some long proofs are moved in the Appendix.

Notations. We consider the space $\mathbb{R}^n$ composed by column vectors. For $x, y \in \mathbb{R}^n$ denote the scalar product $\langle x, y \rangle = x^T y$ and Euclidean norm by $\|x\| = \sqrt{x^T x}$. The projection operator onto the nonempty closed convex set X is denoted by $\Pi_X(\cdot)$ and the distance from a given x to the set X is denoted by $\text{dist}_X(x) = \min_{z \in X} \|x - z\|$. Given any convex set X, the function $\text{dist}_X(\cdot)$ is convex and the squared distance function $\text{dist}_X^2(\cdot)$ has Lipschitz gradient with constant 1. For some function f, we denote by $\partial f(x)$ the subdifferential set

at x. We also use the following definition of the indicator function of a set X:

$$\mathbb{I}_X(x) = \begin{cases} 0, & \text{if } x \in X \\ \infty, & \text{otherwise.} \end{cases}$$

Finally, we define the function $\varphi_\alpha : (0, \infty) \to \mathbb{R}$ as:

$$\varphi_\alpha(x) = \begin{cases} (x^\alpha - 1)/\alpha, & \text{if } \alpha \neq 0 \\ \log(x), & \text{if } \alpha = 0. \end{cases}$$

2. Problem formulation

In many machine learning applications randomness usually enters the problem through the cost function and/or the constraint set. Minimization of problems having complicating constraints can be very challenging. This is usually alleviated by approximating the feasible set by an (in)finite intersection of simple sets (Necoara, 2017; Necoara et al., 2017; Nedic, 2011). Therefore, in this paper we tackle the following stochastic convex constrained optimization problem:

$$\begin{aligned} F^* = \min_{x \in \mathbb{R}^n} & F(x) \quad (:= \mathbb{E}[f(x;S)]) \\ \text{s.t.} \quad & x \in X \quad (:= \cap_{S \in \Omega} X_S)\,, \end{aligned} \tag{4}$$

where $f(\cdot; S) : \mathbb{R}^n \to \mathbb{R}$ are convex functions with full domain $\mathrm{dom} f = \mathbb{R}^n$, X_S are nonempty closed convex sets, and S is a random variable with its associated probability space $(\Omega, \mathbb{P})$. Notice that this formulation allows us to include (in)finite number of constraints. We denote the set of optimal solutions with X^* and x^* any optimal point for (4). For the optimization problem (4) we make the following assumptions.

Assumption 1 *For any $S \in \Omega$, the function $f(\cdot; S)$ is proper, closed, convex and Lipschitz continuous, that is there exists $L_{f,S} > 0$ such that*

$$|f(x;S) - f(y;S)| \leq L_{f,S}\|x - y\| \qquad \forall x, y \in \mathbb{R}^n.$$

Notice that Assumption 1 implies that any subgradient $g_f(x;S) \in \partial f(x;S)$ is bounded, that is $\|g_f(x;S)\| \leq L_{f,S}$ for all $x \in \mathbb{R}^n$ and $S \in \Omega$. For the sets we assume:

Assumption 2 *Given $S \in \Omega$, the following two properties hold:*
(i) X_S are simple convex sets (i.e. projections onto these sets are easy).
(ii) There exists $\zeta > 0$ such that the feasible set X satisfies linear regularity:

$$dist_X^2(x) \ \leq \zeta\, \mathbb{E}[dist_{X_S}^2(x)] \qquad \forall x \in \mathbb{R}^n.$$

Assumption 2 *(ii)* is known in the literature as the *linear regularity property* and it is essential for proving linear convergence for (alternating) projection algorithms, see (Necoara, 2017; Necoara et al., 2017; Nedic, 2011). For example, when X_S are hyperplanes, halfspaces or when X has nonempty interior, then the linear regularity property holds. In particular, if the set X contains a ball of radius $\bar{r}$ and X is contained in a ball of radius $\bar{R}$, then the

ratio $\bar{R}/\bar{r}$ can be taken as the linear regularity constant ζ (Necoara et al., 2017). The linear regularity property is related to the relaxation of strong convexity, the so-called quadratic functional growth condition for an objective function, for smooth convex optimization introduced in (Necoara et al., 2017). In (Necoara et al., 2017) it has been proved that several first order methods converge linearly under functional growth condition and smoothness of the objective function.
Notice that this general optimization model (4) covers a long range of applications from various fields, such as optimization, machine learning, statistics, control, which we discuss in more details below.

2.1 Convex feasibility problem

Let us consider the following objective function and constraints (Necoara, 2017):

$$f(x;S) := \frac{\lambda}{2}\|x\|^2 \quad \forall S \in \Omega \quad \text{and} \quad X = \cap_{S\in\Omega} X_S,$$

where $\lambda > 0$. Then, we obtain the least norm convex feasibility problem:

$$\min_{x\in\mathbb{R}^n} \frac{\lambda}{2}\|x\|^2 \quad \text{s.t. } x \in \cap_{S\in\Omega} X_S.$$

We can also consider another reformulation of the least norm convex feasibility problem:

$$f(x;S) := \frac{\lambda_S}{2}\|x\|^2 + \mathbb{I}_{X_S}(x) \quad \forall S \in \Omega,$$

where $\lambda_S \geq 0$ and $\mathbb{E}[\lambda_S] = \lambda$. Then, this leads to the stochastic optimization model:

$$\min_{x\in\mathbb{R}^n} \mathbb{E}\left[\frac{\lambda_S}{2}\|x\|^2 + \mathbb{I}_{X_S}(x)\right].$$

Finding a point in the intersection of a collection of closed convex sets represents a modeling paradigm for solving important applications such as data compression, neural networks and adaptive filtering, see (Censor et al., 2012) for a complete list.

2.2 Regression problem

Let us consider the matrix $A \in \mathbb{R}^{m\times n}$. For any $S \in \Omega \subseteq \mathbb{R}$, let us define:

$$f(x;S) := \ell(A_S^T x),$$

where ℓ is some loss function and $A_S \in \mathbb{R}^n$. This results in the following constrained optimization model:

$$\min_{x\in\mathbb{R}^n} \quad \mathbb{E}[\ell(A_S^T x)] \quad \text{s.t. } x \in \cap_{S\in\Omega} X_S.$$

Many learning problems can be modeled into this form, see e.g. (Toulis et al., 2016; Shalev-Shwartz and Zhang, 2013). This type of optimization model has been also considered in (Bianchi, 2016; Rosasco et al., 2014).

2.3 Finite sum problem

Let $\Omega = \{1, \cdots, m\}$ and $\mathbb{P}$ be the uniform discrete probability distribution on Ω. Further, we consider convex functions $f(x; i) = \ell_i(x)$. Then, the following constrained finite sum problem is recovered:

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{m} \sum_{i=1}^{m} \ell_i(x) \quad \text{s.t. } x \in \cap_{i=1}^{m} X_i.$$

This constrained optimization model appears often in statistics and machine learning applications, where the functions $\ell_i(\cdot)$ typically represent loss functions associated to a given estimator and the feasible set comes from physical constraints, see e.g. (Defazio et al., 2014; Roux et al., 2012; Vanli et al., 2017; Yurtsever et al., 2016). It is also a particular problem of a more general optimization model considered in (Bianchi, 2016).

2.4 Multiple kernel learning problem

In many classification problems we want to learn a convex combination of kernels $\kappa(x, x') = \sum_{j=1}^{M} \beta_j \kappa_j(x, x')$ (Bach et al., 2004). This approach is useful in complex classification problems, where we use polynomial kernels of different degrees or kernels on different domains. The goal is to learn the weights β_j and they are usually found through SVM optimization:

$$\begin{aligned}
&\min_{(w,\beta,\xi,b)} \ \frac{1}{2} \left(\sum_{j=1}^{M} \beta_j \|w_j\| \right)^2 + C \sum_{i=1}^{N} \xi_i \\
&w = (w_1, \cdots, w_M), \ w_j \in \mathbb{R}^{n_j}, \ \beta = (\beta_1, \cdots, \beta_M), \ \xi = (\xi_1, \cdots, \xi_N) \\
&y_i \left(\sum_{j=1}^{M} \beta_j w_j^T x_{ij} + b \right) \geq 1 - \xi_i \ \forall i = 1 : N, \ \xi \geq 0, \ \beta \geq 0, \ \sum_{j=1}^{M} \beta_j = 1.
\end{aligned}$$

Note that this formulation is equivalent to linear SVM for $M = 1$. We usually obtain a sparse solution in β, where each component β_j corresponds to one kernel κ_j. The dual of this optimization problem takes the form:

$$\begin{aligned}
&\min_{(\gamma,\alpha)} \ \frac{1}{2}\gamma^2 - \sum_{i=1}^{N} \alpha_i \\
&0 \leq \alpha \leq C, \ \sum_{i=1}^{N} \alpha_i y_i = 0, \ \sum_{p=1}^{N} \sum_{q=1}^{N} \alpha_p \alpha_q y_p y_q \kappa_j(x_p, x_q) \leq \gamma^2 \ \ \forall j = 1 : M.
\end{aligned}$$

This convex Quadratic Optimization problem with Quadratic Constraints can be easily reformulated as a Linear Program with infinite number of simple constraints by introducing the notation $Q_j(\alpha) = \sum_{p=1}^{N} \sum_{q=1}^{N} \alpha_p \alpha_q y_p y_q \kappa_j(x_p, x_q)$ (Sonnenburg et al., 2006):

$$\begin{aligned}
&\max_{(\theta,\beta)} \ \theta \\
&\theta \in \mathbb{R}, \ \ \beta \geq 0, \ \ \sum_{j=1}^{M} \beta_j = 1, \ \ \sum_{j=1}^{M} \beta_j \left(\frac{1}{2} Q_j(\alpha) - \sum_{i=1}^{N} \alpha_i \right) \geq \theta \quad \forall \alpha \in \Omega(y),
\end{aligned}$$

where we use the notation

$$\Omega(y) = \left\{ \alpha : \ 0 \leq \alpha \leq C, \ \sum_{i=1}^{N} \alpha_i y_i = 0 \right\}.$$

There are many methods for solving Linear Programs with infinite number of constraints, in particular algorithms related to boosting (Sonnenburg et al., 2006). Note that in this Linear Program formulation the sets X_S are simple hyperplanes.

2.5 Optimal control problem

In this section we briefly present the H_2 optimal control problem for linear systems (see (Karimi and Kammer, 2017) for a detailed exposition). In this application one aims at finding a stabilizing controller K for a linear system which minimize an H_2 performance indicator. This problem can be formulated as:

$$\min_{K(\omega),\Gamma(\omega)} \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \text{trace}[\Gamma(\omega)] d\omega$$
$$\text{s.t.} : W(\omega)[(I_n + G(\omega)K(\omega))^*(I_n + G(\omega)K(\omega))]^{-1}W^*(\omega) \preceq \Gamma(\omega) \quad \forall \omega \in \Omega,$$

where the frequencies ω are taken in the interval $\Omega = \left[-\frac{\pi}{T}, \frac{\pi}{T}\right]$, $G(\omega), W(\omega)$ are the parameters associated with the linear dynamical system under consideration, $\Gamma(\omega)$ is a positive semidefinite matrix and $K(\omega)$ is the controller that needs to be identified. Note that the previous H_2 optimal control problem requires that the constraints, expressed through matrix inequalities, to hold for all frequencies ω in the interval Ω. Moreover, the objective function can be expressed as an expectation over the same interval Ω. In control theory, $\Gamma(\omega)$ and $K(\omega)$ are taken as polynomial matrices in the frequencies ω. Moreover, the previous matrix inequalities are usually convexified using Schur complement and linearization techniques and then the interval Ω is discretized to get a finite number of constraints (linear matrix inequalities) (Karimi and Kammer, 2017).

3. Stochastic Proximal Point algorithm

In this section we propose solving the optimization problem (4) through stochastic proximal point type algorithms. It has been proven in (Necoara et al., 2017) that the optimization problem (4) can be equivalently reformulated under Assumption 2 into the following stochastic optimization problem:

$$\min_{x \in \mathbb{R}^n} \ \mathbb{E}\left[f(x; S) + \mathbb{I}_{X_S}(x) \right]. \tag{5}$$

Since each component of the stochastic objective is nonsmooth, a first possible approach is to apply stochastic subgradient methods (Duchi and Singer, 2009; Moulines and Bach, 2011), which would yield simple algorithms, but having usually a relatively slow sublinear convergence rate. Therefore, for more robustness, one can deal with the nonsmoothness through the Moreau smoothing framework. However, there are multiple potential approaches in

this direction. For a given smoothing parameter $\mu > 0$, we can smooth each functional component and the associated indicator function together to obtain the following smooth approximation for the nonsmooth convex function $f(\cdot; S) + \mathbb{I}_{X_S}$:

$$\bar{f}_\mu(x; S) := \min_{z \in \mathbb{R}^n} \; f(z; S) + \mathbb{I}_{X_S}(z) + \frac{1}{2\mu}\|z - x\|^2.$$

Let us denote the corresponding prox operator by $\bar{z}_\mu(x; S) = \arg\min_{z \in \mathbb{R}^n} f(z; S) + \mathbb{I}_{X_S}(z) + \frac{1}{2\mu}\|z - x\|^2$. It is known that any Moreau approximation $\bar{f}_\mu(\cdot; S)$ is differentiable having the gradient $\nabla \bar{f}_\mu(x; S) = \frac{1}{\mu}(x - \bar{z}_\mu(x; S))$ (Rockafellar and Wets, 1998). Moreover, the gradient is Lipschitz continuous with constants bounded by $\frac{1}{\mu}$. Then, instead of solving the nonsmooth problem (5) we can consider solving the smooth approximation:

$$\min_{x \in \mathbb{R}^n} \quad \bar{F}_\mu(x) \quad \left(:= \mathbb{E}[\bar{f}_\mu(x; S)]\right).$$

Notice that we can easily apply the classical SGD strategy to the newly created smooth objective function, which results in the following iteration:

$$\begin{aligned} x^{k+1} &= x^k - \mu_k \nabla \bar{f}_{\mu_k}(x^k; S_k) = \bar{z}_{\mu_k}(x^k; S_k) \\ &= \arg\min_{z \in \mathbb{R}^n} \; f(z; S_k) + \mathbb{I}_{X_{S_k}}(z) + \frac{1}{2\mu_k}\|z - x^k\|^2. \end{aligned}$$

However, the nonasymptotic analysis technique considered in our paper encounters difficulties with this variant of the algorithm. The main difficulty consists in proving the bound $\|\nabla \bar{f}_\mu(x; S)\| \leq \|g_{f(\cdot;S)+\mathbb{I}_{X_S}}(x)\|$ for all $x \in \mathbb{R}^n$, where $g_{f(\cdot;S)+\mathbb{I}_{X_S}}(x) \in \partial(f(\cdot; S) + \mathbb{I}_{X_S})(x)$. We believe that such a bound is essential in our convergence analysis and we leave for future work the analysis of this iterative scheme. Therefore, we considered a second approach based on a smooth Moreau approximation only for the functional component $f(\cdot; S)$ and keeping the indicator function $\mathbb{I}_{X_S}$ in its original form, that is:

$$f_\mu(x; S) := \min_{z \in \mathbb{R}^n} \; f(z; S) + \frac{1}{2\mu}\|z - x\|^2$$

for some smoothing parameter $\mu > 0$. Then, instead of solving nonsmooth problem (5), we solve the following composite approximation:

$$\min_{x \in \mathbb{R}^n} \quad F_\mu(x) \quad (:= \mathbb{E}[f_\mu(x; S) + \mathbb{I}_{X_S}(x)]). \tag{6}$$

Let us denote the corresponding prox operator by:

$$z_\mu(x; S) = \arg\min_{z \in \mathbb{R}^n} f(z; S) + \frac{1}{2\mu}\|z - x\|^2.$$

Further, on the stochastic composite approximation (6) we can apply the stochastic projected gradient method, which leads to a stochastic proximal point like scheme for solving the original problem (4):

Algorithm SPP $(x_0, \{\mu_k\}_{k\geq 0})$

For $k \geq 1$ compute:
1. Choose randomly $S_k \in \Omega$ w.r.t. probability distribution $\mathbb{P}$
2. Update: $y^k = z_{\mu_k}(x^k; S_k)$ and $x^{k+1} = \Pi_{X_{S_k}}(y^k)$

where $x^0 \in \mathbb{R}^n$ is some initial starting point and $\{\mu_k\}_{k\geq 0}$ is a nonincreasing positive sequence of stepsizes. We assume that the algorithm SPP returns either the last point x^k or the average point $\hat{x}^k = \frac{1}{\sum_{i=0}^{k-1}\mu_i}\sum_{i=0}^{k-1}\mu_i x^i$ when it is called as a subroutine. Since the update rule of the positive smoothing (stepsize) sequence $\{\mu_k\}_{k\geq 0}$ strongly contributes to the convergence of the scheme, we discuss in the following sections the most advantageous choices. We first prove the following useful auxiliary result:

Lemma 3 *Let $\mu > 0$, $S \in \Omega$. Then, for any $g_f(x;S) \in \partial f(x;S)$, the following holds:*

$$\|\nabla f_\mu(x;S)\| \leq \|g_f(x;S)\| \qquad \forall x \in \mathbb{R}^n.$$

Proof The optimality condition of problem $\min_{z\in\mathbb{R}^n} f(z;S) + \frac{1}{2\mu}\|x-z\|^2$ is given by:

$$\frac{1}{\mu}\left(x - z_\mu(x;S)\right) \in \partial f(z_\mu(x;S);S).$$

The above inclusion easily implies that there is $g_f(z_\mu(x;S);S) \in \partial f(z_\mu(x;S);S)$ such that:

$$\begin{aligned}\frac{1}{\mu}\|z_\mu(x;S) - x\|^2 &= \langle g_f(z_\mu(x;S);S), x - z_\mu(x;S)\rangle \\ &= \langle g_f(x;S), x - z_\mu(x;S)\rangle + \langle g_f(z_\mu(x;S);S) - g_f(x;S), x - z_\mu(x;S)\rangle \\ &\leq \langle g_f(x;S), x - z_\mu(x;S)\rangle,\end{aligned}$$

where in the last inequality we used the convexity of f. Lastly, by applying the Cauchy-Schwarz inequality in the right hand side we get the above statement. ■

The following two well-known inequalities, which can be found in (Bullen, 2003), will be also useful in the sequel:

(i) **[Bernoulli]** Let $t \in [0,1]$ and $x \in [-1,\infty)$, then the following holds:

$$(1+x)^t \leq 1 + tx. \tag{7}$$

(ii) **[Minkowski]** Let x and y be two random variables. Then, for any $1 \leq p < \infty$, the following inequality holds:

$$\left(\mathbb{E}[|x+y|^p]\right)^{1/p} \leq \left(\mathbb{E}[|x|^p]\right)^{1/p} + \left(\mathbb{E}[|y|^p]\right)^{1/p}. \tag{8}$$

4. Nonasymptotic complexity of SPP: convex objective function

In this section we analyze, under Assumptions 1 and 2, the iteration complexity of SPP scheme with nonincreasing stepsize rule to approximately solve the optimization problem (4). In order to prove this nonasymptotic result, we define $\hat{\mu}_{1,k} = \sum_{i=0}^{k-1} \mu_i$, $\hat{\mu}_{2,k} = \sum_{i=0}^{k-1} \mu_i^2$ and the averaged sequences $\hat{x}^k = \frac{1}{\hat{\mu}_{1,k}} \sum_{i=0}^{k-1} \mu_i x^i$ and $\hat{y}^k = \frac{1}{\hat{\mu}_{1,k}} \sum_{i=0}^{k-1} \mu_i y^i$. Moreover, denote by $\mathcal{F}_k$ the history of random choices $\{S_k\}_{k\geq 0}$, i.e. $\mathcal{F}_k = \{S_0, \cdots, S_k\}$.

Lemma 4 *Let Assumptions 1 and 2 hold and the sequences $\{x^k, y^k\}_{k\geq 0}$ be generated by SPP scheme with positive stepsize $\{\mu_k\}_{k\geq 0}$. Then the following relation holds:*

$$\mathbb{E}\left[dist^2_{X_{S_k}}(\hat{y}^k)\right] \geq \frac{1}{\zeta}\mathbb{E}\left[dist^2_X(\hat{x}^k)\right] - \frac{\hat{\mu}_{2,k}}{\hat{\mu}_{1,k}}\sqrt{\mathbb{E}[dist^2_X(\hat{x}^k)]}\,\sqrt{\mathbb{E}[L^2_{f,S}]}.$$

Proof See Appendix for the proof. ■

Now, we are ready to derive the convergence rate of SPP in the average sequence $\hat{x}^k$:

Theorem 5 *Under Assumptions 1 and 2, let the sequence $\{x^k\}_{k\geq 0}$ be generated by the algorithm SPP with nonincreasing positive stepsize $\{\mu_k\}_{k\geq 0}$. Define $\mathcal{R}_\mu = \mu_0\zeta(\|x^0 - x^*\|^2 + \mathbb{E}[L^2_{f,S}]\hat{\mu}_{2,k})$, then the following estimates for suboptimality and feasibility violation hold:*

$$\begin{aligned} -\zeta\mathbb{E}[L^2_{f,S}]\left(\frac{\hat{\mu}_{2,k}}{\hat{\mu}_{1,k}} + 2\mu_0\right) - \sqrt{\mathbb{E}[L^2_{f,S}]\,\frac{\mathcal{R}_\mu}{\hat{\mu}_{1,k}}} \leq \mathbb{E}[F(\hat{x}^k)] - F^* \leq \frac{\mathcal{R}_\mu}{2\mu_0\zeta\hat{\mu}_{1,k}} \\ \mathbb{E}[dist^2_X(\hat{x}^k)] \leq 2\zeta^2\mathbb{E}[L^2_{f,S}]\left(\frac{\hat{\mu}_{2,k}}{\hat{\mu}_{1,k}} + 2\mu_0\right)^2 + \frac{2\mathcal{R}_\mu}{\hat{\mu}_{1,k}}. \end{aligned} \tag{9}$$

Proof See Appendix for the proof. ■

Note that the right suboptimality bound (9), obtained for the SPP algorithm, is similar with the one given for the standard subgradient method (Nesterov, 2004). Below we provide the convergence estimates for the algorithm SPP with constant stepsize for a desired accuracy $\epsilon > 0$. For simplicity, assume that $\|x^0 - x^*\| \geq 1$ and $\mathbb{E}[L^2_{f,S}] \geq 2$.

Corollary 6 *Under the assumptions of Theorem 5, let $\{x^k\}_{k\geq 0}$ be the sequence generated by algorithm SPP with constant stepsize $\mu_k = \mu > 0$. Also let $\epsilon > 0$ be the desired accuracy, K be an integer satisfying:*

$$K \geq \frac{\mathbb{E}[L^2_{f,S}]\|x^0 - x^*\|^2}{\epsilon^2}\max\left\{1, (3\zeta + \sqrt{2\zeta})^2\right\},$$

and the stepsize be chosen as:

$$\mu = \frac{\epsilon}{\mathbb{E}[L^2_{f,S}](3\zeta + \sqrt{2\zeta})}.$$

Then, after K iterations, the average point $\hat{x}^K = \frac{1}{K}\sum_{i=0}^{K-1} x^i$ satisfies:

$$\left|\mathbb{E}[F(\hat{x}^K)] - F^*\right| \leq \epsilon \quad \textit{and} \quad \sqrt{\mathbb{E}[dist_X^2(\hat{x}^K)]} \leq \epsilon.$$

Proof We consider $k = K$ in Theorem 5 and, by taking into account that $\mu_k = \mu$ for all $k \geq 0$, we aim to obtain the lowest value of the right hand side of (9) by minimizing over $\mu > 0$. Thus, by denoting that $r_0 = \|x^0 - x^*\|$, we obtain for the optimal smoothing parameter:

$$\mu = \sqrt{\frac{r_0^2}{K\mathbb{E}[L_{f,S}^2]}}$$

the optimal rate

$$\mathbb{E}[F(\hat{x}^K)] - F^* \leq \sqrt{\frac{\mathbb{E}[L_{f,S}^2]r_0^2}{K}}. \tag{10}$$

Also using the optimal parameter $\tilde{\mu}$ into the other relations of Theorem 5 result in:

$$\mathbb{E}[\text{dist}_X^2(\hat{x}^K)] \leq \frac{r_0^2}{K}\left(18\zeta^2 + 4\zeta\right) \tag{11}$$

and

$$\mathbb{E}[F(\hat{x}^K)] - F^* \geq -(3\zeta + \sqrt{2\zeta})\sqrt{\frac{\mathbb{E}[L_{f,S}^2]r_0^2}{K}}. \tag{12}$$

From the upper and lower suboptimality bounds (10), (12) and feasibility bound (11), we deduce the following bound:

$$K \geq \frac{\mathbb{E}[L_{f,S}^2]r_0^2}{\epsilon^2}\max\left\{1, (3\zeta + \sqrt{2\zeta})^2\right\}$$

which confirms our result. ■

In conclusion, Corollary 6 states that for a desired accuracy ϵ, if we choose a constant stepsize $\mu = \mathcal{O}(\epsilon)$ and perform a number of SPP iterations $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ we obtain an ϵ-optimal solution for our original stochastic constrained convex problem (4). Note that for convex problems with objective function having bounded subgradients the previous convergence estimates derived for the SPP algorithm are similar to those corresponding to the classical deterministic proximal point method (Guler, 1991) and subgradient method (Nesterov, 2004).

5. Nonasymptotic complexity of SPP: strongly convex objective function

In this section we analyze the convergence behavior of the SPP scheme under smoothness and strong convexity assumptions on the objective function of constrained problem (4). Therefore, in this section the Assumption 1 is replaced by the following assumptions:

Assumption 7 *Each function $f(\cdot;S)$ is differentiable and $\sigma_{f,S}$-strongly convex, that is there exists strong convexity constant $\sigma_{f,S} \geq 0$ such that:*

$$f(x;S) \geq f(y;S) + \langle \nabla f(y;S), x-y\rangle + \frac{\sigma_{f,S}}{2}\|x-y\|^2 \qquad \forall x,y \in \mathbb{R}^n.$$

Moreover, the strong convexity constants $\sigma_{f,S}$ satisfy $\sigma_F = \mathbb{E}[\sigma_{f,S}] > 0$.

Notice that if for some function $f(\cdot;S)$ the corresponding constant $\sigma_{f,S} = 0$, then $f(\cdot;S)$ is only convex. However, relation $\mathbb{E}[\sigma_{f,S}] = \sigma_F > 0$ implies that the whole objective function F of problem (4) is strongly convex with constant $\sigma_F > 0$. In the sequel we will analyze the SPP scheme under the following additional smoothness assumption:

Assumption 8 *Each function $f(\cdot;S)$ has Lipschitz gradient, that is there exists Lipschitz constant $L_{f,S} > 0$ such that:*

$$\|\nabla f(x;S) - \nabla f(y;S)\| \leq L_{f,S}\|x-y\| \qquad \forall x,y \in \mathbb{R}^n.$$

Note that Assumptions 7 and 8 are standard for the convergence analysis of SPP like schemes, see e.g. (Moulines and Bach, 2011; Ryu and Boyd, 2016). We first present an auxiliary result on the behavior of the proximal mapping $z_\mu(\cdot;S)$.

Lemma 9 *Let $f(\cdot;S)$ satisfy Assumption 7. Further, for any $S \in \Omega$ and $\mu > 0$, we define $\theta_S(\mu) = \frac{1}{1+\mu\sigma_{f,S}}$. Then, the following contraction inequality holds for the prox operator:*

$$\|z_\mu(x;S) - z_\mu(y;S)\| \leq \theta_S(\mu)\|x-y\| \qquad \forall x,y \in \mathbb{R}^n.$$

Proof See Appendix for the proof. ■

Notice that if all the functions $f(\cdot;S)$ are just convex, that is they satisfy Assumption 7 with $\sigma_{f,S} = 0$, then Lemma 9 highlights the nonexpansiveness property of the proximal operator $z_\mu(\cdot;S)$. We will further keep using the notation $\theta_S(\mu)$ for the contraction factor of the operator $z_\mu(\cdot;S)$. Moreover, in all our proofs below, regarding the results in expectation, we use the standard technique of applying first expectation with respect to S_k conditioned on $\mathcal{F}_{k-1}$ and then apply the expectation over the entire history $\mathcal{F}_{k-1}$ (see the proof of Theorem 5). For simplicity of the exposition and for saving space, we omit these details below.

5.1 Linear convergence to noise dominated region for constant stepsize SPP

Next we analyze the sequence generated by the SPP scheme with constant stepsize $\mu > 0$ and provide a nonasymptotic bound on the quadratic mean $\{\mathbb{E}[\|x^k - x^*\|^2]\}_{k\geq 0}$.

Theorem 10 *Under Assumption 7, let the sequence $\{x^k\}_{k\geq 0}$ be generated by the algorithm SPP with constant stepsize $\mu > 0$. Further, assume $\sigma_f^{\max} = \sup_{S\in\Omega} \sigma_{f,S} < \infty$. Then, $\mathbb{E}[\theta_S^2(\mu)] \leq \mathbb{E}[\theta_S(\mu)] < 1$ and the following linear convergence to some region around the optimal point in the quadratic mean holds:*

$$\mathbb{E}[\|x^k - x^*\|^2] \leq 2\left(\mathbb{E}\left[\theta_S^2(\mu)\right]\right)^k \|x^0 - x^*\|^2 + \frac{2\mu^2\mathbb{E}[\|\nabla f(x^*;S)\|^2]}{\left(1 - \sqrt{\mathbb{E}[\theta_S^2(\mu)]}\right)^2}.$$

Proof First, it can be easily seen that for any $\mu > 0$ and $S \in \Omega$ we have $\theta_S^2(\mu) \leq \theta_S(\mu) \leq 1$ and assuming that $\sigma_f^{\max} < \infty$ we obtain:

$$0 \leq \mathbb{E}[\theta_S^2(\mu)] \leq \mathbb{E}[\theta_S(\mu)] = \mathbb{E}\left[\frac{1}{1+\mu\sigma_{f,S}}\right] = 1 - \mathbb{E}\left[\frac{\mu\sigma_{f,S}}{1+\mu\sigma_{f,S}}\right] \leq 1 - \frac{\mu\sigma_F}{1+\mu\sigma_f^{\max}} < 1.$$

Then, by applying Lemma 9 with $S = S_k, x = x^k$ and $z = x^*$, results in:

$$\left\| z_\mu(x^k; S_k) - z_\mu(x^*; S_k) \right\| \leq \theta_{S_k}(\mu)\|x^k - x^*\|,$$

which, by the triangle inequality, further implies:

$$\left\| z_\mu(x^k; S_k) - x^* \right\| \leq \theta_{S_k}(\mu)\|x^k - x^*\| + \|z_\mu(x^*; S_k) - x^*\|.$$

By using the nonexpansiveness property of the projection operator we get that $\|x^{k+1} - x^*\| \leq \|y^k - x^*\|$, then the last inequality leads to the reccurent relation:

$$\left\| x^{k+1} - x^* \right\| \leq \left\| z_\mu(x^k; S_k) - x^* \right\| \leq \theta_{S_k}(\mu)\|x^k - x^*\| + \|z_\mu(x^*; S_k) - x^*\|. \tag{13}$$

The relation (13), Minkowski inequality and Lemma 3 lead to the following recurrence:

$$\begin{aligned}
\sqrt{\mathbb{E}[\|x^{k+1} - x^*\|^2]} &\overset{(13)}{\leq} \sqrt{\mathbb{E}\left[\left(\theta_{S_k}(\mu)\|x^k - x^*\| + \|z_\mu(x^*; S_k) - x^*\|\right)^2\right]} \\
&\overset{(8)}{\leq} \sqrt{\mathbb{E}\left[\theta_{S_k}^2(\mu)\|x^k - x^*\|^2\right]} + \sqrt{\mathbb{E}\left[\|z_\mu(x^*; S_k) - x^*\|^2\right]} \\
&= \sqrt{\mathbb{E}\left[\theta_S^2(\mu)\right]}\sqrt{\mathbb{E}\left[\|x^k - x^*\|^2\right]} + \mu\sqrt{\mathbb{E}\left[\|\nabla f_\mu(x^*; S)\|^2\right]} \\
&\overset{\text{Lemma 3}}{\leq} \sqrt{\mathbb{E}\left[\theta_S^2(\mu)\right]}\sqrt{\mathbb{E}[\|x^k - x^*\|^2]} + \mu\sqrt{\mathbb{E}\left[\|\nabla f(x^*; S)\|^2\right]}.
\end{aligned}$$

This yields the following relation valid for all $\mu > 0$ and $k \geq 0$:

$$\sqrt{\mathbb{E}[\|x^{k+1} - x^*\|^2]} \leq \sqrt{\mathbb{E}\left[\theta_S^2(\mu)\right]}\sqrt{\mathbb{E}[\|x^k - x^*\|^2]} + \mu\sqrt{\mathbb{E}\left[\|\nabla f(x^*; S)\|^2\right]}, \tag{14}$$

Denote $r_k = \sqrt{\mathbb{E}[\|x^k - x^*\|^2]}$, $\eta = \sqrt{\mathbb{E}\left[\|\nabla f(x^*; S)\|^2\right]}$ and $\theta(\mu) = \sqrt{\mathbb{E}\left[\theta_S^2(\mu)\right]}$. Then, we get:

$$r_{k+1} \leq \theta(\mu) r_k + \mu\eta.$$

Finally, a simple inductive argument leads to:

$$\begin{aligned}
r_k &\leq r_0\theta(\mu)^k + \mu\eta\left[1 + \theta(\mu) + \cdots + \theta(\mu)^{k-1}\right] \\
&= r_0\theta(\mu)^k + \mu\eta\frac{1 - \theta(\mu)^k}{1 - \theta(\mu)} \\
&\leq r_0\theta(\mu)^k + \frac{\mu\eta}{1 - \theta(\mu)}.
\end{aligned}$$

By squaring and returning to our basic notations, we recover our statement. ■

Theorem 10 proves a linear convergence rate in expectation, without assuming any kind of smoothness on the objective function, for the sequence $\{x^k\}_{k\geq 0}$ generated by SPP with constant stepsize $\mu > 0$ when the iterates are outside of a *noise dominated* neighborhood of the optimal set of radius $\frac{\mu\sqrt{\mathbb{E}[\|\nabla f(x^*;S)\|^2]}}{1-\sqrt{\mathbb{E}[\theta_S^2(\mu)]}}$. It also establishes the boundedness of the sequence $\{x^k\}_{k\geq 0}$ when the stepsize is constant. Notice that in (Ryu and Boyd, 2016) a similar result has been given for an unconstrained optimization model with the difference that the convergence rate was provided for $\mathbb{E}[\|x^k - x^*\|]$. However, our proof is simpler and more elegant, based on the properties of Moreau approximation, despite the fact that we consider the constrained case.

5.2 Nonasymptotic sublinear convergence rate of variable stepsize SPP

In this section we derive sublinear convergence rate of order $\mathcal{O}(1/k^\gamma)$ for the variable stepsize SPP scheme, in a nonasymptotic fashion. We first prove the boundedness of $\{x^k\}_{k\geq 0}$ when the stepsize is nonincreasing, which will be useful for the subsequent convergence results.

Lemma 11 *Under Assumption 7, let the sequence $\{x^k\}_{k\geq 0}$ be generated by the algorithm SPP with nonincreasing positive stepsize $\{\mu_k\}_{k\geq 0}$. Then, the following relation holds:*

$$\mathbb{E}\left[\|x^k - x^*\|\right] \leq \sqrt{\mathbb{E}[\|x^k - x^*\|^2]} \leq \max\left\{\|x^0 - x^*\|, \frac{\mu_0\sqrt{\mathbb{E}\left[\|\nabla f(x^*;S)\|^2\right]}}{1-\sqrt{\mathbb{E}\left[\theta_S^2(\mu_0)\right]}}\right\}.$$

Proof See Appendix for the proof. ■

Furthermore, we need an upper bound on the sequence $\{\mathbb{E}[\|\nabla f(x^k;S)\|]\}_{k\geq 0}$:

Lemma 12 *Under Assumptions 7 and 8, let the sequence $\{x^k\}_{k\geq 0}$ be generated by the algorithm SPP with nonincreasing positive stepsizes $\{\mu_k\}_{k\geq 0}$. Then, the following holds:*

$$\mathbb{E}[\|\nabla f(x^k;S)\|^2] \leq 2\mathbb{E}[\|\nabla f(x^*;S)\|^2] + 2\mathbb{E}[L_{f,S}^2]\mathcal{A}^2,$$

where $\mathcal{A} = \max\left\{\|x^0 - x^*\|, \frac{\mu_0\sqrt{\mathbb{E}[\|\nabla f(x^*;S)\|^2]}}{1-\sqrt{\mathbb{E}\left[\theta_S^2(\mu_0)\right]}}\right\}$.

Proof From the Lipschitz continuity of $\nabla f(\cdot;S)$ we have that $\|\nabla f(x;S) - \nabla f(x^*;S)\| \leq L_{f,S}\|x - x^*\|$ for all $x \in \mathbb{R}^n$, which implies:

$$\|\nabla f(x^k;S)\|^2 \leq (\|\nabla f(x^*;S)\| + L_{f,S}\|x^k - x^*\|)^2 \leq 2\|\nabla f(x^*;S)\|^2 + 2L_{f,S}^2\|x^k - x^*\|^2.$$

By taking expectation in both sides we get:

$$\mathbb{E}[\|\nabla f(x^k;S)\|^2] \leq 2\mathbb{E}[\|\nabla f(x^*;S)\|^2] + 2\mathbb{E}[L_{f,S}^2]\mathbb{E}[\|x^k - x^*\|^2].$$

Lastly, by using Lemma 11 we obtain our statement. ■

Finally, we provide a non-trivial upper bound on the feasibility gap, which automatically leads to a iterative descent in the distance to the feasible set of the sequence $\{x^k\}_{k\geq 0}$, generated by the SPP scheme with nonincreasing stepsizes.

Lemma 13 *Under Assumptions 2, 7 and 8, let the sequence $\{x^k\}_{k\geq 0}$ be generated by SPP scheme with nonincreasing stepsizes $\{\mu_k\}_{k\geq 0}$. Then, the following relation holds:*

$$\sqrt{\mathbb{E}[dist_X^2(x^k)]} \leq \left(1-\frac{1}{\zeta}\right)^{k/2}\left[dist_X(x^0)+2\mu_0\zeta\mathcal{B}\right]+2\mu_{k-\lceil\frac{k}{2}\rceil}\zeta\mathcal{B},$$

where $\mathcal{B}=\sqrt{2\mathbb{E}[\|\nabla f(x^*;S)\|^2]}+\mathcal{A}\sqrt{2\mathbb{E}[L_{f,S}^2]}$.

Proof See Appendix for the proof. ■

Now, we are ready to derive the nonasymptotic convergence rate of the Algorithm SPP with nonincreasing stepsizes. For simplicity, we denote $\eta=\sqrt{\mathbb{E}[\|\nabla f(x^*;S)\|^2]}$ and keep the notations for $\mathcal{A}$ from Lemma 12 and for $\mathcal{B}$ from Lemma 13.

Theorem 14 *Under Assumptions 2, 7 and 8, let the sequence $\{x^k\}_{k\geq 0}$ be generated by the algorithm SPP with the stepsize $\mu_k=\frac{\mu_0}{k^\gamma}$ for all $k\geq 1$, with $\mu_0>0$ and $\gamma\in(0,1]$, and denote $\theta_0=\mathbb{E}\left[\theta_S^2(\mu_0)\right]=\mathbb{E}\left[\frac{1}{(1+\mu_0\sigma_{f,S})^2}\right]$. Then, the following relations hold:*
(i) If $\gamma\in(0,1)$, then we have the following nonasymptotic convergence rates:

$$\mathbb{E}[\|x^k-x^*\|^2]\leq\theta_0^{\varphi_{1-\gamma}(k)}r_0^2+\mathcal{D}\theta_0^{\varphi_{1-\gamma}(k)-\varphi_{1-\gamma}(\frac{k+1}{2})}\mu_0^2\left[\varphi_{1-2\gamma}\left(\frac{k+1}{2}\right)+2\right]+\frac{\mathcal{D}\mu_0^2 4^\gamma}{(1-\theta_0)k^\gamma}.$$

(ii) If $\gamma=1$, then we have the following nonasymptotic convergence rate:

$$\mathbb{E}[\|x^k-x^*\|^2]\leq\begin{cases}\theta_0^{\varphi_0(k)}r_0^2+\frac{2\mu_0^2}{k\left(\ln\left(\frac{1}{\theta_0}\right)-1\right)} & \text{if } \theta_0<\frac{1}{e}\\ \theta_0^{\varphi_0(k)}r_0^2+\frac{2\mu_0^2\ln k}{k} & \text{if } \theta_0=\frac{1}{e}\\ \theta_0^{\varphi_0(k)}r_0^2+\left(\frac{2}{k}\right)^{\ln\left(\frac{1}{\theta_0}\right)}\frac{\mu_0^2}{1-\ln\left(\frac{1}{\theta_0}\right)} & \text{if } \theta_0>\frac{1}{e},\end{cases}$$

where $\mathcal{D}=4\|\nabla F(x^*)\|\left[\frac{dist_X(x^0)+2\mu_0\zeta\mathcal{B}}{\mu_0\ln(\zeta/(\zeta-1))}+3^\gamma\mathcal{B}\zeta\right]+2\eta\sqrt{2\eta^2+2\mathbb{E}[L_{f,S}^2]\mathcal{A}^2}+2\eta\mathcal{A}\sqrt{\mathbb{E}[L_{f,S}^2]}$.

Proof See Appendix for the proof. ■

For more clear estimates of the convergence rates obtained in Theorem 14, we provide in the next corollary a summary given in terms of the dominant terms:

Corollary 15 *Under the assumptions of Theorem 14 the following convergence rates hold:*
(i) If $\gamma \in (0,1)$, then we have convergence rate of order:

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

(ii) If $\gamma = 1$, then we have convergence rate of order:

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \begin{cases} \mathcal{O}\left(\frac{1}{k}\right) & \text{if } \theta_0 < \frac{1}{e} \\ \mathcal{O}\left(\frac{\ln k}{k}\right) & \text{if } \theta_0 = \frac{1}{e} \\ \mathcal{O}\left(\frac{1}{k}\right)^{2\ln\left(\frac{1}{\theta_0}\right)} & \text{if } \theta_0 > \frac{1}{e}. \end{cases}$$

Proof First assume that $\gamma \in (0, \frac{1}{2})$. This assumption implies that $1 - 2\gamma > 0$ and that:

$$\varphi_{1-2\gamma}\left(\frac{k}{2}+2\right) = \frac{\left(\frac{k}{2}+2\right)^{1-2\gamma} - 1}{1-2\gamma} \leq \frac{\left(\frac{k}{2}+2\right)^{1-2\gamma}}{1-2\gamma}. \tag{15}$$

On the other hand, by using the inequality $e^{-x} \leq \frac{1}{1+x}$ for all $x \geq 0$, we obtain:

$$\begin{aligned}
&\theta_0^{\varphi_{1-\gamma}(k+1)-\varphi_{1-\gamma}(\frac{k+1}{2})}\varphi_{1-2\gamma}\left(\frac{k}{2}+2\right) = e^{(\varphi_{1-\gamma}(k+1)-\varphi_{1-\gamma}(\frac{k+1}{2}))\ln\theta_0}\varphi_{1-2\gamma}\left(\frac{k}{2}+2\right) \\
&\leq \frac{\varphi_{1-2\gamma}\left(\frac{k}{2}+2\right)}{1+[\varphi_{1-\gamma}(k+1)-\varphi_{1-\gamma}(\frac{k}{2}+1)]\ln\frac{1}{\theta_0}} \overset{(15)}{\leq} \frac{\frac{(k+4)^{1-2\gamma}}{2^{1-2\gamma}(1-2\gamma)}}{\frac{1}{1-\gamma}[(k+1)^{1-\gamma}-(\frac{k}{2}+1)^{1-\gamma}]\ln\frac{1}{\theta_0}} \\
&= \frac{\frac{(k+4)^{1-2\gamma}}{2^{1-2\gamma}(1-2\gamma)}}{\frac{(k+2)^{1-\gamma}}{1-\gamma}[(\frac{2}{3})^{1-\gamma}-(\frac{1}{2})^{1-\gamma}]\ln\frac{1}{\theta_0}} = \frac{1-\gamma}{1-2\gamma}\frac{2^\gamma(k+4)^{-\gamma}}{[(\frac{2}{3})^{1-\gamma}-(\frac{1}{2})^{1-\gamma}]\ln\frac{1}{\theta_0}} \approx \mathcal{O}\left(\frac{1}{k^\gamma}\right).
\end{aligned}$$

Therefore, in this case, the overall rate will be given by:

$$r_{k+1}^2 \leq \theta_0^{\mathcal{O}(k^{1-\gamma})} r_0^2 + \mathcal{O}\left(\frac{1}{k^\gamma}\right) \approx \mathcal{O}\left(\frac{1}{k^\gamma}\right).$$

If $\gamma = \frac{1}{2}$, then the definition of $\varphi_{1-2\gamma}(\frac{k}{2}+2)$ provides that:

$$r_{k+1}^2 \leq \theta_0^{\mathcal{O}(\sqrt{k})} r_0^2 + \theta_0^{\mathcal{O}(\sqrt{k})}\mathcal{O}(\ln k) + \mathcal{O}\left(\frac{1}{\sqrt{k}}\right) \approx \mathcal{O}\left(\frac{1}{\sqrt{k}}\right).$$

When $\gamma \in (\frac{1}{2}, 1)$, it is obvious that $\varphi_{1-2\gamma}\left(\frac{k}{2}+2\right) \leq \frac{1}{2\gamma-1}$ and therefore the order of the convergence rate changes into:

$$r_{k+1}^2 \leq \theta_0^{\mathcal{O}(k^{1-\gamma})}[r_0^2 + \mathcal{O}(1)] + \mathcal{O}\left(\frac{1}{k^\gamma}\right) \approx \mathcal{O}\left(\frac{1}{k^\gamma}\right).$$

Lastly, if $\gamma = 1$, by using $\theta_0^{\ln k+1} \leq \left(\frac{1}{k}\right)^{\ln\frac{1}{\theta_0}}$ we obtain the second part of our result. ■

Notice that the above results state that our SPP algorithm with variable stepsize $\frac{\mu_0}{k^\gamma}$ converges with $\mathcal{O}\left(\frac{1}{k^\gamma}\right)$ rate. Similar results have been obtained in (Toulis et al., 2016) for a particular objective function of the form $f(a_S^T x)$ without any constraints and for $\gamma \in (1/2, 1]$. Moreover, for $\gamma = 1$ similar convergence rate, but in asymptotic fashion and for unconstrained problems, has been derived in (Ryu and Boyd, 2016). As we have already mentioned in the introduction section, the convergence rate for the SGD scheme contains an exponential term of the form $\frac{e^{C_2\mu_0^2}}{k^{\alpha\mu_0}}$, which for a given iteration counter k grows exponentially in the initial stepsize μ_0, see (Moulines and Bach, 2011). Thus, although the SGD method achieves a rate $O(\frac{1}{k})$ for a variable stepsize $\frac{\mu_0}{k}$, if μ_0 is chosen too large, then it can induce catastrophic effects in the convergence rate. However, one should notice that for our SPP method, Theorem 14 does not contain this kind of exponential term, therefore SPP is more robust than SGD scheme even in the constrained case. This can be also observed in numerical simulations, see Section 7 below. Clearly, Corollary 15 directly implies the following complexity estimates for attaining a suboptimal point x^k satisfying $\mathbb{E}[\|x^k - x^*\|^2] \leq \epsilon$.

Corollary 16 *Under the assumptions of Theorem 14 and $\epsilon > 0$ the following estimates hold. For $\gamma \in (0, 1)$, if we perform:*

$$\left\lceil \mathcal{O}\left(\frac{1}{\epsilon^{1/\gamma}}\right)\right\rceil$$

iterations of SPP scheme with variable stepsize, then the sequence $\{x^k\}_{k\geq 0}$ satisfies $\mathbb{E}[\|x^k - x^*\|^2] \leq \epsilon$. Moreover, for $\gamma = 1$ and $\theta_0 < \frac{1}{e}$, if we perform:

$$\left\lceil \mathcal{O}\left(\frac{1}{\epsilon}\right)\right\rceil$$

iterations of SPP scheme with variable stepsize, then we have $\mathbb{E}[\|x^k - x^*\|^2] \leq \epsilon$.

Proof The proof follows immediately from Corollary 15. ■

6. A restarted variant of Stochastic Proximal Point algorithm

From previous section we easily notice that an $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ convergence rate is obtained for the SPP algorithm with variable stepsize $\mu_k = \frac{\mu_0}{k}$ only when the initial stepsize μ_0 is chosen sufficiently large such that $\theta_0 = \mathbb{E}\left[\frac{1}{(1+\mu_0\sigma_{f,S})^2}\right] < \frac{1}{\sqrt{e}}$. However, this condition is not easy to check. Therefore, if μ_0 is not chosen adequately, we can encounter the case $\theta_0 > \frac{1}{\sqrt{e}}$, which leads to a worse convergence rate for the SPP scheme of order $\mathcal{O}\left(\epsilon^{-\frac{1}{2\ln(1/\theta_0)}}\right)$, that is implicitly dependent on the choice of the initial stepsize μ_0. In conclusion, in order to remove this dependence on the initial stepsize of the simple SPP scheme, we develop a restarting variant of it. This variant consists of running the SPP algorithm (as a routine) for multiple times (epochs) and restarting it each time after a certain number of iterations. In each epoch t, the SPP scheme runs for an estimated number of iterations K_t, which may vary over the epochs, depending on the assumptions made on the objective function.

More explicitly, the Restarted Stochastic Proximal Point (RSPP) scheme has the following iteration:

Algorithm RSPP

Let $\mu_0 > 0$ and $x^{0,0} \in \mathbb{R}^n$. For $t \geq 1$ do:

1. Compute stepsize μ_t and number of inner iterations K_t
2. Set $x^{K_t,t}$ the average output of SPP($x^{K_{t-1},t-1}, \mu_t$) runned for K_t iterations with constant stepsize μ_t
3. If an outer stopping criterion is satisfied, then **STOP**, otherwise $t := t+1$ and go to step 1.

We analyze below the nonasymptotic convergence rate of the RSPP algorithm under Assumptions 7 and 8.

6.1 Nonasymptotic sublinear convergence of algorithm RSPP

In this section we analyze the convergence rate of the sequence generated by the RSPP scheme, which repeatedly calls the subroutine SPP with a constant stepsize, in multiple epochs. We consider that SPP runs in epoch $t \geq 1$ with the constant stepsize μ_t for K_t iterations. As in previous sections, we first provide a descent lemma for the feasibility gap. For simplicity, we keep the notations of $\mathcal{A}$ from Lemma 12 and $\mathcal{B}$ from Lemma 13.

Lemma 17 *Let Assumptions 2, 7 and 8 hold. Also let the sequence $\{x^{K_t,t}\}_{t\geq 0}$ be generated by RSPP scheme with nonincreasing stepsizes $\{\mu_t\}_{t\geq 0}$ and nondecreasing epoch lengths $\{K_t\}_{t\geq 1}$ such that $K_t \geq 1$ for all $t \geq 1$. Then, the following relation holds:*

$$\sqrt{\mathbb{E}[dist_X^2(x^{K_t,t})]} \leq \left(1-\frac{1}{\zeta}\right)^{\sum_{i=1}^{t}\frac{K_i}{2}} dist_X(x^{0,0}) + 2\left(1-\frac{1}{\zeta}\right)^{\sum_{i=t-\lceil\frac{t}{2}\rceil}^{t}\frac{K_i}{2}} \mu_0\zeta^2\mathcal{B} + 2\mu_{t-\lceil\frac{t}{2}\rceil}\zeta^2\mathcal{B}.$$

Proof See Appendix for the proof. ■

Next, we provide the non-asymptotic bounds on the iteration complexity of RSPP scheme.

Theorem 18 *Let Assumptions 2, 7 and 8 hold and $\epsilon, \mu_0 > 0$. Also let $\gamma > 0$ and $\{x^{K_t,t}\}_{t\geq 0}$ be generated by RSPP scheme with $\mu_t = \frac{\mu_0}{t^\gamma}$ and $K_t = \lceil t^\gamma \rceil$. If we perform the following number of epochs:*

$$T = \left\lceil \max\left\{ \ln\left(\frac{2r_{0,0}^2}{\epsilon}\right)\frac{1}{\ln(1/\theta_0)}, \left(\frac{2^{\gamma+1}\mathcal{D}_r}{\epsilon}\mathcal{C}\right)^{1/\gamma} \right\} \right\rceil,$$

then after a total number of SPP iterations of $\frac{T^{1+\gamma}}{1+\gamma}$*, which is bounded by*

$$\left\lceil \frac{1}{1+\gamma}\max\left\{\ln\left(\frac{2r_{0,0}^2}{\epsilon}\right)^{1+\gamma}\frac{1}{\ln\left(1/\theta_0\right)^{1+\gamma}}, \left(\frac{2^{\gamma+1}\mathcal{D}_r}{\epsilon}\mathcal{C}\right)^{1+\frac{1}{\gamma}}\right\}\right\rceil,$$

where $\mathcal{D}_r = 4\|\nabla F(x^*)\|\left[\frac{dist_X(x^{0,0})+2\mu_0\zeta^2\mathcal{B}}{\mu_0\ln(\zeta/(\zeta-1))}+3^\gamma\mathcal{B}\zeta^2\right]+2\eta\sqrt{2\eta^2+2\mathbb{E}[L_{f,S}^2]\mathcal{A}^2}+2\eta\mathcal{A}\sqrt{\mathbb{E}[L_{f,S}^2]}$ *and* $\mathcal{C}=\frac{1}{2(1-\gamma)\ln 1/\sqrt{\theta_0}}+\frac{\mu_1^2}{(1-\theta_0)^2}$*, we have* $\mathbb{E}[\|x^{K_T,T}-x^*\|^2]\leq\epsilon$*.*

Proof See Appendix for the proof. ■

In conclusion Theorem 18 states that the RSPP algorithm with the choices $(\mu_t, K_t) = \left(\frac{\mu_0}{t^\gamma}, \frac{t^\gamma}{2}\right)$ requires $\mathcal{O}\left(\epsilon^{-\left(1+\frac{1}{\gamma}\right)}\right)$ simple SPP iterations to reach an ϵ optimal point. It is important to observe that this convergence rate is achieved when the stepsize and the epoch length are not dependent on any inaccessible constant, making our restarting scheme easily implementable. Moreover, the parameter γ can be chosen in $(0,\infty)$, i.e. our RSPP scheme allows also stepsizes $\frac{\mu_0}{t^\gamma}$, with $\gamma > 1$. By comparison, an $\mathcal{O}\left(\epsilon^{-1}\right)$ complexity is obtained for SPP with stepsize $\mu_k = \frac{\mu_0}{k}$ only when μ_0 is chosen sufficiently large such that $\theta_0 < \frac{1}{e}$. However, this condition is not easy to check. Moreover, we may fall in the case when $\theta_0 > \frac{1}{e}$, which leads to a complexity of $\mathcal{O}\left(\epsilon^{-\frac{1}{2\ln(1/\theta_0)}}\right)$ of the variable stepsize SPP scheme. Observe that the last convergence rate is implicitly dependent on the constant μ_0 and can be arbitrarily bad, while for $\gamma > 1$ sufficiently large the RSPP scheme achieves the optimal convergence rate $\mathcal{O}\left(\epsilon^{-1}\right)$.

Remark 19 *Notice that there exists a connection between the quadratic mean residual* $\mathbb{E}[\|x^k-x^*\|^2]$ *and the function value residual in a certain point. To obtain this relation, denote* $v^k = [x^k - \frac{1}{L_F}\nabla F(x^k)]_X$ *and observe that for some constant* $L_F \geq \mathbb{E}[L_{f,S}]$ *we have:*

$$\begin{aligned}
F(v^k) &\leq F(x^k) + \langle \nabla F(x^k), v^k - x^k\rangle + \frac{L_F}{2}\|v^k - x^k\|^2\\
&= \min_{y\in X} F(x^k) + \langle \nabla F(x^k), y - x^k\rangle + \frac{L_F}{2}\|y - x^k\|^2\\
&\leq \min_{y\in X} F(y) + \frac{L_F}{2}\|y - x^k\|^2\\
&\leq F(x^*) + \frac{L_F}{2}\|x^k - x^*\|^2,
\end{aligned}$$

where in the second inequality we used the convexity relation. The last relation leads to $F(v^k) - F(x^*) \leq \frac{L_F}{2}\|x^k - x^*\|^2$*.*

7. Numerical experiments

We present numerical evidence to assess the theoretical convergence guarantees of the SPP algorithm. We provide three numerical examples: constrained stochastic least-square with

random generated data (Moulines and Bach, 2011; Toulis et al., 2016), Markowitz portfolio optimization using real data (Brodie et al., 2009; Yurtsever et al., 2016) and logistic regression using real data (Platt, 1998). In all our figures the results are averaged over 20 Monte-Carlo simulations for an algorithm.

7.1 Stochastic least-square problems using random data

In this section we evaluate the practical performance of the SPP schemes on finite large scale least-squares models. To do so, we follow a simple normal (constrained) linear regression example from (Moulines and Bach, 2011; Toulis et al., 2016). Let $m = 10^5$ be the number of observations, and $n = 20$ be the number of features. Let x^* be a randomly a priori chosen ground truth. The feature vectors $a_1, \cdots, a_m \approx \mathcal{N}_n(0, H)$ are i.i.d. normal random variables, and H is a randomly generated symmetric matrix with eigenvalues $1/k$, for $k = 1, \cdots, n$. The outcome b_S is sampled from a normal distribution as $b_S|a_S \approx \mathcal{N}(a_S^T x^*, 1)$, for $S = 1, \cdots, m$. Since the typical loss function is defined as the elementary squared residual $(a_S^T x - b_S)^2$, which is not strongly convex, we consider batches of residuals to form our loss functions, i.e we consider $\ell(x, S)$ of two forms:

$$\ell(x, S) = \|A_{j(S):j(S)+n} x - b_{j(S):j(S)+n}\|^2 \quad \text{or} \quad \ell(x, S) = (a_S^T x - b_S)^2,$$

where a_S is the Sth row of A and $A_{j(S):j(S)+n} \in \mathbb{R}^{n \times n}$ is a submatrix containing n rows of A so that the function $x \mapsto \|A_{j(S):j(S)+n} x - b_{j(S):j(S)+n}\|^2$ is strongly convex. In our tests we used $\texttt{round}\,(m/2n)$ batches of dimension n and we let the rest as elementary residuals, thus having in total $p = m/2 + m/n$ loss functions. Additionally, we impose on the estimator x also p linear inequality constraints $\{x \mid Cx \leq d\}$. This constraints can be found in many applications and they come from physical constraints, see e.g. (Censor et al., 2012; Rosasco et al., 2014). We choose randomly the matrix C for the constraints and $d = C \cdot x^* + [0\ 0\ 0\ v^T]^T$, where $v \geq 0$ is a random vector of appropriate dimension, i.e. three inequalities are active at the solution x^*. Besides the SPP and RSPP algorithms analyzed in the previous sections of our paper, we also implemented SGD and the averaged variant of SPP algorithm (A-SPP), which has the same SPP iteration, but outputs the average of iterates: $\hat{x}^k = (1/\sum_{i=1}^k \mu_i) \sum_{i=1}^k \mu_i x_i$. Convergence behavior of the averaged iterates of stochastic gradient has been initially proposed in the seminal paper (Polyak and Juditsky, 1992).

In Figure 1 we run algorithms SPP, RSPP, A-SPP and SGD for two values of the initial stepsize: $\mu_0 = 0.5$ and $\mu_0 = 1$. Each scheme runs for two stepsize exponents: $\gamma_1 = 1$ (left) and $\gamma_2 = 1/2$ (right). From Figure 1 we can asses one conclusion of Theorem 15: the best performance for SPP is achieved for stepsize exponent $\gamma = 1$. Moreover, we can observe that algorithm RSPP has the fastest behavior, while the averaged variant A-SPP is more robust to changes in the initial stepsize μ_0. The performance of SGD is much worse as exponent γ decreases and it is also sensitive to the learning rate μ_0. Notice that both tests are performed over m iterations (i.e. one pass through data).

In the second set of experiments, we generate random least-square problems of the form $\min_{x:Cx \leq d}\ 1/2\|Ax - b\|^2$, where both matrices A and C have $m = 10^3$ rows and generated randomly. Now, we do not impose the solution x^* to have the form given in the first test. We let SPP and RSPP algorithms to do one pass through data for various stepsize exponents

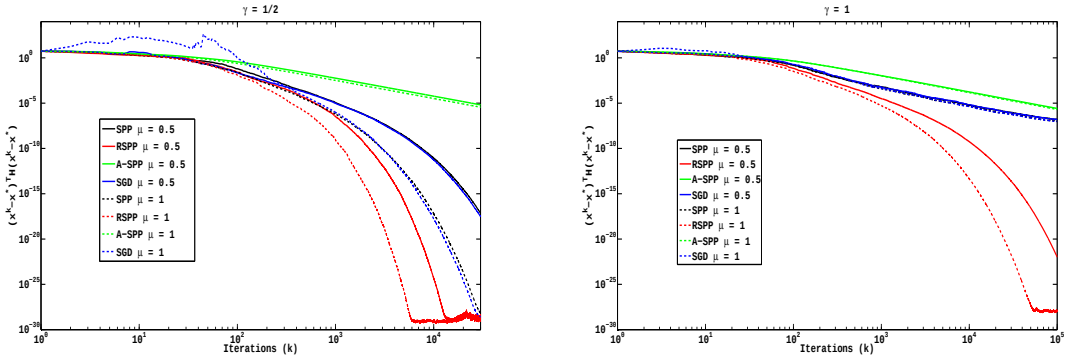

Figure 1: Performance comparison of SPP, A-SPP, RSPP and SGD for two values of initial stepsize $\mu_0 = 0.5$ and $\mu_0 = 1$ and for two values of exponent $\gamma = 1/2$ (left) and $\gamma = 1$ (right).

γ. From Figure 2 we can assess the empirical evidence of the $\mathcal{O}(1/\epsilon^{1/\gamma})$ convergence rate of Theorem 15 for SPP and $\mathcal{O}(1/\epsilon^{1+1/\gamma})$ convergence rate of Theorem 19 for RSPP, by presenting squared relative distance to the optimum solution. Moreover, the simulation results match other conclusions of Theorems 15 and 19 regarding the stepsize exponent γ: (i) the performance of SPP deteriorates with the decrease in the value of the stepsize exponent γ; (ii) from our preliminary numerical experiments we observed that RSPP scheme runs faster for higher values of γ and it has a more robust performance with respect to the variation of γ than SPP algorithm.

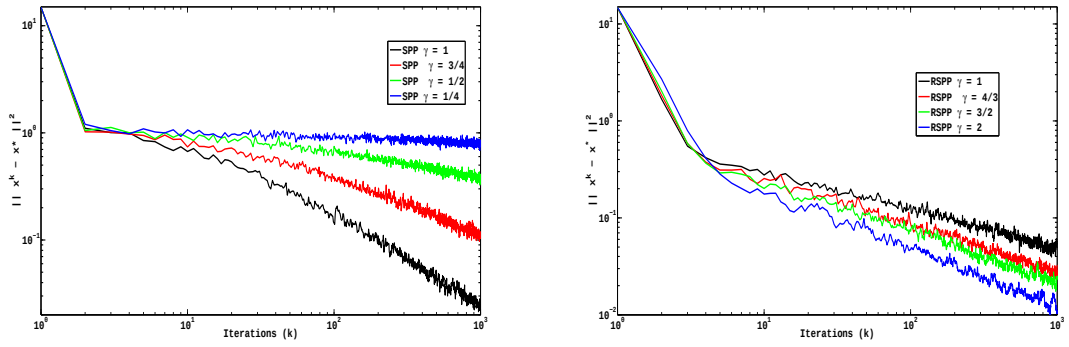

Figure 2: Performance of: SPP for four values of the stepsize exponent $\gamma = 1,\ 3/4,\ 1/2$ and $1/4$ (left); RSPP for four values of the stepsize exponent $\gamma = 1,\ 4/3,\ 3/2$ and 2 (right).

7.2 Markowitz portfolio optimization using real data

Markowitz portfolio optimization aims to reduce the risk by minimizing the variance for a given expected return. This can be mathematically formulated as a convex optimization

problem (Brodie et al., 2009; Yurtsever et al., 2016):

$$\min_{x\in\mathbb{R}^n} \mathbb{E}[(a_S^T x - b)^2] \quad \text{s.t.} \quad x \in X = \{x:\ x \geq 0,\ e^T x \leq 1,\ a_{av}^T x \geq b\},$$

where $a_{av} = \mathbb{E}[a_S]$ is the average returns for each asset that is assumed to be known (or estimated), and b represents a minimum desired return. Since new data points are arriving on-line, one cannot access the entire dataset at any moment of time, which makes the stochastic setting more favorable. For simulations, we approximate the expectation with the empirical mean as follows:

$$\min_{x\in\mathbb{R}^n} \frac{1}{m}\sum_{S=1}^{m}(a_S^T x - b)^2 \quad \text{s.t.} \quad x \in X = X_1 \cap X_2 \cap X_3,$$

where $X_1 = \{x:\ x \geq 0\}$, $X_2 = \{x:\ e^T x \leq 1\}$ and $X_3 = \{x:\ a_{av}^T x \geq b\}$. In this application we have the number of samples m larger than the number of constraints. However, by taking a certain partition of $[m] = \Omega_1 \cup \Omega_2 \cup \Omega_3$, then one can consider: $X_S = X_i$ for all $S \in \Omega_i$, with $i \in \{1, 2, 3\}$. We use 2 different real portfolio datasets: Standard & Poor's 500 (SP500, with 25 stocks for 1276 days) and one dataset by Fama and French (FF100, with 100 portfolios for 23.647 days) that is commonly used in financial literature, see e.g. (Brodie et al., 2009). We split all the datasets into test (10%) and train (90%) partitions randomly. We set the desired return a_{av} as the average return over all assets in the training set and $b = \text{mean}(a_{av})$. The results of this experiment are presented in Figure 3. We plot the value of the objective function over the datapoints in the test partition F_{test} along the iterations. We observe that SGD is very sensitive to both parameters, initial stepsize (μ_0) and stepsize exponent (γ), while SPP is more robust to changes in both parameters and also performs better over one pass through data in the train partition.

7.3 Logistic regression using real data

Finally, we consider the logistic regression problem. In this task we train an estimator over a given dataset (A, b), where $A \in \mathbb{R}^{m\times n}$ is the observations matrix and $b \in \mathbb{R}^m$ is the labels vector. For any $S \in \{1, \cdots, m\}$ we define the logistic loss function:

$$\ell(a_S^T x) = \log\left(1 + e^{-b_S(a_S^T x)}\right),$$

where $a_S \in \mathbb{R}^n$ is the Sth row of matrix A. Notice that the logistic loss function $\ell(a_S^T x)$ is only convex and smooth. However, in logistic regression we also consider a quadratic regularization term (Toulis et al., 2016; Bach, 2010):

$$\min_{x\in\mathbb{R}^n} \frac{1}{m}\sum_{S=1}^{m}\log\left(1 + e^{-b_S(a_S^T x)}\right) + \frac{\lambda}{2}\|x\|^2,$$

where $\lambda > 0$ is taken small, which makes the objective function λ-strongly convex. We have tested the four schemes (SGD, SPP, ASPP and RSPP), on the Adult datasets (a2a with $m = 2265, n = 123$ and a5a with $m = 6414, n = 123$) from LIBSVM/UCI database (Platt, 1998). We set the initial stepsize at value $\mu_0 = 0.6$ and the regularization parameter

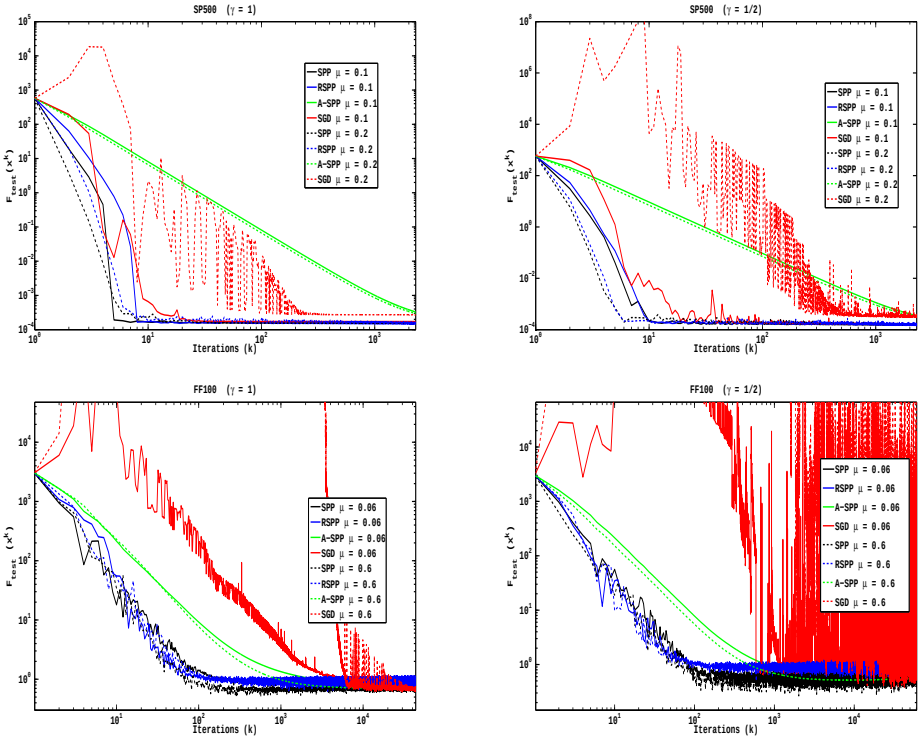

Figure 3: Performance on Markowitz portfolio using real datasets (SP500 - top and FF100 - bottom) for the SPP, A-SPP, RSPP and SGD schemes for several values of the initial stepsize μ_0 and for two values of the exponent ($\gamma = 1/2$ - left and $\gamma = 1$ - right).

$\lambda = 10^{-3}$. Once an approximate solution $\tilde{x}^*$ of the logistic regression problem is obtained, we evaluate the resulted estimator on the test dataset, i.e. $\frac{1}{2p}\sum_{S=1}^{p} |sgn(\tilde{a}_S^T \tilde{x}^*) - \tilde{b}_S|$, where $\tilde{A} \in \mathbb{R}^{p \times n}$ and $\tilde{b}$ are the testing dataset. The results are displayed in Figure 4. We observe that for large stepsize ($\gamma = 1/2$) the performances of all four methods (SGD, SPP, A-SPP and RSPP) are similar. However, when we use a smaller stepsizes ($\gamma = 1$), the RSPP algorithm outperforms the other methods. We also observe that the variation of stepsize exponent γ does not influence too much the performance of RSPP algorithm, showing once more the robustness of this scheme against variations in the stepsize choices μ_0/k^γ.

Since we used different parameter values in our experiments, we want to provide some details on the parameter choices. From the theoretical viewpoint, Theorems 15 and 19 show that the stepsize exponent γ has to be chosen as large as possible to obtain the best convergence rate. Let us consider for simplicity that $\sigma_{f,S} = \sigma > 0$ for all $S \in \Omega$. Then, for the initial stepsize μ_0 Corollary 16 indicates that the best convergence rate is obtained for $\mu_0 > \frac{\sqrt{e}-1}{\sigma}$. Therefore, in the case when σ is known (e.g. regularized logistic regression) we can choose μ_0 appropriately so that we obtain the best convergence. However, when this parameter σ is not known, then there is an inherent need for parameter tuning. From practical point of view, our plots show that the performance of SPP/RSPP deteriorates with the decrease

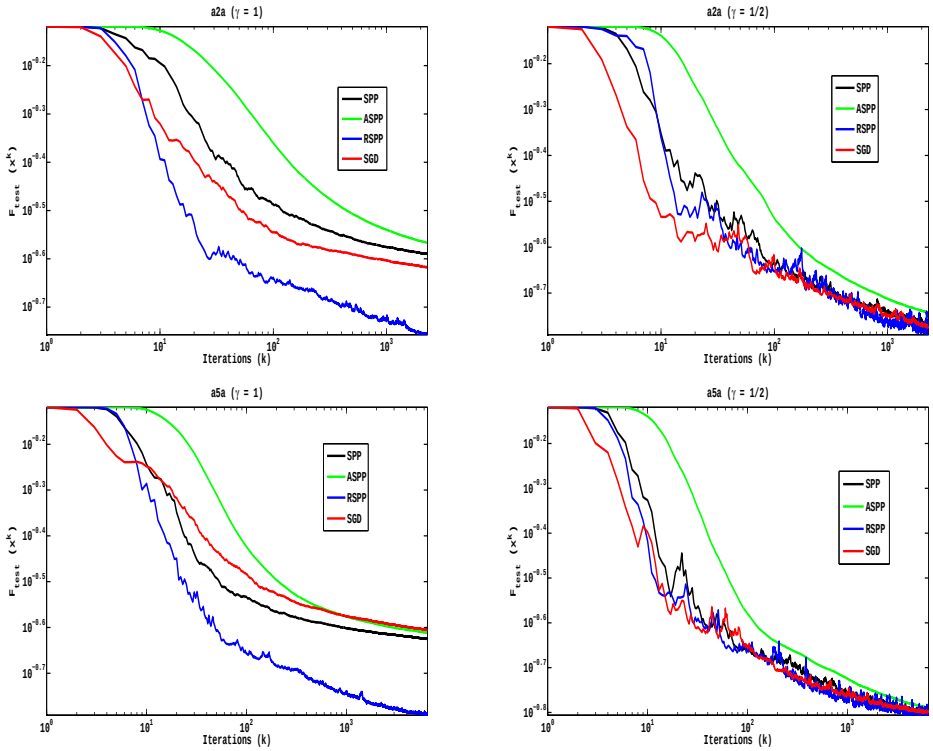

Figure 4: Performance on logistic regression using real datasets (a2a - top and a5a - bottom) for the SPP, A-SPP, RSPP and SGD schemes for two values of the exponent ($\gamma = 1/2$ - left and $\gamma = 1$ - right).

in value of the exponent γ. That is, they indicate that the higher values of this parameter the better performance. However, there is empirical evidence in the literature regarding the SGD performance which shows that values $\gamma < 1$ provide a better performance than the choice $\gamma = 1$. In these cases, the choice of initial stepsize μ_0 requires a detailed tuning procedure. As a general conclusion of our experiments, we can state that both parameters μ_0 and γ are strongly linked to the problem conditioning and, up to some extent, they have to be tuned accordingly to the problem datasets.

8. Appendix

To make the paper more readable, in this Appendix we provide the proofs of some lemmas and theorems.

Proof of Lemma 4:

Proof By using the convexity of the function $\mathbb{I}_{\mu,S}(x) = \frac{1}{2\mu}\text{dist}^2_{X_S}(x)$ and taking the conditional expectation w.r.t. S_k over the history $\mathcal{F}_{k-1} = \{S_0, \cdots, S_{k-1}\}$, we get:

$$\mathbb{E}[\mathbb{I}_{1,S_k}(\hat{y}^k)|\mathcal{F}_{k-1}] \geq \mathbb{E}\left[\mathbb{I}_{1,S_k}(\hat{x}^k) + \langle \nabla \mathbb{I}_{1,S_k}(\hat{x}^k), \hat{y}^k - \hat{x}^k\rangle |\mathcal{F}_{k-1}\right].$$

Taking further the expectation over $\mathcal{F}_{k-1}$ we obtain:

$$\begin{aligned}\mathbb{E}[\mathbb{I}_{1,S_k}(\hat{y}^k)] &\geq \mathbb{E}\left[\mathbb{I}_{1,S_k}(\hat{x}^k)\right] + \mathbb{E}\left[\langle \nabla \mathbb{I}_{1,S_k}(\hat{x}^k), \hat{y}^k - \hat{x}^k\rangle\right] \\ &= \mathbb{E}\left[\mathbb{I}_{1,S_k}(\hat{x}^k)\right] + \frac{\mathbb{E}\left[\langle \nabla \mathbb{I}_{1,S_k}(\hat{x}^k), \sum_{i=0}^{k-1} \mu_i^2 \nabla f_{\mu_i}(x^i;S_i)\rangle\right]}{\hat{\mu}_{1,k}} \\ &\geq \mathbb{E}\left[\mathbb{I}_{1,S_k}(\hat{x}^k)\right] - \mathbb{E}\left[\frac{\hat{\mu}_{2,k}}{\hat{\mu}_{1,k}} \|\nabla \mathbb{I}_{1,S_k}(\hat{x}^k)\| \left\| \sum_{i=0}^{k-1} \frac{\mu_i^2}{\hat{\mu}_{2,k}} \nabla f_{\mu_i}(x^i;S_i)\right\|\right] \\ &\geq \mathbb{E}\left[\mathbb{I}_{1,S_k}(\hat{x}^k)\right] - \frac{\hat{\mu}_{2,k}}{\hat{\mu}_{1,k}} \mathbb{E}\left[\|\nabla \mathbb{I}_{1,S_k}(\hat{x}^k)\| \sum_{i=0}^{k-1} \frac{\mu_i^2}{\hat{\mu}_{2,k}} \left\|\nabla f_{\mu_i}(x^i;S_i)\right\|\right],\end{aligned}$$

where in the second inequality we used the Cauchy-Schwarz inequality and in the third the convexity relation regarding $\|\cdot\|$. Further, using as well Lemma 3, Assumption 2 and Cauchy-Schwarz inequality, we have:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{2}\mathrm{dist}^2_{X_{S_k}}(\hat{y}^k)\right] &\overset{\text{Lemma 3}}{\geq} \mathbb{E}\left[\frac{1}{2}\mathrm{dist}^2_{X_{S_k}}(\hat{x}^k)\right] - \frac{\hat{\mu}_{2,k}}{2\hat{\mu}_{1,k}} \mathbb{E}\left[\mathrm{dist}_{X_{S_k}}(\hat{x}^k) L_{f,S_k}\right] \\ &\overset{\text{Assump. 2}}{\geq} \frac{1}{2\zeta}\mathbb{E}\left[\mathrm{dist}^2_X(\hat{x}^k)\right] - \frac{\hat{\mu}_{2,k}}{2\hat{\mu}_{1,k}} \sqrt{\mathbb{E}[\mathrm{dist}^2_X(\hat{x}^k)]}\sqrt{\mathbb{E}[L^2_{f,S}]},\end{aligned}$$

which proves the statement of the lemma. ■

Proof of Theorem 5:
Proof Since the function $z \to f(z;S) + \frac{1}{2\mu}\|z-x\|^2$ is strongly convex, we have:

$$\begin{aligned}f(z;S) + \frac{1}{2\mu}\|z-x\|^2 &\geq f(z_\mu(x;S);S) + \frac{1}{2\mu}\|z_\mu(x;S)-x\|^2 + \frac{1}{2\mu}\|z_\mu(x;S)-z\|^2 \\ &= f_\mu(x;S) + \frac{1}{2\mu}\|z_\mu(x;S)-z\|^2 \quad \forall z \in \mathbb{R}^n. \end{aligned} \tag{16}$$

By taking $x = x^k, S = S_k, z = x^*, \mu = \mu_k$ in (16) and using the strictly nonexpansive property of the projection operator, see e.g. (Nedic, 2011):

$$\|x - \Pi_{X_{S_k}}(x)\|^2 \leq \|x-z\|^2 - \|z - \Pi_{X_{S_k}}(x)\|^2 \qquad \forall z \in X_{S_k},\ x \in \mathbb{R}^n, \tag{17}$$

then these lead to:

$$\begin{aligned}f(x^*;S_k) + \frac{1}{2\mu_k}\|x^k - x^*\|^2 &\geq f_{\mu_k}(x^k;S_k) + \frac{1}{2\mu_k}\|y^k - x^*\|^2 \\ &\overset{(17)}{\geq} f_{\mu_k}(x^k;S_k) + \frac{1}{2\mu_k}\|\Pi_{X_{S_k}}(y^k) - x^*\|^2 + \frac{1}{2\mu_k}\|y^k - \Pi_{X_{S_k}}(y^k)\|^2 \\ &= f_{\mu_k}(x^k;S_k) + \frac{1}{2\mu_k}\|x^{k+1} - x^*\|^2 + \frac{1}{2\mu_k}\|y^k - x^{k+1}\|^2, \end{aligned} \tag{18}$$

where in the second inequality we used (17) with $x = y^k$ and $z = x^*$. For simplicity we denote $\mathbb{I}_{\mu,S}(x) = \frac{1}{2\mu}\|x - \Pi_{X_S}(x)\|^2$. From relation (18), it can be easily seen that:

$$\begin{aligned}
&\mu_k(f(x^k;S_k) - f(x^*;S_k)) + \mathbb{I}_{1,S_k}(y^k) - \frac{\mu_k^2}{2}L_{f,S_k}^2 \\
&\leq \mu_k(f(x^k;S_k) - f(x^*;S_k)) + \mathbb{I}_{1,S_k}(y^k) - \frac{\mu_k^2}{2}\|\nabla f(x^k;S_k)\|^2 \\
&= \mu_k(f(x^k;S_k) - f(x^*;S_k)) + \mathbb{I}_{1,S_k}(y^k) + \min_{z\in\mathbb{R}^n}\left[\mu_k\langle\nabla f(x^k;S_k), z - x^k\rangle + \frac{1}{2}\|z - x^k\|^2\right] \\
&\leq \mu_k(f(x^k;S_k) - f(x^*;S_k)) + \mathbb{I}_{1,S_k}(y^k) + \mu_k\langle\nabla f(x^k;S_k), y^k - x^k\rangle + \frac{1}{2}\|y^k - x^k\|^2 \\
&= \mu_k(f(x^k;S_k) + \langle\nabla f(x^k;S_k), y^k - x^k\rangle + \frac{1}{2\mu_k}\|y^k - x^k\|^2 - f(x^*;S_k)) + \mathbb{I}_{1,S_k}(y^k) \\
&\overset{\text{conv. f}}{\leq} \mu_k(f_{\mu_k}(x^k;S_k) - f(x^*;S_k)) + \mathbb{I}_{1,S_k}(y^k) \\
&\overset{(18)}{\leq} \frac{1}{2}\|x^k - x^*\|^2 - \frac{1}{2}\|x^{k+1} - x^*\|^2.
\end{aligned}$$

Taking now the conditional expectation in S_k w.r.t. the history $\mathcal{F}_{k-1} = \{S_0, \cdots, S_{k-1}\}$ in the last inequality we have:

$$\begin{aligned}
\mu_k(F(x^k) - F(x^*)) + \mathbb{E}[\mathbb{I}_{1,S_k}(y^k)|\mathcal{F}_{k-1}] - \frac{\mu_k^2}{2}\mathbb{E}[L_{f,S_k}^2] \\
\leq \frac{1}{2}\|x^k - x^*\|^2 - \frac{1}{2}\mathbb{E}[\|x^{k+1} - x^*\|^2|\mathcal{F}_{k-1}].
\end{aligned}$$

Taking further the expectation over $\mathcal{F}_{k-1}$ and summing over $i = 0, \cdots, k-1$, results in:

$$\begin{aligned}
\frac{\|x^0 - x^*\|^2}{2\sum\limits_{i=0}^{k-1}\mu_i} &\geq \frac{1}{\sum\limits_{i=0}^{k-1}\mu_i}\sum_{i=0}^{k-1}\mathbb{E}[\mu_i(F(x^i) - F(x^*))] + \mathbb{E}[\mathbb{I}_{1,S}(y^i)] - \frac{\mu_i^2}{2}\mathbb{E}[L_{f,S}^2] \\
&= \frac{1}{\sum\limits_{i=0}^{k-1}\mu_i}\sum_{i=0}^{k-1}\mathbb{E}[\mu_i(F(x^i) - F(x^*))] + \mu_i\mathbb{E}[\mathbb{I}_{\mu_i,S}(y^i)] - \frac{\mu_i^2}{2}\mathbb{E}[L_{f,S}^2] \\
&\geq \frac{1}{\sum\limits_{i=0}^{k-1}\mu_i}\sum_{i=0}^{k-1}\mathbb{E}[\mu_i(F(x^i) - F(x^*))] + \mu_i\mathbb{E}[\mathbb{I}_{\mu_0,S}(y^i)] - \frac{\mu_i^2}{2}\mathbb{E}[L_{f,S}^2] \\
&\overset{\text{Jensen}}{\geq} \mathbb{E}[F(\hat{x}^k) - F(x^*)] + \mathbb{E}[\mathbb{I}_{\mu_0,S}(\hat{y}^k)] - \frac{\mathbb{E}[L_{f,S}^2]\hat{\mu}_{2,k}}{2\hat{\mu}_{1,k}},
\end{aligned} \tag{19}$$

where in the second inequality we used that $\mathbb{I}_{\mu_i,S}(y) \geq \mathbb{I}_{\mu_0,S}(y)$ for all $S \in \Omega, i \geq 0$. The relation (19) implies the following upper bound on the suboptimality gap:

$$\mathbb{E}[F(\hat{x}^k) - F(x^*)] \leq \frac{\|x^0 - x^*\|^2 + \mathbb{E}[L_{f,S}^2]\hat{\mu}_{2,k}}{2\hat{\mu}_{1,k}}. \tag{20}$$

On the other hand, recalling $\nabla F(x^*) = \mathbb{E}[\nabla f(x^*;S)]$, we use the following fact:

$$\begin{aligned}\mathbb{E}[F(\hat{x}^k)] - F(x^*) &\geq \mathbb{E}[\langle \nabla F(x^*), \hat{x}^k - x^* \rangle] \\ &= \mathbb{E}[\langle \nabla F(x^*), \Pi_X(\hat{x}^k) - x^* \rangle] + \mathbb{E}[\langle \nabla F(x^*), \hat{x}^k - \Pi_X(\hat{x}^k) \rangle] \\ &\geq -\mathbb{E}[L_{f,S}]\mathbb{E}[\mathrm{dist}_X(\hat{x}^k)] \\ &\overset{\text{Jensen}}{\geq} -\sqrt{\mathbb{E}[L_{f,S}^2]\ \mathbb{E}[\mathrm{dist}_X^2(\hat{x}^k)]} \qquad \forall k \geq 0, \end{aligned} \tag{21}$$

which is derived from the optimality conditions $\langle \nabla F(x^*), z - x^* \rangle \geq 0$ for all $z \in X$, the Cauchy-Schwarz and Jensen inequalities. By denoting $r_0 = \|x^0 - x^*\|$ and combining (19) with Lemma 4 and the last inequality (21), we obtain:

$$\begin{aligned}\mathbb{E}[\mathrm{dist}_X^2(\hat{x}^k)] - \zeta\sqrt{\mathbb{E}[L_{f,S}^2]}\left(\frac{\hat{\mu}_{2,k}}{\hat{\mu}_{1,k}} + 2\mu_0\right)\sqrt{\mathbb{E}[\mathrm{dist}_X^2(\hat{x}^k)]} \\ \overset{\text{Lemma 4+(21)}}{\leq} 2\mu_0\zeta\mathbb{E}[F(\hat{x}^k) - F(x^*)] + 2\mu_0\zeta\mathbb{E}[\mathbb{I}_{\mu_0,S}(\hat{y}^k)] \\ \overset{(20)}{\leq} \frac{\mu_0\zeta r_0^2 + \mu_0\zeta\mathbb{E}[L_{f,S}^2]\hat{\mu}_{2,k}}{\hat{\mu}_{1,k}}.\end{aligned}$$

This last relation clearly implies an upper bound on the feasibility residual:

$$\sqrt{\mathbb{E}[\mathrm{dist}_X^2(\hat{x}^k)]} \leq \zeta\sqrt{\mathbb{E}[L_{f,S}^2]}\left(\frac{\hat{\mu}_{2,k}}{\hat{\mu}_{1,k}} + 2\mu_0\right) + \sqrt{\frac{\mu_0\zeta r_0^2 + \mu_0\zeta\mathbb{E}[L_{f,S}^2]\hat{\mu}_{2,k}}{\hat{\mu}_{1,k}}}. \tag{22}$$

Also, combining (21) and (22) we obtain the lower bound on the suboptimality gap:

$$\mathbb{E}[F(\hat{x}^k)] - F^* \geq -\zeta\mathbb{E}[L_{f,S}^2]\left(\frac{\hat{\mu}_{2,k}}{\hat{\mu}_{1,k}} + 2\mu_0\right) - \sqrt{\mathbb{E}[L_{f,S}^2]}\sqrt{\frac{\mu_0\zeta r_0^2 + \mu_0\zeta\mathbb{E}[L_{f,S}^2]\hat{\mu}_{2,k}}{\hat{\mu}_{1,k}}}. \tag{23}$$

From the upper and lower suboptimality bounds (20), (23) and feasibility bound (22), we deduce our convergence rate results. ■

Proof of Lemma 9:

Proof Let $\sigma_{f,S} \geq 0$ be the strong convexity constant of the function $f(\cdot;S)$. Notice that we allow the convex case, that is $\sigma_{f,S} = 0$ for some S. Then, it is known that the Moreau approximation $f_\mu(\cdot;S)$ is also a $\hat{\sigma}_{f,S}$-strongly convex function with strong convexity constant, see e.g. (Rockafellar and Wets, 1998):

$$\hat{\sigma}_{f,S} = \frac{\sigma_{f,S}}{1 + \mu\sigma_{f,S}}.$$

Clearly, in the simple convex case, that is $\sigma_{f,S} = 0$, we also have $\hat{\sigma}_{f,S} = 0$. By denoting $\hat{L}_{f,S} = \frac{1}{\mu}$ the Lipschitz constant of the gradient of $f_\mu(\cdot;S)$, the following well-known relation holds for the smooth and (strongly) convex function $f_\mu(\cdot;S)$, see e.g. (Nesterov, 2004):

$$\begin{aligned}\langle \nabla f_\mu(x;S) - \nabla f_\mu(y;S), x - y \rangle \geq \frac{1}{\hat{\sigma}_{f,S} + \hat{L}_{f,S}}\|\nabla f_\mu(x;S) - \nabla f_\mu(y;S)\|^2 \\ + \frac{\hat{\sigma}_{f,S}\hat{L}_{f,S}}{\hat{L}_{f,S} + \hat{\sigma}_{f,S}}\|x - y\|^2 \quad \forall x, y \in \mathbb{R}^n.\end{aligned} \tag{24}$$

By using Assumption 7, then it can be also obtained that:

$$\|\nabla f_\mu(x;S) - \nabla f_\mu(y;S)\| \geq \hat{\sigma}_{f,S}\|x-y\| \qquad \forall x, y \in \mathbb{R}^n. \tag{25}$$

Using this relation, we further derive that:

$$\begin{aligned}
&\|z_\mu(x;S) - z_\mu(y;S)\|^2 = \|x - y + \mu(\nabla f_\mu(y;S) - \nabla f_\mu(x;S))\|^2 \\
&= \|x-y\|^2 + 2\mu\langle \nabla f_\mu(y;S) - \nabla f_\mu(x;S), x-y\rangle + \mu^2\|\nabla f_\mu(x;S) - \nabla f_\mu(y;S)\|^2 \\
&\overset{(24)}{\leq} \left(1 - \frac{2\mu\hat{\sigma}_{f,S}\hat{L}_{f,S}}{\hat{L}_{f,S}+\hat{\sigma}_{f,S}}\right)\|x-y\|^2 + \mu\left(\mu - \frac{2}{\hat{L}_{f,S}+\hat{\sigma}_{f,S}}\right)\|\nabla f_\mu(x;S) - \nabla f_\mu(y;S)\|^2 \\
&\overset{(25)}{\leq} \left[1 + \hat{\sigma}_{f,S}^2\left(\mu^2 - \frac{2\mu}{\hat{\sigma}_{f,S}+\hat{L}_{f,S}}\right) - \frac{2\mu\hat{\sigma}_{f,S}\hat{L}_{f,S}}{\hat{L}_{f,S}+\hat{\sigma}_{f,S}}\right]\|x-y\|^2 \\
&= (1-\hat{\sigma}_{f,S}\mu)^2\|x-y\|^2 \qquad \forall x, y \in \mathbb{R}^n,
\end{aligned}$$

which implies our result. ■

Proof of Lemma 11:
Proof By taking $\mu = \mu_k$ in relation (14), we obtain:

$$\sqrt{\mathbb{E}[\|x^{k+1} - x^*\|^2]} \leq \sqrt{\mathbb{E}\left[\theta_S^2(\mu_k)\right]}\sqrt{\mathbb{E}[\|x^k - x^*\|^2]} + \mu_k\sqrt{\mathbb{E}\left[\|\nabla f(x^*;S)\|^2\right]}.$$

By using the notations $r_k = \sqrt{\mathbb{E}[\|x^k - x^*\|^2]}, \theta_k = \sqrt{\mathbb{E}\left[\theta_S^2(\mu_k)\right]}$ and $\eta = \sqrt{\mathbb{E}\left[\|\nabla f(x^*;S)\|^2\right]}$, the last inequality leads to:

$$\begin{aligned}
r_{k+1} &\leq \theta_k r_k + (1-\theta_k)\frac{\mu_k}{1-\theta_k}\eta \\
&\leq \max\left\{r_k, \frac{\mu_k}{1-\theta_k}\eta\right\} \leq \max\left\{r_0, \frac{\mu_0}{1-\theta_0}\eta, \cdots, \frac{\mu_k}{1-\theta_k}\eta\right\}.
\end{aligned} \tag{26}$$

By observing the fact that $t \mapsto \mathbb{E}\left[\frac{\sigma_{f,S}}{(1+t\sigma_{f,S})^2} + \frac{\sigma_{f,S}}{1+t\sigma_{f,S}}\right]$ is nonincreasing in t, and implicitly:

$$\begin{aligned}
\frac{\mu_{k-1}}{1-\theta_{k-1}} &= \frac{1}{\mathbb{E}\left[\frac{\sigma_{f,S}}{(1+\mu_{k-1}\sigma_{f,S})^2} + \frac{\sigma_{f,S}}{1+\mu_{k-1}\sigma_{f,S}}\right]} \\
&\geq \frac{1}{\mathbb{E}\left[\frac{\sigma_{f,S}}{(1+\mu_k\sigma_{f,S})^2} + \frac{\sigma_{f,S}}{1+\mu_k\sigma_{f,S}}\right]} = \frac{\mu_k}{1-\theta_k},
\end{aligned}$$

then we have $\max\limits_{0\leq i\leq k}\frac{\mu_i}{1-\theta_i} = \frac{\mu_0}{1-\theta_0}$ and the relation (26) becomes:

$$r_k \leq \max\left\{r_0, \frac{\mu_0}{1-\theta_0}\eta\right\} \quad \forall k \geq 0, \tag{27}$$

which implies our result. ■

We also present the following useful auxiliary result:

Lemma 20 *Let $\gamma \in (0,1]$ and the integers $p, q \in \mathbb{N}$ with $q \geq p \geq 1$. Given the sequence of stepsizes $\mu_k = \frac{\mu_0}{k^\gamma}$ for all $k \geq 1$, where $\mu_0 > 0$, then the following relation holds:*

$$\prod_{i=p}^{q} \mathbb{E}[\theta_S^2(\mu_i)] \leq \left(\mathbb{E}\left[\theta_S^2(\mu_0)\right]\right)^{\varphi_{1-\gamma}(q+1)-\varphi_{1-\gamma}(p)}$$

Proof From definition of $\theta_S(\mu)$ for any $k \geq 1$ we have:

$$\begin{aligned}\mathbb{E}[\theta_S^2(\mu_k)] &= \mathbb{E}\left[\left(\frac{1}{1+\mu_k\sigma_{f,S}}\right)^2\right] = \mathbb{E}\left[\frac{1}{\left(1+\frac{\mu_0}{k^\gamma}\sigma_{f,S}\right)^2}\right] \\ &\overset{(7)}{\leq} \mathbb{E}\left[\left(\frac{1}{1+\mu_0\sigma_{f,S}}\right)^{\frac{2}{k^\gamma}}\right] \leq \left(\mathbb{E}\left[\frac{1}{(1+\mu_0\sigma_{f,S})^2}\right]\right)^{\frac{1}{k^\gamma}} = \left(\mathbb{E}[\theta_S^2(\mu_0)]\right)^{\frac{1}{k^\gamma}}. \end{aligned} \tag{28}$$

By taking into account that $\mathbb{E}[\theta_S^2(\mu_0)] = \mathbb{E}\left[\frac{1}{(1+\mu_0\sigma_{f,S})^2}\right] \leq 1$ and that

$$\sum_{i=p}^{q} \frac{1}{i^\gamma} \geq \varphi_{1-\gamma}(q+1) - \varphi_{1-\gamma}(p) = \int_p^{q+1} \frac{1}{t^\gamma} dt = \begin{cases} \ln\frac{q+1}{p} & \text{if} \quad \gamma = 1 \\ \frac{(q+1)^{1-\gamma}-p^{1-\gamma}}{1-\gamma} & \text{if} \quad \gamma < 1, \end{cases}$$

then the relation (28) implies:

$$\begin{aligned}\prod_{i=p}^{q} \mathbb{E}[\theta_S^2(\mu_i)] &\leq \left(\mathbb{E}\left[\theta_S^2(\mu_0)\right]\right)^{\sum_{i=p}^{q}\frac{1}{i^\gamma}} \leq \left(\mathbb{E}\left[\theta_S^2(\mu_0)\right]\right)^{\varphi_{1-\gamma}(q+1)-\varphi_{1-\gamma}(p)} \\ &= \begin{cases} \left(\mathbb{E}\left[\theta_S^2(\mu_0)\right]\right)^{\ln\frac{q+1}{p}} & \text{if} \quad \gamma = 1 \\ \left(\mathbb{E}\left[\theta_S^2(\mu_0)\right]\right)^{\frac{(q+1)^{1-\gamma}-p^{1-\gamma}}{1-\gamma}} & \text{if} \quad \gamma < 1, \end{cases} \end{aligned} \tag{29}$$

which immediately implies the above statement. ■

Proof of Lemma 13:

Proof By using the strictly nonexpansive property of the projection operator (17), with $z = \Pi_X(y^k), x = y^k$, and the linear regularity assumption, we obtain:

$$\begin{aligned}\mathbb{E}[\text{dist}_X^2(x^{k+1})] &\leq \mathbb{E}[\|x^{k+1} - \Pi_X(y^k)\|^2] \overset{(17)}{\leq} \mathbb{E}[\|y^k - \Pi_X(y^k)\|^2] - \mathbb{E}[\|y^k - x^{k+1}\|^2] \\ &\overset{\text{As. 2}}{\leq} \mathbb{E}[\|y^k - \Pi_X(y^k)\|^2] - \frac{1}{\zeta}\mathbb{E}[\|y^k - \Pi_X(y^k)\|^2] \\ &= \left(1 - \frac{1}{\zeta}\right)\mathbb{E}[\text{dist}_X^2(y^k)]. \end{aligned} \tag{30}$$

On the other hand, from triangle inequality and Minkowski inequality, we obtain:

$$\begin{aligned}\sqrt{\mathbb{E}[\mathrm{dist}_X^2(y^k)]} &\leq \sqrt{\mathbb{E}[\|y^k - \Pi_X(x^k)\|^2]} \leq \sqrt{\mathbb{E}[(\|y^k - x^k\| + \mathrm{dist}_X(x^k))^2]} \\ &\overset{(8)}{\leq} \sqrt{\mathbb{E}[\|z_{\mu_k}(x^k; S_k) - x^k\|^2]} + \sqrt{\mathbb{E}[\mathrm{dist}_X^2(x^k)]} \\ &= \sqrt{\mathbb{E}[\mathrm{dist}_X^2(x^k)]} + \mu_k \sqrt{\mathbb{E}[\|\nabla f_{\mu_k}(x^k; S_k)\|^2]} \\ &\overset{\text{Lemma 3}}{\leq} \sqrt{\mathbb{E}[\mathrm{dist}_X^2(x^k)]} + \mu_k \sqrt{\mathbb{E}[\|\nabla f(x^k; S_k)\|^2]} \\ &\overset{\text{Lemma 12}}{\leq} \sqrt{\mathbb{E}[\mathrm{dist}_X^2(x^k)]} + \mu_k \left(\sqrt{2\mathbb{E}[\|\nabla f(x^*; S)\|^2]} + \mathcal{A}\sqrt{2\mathbb{E}[L_{f,S}^2]} \right). \end{aligned} \tag{31}$$

For simplicity we use notations: $\alpha = \sqrt{1 - \frac{1}{\zeta}}, d_k = \sqrt{\mathbb{E}[\mathrm{dist}_X^2(x^k)]}$ and $\mathcal{B} = \sqrt{2\mathbb{E}[\|\nabla f(x^*; S)\|^2]} + \mathcal{A}\sqrt{2\mathbb{E}[L_{f,S}^2]}$. Combining (30) and (31) yields:

$$d_{k+1} \leq \alpha d_k + \alpha \mu_k \mathcal{B} \leq \alpha^{k+1} d_0 + \mathcal{B} \sum_{i=1}^{k+1} \alpha^i \mu_{k-i+1}. \tag{32}$$

Define $m = \lceil \frac{k+1}{2} \rceil$. By dividing the sum from the right side of (32) in two parts and by taking into account that $\{\mu_k\}_{k \geq 0}$ is nonincreasing, then results in:

$$\begin{aligned}\sum_{i=1}^{k+1} \alpha^i \mu_{k-i+1} &= \sum_{i=1}^{m} \alpha^i \mu_{k-i+1} + \sum_{i=m+1}^{k+1} \alpha^i \mu_{k-i+1} \\ &\leq \mu_{k-m+1} \sum_{i=1}^{m} \alpha^i + \alpha^{m+1} \sum_{i=0}^{k-m} \alpha^i \mu_{k-i-m} \\ &\leq \mu_{k-m+1} \frac{\alpha(1-\alpha^m)}{1-\alpha} + \mu_0 \alpha^{m+1} \frac{1-\alpha^{k-m+1}}{1-\alpha} \\ &\leq \mu_{k-m+1} \frac{\alpha}{1-\alpha} + \alpha^{m+1} \frac{\mu_0}{1-\alpha}. \end{aligned}$$

By using the last inequality into (32) and using the bound $\frac{\alpha}{1-\alpha} \leq 2\zeta$, then these facts imply the statement of the lemma. ■

Proof of Theorem 14:

Proof Let $\mu > 0, x \in \mathbb{R}^n$ and $S \in \Omega$, then we have:

$$\begin{aligned}&\frac{1}{2}\|z_\mu(x; S) - x^*\|^2 \\ &= \frac{1}{2}\|z_\mu(x; S) - z_\mu(x^*; S)\|^2 + \langle z_\mu(x; S) - z_\mu(x^*; S), z_\mu(x^*; S) - x^* \rangle + \frac{1}{2}\|z_\mu(x^*; S) - x^*\|^2 \\ &\leq \frac{\theta_S^2(\mu)}{2}\|x - x^*\|^2 - \mu\langle \nabla f(x^*; S), x - x^* \rangle + \langle z_\mu(x^*; S) - x^* + \mu \nabla f(x^*; S), x - x^* \rangle \\ &\qquad + \langle z_\mu(x; S) - x, z_\mu(x^*; S) - x^* \rangle - \frac{\mu^2}{2}\|\nabla f_\mu(x^*; S)\|^2. \end{aligned} \tag{33}$$

Now we take expectation in both sides and consider $x = x^k$ and $\mu = \mu_k$. We thus seek a bound for each term from the right hand side in (33). For the second term, by using the optimality conditions $\langle \nabla F(x^*), z - x^* \rangle \geq 0$ for all $z \in X$, we have:

$$\begin{aligned}
\mathbb{E}[\langle \nabla f(x^*;S), x^* - x^k \rangle] &= \mathbb{E}[\langle \nabla F(x^*), x^* - \Pi_X(x^k) \rangle] + \mathbb{E}[\langle \nabla F(x^*), \Pi_X(x^k) - x^k \rangle] \\
&\leq \mathbb{E}[\langle \nabla F(x^*), \Pi_X(x^k) - x^k \rangle] \\
&\overset{\text{C.-S.}}{\leq} \|\nabla F(x^*)\| \, \mathbb{E}[\text{dist}_X(x^k)] \leq \|\nabla F(x^*)\| \, \sqrt{\mathbb{E}[\text{dist}_X^2(x^k)]} \\
&\overset{\text{Lemma 13}}{\leq} \|\nabla F(x^*)\| \left[\left(1 - \frac{1}{\zeta}\right)^{\frac{k}{2}} \left(\text{dist}_X(x^0) + 2\mu_0 \zeta \mathcal{B}\right) + 2\mu_{k-\lceil \frac{k}{2} \rceil} \zeta \mathcal{B} \right],
\end{aligned}$$

where in the second inequality we used the Cauchy-Schwarz inequality. By using that $e^x \geq 1 + x$, for all $x \geq 0$, and the fact that $\frac{1}{k} \leq \frac{1}{k^\gamma}$ when $k \geq 1$ and $\gamma \in (0, 1]$, then the last inequality implies:

$$\begin{aligned}
\mathbb{E}[\langle \nabla f(x^*;S), x^* - x^k \rangle] &\leq \|\nabla F(x^*)\| \left[\frac{2\text{dist}_X(x^0) + 4\mu_0 \zeta \mathcal{B}}{k \ln\left(\zeta/(\zeta-1)\right)} + 2\mu_{k-\lceil \frac{k}{2} \rceil} \mathcal{B} \zeta \right] \\
&\leq \mu_k \|\nabla F(x^*)\| \left[\frac{2\text{dist}_X(x^0) + 4\mu_0 \zeta \mathcal{B}}{\mu_0 \ln\left(\zeta/(\zeta-1)\right)} + \frac{2\mu_{k-\lceil \frac{k}{2} \rceil} \mathcal{B} \zeta}{\mu_k} \right]. \qquad (34)
\end{aligned}$$

For the third term in (33) we observe from the optimality conditions for $z_{\mu_k}(x^*;S)$ that:

$$\begin{aligned}
\left\| \frac{1}{\mu_k} \left(z_{\mu_k}(x^*;S) - x^* \right) + \nabla f(x^*;S) \right\| &= \|\nabla f(z_{\mu_k}(x^*;S);S) - \nabla f(x^*;S)\| \\
&\overset{\text{As.1}}{\leq} L_{f,S} \|z_{\mu_k}(x^*;S) - x^*\| = \mu_k L_{f,S} \|\nabla f_{\mu_k}(x^*;S)\| \\
&\overset{\text{Lemma 3}}{\leq} \mu_k L_{f,S} \|\nabla f(x^*;S)\|,
\end{aligned}$$

which yields the following bound:

$$\begin{aligned}
&\langle z_{\mu_k}(x^*;S) - x^* + \mu_k \nabla f(x^*;S), x^k - x^* \rangle \leq \|z_{\mu_k}(x^*;S) - x^* + \mu_k \nabla f(x^*;S)\| \cdot \|x^k - x^*\| \\
&\leq \|\mu_k \nabla f(x^*;S) - \mu_k \nabla f(z_{\mu_k}(x^*;S);S)\| \cdot \|x^k - x^*\| \overset{\text{As.1}}{\leq} \mu_k L_{f,S} \|x^* - z_{\mu_k}(x^*;S)\| \cdot \|x^k - x^*\| \\
&\leq \mu_k^2 L_{f,S} \|\nabla f_{\mu_k}(x^*;S)\| \cdot \|x^k - x^*\| \overset{\text{Lemma 3}}{\leq} \mu_k^2 L_{f,S} \|\nabla f(x^*;S)\| \cdot \|x^k - x^*\|,
\end{aligned}$$

where in the first inequality we used the Cauchy-Schwarz. By taking expectation in both sides and using Lemma 11, we obtain the refinement:

$$\begin{aligned}
&\mathbb{E}[\langle z_{\mu_k}(x^*;S) - x^* + \mu_k \nabla f(x^*;S), x^k - x^* \rangle] \\
&\qquad = \mu_k \mathbb{E}[\langle \nabla f(x^*;S) - \nabla f(z_{\mu_k}(x^*;S);S), x^k - x^* \rangle] \\
&\qquad \leq \mu_k \mathbb{E}[\|\nabla f(x^*;S) - \nabla f(z_{\mu_k}(x^*;S);S)\| \|x^k - x^*\|] \qquad (35)
\end{aligned}$$

$$\begin{aligned}
&\overset{\text{As. 1}}{\leq} \mu_k \mathbb{E}[L_{f,S}\|x^* - z_{\mu_k}(x^*;S)\|\|x^k - x^*\|] \\
&= \mu_k \mathbb{E}[L_{f,S}\|\nabla f_{\mu_k}(x^*;S)\|\|x^k - x^*\|] \\
&\overset{\text{Lemma 3}}{\leq} \mu_k \mathbb{E}[L_{f,S}\|\nabla f(x^*;S)\|\|x^k - x^*\|] \\
&\leq \mu_k^2 \sqrt{\mathbb{E}[L_{f,S}^2]}\sqrt{\mathbb{E}[\|\nabla f(x^*;S)\|^2]}\mathbb{E}[\|x^k - x^*\|] \\
&\leq \mu_k^2 \sqrt{\mathbb{E}[L_{f,S}^2]}\sqrt{\mathbb{E}[\|\nabla f(x^*;S)\|^2]}\mathbb{E}[\|x^k - x^*\|] \\
&\overset{\text{Lemma 11}}{\leq} \mu_k^2 \sqrt{\mathbb{E}[L_{f,S}^2]}\eta\mathcal{A},
\end{aligned} \tag{36}$$

where in the first inequality we again used Cauchy-Schwarz relation and in the second we used Assumption 1. Finally, for the fourth term in (33) we use Lemma 12:

$$\begin{aligned}
&\mathbb{E}[\langle z_{\mu_k}(x^k;S) - x^k, z_{\mu_k}(x^*;S) - x^*\rangle] = \mu_k^2 \mathbb{E}[\langle \nabla f_{\mu_k}(x^k;S), \nabla f_{\mu_k}(x^*;S)\rangle] \\
&\leq \mu_k^2 \mathbb{E}[\|\nabla f_{\mu_k}(x^k;S)\|\|\nabla f_{\mu_k}(x^*;S)\|] \\
&\overset{\text{Lemma 3}}{\leq} \mu_k^2 \mathbb{E}[\|\nabla f(x^k;S)\|\|\nabla f(x^*;S)\|] \leq \mu_k^2 \sqrt{\mathbb{E}[\|\nabla f(x^k;S)\|^2]\mathbb{E}[\|\nabla f(x^*;S)\|^2]} \\
&\overset{\text{Lemma 12}}{\leq} \mu_k^2 \eta \sqrt{2\eta^2 + 2\mathbb{E}[L_{f,S}^2]\mathcal{A}^2},
\end{aligned} \tag{37}$$

where in the first inequality we used Cauchy-Schwarz. By taking expectation in (33), using the relations (34)-(37) and taking into account that $\frac{\mu_k}{\mu_{k-\lceil \frac{k}{2} \rceil}} \leq 3^\gamma$ for all $k \geq 1$, we obtain:

$$\begin{aligned}
&\mathbb{E}[\|z_{\mu_k}(x^k;S) - x^*\|^2] \\
&\leq \mathbb{E}\left[\theta_S^2(\mu_k)\|x^k - x^*\|^2\right] + 4\mu_k^2\|F(x^*)\|\left[\frac{\mathrm{dist}_X(x^0) + 2\mu_0\zeta\mathcal{B}}{\mu_0 \ln(\zeta/(\zeta-1))} + 3^\gamma\mathcal{B}\zeta\right] \\
&\qquad + 2\mu_k^2\eta\sqrt{2\eta^2 + 2\mathbb{E}[L_{f,S}^2]\mathcal{A}^2} + 2\mu_k^2\eta\mathcal{A}\sqrt{\mathbb{E}[L_{f,S}^2]} \\
&= \mathbb{E}\left[\theta_S^2(\mu_k)\right]\mathbb{E}[\|x^k - x^*\|^2] + \mu_k^2\mathcal{D}.
\end{aligned}$$

For simplicity, we use further in the proof the following notations: $r_k = \sqrt{\mathbb{E}[\|x^k - x^*\|^2]}$ and $\theta_k = \mathbb{E}[\theta_S^2(\mu_k)]$. Then, through the nonexpansiveness property of the projection operator, the previous inequality turns into:

$$\begin{aligned}
r_{k+1}^2 &\leq \mathbb{E}[\|z_{\mu_k}(x^k;S) - x^*\|^2] \leq \theta_k r_k^2 + \mu_k^2\mathcal{D} \\
&\leq r_0^2 \prod_{i=0}^{k}\theta_i + \mathcal{D}\sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}\theta_j\right)\mu_i^2.
\end{aligned} \tag{38}$$

To further refine the right hand side in (38), we first notice from Lemma 20 that we have $\prod_{i=0}^{k}\theta_i \leq \theta_0^{\varphi_{1-\gamma}(k+1)}$. Then, from (38) we can derive different upper bounds for the two cases of the parameter γ: $\gamma < 1$ and $\gamma = 1$.

Case (i) $\gamma < 1$. From Lemma 20, we derive an upper approximation for the second term in the right hand side of (38). Therefore, if we let $m = \lceil \frac{k}{2} \rceil$ we obtain:

$$\begin{aligned}
\sum_{i=0}^{k} \mu_i^2 \left(\prod_{j=i+1}^{k} \theta_j \right) &= \sum_{i=0}^{m} \mu_i^2 \left(\prod_{j=i+1}^{k} \theta_j \right) + \sum_{i=m+1}^{k} \mu_i^2 \left(\prod_{j=i+1}^{k} \theta_j \right) \\
&\overset{\text{Lemma 20}}{\leq} \sum_{i=0}^{m} \mu_i^2 \theta_0^{\varphi_{1-\gamma}(k+1)-\varphi_{1-\gamma}(i+1)} + \mu_{m+1} \sum_{i=m+1}^{k} \mu_i \left(\prod_{j=i+1}^{k} \theta_j \right) \\
&\leq \theta_0^{\varphi_{1-\gamma}(k+1)-\varphi_{1-\gamma}(m+1)} \sum_{i=0}^{m} \mu_i^2 + \mu_{m+1} \sum_{i=m+1}^{k} \mu_i \left(\prod_{j=i+1}^{k} \theta_j \right) \\
&= \theta_0^{\varphi_{1-\gamma}(k+1)-\varphi_{1-\gamma}(m+1)} \sum_{i=0}^{m} \mu_i^2 + \mu_{m+1} \sum_{i=m+1}^{k} \frac{\mu_i}{1-\theta_i} (1-\theta_i) \left(\prod_{j=i+1}^{k} \theta_j \right). \qquad (39)
\end{aligned}$$

We will further refine the right hand side of (39) by noticing the following two facts. First, the constant $\frac{\mu_i}{1-\theta_i}$ can be upper bounded by:

$$\frac{\mu_i}{1-\theta_i} = \frac{1}{\mathbb{E}\left[\frac{\sigma_S}{(1+\mu_i\sigma_S)^2} + \frac{\sigma_S}{1+\mu_i\sigma_S} \right]} \leq \frac{\mu_{i-1}}{1-\theta_{i-1}} \leq \cdots \leq \frac{\mu_0}{1-\theta_0}.$$

Second, the sum of products is upper bounded as:

$$\sum_{i=m+1}^{k} (1-\theta_i) \left(\prod_{j=i+1}^{k} \theta_j \right) = \sum_{i=m+1}^{k} \left(\prod_{j=i+1}^{k} \theta_j - \prod_{j=i}^{k} \theta_j \right) = 1 - \prod_{j=m+1}^{k} \theta_j \leq 1.$$

By using the last two inequalities into (39), we have:

$$\sum_{i=0}^{k} \mu_i^2 \left(\prod_{j=i+1}^{k} \theta_j \right) \leq \theta_0^{\varphi_{1-\gamma}(k+1)-\varphi_{1-\gamma}(m+1)} \sum_{i=0}^{m} \mu_i^2 + \mu_{m+1} \frac{\mu_0}{1-\theta_0}. \qquad (40)$$

Since $\sum_{i=0}^{m} \mu_i^2 \leq \mu_0^2(\varphi_{1-2\gamma}(m)+2) \leq \mu_0^2(\varphi_{1-2\gamma}(m)+2) \leq \mu_0^2[\varphi_{1-2\gamma}\left(\frac{k}{2}+1\right)+2]$ and using (40) into (38), we obtain the above result.

Case (ii) $\gamma = 1$. In this case we have:

$$\begin{aligned}
&\sum_{i=1}^{k} \mu_i^2 \left(\prod_{j=i+1}^{k} \theta_j \right) \overset{\text{Lemma 20}}{\leq} \sum_{i=1}^{k} \mu_i^2 \theta_0^{\varphi_0(k+1)-\varphi_0(i+1)} \\
&= \sum_{i=1}^{k} \frac{\mu_1^2}{i^2} \theta_0^{\ln \frac{k+1}{i+1}} = \sum_{i=1}^{k} \frac{\mu_1^2}{i^2} \left(\frac{k+1}{i+1} \right)^{\ln \theta_0} \leq \left(\frac{1}{k} \right)^{\ln\left(\frac{1}{\theta_0}\right)} \sum_{i=1}^{k} \frac{\mu_1^2}{i^{2-\ln\frac{1}{\theta_0}}} \\
&\leq \left(\frac{1}{k} \right)^{\ln\left(\frac{1}{\theta_0}\right)} \mu_0^2 \varphi_{\ln\frac{1}{\theta_0}-1}(k).
\end{aligned}$$

Therefore, the variation of θ_0 leads to the following cases:

$$\sum_{i=1}^{k}\mu_i^2\left(\prod_{j=i+1}^{k}\theta_j\right)\leq\begin{cases}\frac{\mu_0^2}{k\left(\ln\left(\frac{1}{\theta_0}\right)-1\right)} & \text{if } \theta_0<\frac{1}{e}\\ \frac{\mu_0^2\ln k}{k} & \text{if } \theta_0=\frac{1}{e}\\ \left(\frac{1}{k}\right)^{\ln\left(\frac{1}{\theta_0}\right)}\frac{\mu_0^2}{1-\ln\left(\frac{1}{\theta_0}\right)} & \text{if } \theta_0>\frac{1}{e},\end{cases}$$

which leads to the second part of the result. ■

Proof of Lemma 17:

Proof The proof follows similar lines with the one of Lemma 30. Therefore, by using notations: $\alpha=\sqrt{1-\frac{1}{\zeta}}$ and $d_{k,t}=\sqrt{\mathbb{E}[\mathrm{dist}_X^2(x^{k,t})]}$ results in:

$$\begin{aligned}d_{k+1,t}&\leq\alpha d_{k,t}+\alpha\mu_t\mathcal{B}\leq\alpha^{k+1}d_{0,t}+\mu_t\mathcal{B}\sum_{i=1}^{k+1}\alpha^i\\&\leq\alpha^{k+1}d_{0,t}+\mu_t\mathcal{B}\frac{\alpha}{1-\alpha}.\end{aligned}$$

By setting $k=K_t-1$, then the last inequality implies:

$$\begin{aligned}d_{K_t,t}&\leq\alpha^{K_t}d_{K_{t-1},t-1}+\mu_t\mathcal{B}\frac{\alpha}{1-\alpha}\\&\leq\alpha^{\sum_{i=1}^{t}K_i}d_{0,0}+\mathcal{B}\frac{\alpha}{1-\alpha}\sum_{j=0}^{t-1}\alpha^{\sum_{i=t-j+1}^{t}K_i}\mu_{t-j}.\end{aligned}$$

Now set $m=\lceil\frac{t}{2}\rceil$. By dividing the sum from the right side of (32) in two parts, by taking into account that $\{\mu_t\}_{t\geq0}$ is nonincreasing and $\{K_t\}_{t\geq0}$ is nondecreasing, then results in:

$$\begin{aligned}\sum_{j=0}^{t-1}\alpha^{\sum_{i=t-j+1}^{t}K_i}\mu_{t-j}&=\sum_{j=0}^{m}\alpha^{\sum_{i=t-j+1}^{t}K_i}\mu_{t-j}+\sum_{j=m+1}^{t-1}\alpha^{\sum_{i=t-j+1}^{t}K_i}\mu_{t-j}\\&\leq\mu_{t-m}\sum_{j=0}^{m}\alpha^{\sum_{i=t-j+1}^{t}K_i}+\mu_0\alpha^{K_t}\sum_{j=m+1}^{t-1}\alpha^{\sum_{i=t-j+1}^{t-1}K_i}\\&\leq\mu_{t-m}\frac{1-\alpha^{m+1}}{1-\alpha}+\mu_0\alpha^{\sum_{i=t-m}^{t}K_i}\frac{1-\alpha^{t-m+2}}{1-\alpha}\\&\leq\frac{\mu_{t-m}}{1-\alpha}+\frac{\mu_0\alpha^{\sum_{i=t-m}^{t}K_i}}{1-\alpha}.\end{aligned}$$

By using the last inequality into (32) and using the bound $\frac{\alpha}{1-\alpha}\leq2\zeta$, then these facts imply the statement of the lemma. ■

Proof of Theorem 18:

Proof First notice that from $e^x \geq 1+x$ for all $x \geq 0$, we have $\left(1-\frac{1}{\zeta}\right)^{\sum_{i=1}^{t}\frac{K_i}{2}} \leq \left(1-\frac{1}{\zeta}\right)^{\frac{K_t}{2}} \leq \frac{2}{K_t \ln(\zeta/\zeta-1)}$ and $\left(1-\frac{1}{\zeta}\right)^{\sum_{i=t-\lceil \frac{t}{2} \rceil}^{t}\frac{K_i}{2}} \leq \left(1-\frac{1}{\zeta}\right)^{\frac{K_t}{2}} \leq \frac{2}{K_t \ln(\zeta/\zeta-1)}$, which imply that Lemma 2 becomes

$$\sqrt{\mathbb{E}[\mathrm{dist}_X^2(x^{K_t,t})]} \leq \mu_t \frac{2\mathrm{dist}_X(x^{0,0})}{\mu_0 \ln(\zeta/\zeta-1)} + \mu_t \frac{4\zeta^2\mathcal{B}}{\ln(\zeta/\zeta-1)} + 2\mu_{t-\lceil \frac{t}{2} \rceil}\zeta^2\mathcal{B}. \tag{41}$$

It can be seen that by combining (41) with a similar argument as in Theorem 14 we obtain a similar descent as (38). Therefore, let $k \geq 0$ and $x^{k,t}$ be the kth iterate from the tth epoch. Then, by denoting $r_{k,t}^2 = \mathbb{E}[\|x^{k,t} - x^*\|^2]$, results in:

$$r_{k+1,t}^2 \leq \mathbb{E}[\theta_S(\mu_t)^2] r_{k,t}^2 + \mu_t^2 \mathcal{D}_r.$$

Now taking $k = K_t$ results in:

$$r_{0,t+1}^2 = r_{K_t,t}^2 \leq r_{0,t}^2 \theta_t^{K_t} + \mathcal{D}_r \mu_t^2 \sum_{i=0}^{k} \theta_t^i \leq r_{0,t}^2 \theta_t^{K_t} + \frac{\mathcal{D}_r \mu_t^2}{1-\theta_t}. \tag{42}$$

Recalling that we chose $\mu_t = \frac{\mu_0}{t^\gamma}$ and $K_t = \lceil t^\gamma \rceil$, then (7) leads to:

$$\theta_t^{K_t} \leq \left(\mathbb{E}\left[\frac{1}{(1+\mu_0\sigma_{f,S})^2}\right]\right)^{\frac{K_t}{t^\gamma}} \leq \theta_0.$$

Therefore, (42) leads to:

$$r_{0,t+1} \overset{(42)}{\leq} \theta_0 r_{0,t}^2 + \frac{\mathcal{D}_r \mu_t^2}{1-\theta_t} \leq \theta_0^t r_{0,1}^2 + \mathcal{D}_r \sum_{i=1}^{t} \frac{\mu_i^2 \theta_0^{t-i}}{1-\theta_i}. \tag{43}$$

Note that $\frac{\mu_i^2}{1-\theta_i}$ is nonincreasing in i. Then, if we fix $m = \lceil \frac{t}{2} \rceil$, then the sum $\sum_{i=1}^{t} \frac{\mu_i^2 \theta_0^{t-i}}{1-\theta_i}$ can be bounded as follows:

$$\begin{aligned}
\sum_{i=1}^{t} \frac{\mu_i^2 \theta_0^{t-i}}{1-\theta_i} &\leq \theta_0^m \sum_{i=1}^{m} \frac{\mu_i^2}{1-\theta_i} + \sum_{i=m}^{t} \frac{\mu_i^2 \theta_0^{t-i}}{1-\theta_i} \\
&\leq \theta_0^m \sum_{i=1}^{m} \frac{\mu_i^2}{1-\theta_i} + \frac{\mu_m^2}{1-\theta_m} \sum_{i=1}^{t-m} \theta_0^i \\
&\leq \frac{\theta_0^m \mu_1}{1-\theta_0} \left(\sum_{i=1}^{m} \mu_i\right) + \frac{\mu_m^2}{(1-\theta_m)(1-\theta_0)} \\
&\leq \frac{\theta_0^m \mu_1}{1-\theta_0} \left(\sum_{i=1}^{m} \mu_i\right) + \mu_m \frac{\mu_1}{(1-\theta_0)^2}.
\end{aligned} \tag{44}$$

Taking into account that $\sum_{i=1}^{m}\mu_i \leq \int_1^m \frac{1}{s^\gamma}ds \leq \frac{2^{\gamma-1}}{(1-\gamma)t^{\gamma-1}}$ and that $\theta_0^m \leq \frac{1}{1+\frac{t}{2}\ln\frac{1}{\theta_0}}$, the previous relation (44) implies:

$$\sum_{i=1}^{t}\frac{\mu_i^2\theta_0^{t-i}}{1-\theta_i} \leq \left(\frac{2}{t}\right)^{\gamma}\left[\frac{1}{2(1-\gamma)\ln 1/\sqrt{\theta_0}}+\frac{\mu_1^2}{(1-\theta_0)^2}\right]. \tag{45}$$

By using this bound in relation (44), then in order to obtain $r_{0,t+1}^2 \leq \epsilon$ it is sufficient that the number of epochs t to satisfy:

$$t \geq \max\left\{\ln\left(\frac{2r_{0,0}^2}{\epsilon}\right)\frac{1}{\ln(1/\theta_0)}, \left(\frac{2^{\gamma+1}\mathcal{D}_r}{\epsilon}\mathcal{C}\right)^{1/\gamma}\right\}. \tag{46}$$

Finally, the total number of SPP iterations performed by RSPP algorithm satisfies:

$$\begin{aligned}\sum_{i=1}^{t}K_t &\geq \sum_{i=1}^{t}i^\gamma \geq \int_0^t s^\gamma ds = \frac{t^{1+\gamma}}{1+\gamma}\\ &\geq \frac{1}{1+\gamma}\max\left\{\ln\left(\frac{2r_{0,0}^2}{\epsilon}\right)^{1+\gamma}\frac{1}{\ln(1/\theta_0)^{1+\gamma}}, \left(\frac{2^{\gamma+1}\mathcal{D}_r}{\epsilon}\mathcal{C}\right)^{1+\frac{1}{\gamma}}\right\},\end{aligned}$$

which proves the statement of the theorem. ■

9. Acknowledgments

The research leading to these results has received funding from the Executive Agency for Higher Education, Research and Innovation Funding (UEFISCDI), Romania: PNIII-P4-PCE-2016-0731, project ScaleFreeNet, no. 39/2017.

References

Y.F. Atchade, G. Fort and E. Moulines, *On perturbed proximal gradient algorithms*, Journal of Machine Learning Research, 18(1):310–342, 2014.

F. Bach, *Self-concordant analysis for logistic regression*, Electronic Journal of Statistics, 4:384–414, 2010.

F. Bach, G. Lanckriet and M. Jordan, *Multiple kernel learning, conic duality, and the SMO algorithm*, International Conference on Machine Learning (ICML), 2004.

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

P. Bianchi, *Ergodic convergence of a stochastic proximal point algorithm*, SIAM Journal on Optimization, 26(4):2235–2260, 2016.

D. Blatt and A.O. Hero, *III: Energy based sensor network source localization via projection onto convex sets (POCS)*, IEEE Transactions on Signal Processing, 54(9):3614–3619, 2006.

L. Bottou, F.E. Curtis and J. Nocedal, *Optimization methods for large-scale machine learning*, arXiv:1606.04838, 2016.

J. Brodie, I. Daubechies, C. de Mol, D. Giannone and I. Loris, *Sparse and stable Markowitz portfolios*, Proc. Natl. Acad. Sci, 106:12267–12272, 2009.

P.S. Bullen, *Handbook of means and their inequalities*, Kluwer Academic Publisher, Dordercht, 2003.

Y. Censor, W. Chen, P. L. Combettes, R. Davidi and G. T. Herman, *On the effectiveness of projection methods for convex feasibility problems with linear inequality constraints*, Computational Optimization and Applications, 51(3):1065–1088, 2012.

P.L. Combettes and J.C. Pesquet, *Stochastic approximations and perturbations in forward-backward splitting for monotone operators*, Pure and Applied Functional Analysis, 1(1):13–37, 2016.

A. Defazio, F. Bach and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, Advances in Neural Information Processing Systems (NIPS), 1646–1654, 2014.

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

O. Guler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM Journal on Control and Optimization, 29(2):403–419, 1991.

R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Advances in Neural Information Processing Systems (NIPS), 315–323, 2013.

A. Karimi and C. Kammer, *A data-driven approach to robust control of multivariable systems by convex optimization*, Automatica, 85:227–233, 2017.

J. Koshal, A. Nedic and U.V. Shanbhag, *Regularized iterative stochastic approximation methods for stochastic variational inequality problems*, IEEE Transactions on Automatic Control, 58(3):594–609, 2013.

A. Mokhtari and A. Ribeiro, *RES: Regularized stochastic BFGS algorithm*, IEEE Transactions on Signal Processing, 62(23):6089–6104, 2014.

E. Moulines and F.R. Bach, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, Advances in Neural Information Processing Systems (NIPS), 451–459, 2011.

I. Necoara, *Random algorithms for convex minimization over intersection of simple sets*, submitted to European Control Conference (ECC18), 2017.

I. Necoara, V. Nedelcu and I. Dumitrache, *Parallel and distributed optimization methods for estimation and control in networks*, Journal of Process Control, 21(5):756–766, 2011.

I. Necoara, Yu. Nesterov and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Mathematical Programming, in press:135, 2017a.

I. Necoara, P. Richtarik, and A. Patrascu, *Randomized projection methods for convex feasibility problems*, Technical Report, UPB:1–30, 2017b.

A. Nedic, *Random algorithms for convex minimization problems*. Mathematical Programming, 129(2):225253, 2011.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization, 19(4):15741609, 2009.

Yu. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publisher, Boston, 2004.

F. Niu, B. Recht, C. Re, and S. J. Wright, *HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent*, In Advances in Neural Information Processing Systems (NIPS), pages 693–701, 2011.

J. C. Platt, *Fast training of support vector machines using sequential minimal optimization*, In Advances in Kernel Methods - Support Vector Learning, Cambridge, MA, 1998.

B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM Journal on Control and Optimization, 30(4):838–855, 1992.

R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin Heidelberg, 1998.

L. Rosasco, S. Villa, and B. C. Vu, *Convergence of stochastic proximal gradient algorithm*, arXiv:1403.5074, 2014.

L. Rosasco, S. Villa, and B. C. Vu, *A first-order stochastic primal-dual algorithm with correction step*, Numerical Functional Analysis and Optimization, 38(5):602–626, 2017.

N. L. Roux, M. Schmidt, and F. Bach, *A stochastic gradient method with an exponential convergence rate for finite training sets*, In Advances in Neural Information Processing Systems (NIPS), 2672–2680, 2012.

E. Ryu and S. Boyd, *Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent*, www.math.ucla.edu/eryu/papers/spi.pdf, 2016.

S. Shalev-Shwartz and T. Zhang, *Stochastic dual coordinate ascent methods for regularized loss*, Journal of Machine Learning Research, 14(1):567–599, 2013.

S. Sonnenburg, G. Ratscha, C. Schafer, and B. Scholkopf, *Large scale multiple kernel learning*, Journal of Machine Learning Research, 7:15311565, 2006.

P. Toulis, D. Tran, and E. M. Airoldi, *Towards stability and optimality in stochastic gradient descent*, In International Conference on Artificial Intelligence and Statistics (AISTATS), 1290–1298, 2016.

N. Denizcan Vanli, M. Gurbuzbalaban, and A. Ozdaglar, *Global convergence rate of proximal incremental aggregated gradient methods*, arXiv:1608.01713, 2016.

W. Xu, *Towards optimal one pass large scale learning with averaged stochastic gradient descent*, CoRR, abs/1107.2490, 2011.

T. Yang and Q. Lin, *Stochastic subgradient methods with linear convergence for polyhedral convex optimization*, arXiv:1510.01444, 2016.

A. Yurtsever, B. C. Vu, and V. Cevher, *Stochastic three-composite convex minimization*, In Advances in Neural Information Processing Systems (NIPS), 4322–4330, 2016.