Randomized projection methods for convex feasibility problems: conditioning and convergence rates

Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0364-765X | EISSN 1526-5471 | 00 | 0000 | 0001
DOI 10.1287/xxxx.0000.0000

Randomized projection methods for convex feasibility problems: conditioning and convergence rates

Ion Necoara and Andrei Patrascu
Automatic Control and Systems Engineering Department, University Politehnica Bucharest, 060042 Bucharest, Romania,
ion.necoara@acse.pub.ro.

Peter Richtarik
School of Mathematics, The Maxwell Institute for Mathematical Sciences, University of Edinburgh, United Kingdom,
peter.richtarik@ed.ac.uk.

Finding a point in the intersection of a collection of closed convex sets, that is the convex feasibility problem, represents the main modeling strategy for many computational problems. In this paper we analyze new stochastic reformulations of the convex feasibility problem in order to facilitate the development of new algorithmic schemes. We also analyze the conditioning problem parameters using certain (linear) regularity assumptions on the individual convex sets. Then, we introduce a general random projection algorithmic framework, which extends to the random settings many existing projection schemes, designed for the general convex feasibility problem. Our general random projection algorithm allows to project simultaneously on several sets, thus providing great flexibility in matching the implementation of the algorithm on the parallel architecture at hand. Based on the conditioning parameters, besides the asymptotic convergence results, we also derive explicit sublinear and linear convergence rates for this general algorithmic framework.

History: First version: March 2017.

1. Introduction Finding a point in the intersection of a collection of closed convex sets, that is *the convex feasibility problem*, represents a modeling paradigm which has been used for many decades for posing and solving engineering and physics problems. Among the most important applications modeled by the convex feasibility formalism are: radiation therapy treatment planning [20], computerized tomography [19] and magnetic resonance imaging [33]; wavelet-based denoising [13], color imaging [34] and demosaicking [24]; antenna design [17] and sensor networks problems [8]; data compression [23], neural networks [35] and adaptive filtering [38].
Convex feasibility problems have various formulations, such as finding the fixed points of a non-expansive operator, the set of optimal solutions of a specific optimization problem or the set of solutions to some convex inequalities. Projection methods were first used for solving systems of linear equalities [21] or linear inequalities [25], and then extended to general convex feasibility problems, e.g. in [14]. Projection methods are very attractive in applications since they are able to handle problems of huge dimensions with a very large number of convex sets in the intersection. For instance, the *projection algorithm* which represents one of the first iterative algorithms for feasibility problems, rely at each iteration on orthogonal projections onto given individual sets. Its simple algorithmic structure supports the current large scale setting and can be easily adapted to parallel environments, making such schemes adequate to modern computational architectures. If the iteration of a given projection algorithm rely on an alternating sequence of projections onto sets over the iterations, then it belongs to an *alternating projection* schemes [5, 6, 27, 31]. Furthermore, depending on the variant of the alternating projection algorithm, the current set (or sets) on which the projection is made can be chosen, for example, in a random, cyclic or greedy manner.

Otherwise, if the scheme uses at current iteration an average of multiple projections of the current iterate onto various sets, then also it can be viewed as an *average projection* algorithm [11,12].
The convergence properties, the iteration complexity and even the inherent limitations of the class of projection schemes has been intensely analyzed over the last decades, as it can be seen in [1–6,11,12,14,27,28,31] and the references therein. In [1] a barycenter type projection algorithm is developed, which allows the efficient handling of feasibility problems arising in the nonnegative orthant. The proposed method uses approximate projections on the sets and is proven globally convergent under the sole assumption that the given intersection is nonempty and the errors are controllable. An important contribution is made in [3], where the rates of convergence of some projections algorithms are analyzed for solving the general convex feasibility problem. Besides revealing some connections between the Slater's condition and the classical linear regularity property, the authors show that if the Slater's condition does not hold, the projection algorithms can behave quite badly, i.e. with a rate of convergence which is not bounded. Moreover, the authors also propose an alternative local linear regularity bound to derive further convergence rate results. Linear convergence of the conditional gradient method applied on the equivalent optimization formulation of the problem of finding a point in the intersection of an affine set with a compact convex set is derived in [4]. In a more general setting, [2] studies the problem of finding a point in the intersection of affine constraints with a nonconvex closed set and a simple gradient projection scheme is developed. The scheme is proven to converge to a unique solution of the problem, at a linear rate, under a natural assumption defined in terms of the problem's data.

Contributions. Below, we clarify the relationship and differences between our work and earlier research in this direction. In particular, the main contributions of this paper consist in unifying and extending existing projection methods in several aspects:
(i) The classical convex feasibility problem was usually formulated for a finite intersection of simple convex sets. While finding a point in the intersection of a finite number of convex sets is a problem with its own challenges, it does not cover many interesting applications modeled by an intersection of (infinite) countable/uncountable number of simple convex sets (see e.g. [29]). In this paper we present several new equivalent stochastic formulations of the convex feasibility problem, which allow us to deal with intersections of families of convex sets that may be even uncountable.
(ii) From an algorithmic point of view, most of the previous approaches are limited to cycle based alternating projection schemes. Moreover, for this strategy it is difficult to prove asymptotic convergence and to estimate the rate of convergence in the general convex feasibility case. Therefore, we introduce a general random projection algorithmic framework, which covers or extends to the random settings many existing projection schemes, designed for the general convex feasibility problem. Besides asymptotic convergence results, we also derive explicit convergence rates for this general algorithm. It is worth to mention that our convergence rates depend explicitly on the number of computed projections per iteration. Moreover, our general framework generates new algorithms, that are not analyzed in the literature, with possible better convergence rates than the existing ones.
(iii) From our convergence analysis it follows that we can use large step-sizes, besides the usual naturally arisen constant step-size policy. Thus, we prove theoretically, what is empirically known in numerical applications for a long time, namely that these over-relaxations accelerate significantly the convergence of projection methods.
(iv) Our general random projection algorithm allows to project simultaneously onto several sets, thus providing great flexibility in matching the implementation of the algorithms on the parallel architecture at hand.

Notations. For given $m \in \mathbb{N}\backslash\{0\}$, we denote the set $[m] = \{1, \dots, m\}$. We consider the space $\mathbb{R}^n$ composed by column vectors. For $x, y \in \mathbb{R}^n$ denote the scalar product by $\langle x, y\rangle = x^T y$ and the Euclidean norm by $\|x\| = \sqrt{x^T x}$. We use the notation x_i for the ith component of the vector x

and e_i for the ith column of the identity matrix. The projection operator onto the closed convex set X is denoted by $\Pi_X(\cdot)$ and the distance from a given x to set X is denoted by $\text{dist}_X(x)$. Let $Q \in \mathbb{R}^{n\times n}$, then we use notation Q_i for the ith row of the matrix Q. The minimal non-zero singular value and the minimal nonzero eigenvalue of the matrix Q are represented by $\sigma^{\text{nz}}_{\min}(Q)$ and $\lambda^{\text{nz}}_{\min}(Q)$, respectively. Similarly, $\sigma_{\max}(Q)$ and $\lambda_{\max}(Q)$ denote the largest singular value and the largest eigenvalue of the matrix Q, respectively. Also $\|Q\|_F$ denotes its Frobenius norm.

2. Problem formulation In this paper we consider the convex feasibility problem:

$$\text{Find} \quad x \in \mathcal{X}, \tag{1}$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set. We assume that $\mathcal{X}$ is nonempty. In general, in most convex feasibility problems one should seek scalable algorithms with simple iterations which are able to find an approximation of a point from the set $\mathcal{X}$. For this purpose, we usually assume that $\mathcal{X}$ can be represented as the intersections of finitely/infinitely many simple closed convex sets. Then, a simple and widely known idea for solving the convex feasibility problem is to project successively onto the individual sets in a certain fashion, e.g. cyclic or random. These projection algorithms are most efficient when the projections onto the individual sets are computationally cheap. However, in many cases, it is difficult to find an explicit representation of the set $\mathcal{X}$ as intersection of simple sets. That is why in the sequel we consider different relaxations of (1), based on several representations for the individual sets, and we investigate when this relaxations are exact.

2.1. Stochastic reformulations In many applications the set $\mathcal{X}$ has explicit representations, while in others this set it is not known explicitly. Therefore, below we present several representations or approximations for the set $\mathcal{X}$. For that, we introduce the concept of *stochastic approximation* of $\mathcal{X}$. Given a probability distribution $\mathbf{P}$, we consider a random variable $S \sim \mathbf{P}$ from a probability space Ω.

Definition 1 (Stochastic approximation of sets). For any $S \in \Omega$ let $\mathcal{X}_S$ be a random closed convex subset of $\mathbb{R}^n$. We say that $\mathcal{X}_S$ is a *stochastic approximation* of $\mathcal{X}$ if $\mathcal{X} \subseteq \mathcal{X}_S$ for all $S \in \Omega$.

We will henceforth consider stochastic approximation sets $\mathcal{X}_S$ arising as a function of some random variable S from a probability space $(\Omega, \mathbf{P})$. Therefore, the set $\mathcal{X}$ may be represented as an exact countable/uncountable intersection of stochastic approximation sets $\mathcal{X}_S$, that is $\mathcal{X} = \cap_{S\in\Omega}\mathcal{X}_S$, or approximated by this intersection, that is $\mathcal{X} \subseteq \cap_{S\in\Omega}\mathcal{X}_S$. Clearly, having a family of stochastic approximation sets $(\mathcal{X}_S)_{S\in\Omega}$, we have the first relaxation of (1):

$$\mathcal{X} \subseteq \bigcap_{S\in\Omega} \mathcal{X}_S. \tag{2}$$

Then, we consider the following convex feasibility problem, which may be a relaxation of the potentially difficult original problem (1):

$$\text{Find} \quad x \in \bigcap_{S\in\Omega} \mathcal{X}_S. \tag{3}$$

In this paper, we propose several stochastic reformulations of the convex feasibility problem (3).

1. **Stochastic fixed point problem.**

$$\text{Find a fixed point of the mapping} \quad x \mapsto \mathbf{E}_{S\sim\mathbf{P}}\left[\Pi_{\mathcal{X}_S}(x)\right]. \tag{4}$$

2. **Stochastic non-smooth optimization problem.**

$$\text{Minimize} \quad \left\{ f(x) \stackrel{\text{def}}{=} \mathbf{E}_{S\sim\mathbf{P}}\left[\mathbb{I}_{\mathcal{X}_S}(x)\right] \right\} \quad \text{subject to} \quad x \in \mathbb{R}^n. \tag{5}$$

3. **Stochastic smooth optimization problem.**

$$\text{Minimize} \quad \left\{ F(x) \stackrel{\text{def}}{=} \frac{1}{2}\mathbf{E}_{S\sim\mathbf{P}}\left[\|x-\Pi_{\mathcal{X}_S}(x)\|^2\right] \right\} \quad \text{subject to} \quad x\in\mathbb{R}^n. \tag{6}$$

4. **Stochastic intersection problem.**

$$\text{Find} \quad x\in\mathbb{R}^n \quad \text{such that} \quad \mathbf{P}(x\in\mathcal{X}_S)=1. \tag{7}$$

Equivalence of the above reformulations is captured by the following lemma:

LEMMA 1 **(Equivalence)**. *Assume* $\cap_{S\sim\mathbf{P}}\mathcal{X}_S\neq\emptyset$. *The stochastic reformulations* (4), (5), (6) *and* (7) *of the convex feasibility problem* (3) *are equivalent. That is, the set of fixed points of* $x\mapsto\mathbf{E}_{S\sim\mathbf{P}}[\Pi_{\mathcal{X}_S}(x)]$ *is equal to the set of minimizers of the objective functions* f *or* F, *and to the set* $\{x:\mathbf{P}(x\in\mathcal{X}_S)=1\}$. *We shall use the symbol* $\mathcal{Y}$ *to denote this set.*

Proof: An elementary property of the Lebesgue integral states that if $\phi\geq 0$, then $\mathbf{E}[\phi]=0$ if and only if $\phi=0$ almost sure (a.s.). Using this classical result, we can prove the following equivalences:

(5)$\Leftrightarrow$ (7). The $\mathbf{P}$-measurable function $f_S(x)=\mathbb{I}_{\mathcal{X}_S}(x)$ is non-negative and thus the set of minimizers in (5) are those x for which $\mathbf{E}_{S\sim\mathbf{P}}[\mathbb{I}_{\mathcal{X}_S}(x)]=0$, which is equivalent to $\mathbb{I}_{\mathcal{X}_S}(x)=0$ a.s., that is $x\in\mathcal{X}_S$ a.s., or equivalent to $\mathbf{P}(x\in\mathcal{X}_S)=1$.

(6)$\Leftrightarrow$ (7). The function $F_S(x)=\|x-\Pi_{\mathcal{X}_S}(x)\|^2$ is non-negative and thus the set of minimizers in (6) are those x for which $\mathbf{E}_{S\sim\mathbf{P}}[\|x-\Pi_{\mathcal{X}_S}(x)\|^2]=0$, which is equivalent to $\|x-\Pi_{\mathcal{X}_S}(x)\|=0$ a.s. or equivalently $x=\Pi_{\mathcal{X}_S}(x)$ a.s. or equivalently $x\in\mathcal{X}_S$ a.s., or equivalent to $\mathbf{P}(x\in\mathcal{X}_S)=1$.

(6)$\Rightarrow$(4). Since $\|\mathbf{E}_{S\sim\mathbf{P}}[x-\Pi_{\mathcal{X}_S}(x)]\|^2\leq\mathbf{E}_{S\sim\mathbf{P}}[\|x-\Pi_{\mathcal{X}_S}(x)\|^2]$, then it follows that the set of minimizers of (6) are included in the set of fixed points of the average projection operator $\Pi(x)=\mathbf{E}_{S\sim\mathbf{P}}[\Pi_{\mathcal{X}_S}(x)]$ defined in (4).

It remains to prove the other inclusion (4)$\Rightarrow$(6). Let x be a fixed point of the average projection operator, that is $x=\mathbf{E}_{S\sim\mathbf{P}}[\Pi_{\mathcal{X}_S}(x)]$. Then, for any $z\in\cap_{S\sim\mathbf{P}}\mathcal{X}_S$, it follows that $z\in\mathcal{X}_S$ for all S and from the optimality condition for the projection onto $\mathcal{X}_S$ we have $\langle x-\Pi_{\mathcal{X}_S}(x),\Pi_{\mathcal{X}_S}(x)-z\rangle\geq 0$. This leads to:

$$\begin{aligned}
0&=\langle\mathbf{E}_{S\sim\mathbf{P}}[x-\Pi_{\mathcal{X}_S}(x)],x-z\rangle=\mathbf{E}_{S\sim\mathbf{P}}[\langle x-\Pi_{\mathcal{X}_S}(x),x-z\rangle]\\
&=\mathbf{E}_{S\sim\mathbf{P}}[\langle x-\Pi_{\mathcal{X}_S}(x),x-\Pi_{\mathcal{X}_S}(x)+\Pi_{\mathcal{X}_S}(x)-z\rangle]\\
&=\mathbf{E}_{S\sim\mathbf{P}}\left[\|x-\Pi_{\mathcal{X}_S}(x)\|^2\right]+\mathbf{E}_{S\sim\mathbf{P}}\left[\underbrace{\langle x-\Pi_{\mathcal{X}_S}(x),\Pi_{\mathcal{X}_S}(x)-z\rangle}_{\geq 0}\right],
\end{aligned}$$

for all $z\in\cap_{S\in\Omega}\mathcal{X}_S$. Thus, sum of two non-negative scalars is zero implies that each term is zero, that is $\mathbf{E}_{S\sim\mathbf{P}}[\|x-\Pi_{\mathcal{X}_S}(x)\|^2]=0$ and therefore the set of fixed points of (4) are included into the set of minimizers of (6). Q.E.D.

2.2. Discussion The proof of Lemma 1 provides several connections between (4), (5), (6) and (7). There is also an interesting interpretation between (5) and (6). Notice that for any given nonempty closed convex set Y, the indicator function $\mathbb{I}_Y$ is convex, lower semi-continuous, that is not identically $+\infty$. Therefore, the value function:

$$\frac{1}{2}\|x-\Pi_Y(x)\|^2=\min_{z\in\mathbb{R}^n}\mathbb{I}_Y(z)+\frac{1}{2}\|z-x\|^2$$

is known to be well-defined and finite everywhere [7] (Chapter 12). Moreover, the function $x \mapsto \|x - \Pi_Y(x)\|^2$ is the Moreau approximation of the non-smooth indicator function $\mathbb{I}_Y$, thus it has Lipschitz continuous gradient with constant 1, see [27]. This implies that the function F has Lipschitz continuous gradient with constant $L_F = 1$. Observe that the smooth optimization problem (6) is obtained from the Moreau approximation $F_S(x) = 1/2\|x - \Pi_{\mathcal{X}_S}(x)\|^2$ of each indicator function $f_S(x) = \mathbb{I}_{\mathcal{X}_S}(x)$ of the non-smooth optimization problem (5), that is:

$$\min_{x\in\mathbb{R}^n} F(x) = \min_{x\in\mathbb{R}^n} \mathbf{E}_{S\sim\mathbf{P}}\left[F_S(x)\right] = \min_{x\in\mathbb{R}^n} \mathbf{E}_{S\sim\mathbf{P}}\left[\min_{z\in\mathbb{R}^n} f_S(z) + \frac{1}{2}\|z-x\|^2\right] \tag{8}$$
$$= \min_{x\in\mathbb{R}^n} \mathbf{E}_{S\sim\mathbf{P}}\left[\underbrace{\min_{z\in\mathbb{R}^n} \mathbb{I}_{\mathcal{X}_S}(z) + \frac{1}{2}\|z-x\|^2}_{\|x-\Pi_{\mathcal{X}_S}(x)\|^2}\right].$$

Note that, for general functions f_S, there are no connections between the two problems (5) and (6) as expressed in (8). However, for indicator functions $f_S(x) = \mathbb{I}_{\mathcal{X}_S}(x)$ we have $\arg\min_x f(x) = \arg\min_x F(x)$, according to previous lemma.
For the convex feasibility problem (3), with Ω having finite support, the following basic alternating projection algorithm has been extensively studied in the literature [18, 27]:

$$(\textbf{B-AP}):\ \text{choose } S_k \text{ cyclic/random } \&\ \text{update } x^{k+1} = \Pi_{\mathcal{X}_{S_k}}(x^k).$$

The (B-AP) algorithm can be interpreted in several ways depending on the reformulations (4)-(7):

1. For example, when solving the stochastic fixed point problem (4), we do not have an explicit access to the average projection map $x \to \mathbf{E}_{S\sim\mathbf{P}}\left[\Pi_{\mathcal{X}_S}(x)\right]$. Instead, we are able to repeatedly sample $S \sim \mathbf{P}$ and use the stochastic projection map $x \to \Pi_{\mathcal{X}_S}(x)$, which leads to the random variant of (B-AP) algorithm.
2. Since the stochastic optimization problem (5) with $f_S = \mathbb{I}_{\mathcal{X}_S}$:

$$\min_x f(x) = \mathbf{E}_{S\sim\mathbf{P}}\left[f_S(x)\right],$$

is non-smooth, then we approximate each indicator function $f_S = \mathbb{I}_{\mathcal{X}_S}$ with its Moreau approximation F_S and we can apply gradient method on the resulting expected approximation which leads to the proximal point method. Since we do not have access to the function $\mathbf{E}_{S\sim\mathbf{P}}\left[\mathbb{I}_{\mathcal{X}_S}(z) + \frac{1}{2}\|z-x\|^2\right]$ for some fixed x, but we can repeatedly sample $S \sim \mathbf{P}$ we can apply stochastic proximal point:

$$x^+ = \arg\min_z \mathbb{I}_{\mathcal{X}_S}(z) + \frac{1}{2}\|z-x\|^2 = \Pi_{\mathcal{X}_S}(x).$$

3. When solving the stochastic optimization problem (6):

$$\min_x F(x) = \mathbf{E}_{S\sim\mathbf{P}}\left[F_S(x)\right],$$

where

$$F_S(x) = \frac{1}{2}\|x - \Pi_{\mathcal{X}_S}(x)\|^2$$

we do not have access to the gradient of F:

$$\nabla F(x) = \mathbf{E}_{S\sim\mathbf{P}}\left[\nabla F_S(x)\right] = \mathbf{E}_{S\sim\mathbf{P}}\left[x - \Pi_{\mathcal{X}_S}(x)\right].$$

Instead, we can repeatedly sample $S \sim \mathbf{P}$ and receive unbiased samples of this gradient at points of interest, that is $\nabla F_S(x) = x - \Pi_{\mathcal{X}_S}(x)$. Then, applying the stochastic gradient method with stepsize 1 leads to the random variant of (B-AP).

4. We observe that (7) can be written equivalently as:

$$\text{Find} \quad x \in \{x : \mathbf{P}(x \in \mathcal{X}_S) = 1\} := \bigcap_{S \sim \mathbf{P}} \mathcal{X}_S.$$

Then, when solving the previous stochastic intersection problem we typically do not have explicit access to the stochastic intersection $\cap_{S\sim\mathbf{P}}\mathcal{X}_S$. Rather, we can sample $S \sim \mathbf{P}$ and utilize the simple form of $\mathcal{X}_S$ to derive (B-AP) algorithm. Notice that if Ω is finite/countable, then the stochastic intersection problem reduces to the standard intersection problem (3).
However, in Section 6 we will give a more general algorithmic framework for solving the four equivalent problems with larger stepsize and better performances than (B-AP).
As Lemma 1 claims, our four stochastic reformulations are equivalent and they are all relaxations of the convex feasibility problem (3), that is:

$$\cap_{S\in\Omega}\mathcal{X}_S \subseteq \mathcal{Y}.$$

Therefore, for any family of stochastic approximation sets $(\mathcal{X}_S)_{S\sim\mathbf{P}}$ over a probability space $(\Omega, \mathbf{P})$, we clearly have:

$$\mathcal{X} \subseteq \bigcap_{S\in\Omega} \mathcal{X}_S \subseteq \mathcal{Y}. \tag{9}$$

Therefore, it is natural to investigate when these inclusions hold with equality.

3. Exactness For simplicity we further redenote $\mathbf{E}_{S\sim\mathbf{P}}[\cdot]$ with the simpler notation $\mathbf{E}[\cdot]$. From previous discussion we note that for any family of stochastic approximations $(\mathcal{X}_S)_{S\sim\mathbf{P}}$ over a probability space $(\Omega, \mathbf{P})$ we trivially have the inclusion $\mathcal{X} \subseteq \mathcal{Y}$. If $\mathcal{X} = \mathcal{Y}$, then the stochastic reformulations (4), (5), (6) and (7) are equivalent to the convex feasibility problems (1) and (3). However, this need not be the case, not without additional assumptions. To see this, consider $\mathcal{X} = \bigcap_{i=1}^m \mathcal{X}_i$, that is finite intersection of closed convex sets $\mathcal{X}_i$, and the random set $\mathcal{X}_S = \mathcal{X}_1$. Since $\mathcal{X} \subseteq \mathcal{X}_1$, this constitutes a stochastic approximation of $\mathcal{X}$, as defined in Definition 1. However, $\mathcal{Y} = \mathcal{X}_1$, which is not necessarily equal to $\mathcal{X}$. In view of the above, we need to enforce a regularity assumption, which we call *exactness*.

Assumption 1 **(Exactness)**. *Stochastic reformulations* (4), (5), (6) *and* (7) *of the convex feasibility problems* (1) *and* (3) *are exact. That is,* $\mathcal{X} = \mathcal{Y}$.

In the next result we give a sufficient condition for exactness:

Lemma 2. *The following statement hold: If there exists* $\kappa < \infty$ *such that the following inequality (a.k.a. "linear regularity property") holds for all* $x \in \mathbb{R}^n$*:*

$$\text{dist}^2_{\mathcal{X}}(x) \leq \kappa\, \mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right], \tag{10}$$

then $\mathcal{X} = \mathcal{Y}$ *(i.e., exactness holds).*

Proof: The set $\mathcal{Y}$ of optimal points of the stochastic smooth optimization problem (6) satisfies: $F(x) = 0$ for all $x \in \mathcal{Y}$. Moreover, the relation $F(x) = \mathbf{E}[\|x - \Pi_{\mathcal{X}_S}(x)\|^2] = \mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right]$ holds. Therefore, for any $x \in \mathcal{Y}$ we have $\mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right] = 0$. From (10) we conclude that $\text{dist}^2_{\mathcal{X}}(x) = 0$, which means that $x \in \mathcal{X}$. Combined with (9), this implies that $\mathcal{X} = \mathcal{Y}$ holds. Q.E.D.
Since $\text{dist}_{\mathcal{X}_S}(x) \leq \text{dist}_{\mathcal{X}}(x)$ it follows immediately from (10) that $\kappa \geq 1$. The feasibility problem is ill-conditioned when κ is large. Notice that linear regularity is a very conservative condition for exactness. We can see that if Ω is finite/countable, then the stochastic intersection problem reduces to the standard intersection problem (3), i.e. we have exactness. Note that linear regularity property does not hold for any collection of closed convex sets as the following example shows:

EXAMPLE 1. Let $\mathcal{X}_1 = \{x : |x_1|^p \leq x_2\}$ with $p > 1$, and $\mathcal{X}_2 = \{x : x_2 = 0\}$. These two sets are convex and $\mathcal{X} = \mathcal{X}_1 \cap \mathcal{X}_2 = \{0\}$. Then, for any $x \in \mathcal{X}_1$, satisfying $|x_1|^p = x_2$, we have:

$$\text{dist}^2_{\mathcal{X}}(x) = x_1^2 + x_2^2 \quad \text{and} \quad \text{dist}^2_{\mathcal{X}_1}(x) + \text{dist}^2_{\mathcal{X}_2}(x) = x_2^2.$$

Then, clearly there is no finite $\kappa > 0$ such that:

$$x_1^2 + x_2^2 \leq \kappa x_2^2 \quad \forall |x_1|^p = x_2,\ x_1 \geq 0,$$

since by replacing x_2 and obtaining

$$x_1^2 + x_1^{2p} \leq \kappa x_1^{2p} \Rightarrow \frac{1}{x_1^{2p-2}} + 1 \leq \kappa,$$

we can take x_1 very small (close to zero). Q.E.D.
Notice that linear regularity is related to Slater's condition, as discussed in [3]. Moreover, this property is directly related to the stochastic formulations (4)-(7), as we will show below.

3.1. Properties of the smooth function F If we consider the smooth stochastic optimization problem (6), then we have the following important relation:

$$F(x) = \frac{1}{2}\mathbf{E}\left[\|\nabla F_S(x)\|^2\right] \quad \forall x \in \mathbb{R}^n, \tag{11}$$

since we recall that $\nabla F_S(x) = x - \Pi_{\mathcal{X}_S}(x)$. Moreover, the linear regularity property (10) is equivalent with the quadratic functional growth condition on F introduced in [26], which was defined as a relaxation of strong convexity. Indeed, under the exactness assumption, we have $\mathcal{X} = \mathcal{Y} = \arg\min_x F(x)$ and the optimal value $F^* = 0$. Moreover, we have $F(x) = \mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right]$. Then, the property (10) can be rewritten equivalently as:

$$F(x) - F^* \geq \frac{1}{2\kappa}\|x - \Pi_{\mathcal{X}}(x)\|^2 \quad \forall x \in \mathbb{R}^n, \tag{12}$$

which is exactly the definition of the quadratic functional growth condition introduced in [26]. Typically, the standard assumption for proving linear convergence of first order methods for smooth convex optimization is the strong convexity of the objective function, an assumption which does not hold for many practical applications, including the one presented in this paper. In [26] it has been proved that we can still achieve linear convergence rates of several first order methods for solving smooth non-strongly convex constrained optimization problems, i.e. involving an objective function with a Lipschitz continuous gradient that satisfies the relaxed strong convexity condition (12). Moreover, in [26] it has been shown that the quadratic functional growth condition (12) is equivalent with the so-called error bound condition for unconstrained problem (6).
Further, let $\gamma \geq 0$ be the smallest constant satisfying the inequality:

$$\|\mathbf{E}\left[x - \Pi_{\mathcal{X}_S}(x)\right]\|^2 \leq \gamma \cdot \mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right] \qquad \forall x \in \mathbb{R}^n. \tag{13}$$

By Jensen's inequality, $\gamma \leq 1$. However, for specific sets and distributions $\mathbf{P}$, it is possible for γ to be strictly smaller than 1, as the following examples show. For example, we can consider finding a solution of a linear system $\mathcal{X} = \{x : Ax = b\}$, where $A \in \mathbb{R}^{m\times n}$. For this set we can easily construct stochastic approximations sets $\mathcal{X}_S = \{x :\ S^T Ax = S^T b\}$ taking any matrix $S \in \mathbb{R}^{m\times q}$. Clearly, for any matrix S we have $\mathcal{X} \subseteq \mathcal{X}_S$. Then, we have the following characterization for γ:

THEOREM 1. *Let us consider finding a solution of the linear system $\mathcal{X} = \{x : Ax = b\}$, where $A \in \mathbb{R}^{m\times n}$. Further, let us consider the stochastic approximation sets $\mathcal{X}_S = \{x : S^T Ax = S^T b\}$, where $S \in \Omega = \mathbb{R}^{m\times q}$ and a probability distribution $\mathbf{P}$ on Ω. Then, (13) holds with:*

$$\gamma = \lambda_{max}(A^T \mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right] A) \leq 1.$$

Proof: Clearly, for x satisfying $Ax = b$ the inequality (13) holds for any $\gamma \leq 1$. It remains to prove for x satisfying $Ax - b \neq 0$. However, since $\mathcal{X}_S = \{x : S^T Ax = S^T b\}$, then the projection of x onto $\mathcal{X}_S$ can be computed explicitly:

$$\Pi_{\mathcal{X}_S}(x) = x - A^T S(S^T AA^T S)^\dagger S^T (Ax - b)$$

and the relation we need to prove becomes as follows:

$$\left\|\mathbf{E}\left[A^T S(S^T AA^T S)^\dagger S^T (Ax - b)\right]\right\|^2 \leq \gamma \mathbf{E}\left[\|A^T S(S^T AA^T S)^\dagger S^T (Ax - b)\|^2\right].$$

Using the standard properties of the pseudoinverse, that is $Q^\dagger QQ^\dagger = Q^\dagger$ for any matrix Q, the previous relation is equivalent to:

$$\left\|A^T \mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right](Ax - b)\right\|^2 \leq \gamma (Ax - b)^T \mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right](Ax - b).$$

For simplicity, let us denote $E = \mathbf{E}\left[S(S^T AA^T S)^{-1} S^T\right]$. Then E is a positive semidefinite matrix and thus there exists $E^{1/2}$. Clearly, for $Ax - b \in \text{Null}(E)$ the previous inequality holds for any γ. Therefore, γ is defined as:

$$\begin{aligned}
\gamma &= \max_{x: Ax-b \notin \text{Null}(E)} \frac{\|A^T E(Ax - b)\|^2}{(Ax - b)^T E(Ax - b)} \\
&= \max_{x: Ax-b \notin \text{Null}(E^{1/2})} \frac{\|A^T E^{1/2} E^{1/2}(Ax - b)\|^2}{\|E^{1/2}(Ax - b)\|^2} \\
&= \max_{z \neq 0} \frac{\|A^T E^{1/2} z\|^2}{\|z\|^2} = \sigma^2_{\max}(A^T E^{1/2}) = \lambda_{\max}(A^T EA).
\end{aligned}$$

Therefore, we have $\gamma = \lambda_{\max}(A^T \mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right] A)$. Since the function $W \mapsto \lambda_{\max}(W)$ is convex over the space of positive semidefinite matrices, then using Jensen's inequality we have:

$$\lambda_{\max}(A^T \mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right] A) \leq \mathbf{E}\left[\lambda_{\max}(A^T S(S^T AA^T S)^\dagger S^T A)\right].$$

Furthermore, the matrix $P_S = A^T S(S^T AA^T S)^\dagger S^T A$ is idempotent, that is $P_S^2 = P_S$. Therefore, all the eigenvalues of P_S are either 0 or 1. Then, we get:

$$\gamma = \lambda_{\max}(A^T \mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right] A) \leq \mathbf{E}\left[\lambda_{\max}(A^T S(S^T AA^T S)^\dagger S^T A)\right] \leq 1,$$

which proves the statement of the theorem. Q.E.D.
Based on the previous theorem we can prove that for particular choices of the probability distribution $\mathbf{P}$ we have $\gamma < 1$, see e.g. the next corollary:

COROLLARY 1. *Let us consider finding a solution of the linear system $\mathcal{X} = \{x : Ax = b\}$, where $A \in \mathbb{R}^{m\times n}$ having rang$(A) \geq 2$. Further, let us consider $\Omega = \{e_1, \cdots, e_m\}$, the standard basis of $\mathbb{R}^m$, and the corresponding stochastic approximation sets $\mathcal{X}_{e_i} = \{x : A_i^T x = b_i\}$ for all $i \in [m]$. Then, for two choices of the probability distribution $\mathbf{P}$ on Ω, inequality (13) holds with:*

$$\gamma = \begin{cases} \frac{\lambda_{max}(A^T A)}{\|A\|_F^2} & \text{if} \quad \mathbf{P}(S = e_i) = \frac{\|A_i\|^2}{\|A\|_F^2} \\ \frac{\lambda_{max}(A^T DA)}{m} & \text{if} \quad \mathbf{P}(S = e_i) = \frac{1}{m} \end{cases} < 1, \tag{14}$$

where the diagonal matrix $D \overset{def}{=} diag(\|A_1\|^{-2}, \cdots, \|A_m\|^{-2})$.

Proof: In this case we have the following expression:

$$S(S^TAA^TS)^\dagger S^T = e_i(e_i^TAA^Te_i)^\dagger e_i^T = \frac{1}{\|A_i\|^2}e_ie_i^T.$$

Then, from Theorem 1 we get for probability distribution $\mathbf{P}(S=e_i) = \frac{\|A_i\|^2}{\|A\|_F^2}$:

$$\gamma = \lambda_{\max}\left(A^T\mathbf{E}\left[\frac{1}{\|A_i\|^2}e_ie_i^T\right]A\right) = \lambda_{\max}\left(A^T\sum_{i=1}^m\frac{\|A_i\|^2}{\|A\|_F^2}\frac{1}{\|A_i\|^2}e_ie_i^TA\right)$$
$$= \lambda_{\max}\left(\frac{A^TA}{\|A\|_F^2}\right) = \lambda_{\max}\left(\frac{AA^T}{\|A\|_F^2}\right).$$

In the last equality we used the fact that the maximum eigenvalues of the matrices A^TA and AA^T coincides. But we can easily see that the trace of the matrix $\frac{AA^T}{\|A\|_F^2}$ is equal to 1 and thus:

$$\sum_{i=1}^m\lambda_i\left(\frac{AA^T}{\|A\|_F^2}\right) = \text{Trace}\left(\frac{AA^T}{\|A\|_F^2}\right) = 1.$$

Therefore, if $\text{rang}(A)\geq 2$, then $\gamma = \lambda_{\max}\left(\frac{AA^T}{\|A\|_F^2}\right) < 1$. Similarly, from Theorem 1 we obtain for the uniform probability distribution $\mathbf{P}(S=e_i) = \frac{1}{m}$:

$$\gamma = \lambda_{\max}\left(A^T\mathbf{E}\left[\frac{1}{\|A_i\|^2}e_ie_i^T\right]A\right) = \lambda_{\max}\left(A^T\sum_{i=1}^m\frac{1}{m}\frac{1}{\|A_i\|^2}e_ie_i^TA\right)$$
$$= \lambda_{\max}\left(\frac{A^TDA}{m}\right) = \lambda_{\max}\left(\frac{AA^TD}{m}\right),$$

where $D = \text{diag}(\|A_1\|^{-2},\cdots,\|A_m\|^{-2})$ and we used the fact that the sets of nonzero eigenvalues of the matrices UV and VU are the same for any two matrices U and V of appropriate dimensions, in particular $U = A^TD$ and $V = A$. Moreover, the trace of the matrix $\frac{AA^TD}{m}$ is equal to 1 and thus:

$$\sum_{i=1}^m\lambda_i\left(\frac{AA^TD}{m}\right) = \text{Trace}\left(\frac{AA^TD}{m}\right) = 1.$$

If $\text{rang}(A)\geq 2$, then $\gamma = \lambda_{\max}\left(\frac{AA^TD}{m}\right) < 1$ for uniform distribution. Q.E.D.

For systems of linear inequalities we can obtain similar statements. For example, we can consider finding a feasible point for a system of linear inequalities $\mathcal{X} = \{x : Ax\leq b\}$, where $A\in\mathbb{R}^{m\times n}$. For this set we can easily construct stochastic approximations sets $\mathcal{X}_S = \{x:\ S^TAx\leq S^Tb\}$, where S is a vector with nonnegative entries, i.e. $S\in\mathbb{R}_+^m$. Clearly, if the vector S has nonnegative entries, we have $\mathcal{X}\subseteq\mathcal{X}_S$. Then, we have the following characterization for γ:

THEOREM 2. *Let us consider finding a solution of a system of linear inequalities $\mathcal{X} = \{x : Ax\leq b\}$, where $A\in\mathbb{R}^{m\times n}$. Further, let us consider the stochastic approximation sets $\mathcal{X}_S = \{x:\ S^TAx\leq S^Tb\}$, where $S\in\Omega = \mathbb{R}_+^m$ and a probability distribution $\mathbf{P}$ on Ω. Then, (13) holds with:*

$$\gamma = \lambda_{max}\left(A^T\mathbf{E}\left[S(S^TAA^TS)^{-1}S\right]A\right)\leq 1.$$

Proof: Clearly, for x satisfying $Ax\leq b$ the inequality (13) holds for any $\gamma\leq 1$. It remains to prove for x satisfying $Ax\not\leq b$. However, since $\mathcal{X}_S = \{x:\ S^TAx\leq S^Tb\}$, then the projection of x onto $\mathcal{X}_S$ can be computed explicitly:

$$\Pi_{\mathcal{X}_S}(x) = x - \frac{\max(0, S^T(Ax-b))}{\|A^TS\|^2}A^TS = x - \frac{\Pi_+(S^T(Ax-b))}{\|A^TS\|^2}A^TS$$

and the relation we need to prove becomes as follows:

$$\left\|\mathbf{E}\left[A^TS(S^TAA^TS)^{-1}\Pi_+(S^T(Ax-b))\right]\right\|^2 \leq \gamma\mathbf{E}\left[\|A^TS(S^TAA^TS)^{-1}\Pi_+(S^T(Ax-b))\|^2\right]$$

or equivalently

$$\left\|A^T\mathbf{E}\left[S(S^TAA^TS)^{-1}\Pi_+(S^T(Ax-b))\right]\right\|^2 \leq \gamma\mathbf{E}\left[\Pi_+(S^T(Ax-b))(S^TAA^TS)^{-1}\Pi_+(S^T(Ax-b))\right].$$

Moreover, if we define the event $\mathcal{I}(x)=\{S\in\Omega : S^T(Ax-b)>0\}$, then the previous relation can be written as follows:

$$\left\|A^T\left(\int_{\mathcal{I}(x)} S(S^TAA^TS)^{-1}S^TdP\right)(Ax-b)\right\|^2 \leq \gamma(Ax-b)^T\left(\int_{\mathcal{I}(x)} S(S^TAA^TS)^{-1}S^TdP\right)(Ax-b).$$

Let us define $E(x)=\int_{\mathcal{I}(x)} S(S^TAA^TS)^{-1}SdP$ and $E=\int_\Omega S(S^TAA^TS)^{-1}SdP$. Then both matrices are positive semidefinite and $E(x)\preceq E$ for all x such that $Ax\not\leq b$. It follows that γ is an upper bound on the following function:

$$\mathcal{R}(x)=\frac{\|A^TE(x)(Ax-b)\|^2}{(Ax-b)^TE(x)(Ax-b)}\leq\gamma \quad \forall x:\ Ax\not\leq b.$$

However, it is easy to find an upper bound for this function $\mathcal{R}(x)$ for each fixed x, namely:

$$\mathcal{R}(x)\leq\lambda_{\max}(A^TE(x)A) \quad \forall x:\ Ax\not\leq b.$$

Since $E(x)\preceq E$, then $A^TE(x)A\preceq A^TEA$ and consequently $\lambda_{\max}(A^TE(x)A)\leq\lambda_{\max}(A^TEA)$. Moreover, there exists x such that $\mathcal{I}(x)=\Omega$. Thus, we have:

$$\gamma=\lambda_{\max}(A^TEA)=\lambda_{\max}\left(A^T\mathbf{E}\left[S(S^TAA^TS)^{-1}S^T\right]A\right).$$

Since the function $W\mapsto\lambda_{\max}(W)$ is convex over the space of positive semidefinite matrices, then using Jensen's inequality we have:

$$\lambda_{\max}\left(A^T\mathbf{E}\left[S(S^TAA^TS)^{-1}S^T\right]A\right)\leq\mathbf{E}\left[\lambda_{\max}\left(A^TS(S^TAA^TS)^{-1}S^TA\right)\right].$$

Furthermore, the matrix $P_S=A^TS(S^TAA^TS)^{-1}S^TA$ is idempotent, that is $P_S^2=P_S$. Therefore, all the eigenvalues of P_S are either 0 or 1. Then, we get:

$$\gamma=\lambda_{\max}\left(A^T\mathbf{E}\left[S(S^TAA^TS)^{-1}S^T\right]A\right)\leq\mathbf{E}\left[\lambda_{\max}\left(A^TS(S^TAA^TS)^{-1}S^TA\right)\right]\leq 1,$$

which proves the statement of the theorem. Q.E.D.

Based on the previous theorem we can prove that for particular choices of the probability distribution $\mathbf{P}$ we have $\gamma<1$, see e.g. the next corollary:

COROLLARY 2. *Let us consider solving a system of linear inequalities $\mathcal{X}=\{x : Ax\leq b\}$, where $A\in\mathbb{R}^{m\times n}$ having rang$(A)\geq 2$. Further, let us consider $\Omega=\{e_1,\cdots,e_m\}$, the standard basis of $\mathbb{R}^m$, and the corresponding stochastic approximation sets $\mathcal{X}_{e_i}=\{x:\ A_i^Tx\leq b_i\}$ for all $i\in[m]$. Then, for two choices of the probability distribution $\mathbf{P}$ on Ω, inequality (13) holds with:*

$$\gamma=\begin{cases}\frac{\lambda_{max}\left(A^TA\right)}{\|A\|_F^2} & \text{if} \quad \mathbf{P}(S=e_i)=\frac{\|A_i\|^2}{\|A\|_F^2}\\ \frac{\lambda_{max}\left(A^TDA\right)}{m} & \text{if} \quad \mathbf{P}(S=e_i)=\frac{1}{m}\end{cases}<1, \tag{15}$$

where the diagonal matrix $D\overset{def}{=}diag(\|A_1\|^{-2},\cdots,\|A_m\|^{-2})$.

Proof: The proof is similar to the one given in Corollary 1. Q.E.D.
The reader can easily find other examples of convex feasibility problems with $\gamma < 1$. The linear regularity inequality (10) and the Jensen type inequality (13) impose strong conditions on the shape of the function F:

THEOREM 3. *Let the linear regularity condition (10) hold. Then, the following bounds are valid for the smooth objective function F:*

$$\frac{1}{2\kappa}\|x-\Pi_{\mathcal{X}}(x)\|^2 \leq F(x)-F^* \leq \frac{\gamma}{2}\|x-\Pi_{\mathcal{X}}(x)\|^2 \quad \forall x \in \mathbb{R}^n, \tag{16}$$

and their dual formulations

$$\frac{1}{2\gamma}\|\nabla F(x)\|^2 \leq F(x)-F^* \leq \frac{\kappa}{2}\|\nabla F(x)\|^2 \quad \forall x \in \mathbb{R}^n. \tag{17}$$

Proof: Under the linear regularity condition (10) we have (12), which represents the left hand side inequality in (16). For proving the right hand side inequality in (16) we use a well-known property of the projection:

$$\|x-\Pi_{\mathcal{X}_S}(x)\|^2 \leq \|x-z\|^2 - \|\Pi_{\mathcal{X}_S}(x)-z\|^2 \quad \forall z \in \mathcal{X}_S. \tag{18}$$

Then, using that $\Pi_{\mathcal{X}}(x) \in \mathcal{X}_S$ we have:

$$\begin{aligned}\mathbf{E}\left[\|x-\Pi_{\mathcal{X}_S}(x)\|^2\right] &= \|x-\Pi_{\mathcal{X}}(x)\|^2 + \mathbf{E}\left[\|\Pi_{\mathcal{X}}(x)-\Pi_{\mathcal{X}_S}(x)\|^2\right] + 2\langle x-\Pi_{\mathcal{X}}(x), \mathbf{E}\left[\Pi_{\mathcal{X}}(x)-\Pi_{\mathcal{X}_S}(x)\right]\rangle \\ &\overset{(18)}{\leq} 2\|x-\Pi_{\mathcal{X}}(x)\|^2 - \mathbf{E}\left[\|x-\Pi_{\mathcal{X}_S}(x)\|^2\right] + 2\langle x-\Pi_{\mathcal{X}}(x), \mathbf{E}\left[\Pi_{\mathcal{X}}(x)-\Pi_{\mathcal{X}_S}(x)\right]\rangle \\ &= -\mathbf{E}\left[\|x-\Pi_{\mathcal{X}_S}(x)\|^2\right] + 2\langle x-\Pi_{\mathcal{X}}(x), \mathbf{E}\left[x-\Pi_{\mathcal{X}_S}(x)\right]\rangle,\end{aligned}$$

where in the first inequality we used (18). In conclusion, we get:

$$\mathbf{E}\left[\|x-\Pi_{\mathcal{X}_S}(x)\|^2\right] \leq \langle x-\Pi_{\mathcal{X}}(x), \mathbf{E}\left[x-\Pi_{\mathcal{X}_S}(x)\right]\rangle. \tag{19}$$

Furthermore, using Cauchy-Schwartz inequality and (13) in (19) we get:

$$\begin{aligned}\mathbf{E}\left[\|x-\Pi_{\mathcal{X}_S}(x)\|^2\right] &\leq \|x-\Pi_{\mathcal{X}}(x)\|\|\mathbf{E}\left[x-\Pi_{\mathcal{X}_S}(x)\right]\| \\ &\leq \|x-\Pi_{\mathcal{X}}(x)\|\sqrt{\gamma\mathbf{E}\left[\|x-\Pi_{\mathcal{X}_S}(x)\|^2\right]}\end{aligned}$$

which leads to:

$$F(x)-F^* \leq \frac{\gamma}{2}\|x-\Pi_{\mathcal{X}}(x)\|^2,$$

that is, the right hand side inequality in (16) holds. This proves the first statement of the theorem, i.e. (16).
For proving the second statement, (17), we first notice that since the Jensen type inequality (13) always holds for some $\gamma \leq 1$ and using the expression of F and that $F^* = 0$, then we can easily find the left hand side inequality in (17):

$$\frac{1}{2}\|\nabla F(x)\|^2 \leq \gamma(F(x)-F^*) \quad \forall x \in \mathbb{R}^n. \tag{20}$$

Then, combining (10) and (19) and using Cauchy-Schwartz inequality, we get:

$$\begin{aligned}\mathbf{E}\left[\|x-\Pi_{\mathcal{X}_S}(x)\|^2\right] &\leq \|x-\Pi_{\mathcal{X}}(x)\|\|\mathbf{E}\left[x-\Pi_{\mathcal{X}_S}(x)\right]\| \\ &\leq \sqrt{\kappa\mathbf{E}\left[\|x-\Pi_{\mathcal{X}_S}(x)\|^2\right]}\|\mathbf{E}\left[x-\Pi_{\mathcal{X}_S}(x)\right]\|,\end{aligned}$$

which leads to

$$\frac{1}{\kappa}(F(x)-F^*)\leq\frac{1}{2}\|\nabla F(x)\|^2.$$

Combining the previous inequality with (20) we obtain the second statement of the theorem, i.e. (17). Q.E.D.

Theorem 3 states that F is strongly convex with constant $\frac{1}{k}$ and has Lipschitz continuous gradient with constant γ when restricted along any segment $[x,\Pi_{\mathcal{X}}(x)]$. Indeed, since $\nabla F(\Pi_{\mathcal{X}}(x))=0$, then from (16)-(17) we obtain:

$$\begin{aligned}\frac{1}{2\kappa}\|x-\Pi_{\mathcal{X}}(x)\|^2+\langle\nabla F(\Pi_{\mathcal{X}}(x)),x-\Pi_{\mathcal{X}}(x)\rangle+F^*\\ \leq F(x)\leq\frac{\gamma}{2}\|x-\Pi_{\mathcal{X}}(x)\|^2+\langle\nabla F(\Pi_{\mathcal{X}}(x)),x-\Pi_{\mathcal{X}}(x)\rangle+F^*\end{aligned}$$

which are exactly the strong convexity condition and the Lipschitz continuity condition, respectively, along any segment $[x,\Pi_{\mathcal{X}}(x)]$, see [26] for more details. It follows that $\kappa\gamma\geq 1$ and $\kappa\gamma$ represents the condition number of the convex feasibility problem (3). Note that F has global Lipschitz continuous gradient with constant $L_F=1$.

3.2. Properties of the operator $\Pi=\mathbf{E}\left[\Pi_{\mathcal{X}_S}\right]$ It is well-known that the projection operator is firmly nonexpansive:

$$\langle\Pi_{\mathcal{X}_S}(x)-\Pi_{\mathcal{X}_S}(y),x-y\rangle\geq\|\Pi_{\mathcal{X}_S}(x)-\Pi_{\mathcal{X}_S}(y)\|^2\quad\forall x,y\in\mathbb{R}^n.$$

Taking the expectation in the previous relation, we get that average projection operator $\Pi(x)=\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]$ is also firmly nonexpansive:

$$\langle\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]-\mathbf{E}\left[\Pi_{\mathcal{X}_S}(y)\right],x-y\rangle\geq\mathbf{E}\left[\|\Pi_{\mathcal{X}_S}(x)-\Pi_{\mathcal{X}_S}(y)\|^2\right]\geq\|\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]-\mathbf{E}\left[\Pi_{\mathcal{X}_S}(y)\right]\|^2\quad(21)$$

for all $x,y\in\mathbb{R}^n$. Similar to Theorem 3 we can derive some contraction inequalities for the average projection operator Π.

Theorem 4. *Let the linear regularity condition (10) hold. Then, the following bounds are valid for the average projection operator $\Pi(x)=\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]$:*

$$(1-\gamma)\|x-x^*\|^2\leq\langle\Pi(x)-\Pi(x^*),x-x^*\rangle\leq\left(1-\frac{1}{\kappa}\right)\|x-x^*\|^2\quad(22)$$

for all $x\in\mathbb{R}^n$ and the corresponding fixed point $x^*=\Pi_{\mathcal{X}}(x)$.

Proof: In order to prove the right hand side inequality, we choose in (21) the fixed point $y=\Pi_{\mathcal{X}}(x)$, which leads to:

$$\begin{aligned}&\langle\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]-\Pi_{\mathcal{X}}(x),x-\Pi_{\mathcal{X}}(x)\rangle\geq\mathbf{E}\left[\|\Pi_{\mathcal{X}_S}(x)-\Pi_{\mathcal{X}}(x)\|^2\right]\\&=\mathbf{E}\left[\|\Pi_{\mathcal{X}_S}(x)-x\|^2\right]-\|x-\Pi_{\mathcal{X}}(x)\|^2+2\langle\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]-\Pi_{\mathcal{X}}(x),x-\Pi_{\mathcal{X}}(x)\rangle,\end{aligned}$$

which combined with (10) leads to

$$\begin{aligned}\langle\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]-\Pi_{\mathcal{X}}(x),x-\Pi_{\mathcal{X}}(x)\rangle&\leq\|x-\Pi_{\mathcal{X}}(x)\|^2-\mathbf{E}\left[\|\Pi_{\mathcal{X}_S}(x)-x\|^2\right]\\&\overset{(10)}{\leq}\left(1-\frac{1}{\kappa}\right)\|x-\Pi_{\mathcal{X}}(x)\|^2.\end{aligned}$$

For the left hand side inequality we proceed as follows:

$$\begin{aligned}\langle \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]-\Pi_{\mathcal{X}}(x), x-\Pi_{\mathcal{X}}(x)\rangle &= \|x-\Pi_{\mathcal{X}}(x)\|^2+\langle \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]-x, x-\Pi_{\mathcal{X}}(x)\rangle\\ &\geq \|x-\Pi_{\mathcal{X}}(x)\|^2-\|x-\Pi_{\mathcal{X}}(x)\|\|\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]-x\|\\ &\geq \|x-\Pi_{\mathcal{X}}(x)\|^2-\|x-\Pi_{\mathcal{X}}(x)\|\sqrt{\gamma\mathbf{E}\left[\|\Pi_{\mathcal{X}_S}(x)-x\|^2\right]}\\ &= \|x-\Pi_{\mathcal{X}}(x)\|^2-\|x-\Pi_{\mathcal{X}}(x)\|\sqrt{2\gamma(F(x)-F^*)}\\ &\overset{(16)}{\geq} \|x-\Pi_{\mathcal{X}}(x)\|^2-\|x-\Pi_{\mathcal{X}}(x)\|\gamma\sqrt{\|x-\Pi_{\mathcal{X}}(x)\|^2}\\ &= (1-\gamma)\|x-\Pi_{\mathcal{X}}(x)\|^2,\end{aligned}$$

where in the first inequality we used Cauchy-Schwartz inequality, in the second inequality we used (13) and in the third inequality we used relation (16). Q.E.D.
Theorem 4 shows that the operator Π is a contraction with contraction constant $c=1-\frac{1}{k}<1$ when restricted along any segment $[x,\Pi_{\mathcal{X}}(x)]$.

4. Examples: finite intersection We consider X represented as the intersection of a finite family of convex sets:

$$\mathcal{X}=\bigcap_{i=1}^{m}\mathcal{X}_i,$$

where $\mathcal{X}_i$ are nonempty closed convex sets. We also assume that $\mathcal{X}\neq\emptyset$. In several papers, such as [5,27], the authors introduced a *linear regularity property* for the set $\mathcal{X}=\bigcap_{i=1}^{m}\mathcal{X}_i$. That is, there exists $\kappa_{\max}<\infty$ such that:

$$\operatorname{dist}^2_{\mathcal{X}}(x)\leq\kappa_{\max}\max_{i\in[m]}\operatorname{dist}^2_{\mathcal{X}_i}(x)\qquad\forall x\in\mathbb{R}^n. \tag{23}$$

Based on this condition, linear convergence rate, depending on the constant $\kappa_{\max}$, has been derived for the alternating projection algorithm (B-AP). Note that our definition of linear regularity (10) extends the one given in (23) for finite intersection to the more general convex feasibility problem (3). More precisely, in order to show linear convergence for our general algorithmic framework introduced in this paper, we require the linear regularity property for the set $\mathcal{X}=\cap_{S\in\Omega}\mathcal{X}_S$ defined in (10). For a uniform probability over the set $\Omega=[m]\overset{\text{def}}{=}\{1,2,\ldots,m\}$ we have:

$$\begin{aligned}\operatorname{dist}^2_{\mathcal{X}}(x)&\leq\kappa_{\max}\max_{i\in[m]}\operatorname{dist}^2_{\mathcal{X}_i}(x)\leq\kappa_{\max}\sum_{i=1}^{m}\operatorname{dist}^2_{\mathcal{X}_i}(x)\\ &=m\cdot\kappa_{\max}\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right].\end{aligned}$$

This shows that:

$$\kappa\leq m\cdot\kappa_{\max}.$$

Thus, condition (23) is a relaxation of our more general condition (10), also analyzed in [28]. Further, we analyze this property (10) and estimate the constant κ for several representative cases of stochastic approximation sets for $\mathcal{X}$.

4.1. Standard Let $\mathcal{X}_S=\mathcal{X}_i$ for all $i\in\Omega=[m]$, endowed with some probability $p_i\geq 0$. Since $\cap_{i=1}^{m}\mathcal{X}_i=\mathcal{X}\subseteq\mathcal{X}_S$, then $\mathcal{X}_S$ is a stochastic approximation of $\mathcal{X}$. Note that:

$$\mathcal{Y}=\left\{x:\sum_{i=1}^{m}p_i\mathbb{I}_{\mathcal{X}_S}(x)=0\right\}=\bigcap_{i:\,p_i>0}\mathcal{X}_i.$$

Hence, a sufficient condition for exactness is to require $p_i > 0$ for all $i \in [m]$. Moreover, under this condition and (23) it follows that linear regularity (10) holds with $\kappa = \frac{\kappa_{\max}}{p_{\min}}$, where $p_{\min} = \min_{i\in[m]} p_i$. Indeed, we can use the following inequality:

$$p_{\min} \max_{i\in[m]} \mathrm{dist}^2_{\mathcal{X}_i}(x) \leq \sum_{i=1}^m p_{\min} \mathrm{dist}^2_{\mathcal{X}_i}(x) \leq \sum_{i=1}^m p_i \mathrm{dist}^2_{\mathcal{X}_i}(x) = \mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}_S}(x)\right].$$

4.2. Subsets With each nonempty subset $S \subseteq [m]$ we associate a probability $p_S \geq 0$, such that $\sum_{S\subseteq[m]} p_S = 1$. We then define $\mathcal{X}_S = \cap_{i\in S} \mathcal{X}_i$ with probability p_S. Since $\mathcal{X} \subseteq \mathcal{X}_S$, then this is a stochastic approximation. Moreover,

$$\mathcal{Y} = \left\{ x \,:\, \sum_S p_S \mathbb{I}_{\mathcal{X}_S}(x) = 0 \right\} = \bigcap_{S\,:\,p_S>0} \mathcal{X}_S.$$

A sufficient condition for the last set to be equal to $\mathcal{X}$ (i.e., a sufficient condition for exactness) is

$$[m] = \bigcup_{S\,:\,p_S>0} S.$$

In words, this condition requires us to assign positive probabilities to some collection of subsets covering $[m]$. If we only assign positive probabilities to singletons, we recover example 1. Moreover, under this condition and (23) it follows that linear regularity (10) holds with $\kappa = \frac{\kappa_{\max}}{p_{\min}}$, where $p_{\min} = \min_{S:p_S>0} p_S$. This is due to the fact that $\max_{i\in[m]} \mathrm{dist}^2_{\mathcal{X}_i}(x) \leq \sum_{i=1}^m \mathrm{dist}^2_{\mathcal{X}_i}(x)$, that $\mathrm{dist}^2_{\mathcal{X}_i}(x) \leq \mathrm{dist}^2_{\cap_{j\in S}\mathcal{X}_j}(x) = \mathrm{dist}^2_{\mathcal{X}_S}(x), \forall i \in S$ and that we assume there is a collection of subsets S covering $[m]$.

4.3. Convex combination Fix $r \in [m]$, and let us consider a countable subset Ω_r defined as follows:

$$\Omega_r \subset \left\{ S \in \mathbb{R}^m : \sum_{i=1}^m S_i = 1,\ S \geq 0,\ \|S\|_0 \leq r \right\}.$$

Let us consider a discrete probability distribution $\mathbf{P}$ on Ω_r. We then choose $S \sim \mathbf{P}$ and define the stochastic approximation set as:

$$\mathcal{X}_S = \sum_{i=1}^m S_i \mathcal{X}_i \stackrel{\mathrm{def}}{=} \left\{ \sum_{i=1}^m S_i x_i \,:\, x_i \in \mathcal{X}_i \right\}.$$

This is clearly a stochastic approximation, that is $\mathcal{X} \subseteq \mathcal{X}_S$, since $\sum_{i=1}^m S_i = 1$ and for any $x \in \mathcal{X}$ it follows that $x \in \mathcal{X}_i$ for all $i \in [m]$ and thus $x = \sum_i S_i x \in \mathcal{X}_S$. For $r = 1$ we recover the standard example from Section 4.1. If additionally, we assume that Ω_r contains the basic vectors, i.e. $\{e_1, \cdots, e_m\} \subseteq \Omega_r$, and $\mathcal{X}_S$ defined as above, then exactness holds when $p_i = \mathbf{P}(S = e_i) > 0$ for all $i \in [m]$. Indeed, if $x \in \mathcal{Y}$, then:

$$0 = \mathbf{E}\left[\mathbb{I}_{\mathcal{X}_S}(x)\right] = \sum_{S\in\Omega} p_S \mathbb{I}_{\mathcal{X}_S}(x) \geq \sum_{S\in\{e_1,\cdots,e_m\}} p_S \mathbb{I}_{\mathcal{X}_S}(x),$$

which implies $x \in \mathcal{X}_i$, provided that $p_i > 0$, for all $i \in [m]$. Moreover, under this condition and (23) it follows that linear regularity (10) holds with $\kappa = \frac{\kappa_{\max}}{p_{\min}}$, where $p_{\min} = \min_{i\in[m]} p_i$. This is due to the fact that $\mathcal{X}_{e_i} = \mathcal{X}_i$ and that:

$$p_{\min} \max_{i\in[m]} \mathrm{dist}^2_{\mathcal{X}_i}(x) \leq \sum_{i=1}^m p_i \mathrm{dist}^2_{\mathcal{X}_i}(x) \leq \sum_{S\in\Omega} p_S \mathrm{dist}^2_{\mathcal{X}_S}(x) = \mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}_S}(x)\right].$$

4.4. Equality constraints Assume a linear representation for the set $\mathcal{X}$, that is $\mathcal{X} = \{x \in \mathbb{R}^n : Ax = b\}$, where the matrix $A \in \mathbb{R}^{m\times n}$. In this case we have $\mathcal{X}_i = \{x \in \mathbb{R}^n : A_i^T x - b_i = 0\}$, where A_i is the ith row of matrix A. Let $q \leq m$, $\Omega \subseteq \mathbb{R}^{m\times q}$ and a probability distribution $\mathbf{P}$ on Ω. Thus, we define the stochastic approximation:

$$\mathcal{X}_S = \{x \in \mathbb{R}^n : \ S^T Ax = S^T b\} \quad \forall S \in \Omega.$$

We notice that $\cap_{S\in\Omega}\mathcal{X}_S = \{x : SAx = Sb \, \forall S \in \Omega\}$. If we can find m linearly independent columns in the family of matrices $(S)_{S\in\Omega}$, then $\mathcal{X} = \cap_{S\in\Omega}\mathcal{X}_S$. Next we derive sufficient conditions for exactness, that is conditions that guarantee $\mathcal{X} = \mathcal{Y}$, and we also provide an estimate for κ.

THEOREM 5. *Let $\mathcal{X} = \{x \in \mathbb{R}^n : Ax = b\}$, with $A \in \mathbb{R}^{m\times n}$ and consider the stochastic approximation $\mathcal{X}_S = \{x \in \mathbb{R}^n : S^T Ax = S^T b\}$, where $S \in \mathbb{R}^{m\times q}$ is a random matrix in the probability space $(\Omega, \mathbf{P})$. Furthermore, assume that S satisfies $\mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right] \succ 0$. Then, we have exactness and the linear regularity property (10) holds with constant:*

$$\kappa = \frac{1}{\lambda_{\min}^{nz}(A^T \mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right] A)} > 0. \tag{24}$$

Proof: Notice that the projection $\Pi_{\mathcal{X}_S}(x)$ of x onto $\mathcal{X}_S$ can be expressed as:

$$\Pi_{\mathcal{X}_S}(x) = x - A^T S(S^T AA^T S)^\dagger S^T (Ax - b),$$

thus the local distance $\text{dist}_{\mathcal{X}_S}(x)$ from x to the set $\mathcal{X}_S$ is given by:

$$\begin{aligned} \text{dist}_{\mathcal{X}_S}(x) = \|x - \Pi_{\mathcal{X}_S}(x)\| &= \|A^T S(S^T AA^T S)^\dagger S^T (Ax - b)\| \\ &= \|A^T S(S^T AA^T S)^\dagger S^T A(x - \Pi_{\mathcal{X}}(x))\|. \end{aligned} \tag{25}$$

Further, the matrix $P_S = A^T S(S^T AA^T S)^\dagger S^T A$ is idempotent, that is $P_S^2 = P_S$, which implies that $\|P_S z\|^2 = z^T P_S z$ for any $z \in \mathbb{R}^n$. By squaring and taking expectation in both sides of (25) and also using the previous property of P_S, we further obtain:

$$\begin{aligned} \mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right] \overset{(25)}{=} &\mathbf{E}\left[\|P_S(x - \Pi_{\mathcal{X}}(x))\|^2\right] \\ =&\mathbf{E}\left[(x - \Pi_{\mathcal{X}}(x))^T P_S (x - \Pi_{\mathcal{X}}(x))\right] \\ =&(x - \Pi_{\mathcal{X}}(x))^T \mathbf{E}\left[P_S\right](x - \Pi_{\mathcal{X}}(x)). \end{aligned} \tag{26}$$

On the other hand, it is well known from the Courant-Fischer theorem [30], that for any $C \in \mathbb{R}^{m\times n}$ we have:

$$\|Cz\| \geq \sigma_{\min}^{\text{nz}}(C)\|z\| \qquad \forall z \in \text{Im}(C^T),$$

where recall that $\sigma_{\min}^{\text{nz}}$ denotes the smallest nonzero singular value of a matrix. If we define the matrix $E = \mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right]$ and take $C = E^{1/2}A$, then the above relation leads to:

$$\|E^{1/2}Az\| \geq \sigma_{\min}^{\text{nz}}(E^{1/2}A)\|z\| \qquad \forall z \in \text{Im}(A^T E^{1/2}). \tag{27}$$

Further, since we assume that $E = \mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right] \succ 0$, then $E^{1/2} \succ 0$ and $\text{Im}(A^T) = \text{Im}(A^T E^{1/2})$. Moreover, we have the fact that $x - \Pi_{\mathcal{X}}(x) \in \text{Im}(A^T)$. Therefore, by applying the relation (27) for $z = x - \Pi_{\mathcal{X}}(x)$, observing that $\mathbf{E}\left[P_S\right] = A^T EA$, and by combining relations (26) and (27), we have:

$$\begin{aligned} \mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right] &= \|E^{1/2}A(x - \Pi_{\mathcal{X}}(x))\|^2 \\ &\overset{(27)}{\geq} \left(\sigma_{\min}^{\text{nz}}(E^{1/2}A)\right)^2 \text{dist}^2_{\mathcal{X}}(x) \\ &= \lambda_{\min}^{\text{nz}}(A^T EA)\text{dist}^2_{\mathcal{X}}(x) \\ &= \lambda_{\min}^{\text{nz}}(\mathbf{E}\left[P_S\right])\text{dist}^2_{\mathcal{X}}(x) \\ &= \lambda_{\min}^{\text{nz}}(\mathbf{E}\left[A^T S(S^T AA^T S)^\dagger S^T A\right])\text{dist}^2_{\mathcal{X}}(x) \end{aligned}$$

for all $x \in \mathbb{R}^n$. This final relation implies our statement. Q.E.D.
In [31] it has been proved that, when we consider discrete samplings, such as $S \in \Omega = \{e^1, \cdots, e^m\}$, and full row rank matrices A with no strictly zero rows, the matrix $\mathbf{E}\left[S^T(SAA^TS^T)^\dagger S\right]$ is positive definite, that is it satisfies our assumption considered in the previous theorem. A simple consequence of previous theorem is the following:

COROLLARY 3. *If we consider $\Omega = \{e_1, \cdots, e_m\}$, then for two choices of the probability distribution $\mathbf{P}$ on Ω the linear regularity constant takes the form:*

$$\kappa = \begin{cases} \frac{\|A\|_F^2}{\lambda_{\min}^{nz}\left(A^TA\right)} & \text{if} \quad \mathbf{P}(S = e_i) = \frac{\|A_i\|^2}{\|A\|_F^2} \\ \frac{m}{\lambda_{\min}^{nz}\left(A^TDA\right)} & \text{if} \quad \mathbf{P}(S = e_i) = \frac{1}{m} \end{cases} \geq 1, \tag{28}$$

where the diagonal matrix $D \overset{def}{=} diag(\|A_1\|^{-2}, \cdots, \|A_m\|^{-2})$.

Proof: If $\Omega = \{e_1, \cdots, e_m\}$ and the probability $\mathbf{P}(S = e_i) = \|A_i\|^2/\|A\|_F^2$, then the stochastic approximation set $\mathcal{X}_{e_i}$ is given by a linear hyperplane, i.e. $\mathcal{X}_{e_i} = \{x \in \mathbb{R}^n : A_i^T x = b_i\}$, and the expression in (24) becomes:

$$\lambda_{\min}^{\text{nz}}(A^T\mathbf{E}\left[S(S^TAA^TS)^\dagger S^T\right]A) = \lambda_{\min}^{\text{nz}}\left(A^T\mathbf{E}\left[\frac{e_ie_i^T}{\|A_i\|^2}\right]A\right)$$
$$= \lambda_{\min}^{\text{nz}}\left(A^T\sum_{i=1}^m \frac{\|A_i\|^2}{\|A\|_F^2}\frac{e_ie_i^T}{\|A_i\|^2}A\right) = \lambda_{\min}^{\text{nz}}\left(A^T\frac{I_m}{\|A\|_F^2}A\right) = \frac{\lambda_{\min}^{\text{nz}}(A^TA)}{\|A\|_F^2}.$$

Thus, in this case, the linear regularity constant is given by:

$$\kappa \overset{(24)}{=} \frac{\|A\|_F^2}{\lambda_{\min}^{\text{nz}}\left(A^TA\right)} = \left(\frac{\|A\|_F}{\sigma_{\min}^{\text{nz}}(A)}\right)^2 \geq 1.$$

For the uniform probability $\mathbf{P}(S = e_i) = 1/m$, the expression in (24) becomes:

$$\lambda_{\min}^{\text{nz}}(A^T\mathbf{E}\left[S(S^TAA^TS)^\dagger S^T\right]A) = \lambda_{\min}^{\text{nz}}\left(A^T\mathbf{E}\left[\frac{e_ie_i^T}{\|A_i\|^2}\right]A\right)$$
$$= \lambda_{\min}^{\text{nz}}\left(A^T\sum_{i=1}^m \frac{1}{m}\frac{e_ie_i^T}{\|A_i\|^2}A\right) = \frac{\lambda_{\min}^{\text{nz}}(A^TDA)}{m},$$

where the diagonal matrix $D = \text{diag}(\|A_1\|^{-2}, \cdots, \|A_m\|^{-2})$. This proves our statement. Q.E.D.

4.5. Inequality constraints Let $q \leq m$, $\Omega \subseteq \mathbb{R}_+^{m\times q}$ the set of matrices with nonnegative entries, i.e., $\mathbb{R}_+^{m\times q} = \{S \in \mathbb{R}^{m\times q} : S_{ij} \geq 0 \ \forall i \in [m], j \in [q]\}$, and a probability distribution $\mathbf{P}$ on Ω. Assume a functional representation for the set $\mathcal{X}$, that is $\mathcal{X} = \{x \in \mathbb{R}^n : \mathcal{F}(x) \leq 0\}$, where $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}^m$ is a vector of convex closed functions, that is $\mathcal{F} = (\mathcal{F}_1, \cdots, \mathcal{F}_m)$. In this case we have $\mathcal{X}_i = \{x \in \mathbb{R}^n : \mathcal{F}_i(x) \leq 0\}$. Thus, we define the stochastic approximation:

$$\mathcal{X}_S = \{x \in \mathbb{R}^n : S^T\mathcal{F}(x) \leq 0\} \quad \forall S \in \Omega.$$

We notice that $\cap_{S\in\Omega}\mathcal{X}_S = \{x : S^T\mathcal{F}(x) \leq 0 \, \forall S \in \Omega\}$. If there exist m linearly independent columns in the family of matrices $(S)_{S\in\Omega}$, then $\mathcal{X} = \cap_S\mathcal{X}_S$. Moreover, if the probability space is finite, then we also have exactness. Next, we provide estimates for the linear regularity constant κ for some particular sets. First, we consider finding a point in the intersection of halfspaces, that is $\mathcal{X} = \{x \in \mathbb{R}^n : Ax \leq b\}$.

THEOREM 6. *Let $\mathcal{X} = \{x \in \mathbb{R}^n : Ax \leq b\}$ and consider stochastic approximation halfspaces $\mathcal{X}_S = \{x \in \mathbb{R}^n : S^T Ax \leq S^T b\}$, where S is a random vector from the finite probability space $\Omega_r \subset \{S \in \mathbb{R}^m : S \geq 0, \|S\|_0 \leq r\}$ for some given $r \in [m]$ endowed with a probability distribution $\mathbf{P} = (p_S)_{S \in \Omega_r}$. We further denote the Hoffman constant for the polyhedral set $\mathcal{X}$ with $\tilde{\kappa}$. Then, under exactness the linear regularity property (10) holds with constant:*

$$\kappa = \frac{\max\limits_{S \in \Omega_r} \|A^T S\|^2}{\min\limits_{S \in \Omega_r} p_S} \tilde{\kappa}. \tag{29}$$

Proof: Notice that in this case we have an explicit projection onto $\mathcal{X}_S$ given by $\Pi_{\mathcal{X}_S}(x) = x - \frac{\Pi_+(S^T Ax - S^T b)}{\|A^T S\|^2} A^T S$, which implies that:

$$\operatorname{dist}_{\mathcal{X}_S}(x) = \frac{\Pi_+(S^T Ax - S^T b)}{\|A^T S\|} \geq \frac{\Pi_+(S^T Ax - S^T b)}{\max\limits_{S \in \Omega_r} \|A^T S\|}. \tag{30}$$

From Markov inequality we have:

$$\frac{\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right]}{\max\limits_{S \in \Omega_r} \operatorname{dist}^2_{\mathcal{X}_S}(x)} \geq \mathbf{P}\left(\operatorname{dist}^2_{\mathcal{X}_S}(x) \geq \max_{S \in \Omega_r} \operatorname{dist}^2_{\mathcal{X}_S}(x)\right).$$

Combining the previous inequality with (30), we obtain:

$$\begin{aligned}
\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right] &\geq \mathbf{P}(\operatorname{dist}^2_{\mathcal{X}_S}(x) \geq \max_{S \in \Omega_r} \operatorname{dist}^2_{\mathcal{X}_S}(x)) \cdot \max_{S \in \Omega_r} \operatorname{dist}^2_{\mathcal{X}_S}(x) \\
&= \mathbf{P}(\operatorname{dist}_{\mathcal{X}_S}(x) \geq \max_{S \in \Omega_r} \operatorname{dist}_{\mathcal{X}_S}(x)) \cdot \max_{S \in \Omega_r} \operatorname{dist}^2_{\mathcal{X}_S}(x) \\
&\geq \min_{S \in \Omega_r} p_S \cdot \max_{S \in \Omega_r} \operatorname{dist}^2_{\mathcal{X}_S}(x) \\
&\overset{(30)}{\geq} \min_{S \in \Omega_r} p_S \cdot \frac{\max\limits_{S \in \Omega_r} \Pi^2_+(S^T Ax - S^T b)}{\max\limits_{S \in \Omega_r} \|A^T S\|^2}.
\end{aligned} \tag{31}$$

On the other hand it is well know that for a polyhedral set the Hoffman inequality is valid, see [9]. Since we assume exactness and that Ω_r has a finite number of elements, then there exists some positive constant $\tilde{\kappa} > 0$ such that:

$$\operatorname{dist}^2_{\mathcal{X}}(x) \leq \tilde{\kappa} \max_{S \in \Omega_r} \Pi^2_+(S^T Ax - S^T b) \quad \forall x \in \mathbb{R}^n.$$

Using this Hoffman inequality in (31) leads to the relation:

$$\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right] \geq \frac{\min\limits_{S \in \Omega_r} p_S}{\tilde{\kappa} \max\limits_{S \in \Omega_r} \|A^T S\|^2} \operatorname{dist}^2_{\mathcal{X}}(x) \quad \forall x \in \mathbb{R}^n,$$

which proves our statement. Q.E.D.
However, for a specific choice of the probability distribution we can get better estimate for κ, as the next corollary shows:

COROLLARY 4. *Let $\mathcal{X} = \{x \in \mathbb{R}^n : Ax \leq b\}$ and consider stochastic approximation halfspaces $\mathcal{X}_S = \{x \in \mathbb{R}^n : S^T Ax \leq S^T b\}$, where S is a random vector from the finite probability space $\Omega_r \subset \{S \in \mathbb{R}^m : S \geq 0, \|S\|_0 \leq r\}$ for some given $r \in [m]$ endowed with the probability distribution $\mathbf{P} = (p_S)_{S \in \Omega_r}$ given by $p_S = \|A^T S\|^2 / \sum_{S \in \Omega_r} \|A^T S\|^2$. We further denote the Hoffman constant for the polyhedral set $\mathcal{X}$ with $\tilde{\kappa}$. Then, under exactness the linear regularity property (10) holds with constant:*

$$\kappa = \tilde{\kappa} \sum_{S \in \Omega_r} \|A^T S\|^2. \tag{32}$$

Proof: Since $\Pi_{\mathcal{X}_S}(x) = x - \frac{\Pi_+(S^TAx - S^Tb)}{\|A^TS\|^2}A^TS$, then we have:

$$\text{dist}_{\mathcal{X}_S}(x) = \frac{\Pi_+(S^TAx - S^Tb)}{\|A^TS\|}.$$

Using the expressions for the distance and for the probability, we further have:

$$\begin{aligned}\mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right] &= \sum_{S\in\Omega_r} p_S \text{dist}^2_{\mathcal{X}_S}(x)\\ &= \sum_{S\in\Omega_r} \frac{\|A^TS\|^2}{\sum_{S\in\Omega_r}\|A^TS\|^2}\cdot\frac{\Pi^2_+(S^TAx - S^Tb)}{\|A^TS\|^2}\\ &= \frac{1}{\sum_{S\in\Omega_r}\|A^TS\|^2}\sum_{S\in\Omega_r}\Pi^2_+(S^TAx - S^Tb). \end{aligned} \tag{33}$$

On the other hand, under exactness and Ω_r has a finite number of elements there exists some positive Hoffman constant $\tilde{\kappa} > 0$ such that:

$$\text{dist}^2_{\mathcal{X}}(x) \le \tilde{\kappa}\sum_{S\in\Omega_r}\Pi^2_+(S^TAx - S^Tb) \quad \forall x\in\mathbb{R}^n.$$

Using the Hoffman inequality in (33) leads to the relation:

$$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right] \ge \frac{1}{\tilde{\kappa}\sum_{S\in\Omega_r}\|A^TS\|^2}\text{dist}^2_{\mathcal{X}}(x) \quad \forall x\in\mathbb{R}^n,$$

which proves our statement. Q.E.D.
Second, following similar ideas as in [18,27], we consider the general case of a convex set $\mathcal{X}$ with nonempty interior, that is, there exists a ball of radius $\delta > 0$ and center $\bar{x}\in\mathcal{X}$ such that:

$$\{x\in\mathbb{R}^n : \|\bar{x} - x\| \le \delta\} \subseteq \mathcal{X}.$$

THEOREM 7. *Let $\mathcal{X}$ be a convex set with nonempty interior, that is there exists $\delta > 0$ and $\bar{x}\in\mathcal{X}$ such that $\{x\in\mathbb{R}^n : \|\bar{x} - x\| \le \delta\} \subseteq \mathcal{X}$. Consider any family of stochastic approximations $\mathcal{X}_S$, where S is a random variable from the finite probability space Ω endowed with a probability distribution $\mathbf{P} = (p_S)_{S\in\Omega}$. Then, under exactness the linear regularity property (10) holds over any bounded set Q with constant:*

$$\kappa = \frac{\max\limits_{x\in Q}\|x - \bar{x}\|^2}{\delta^2\min\limits_{S\in\Omega} p_S} \qquad \forall x\in Q. \tag{34}$$

Proof: Let us define for some $\alpha > 0$ and $x\in\mathbb{R}^n$ the vector:

$$y_\alpha(x) = \frac{\alpha}{\alpha+\delta}\bar{x} + \frac{\delta}{\alpha+\delta}x.$$

Now we show that by choosing $\tilde{\alpha} = \max\limits_{S\in\Omega}\text{dist}_{\mathcal{X}_S}(x)$, then $y_{\tilde{\alpha}}(x)\in\mathcal{X}$ for all $x\in\mathbb{R}^n$. Indeed, we first rewrite $y_{\tilde{\alpha}}(x)$ as:

$$y_{\tilde{\alpha}}(x) = \frac{\tilde{\alpha}}{\tilde{\alpha}+\delta}z + \frac{\delta}{\tilde{\alpha}+\delta}\Pi_{\mathcal{X}_S}(x),$$

where $z = \bar{x} + \frac{\delta}{\tilde{\alpha}}(x - \Pi_{\mathcal{X}_S}(x))$. Notice that:

$$\|z - \bar{x}\| = \frac{\delta}{\alpha}\|x - \Pi_{\mathcal{X}_S}(x)\| = \delta\frac{\text{dist}_{\mathcal{X}_S}(x)}{\max\limits_{S\in\Omega}\text{dist}_{\mathcal{X}_S}(x)} \le \delta.$$

Thus, we have $z \in \mathcal{X}$, which implies $z \in \mathcal{X}_S$ for all $S \in \Omega$. Since $z \in \mathcal{X}_S$, then further we conclude that also $y_\alpha(x) \in \mathcal{X}_S$ for all $\mathcal{X}_S$, which finally confirms that $y_\alpha(x) \in \mathcal{X}$, due to exactness. By using this fact, results that:

$$\begin{aligned} \text{dist}_{\mathcal{X}}(x) &\leq \|y_{\tilde{\alpha}}(x) - x\| = \frac{\tilde{\alpha}}{\tilde{\alpha}+\delta}\|x - \bar{x}\| \\ &\leq \frac{\tilde{\alpha}}{\delta}\|x - \bar{x}\| = \frac{\|x-\bar{x}\|}{\delta} \max_{S\in\Omega} \text{dist}_{\mathcal{X}_S}(x). \end{aligned} \tag{35}$$

From the Markov inequality we get the bound:

$$\min_{S\in\Omega} p_S \leq \mathbf{P}(\text{dist}^2_{\mathcal{X}_S}(x) \geq \max_{S\in\Omega} \text{dist}^2_{\mathcal{X}_S}(x)) \leq \frac{\mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right]}{\max_{S\in\Omega} \text{dist}^2_{\mathcal{X}_S}(x)}. \tag{36}$$

Using (35) and (36), we obtain for any $x \in Q$:

$$\text{dist}^2_{\mathcal{X}}(x) \overset{(35)+(36)}{\leq} \frac{\|x-\bar{x}\|^2}{\delta^2 \min_{S\in\Omega} p_S} \mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right] \leq \frac{\max_{x\in Q}\|x-\bar{x}\|^2}{\delta^2 \min_{S\in\Omega} p_S} \mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right],$$

which confirms our result. Q.E.D.

5. Examples: infinite intersection Assume $\mathcal{X} = \cap_{S\in\Omega}\mathcal{X}_S$, for some (possibly infinite) index set Ω and sets $\mathcal{X}_S \subseteq \mathbb{R}^n$. Many interesting applications can be modeled as the intersection of infinite (countable/uncountable) number of simple convex sets, see e.g. [29] for some control and machine learning applications. Let $\mathbf{P}$ be a probability measure on Ω. Then, if we choose $S \sim \mathbf{P}$, $\mathcal{X}_S$ is a stochastic approximation of $\mathcal{X}$. Note that

$$\mathcal{Y} = \{x : \mathbf{P}(x \in \mathcal{X}_S) = 1\}.$$

5.1. Separation oracle Assume that we have access to a *separation oracle* for $\mathcal{X}$. That is, for each $S \in \mathbb{R}^n$, the oracle either confirms that $S \in \mathcal{X}$, or outputs a vector $g = g(S) \in \mathbb{R}^n$ such that $\langle g, z - S\rangle \leq 0$ for all $z \in \mathcal{X}$. If we let

$$\mathcal{X}_S \overset{\text{def}}{=} \begin{cases} \mathbb{R}^n & S \in \mathcal{X} \\ \{x : \langle g, x - S\rangle \leq 0\} & S \notin \mathcal{X}, \end{cases}$$

then clearly $\mathcal{X} \subseteq \mathcal{X}_S$ for all $S \in \mathbb{R}^n$. Given any distribution $\mathbf{P}$ over $\mathbb{R}^n$, $\mathcal{X}_S$ is a stochastic approximation of $\mathcal{X}$. In this case we can only guarantee:

$$\mathcal{X} \subseteq \cap_{S\in\mathbb{R}^n}\mathcal{X}_S.$$

5.2. Supporting halfspaces A particular case of the convex feasibility problem is the so-called split feasibility problem [10]:

$$\text{Find } x \in \mathcal{X} = \{x \in \mathbb{R}^n : Ax \in \mathcal{Z}\},$$

i.e., $\mathcal{X}$ is defined by imposing convex constraints defined by the set $\mathcal{Z}$ in the range of the matrix $A \in \mathbb{R}^{m\times n}$. Then, if we choose any $S \in \mathbb{R}^n$ we can define a stochastic approximation as the entire space or the following halfspace:

$$\mathcal{X}_S \overset{\text{def}}{=} \begin{cases} \mathbb{R}^n & S \in \mathcal{X} \\ \{x : c_S^T x \leq b_S\} & S \notin \mathcal{X}, \end{cases}$$

where $c_S \neq 0$ and b_S are defined as follows:

$$c_S = A^T(AS - \Pi_{\mathcal{Z}}(AS)) \text{ and } b_S = \|AS\|^2 - (\Pi_{\mathcal{Z}}(AS))^T AS - \|AS - \Pi_{\mathcal{Z}}(AS)\|^2.$$

Note that the halfspace $\mathcal{X}_S = \{x : c_S^T x \leq b_S\}$ can be written equivalently as:

$$\mathcal{X}_S = \{x : \langle AS - \Pi_{\mathcal{Z}}(AS), Ax - \Pi_{\mathcal{Z}}(AS) \leq 0\}.$$

It is easy to check using the optimality conditions for the projection onto $\mathcal{Z}$ that for any $S \notin \mathcal{X}$ the hyperplane $c_S^T x = b_S$ separates S from $\mathcal{X}$, that is:

$$\mathcal{X} \subseteq \mathcal{X}_S \quad \forall S \in \mathbb{R}^n.$$

Therefore, given any distribution $\mathbf{P}$ over $\mathbb{R}^n$, the halfspace $\mathcal{X}_S$ is a stochastic approximation of $\mathcal{X}$. In fact, in this case we have:

$$\mathcal{X} = \cap_{S \in \mathbb{R}^n} \mathcal{X}_S.$$

Indeed, it is straightforward that we have $\mathcal{X} \subseteq \cap_{S \in \mathbb{R}^n} \mathcal{X}_S$. For the other inclusion, let us take any $x \in \cap_{S \in \mathbb{R}^n} \mathcal{X}_S$. Then, $x \in \mathcal{X}_S$ for any fixed S. Now, if we make the particular choice $S = x$, then $x \in \mathcal{X}_x$, that is it satisfies:

$$\langle Ax - \Pi_{\mathcal{Z}}(Ax), Ax - \Pi_{\mathcal{Z}}(Ax) \leq 0$$

which holds if and only if $Ax = \Pi_{\mathcal{Z}}(Ax)$, that is $x \in \mathcal{X}$.

5.3. Normal cone Let $\Omega \in \mathbb{R}^n$ be a closed convex set and fix $\bar{x} \in \Omega$. Consider $\mathcal{X}$ to be the normal cone of the convex set Ω at some fixed point $\bar{x} \in \Omega$:

$$\mathcal{X} = \{x : (x - \bar{x})^T (S - \bar{x}) \leq 0 \text{ for all } S \in \Omega\} = \bigcap_{S \in \Omega} \mathcal{X}_S,$$

where $\mathcal{X}_S \stackrel{\text{def}}{=} \{x : (x - \bar{x})^T (S - \bar{x}) \leq 0\}$. If $\mathbf{P}$ is a probability distribution over Ω, and $S \sim \mathbf{P}$, then $\mathcal{X}_S$ is a stochastic approximation of $\mathcal{X}$. Moreover, in this case we have $\mathcal{X} = \bigcap_{S \in \Omega} \mathcal{X}_S$.

6. Stochastic Projection Algorithm In this section we propose the following parallel stochastic projection method:

Algorithm SPA (general case)
Choose $x^0 \in \mathbb{R}^n$, minibatch size $N \geq 1$, and positive stepsizes $\{\alpha_k\}_{k \geq 0}$. For $k \geq 0$ repeat:
1. Draw N independent samples, $S_1^k, S_2^k, \cdots, S_N^k \sim \mathbf{P}$
2. Compute $x^{k+1} = x^k - \alpha_k \left(x^k - \frac{1}{N} \sum_{i=1}^{N} \Pi_{\mathcal{X}_{S_i^k}}(x^k) \right)$

This algorithm can be viewed as a random implementation of the extrapolated method of parallel projections from [15], which generates a sequence by extrapolation of convex combinations of projections onto subfamilies of sets cyclically.

6.1. Interpretation The minibatch algorithm **SPA** performs at each iteration k a number of N projections onto the simple sets $\mathcal{X}_{S_1^k}, \cdots, \mathcal{X}_{S_N^k}$ in parallel and then the new iterate is computed taking a linear combination between the previous iterate and the average of those projections. Such minibatch strategy has several interpretations. For example, when we consider the stochastic smooth optimization problem (6):

$$\min_{x\in\mathbb{R}^n} F(x) = \mathbf{E}\left[F_S(x)\right],$$

where $F_S(x) = 1/2\|x - \Pi_{\mathcal{X}_S}(x)\|^2$, usually a Monte Carlo simulation-based approach is used for solving it. It consists in generating random samples of S and the expected value function F is approximated by the corresponding sample average function. That is, let $S_1, \cdots, S_N$ be independently and identically distributed random sample of N realizations of the random variable S. Then, we consider the sample average function $\hat{F}_N = 1/N \sum_{i=1}^N F_{S_i}$ and the associated problem:

$$\min_{x\in\mathbb{R}^n} \hat{F}_N(x).$$

Finally, this sample average optimization problem is solved. The idea of using sample average approximations for solving stochastic programs is a natural one and was used by various authors over the years [32]. However, the solution $\hat{x}_N^*$ of the sample average optimization problem converges to the true solution x^* of the stochastic optimization problem only for large enough number of samples $N \to \infty$. On the other hand, in our minibatch algorithm **SPA** the approach is different. First, we fix the number of samples N. Then, at each iteration k we draw N independent samples $S_1^k, S_2^k, \cdots, S_N^k$ to also form a sample average function $\hat{F}_N^k = 1/N \sum_{i=1}^N F_{S_i^k}$. Finally, we do not solve the sample overage optimization problem:

$$\min_{x\in\mathbb{R}^n} \hat{F}_N^k(x),$$

instead we only perform one gradient step for this problem with stepsize α_k

$$x^{k+1} = x^k - \alpha_k \nabla \hat{F}_N^k(x^k),$$

and then repeat the procedure. In this case we are not forced to take N large in order to obtain an approximative solution of the original problem. In fact, we can even consider $N = 1$.
The minibatch algorithm **SPA** can be also interpreted in terms of the stochastic non-smooth optimization problem (5):

$$\min_{x\in\mathbb{R}^n} f(x) = \mathbf{E}\left[f_S(x)\right],$$

where $f_S(x) = \mathbb{I}_{\mathcal{X}_S}(x)$. If we fix the number of samples N, then at each iteration k we draw N independent samples $S_1^k, S_2^k, \cdots, S_N^k$ to form the same sample average function:

$$\hat{F}_N^k(x) = \frac{1}{N}\sum_{i=1}^N \left(\min_{z\in\mathbb{R}^n} f_{S_i^k}(z) + \frac{1}{2}\|x - z\|^2\right),$$

and then consider solving the sample overage optimization problem

$$\min_{x\in\mathbb{R}^n} \hat{F}_N^k(x),$$

which can be rewritten using the notation $z = [z_1 \ \cdots \ z_N]^T$ as follows:

$$\min_{x\in\mathbb{R}^n, z_i\in\mathbb{R}^n} \hat{F}_N^k(x,z) \quad \left(:= \frac{1}{N}\sum_{i=1}^N \left[\mathbb{I}_{\mathcal{X}_{S_i^k}}(z_i) + \frac{1}{2}\|x - z_i\|^2\right]\right).$$

However, we do not solve the previous average optimization problem in the variables (x,z), instead we only perform one step of Relaxed Block Alternating Minimization Method. That is, given x^k, we compute:

$$z^{k+1} = \arg\min_{z\in\mathbb{R}^{Nn}} \hat{F}_N^k(x^k,z), \quad \tilde{x}^{k+1} = \arg\min_{x\in\mathbb{R}^n} \hat{F}_N^k(x,z^{k+1})$$
$$x^{k+1} = (1-\alpha_k)x^k + \alpha_k\tilde{x}^{k+1},$$

and repeat the whole procedure. Again, this strategy allows us to work also with N small, including $N=1$.

6.2. Convergence analysis Our convergence analysis is based on two important properties of the family of convex sets $(\mathcal{X}_S)_{S\in\Omega}$. For simplicity, we recall them once more here. First, there exists $\gamma \leq 1$ satisfying the inequality (13), i.e.:

$$\|\mathbf{E}\left[x-\Pi_{\mathcal{X}_S}(x)\right]\|^2 \leq \gamma\cdot\mathbf{E}\left[\|x-\Pi_{\mathcal{X}_S}(x)\|^2\right] \qquad \forall x\in\mathbb{R}^n. \tag{37}$$

However, for specific sets and distributions $\mathbf{P}$, we proved in Section 3.1 that γ can be much smaller than 1. Second, there exists $\kappa\leq\infty$ such that the family of convex sets $(\mathcal{X}_S)_{S\in\Omega}$ satisfies the linear regularity property (10), i.e.:

$$\text{dist}^2_{\mathcal{X}}(x) \leq \kappa\,\mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right] \qquad \forall x\in\mathbb{R}^n. \tag{38}$$

However, we have proved in Section 4 that for specific sets and distributions $\mathbf{P}$, the constant κ can be finite, that is $\kappa<\infty$. Based on the properties (37) and (38) the smooth objective function F of the stochastic optimization problem (6) satisfies Theorem 3, in particular we have:

$$\frac{1}{2\kappa}\|x-\Pi_{\mathcal{X}}(x)\|^2 \leq F(x)-F^* \leq \frac{\gamma}{2}\|x-\Pi_{\mathcal{X}}(x)\|^2 \qquad \forall x\in\mathbb{R}^n. \tag{39}$$

There is an interesting interpretation of inequality (39), that is the objective function F is strongly convex with constant $\frac{1}{\kappa}$ and has Lipschitz continuous gradient with constant $\gamma\leq 1$ when restricted along any segment $[x,\Pi_{\mathcal{X}}(x)]$. Thus, $\kappa\gamma$ represents the condition number of the convex feasibility problem (3). Using the inequalities (37)-(39) we can prove not only asymptotic convergence of the sequence $\{x^k\}_{k\geq 0}$ generated by algorithm **SPA**, but also rates of convergence. We start with a basic result from probability theory, see e.g. [28]:

LEMMA 3 **(Supermartingale Convergence Lemma)**. *Let v^k and u^k be sequences of non-negative random variables such that:*

$$\mathbf{E}\left[v^{k+1}|F_k\right] \leq v^k - u^k \quad a.s. \quad \forall k\geq 0,$$

where F_k denotes the collection $\{v^0,\cdots,v^k,u^0,\cdots,u^k\}$. Then, we have v^k convergent to a random variable v a.s. and $\sum_{k=0}^{\infty}u^k<\infty$ a.s.

Then, we obtain the following asymptotic convergence result:

THEOREM 8. *Assume that the set $\mathcal{X}$ is nonempty and define $\gamma_N \overset{def}{=} \frac{1}{N}+\left(1-\frac{1}{N}\right)\gamma\leq 1$. Let $\{x^k\}_{k\geq 0}$ be generated by algorithm **SPA** with stepsizes $0<\alpha_k<\frac{2}{\gamma_N}$. Then, we have the following average decrease:*

$$\mathbf{E}\left[\|x^{k+1}-x^*\|^2 \mid x^k\right] \leq \|x^k-x^*\|^2 - 2(2\alpha_k-\alpha_k^2\gamma_N)F(x^k) \tag{40}$$

for all $k\geq 0$ and $x^*\in\mathcal{X}$. Moreover, the fastest decrease is given by the constant stepsize $\alpha_k=1/\gamma_N$. If additionally, exactness holds and the stepsize satisfies $\delta\leq\alpha_k\leq\frac{2}{\gamma_N}-\delta$ for some $0<\delta\leq\frac{1}{\gamma_N}$, then the sequence x^k converges almost sure to a random point in the set $\mathcal{X}$ and $\lim_{k\to\infty}F(x^k)=0$ almost sure.

Proof: For simplicity, we shall write $\Pi_i^k = \Pi_{\mathcal{X}_{S_i^k}}(x^k)$. Let x^* be any element of $\mathcal{X}$. Then, we have the following:

$$\begin{aligned}
\|x^{k+1}-x^*\|^2 &= \left\| x^k - x^* - \alpha_k \left(x^k - \frac{1}{N}\sum_{i=1}^N \Pi_i^k \right) \right\|^2 \\
&= \left\| x^k - x^* - \alpha_k \frac{1}{N}\sum_{i=1}^N (x^k - \Pi_i^k) \right\|^2 \\
&= \|x^k - x^*\|^2 - \frac{2\alpha_k}{N}\sum_{i=1}^N \langle x^k - x^*, x^k - \Pi_i^k\rangle + \frac{\alpha_k^2}{N^2}\left\| \sum_{i=1}^N (x^k - \Pi_i^k) \right\|^2 \\
&\leq \|x^k - x^*\|^2 - \frac{2\alpha_k}{N}\sum_{i=1}^N \left\|x^k - \Pi_i^k\right\|^2 + \frac{\alpha_k^2}{N^2}\left\| \sum_{i=1}^N (x^k - \Pi_i^k) \right\|^2 \\
&= \|x^k - x^*\|^2 - \frac{2\alpha_k}{N}\sum_{i=1}^N \left\|x^k - \Pi_i^k\right\|^2 \\
&\qquad + \frac{\alpha_k^2}{N^2}\left(\sum_{i=1}^N \|x^k - \Pi_i^k\|^2 + \sum_{i\neq j} \langle x^k - \Pi_i^k, x^k - \Pi_j^k\rangle \right),
\end{aligned} \tag{41}$$

where the inequality follows from the bound:

$$\langle x^k - x^*, x^k - \Pi_i^k\rangle = \langle x^k - \Pi_i^k, x^k - \Pi_i^k\rangle + \langle \Pi_i^k - x^*, x^k - \Pi_i^k\rangle \geq \|x^k - \Pi_i^k\|^2,$$

since $\langle \Pi_i^k - x^*, x^k - \Pi_i^k\rangle \geq 0$ for all $x^* \in \mathcal{X} \subseteq \mathcal{X}_{S_i^k}$. Taking expectations conditioned on x^k and using the definition of F:

$$\begin{aligned}
F(x^k) &= \frac{1}{2}\mathbf{E}\left[\|x^k - \Pi_i^k\|^2 \mid x^k\right] = \frac{1}{2}\mathbf{E}\left[\|x^k - \Pi_{\mathcal{X}_{S_i^k}}(x^k)\|^2 \mid x^k\right] \\
&= \frac{1}{2}\mathbf{E}\left[\|x^k - \Pi_{\mathcal{X}_S}(x^k)\|^2 \mid x^k\right],
\end{aligned}$$

and invoking conditional independence of Π_i^k and Π_j^k for $i \neq j$ (inherited from independence of S_i^k and S_j^k), we obtain:

$$\begin{aligned}
&\mathbf{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] \leq \|x^k - x^*\|^2 - 4\alpha_k F(x^k) \\
&\qquad + \frac{\alpha_k^2}{N^2}\left(2NF(x^k) + \sum_{i\neq j} \left\langle \mathbf{E}\left[x^k - \Pi_i^k \mid x^k\right], \mathbf{E}\left[x^k - \Pi_j^k \mid x^k\right] \right\rangle \right) \\
&= \|x^k - x^*\|^2 - 4\alpha_k F(x^k) + \frac{2\alpha_k^2}{N}F(x^k) + \frac{\alpha_k^2(N^2-N)}{N^2}\left\|\mathbf{E}\left[x^k - \Pi_{\mathcal{X}_S}(x^k) \mid x^k\right]\right\|^2 \\
&\overset{(37)}{\leq} \|x^k - x^*\|^2 - 4\alpha_k F(x^k) + \frac{2\alpha_k^2}{N}F(x^k) + \frac{\alpha_k^2(N-1)}{N}\gamma\mathbf{E}\left[\|x^k - \Pi_{\mathcal{X}_S}(x^k)\|^2 \mid x^k\right] \\
&= \|x^k - x^*\|^2 - 4\alpha_k F(x^k) + \frac{2\alpha_k^2}{N}F(x^k) + \frac{2\alpha_k^2(N-1)}{N}\gamma F(x^k) \\
&= \|x^k - x^*\|^2 - 2(2\alpha_k - \alpha_k^2\gamma_N)F(x^k).
\end{aligned} \tag{42}$$

Thus, we have obtained for all $k \geq 0$ and $x^* \in \mathcal{X}$:

$$\mathbf{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] \leq \|x^k - x^*\|^2 - 2(2\alpha_k - \alpha_k^2\gamma_N)F(x^k).$$

Clearly, the fastest decrease is obtained by maximizing $2\alpha_k - \alpha_k^2\gamma_N$ in α_k, that is the maximum is obtained for constant stepsize $\alpha_k = 1/\gamma_N$. Further, for the stepsizes satisfying $\delta \leq \alpha_k \leq \frac{2}{\gamma_N} - \delta$

we have $2\alpha_k - \alpha_k^2\gamma_N \geq \delta^2\gamma_N > 0$. Then, from Supermartingale Convergence Lemma we have that $\|x^k - x^*\|^2$ converges a.s. for every $x^* \in \mathcal{X}$ and thus the sequence x^k is bounded a.s. This implies that x^k has a limit point $\tilde{x}^*$. Since we also have $\sum_{k=0}^{\infty} F(x^k) < \infty$ a.s., it follows that $F(x^k) \to 0$ a.s. Therefore, for any accumulation point $\tilde{x}^*$ of x^k we have $F(\tilde{x}^*) = 0$ a.s. (by continuity of F). This leads to $\tilde{x}^* \in \mathcal{Y}$ a.s. When exactness holds (i.e. $\mathcal{X} = \mathcal{Y}$), it follows that at least a subsequence of x^k converges almost surely to a random point $\tilde{x}^*$ from the set $\mathcal{X}$. Q.E.D.
The previous theorem clearly shows that in order to have decrease in average distances (see (44)) the stepsize α_k has to satisfy:

$$0 < \alpha_k < \frac{2}{\gamma_N} \qquad \forall k \geq 0. \tag{43}$$

This shows that we can use large stepsizes α_k. Thus, we prove theoretically, what is known in numerical applications for a long time, namely that this overrelaxation $\alpha_k \approx \frac{2}{\gamma_N} > 1$ accelerates significantly in practice the convergence of projection methods as compared to its basic counterpart $\alpha_k = 1$, see [12, 15]. For several important sets we can estimate the "Lipschitz" constant γ and consequently γ_N, see Section 3.1. For other sets however, it is difficult to compute γ. In this case we propose an adaptive estimation of γ at each iteration k as follows:

$$\gamma^k = \frac{\|\mathbf{E}\left[x^k - \Pi_{\mathcal{X}_S}(x^k)\right]\|^2}{\mathbf{E}\left[\|x^k - \Pi_{\mathcal{X}_S}(x^k)\|^2\right]}.$$

This choice has the following interpretation. From Theorem 3 we have that F has Lipschitz continuous gradient with constant γ on any segment $[x^k,\ \Pi_{\mathcal{X}}(x^k)]$:

$$F(x^k) \overset{(17)}{\geq} F^* + \langle \nabla F(\Pi_{\mathcal{X}}(x)), x^k - \Pi_{\mathcal{X}}(x)\rangle + \frac{1}{2\gamma}\|\nabla F(x^k) - \nabla F(\Pi_{\mathcal{X}}(x))\|^2,$$

which, using $F^* = F(\Pi_{\mathcal{X}}(x^k)) = 0$ and $\nabla F(\Pi_{\mathcal{X}}(x^k)) = 0$, is equivalent to:

$$\gamma \geq \frac{1/2\|\nabla F(x^k)\|^2}{F(x^k)} = \gamma^k.$$

Using arguments of Theorem 8, it is straightforward to obtain the following descent.

COROLLARY 5. *Assume that the set $\mathcal{X}$ is nonempty and define $\gamma_N^k \overset{def}{=} \frac{1}{N} + \left(1 - \frac{1}{N}\right)\gamma^k \leq 1$. Let $\{x^k\}_{k\geq 0}$ be generated by algorithm **SPA** with stepsizes $0 < \alpha_k < \frac{2}{\gamma_N^k}$. Then, we have the following average decrease:*

$$\mathbf{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] \leq \|x^k - x^*\|^2 - 2(2\alpha_k - \alpha_k^2\gamma_N^k)F(x^k) \tag{44}$$

for all $k \geq 0$ and $x^* \in \mathcal{X}$.

Proof: From the relation (42) we have:

$$\begin{aligned}
\mathbf{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] &\leq \|x^k - x^*\|^2 - 4\alpha_k F(x^k) + \frac{2\alpha_k^2}{N}F(x^k) + \frac{\alpha_k^2(N^2-N)}{N^2}\left\|\mathbf{E}\left[x^k - \Pi_{\mathcal{X}_S}(x^k) \mid x^k\right]\right\|^2 \\
&= \|x^k - x^*\|^2 - 4\alpha_k F(x^k) + \frac{2\alpha_k^2}{N}F(x^k) + \frac{\alpha_k^2(N-1)}{N}\gamma^k\mathbf{E}\left[\|x^k - \Pi_{\mathcal{X}_S}(x^k)\|^2 \mid x^k\right] \\
&= \|x^k - x^*\|^2 - 4\alpha_k F(x^k) + \frac{2\alpha_k^2}{N}F(x^k) + \frac{2\alpha_k^2(N-1)}{N}\gamma^k F(x^k) \\
&= \|x^k - x^*\|^2 - 2(2\alpha_k - \alpha_k^2\gamma_N^k)F(x^k),
\end{aligned}$$

which confirms the results. Q.E.D.

When even the expectation is difficult to compute for finding γ^k, then, inspired by [15], we propose to use the following approximation for the previous ratio:

$$\gamma^k = \frac{\| \sum_{i=1}^N w_i^k (x^k - \Pi_{\mathcal{X}_{S_i^k}}(x^k))\|^2}{\sum_{i=1}^N w_i^k \|x^k - \Pi_{\mathcal{X}_{S_i^k}}(x^k)\|^2},$$

where the weights w_i^k satisfy $\sum_{i=1}^N w_i^k = 1$ and $w_i^k > 0$. For these situations we take the stepsize:

$$\alpha_k = \frac{\alpha}{\gamma_N^k}, \quad \text{with } \alpha \in (0,\ 2).$$

The effectiveness of this choice for the stepsize has been shown in many practical applications, see e.g. [12, 15]. Next theorem provides rates of convergence for the sequence x^k generated by **SPA**:

THEOREM 9. *Assume that the set $\mathcal{X}$ is nonempty and define $\gamma_N = \frac{1}{N} + \left(1 - \frac{1}{N}\right)\gamma$. Let $\{x^k\}_{k\geq 0}$ be generated by algorithm **SPA** with stepsizes satisfying $\delta \leq \alpha_k \leq \frac{2}{\gamma_N} - \delta$ for some $0 < \delta \leq \frac{1}{\gamma_N}$. Then:*

(i) For the average point $\hat{x}^k = \frac{1}{\Sigma_k} \sum_{i=0}^{k-1} \alpha_i x^i$, where $\Sigma_k = \sum_{i=0}^{k-1} \alpha_i$, we have the following sublinear convergence rate:

$$\mathbf{E}\left[F(\hat{x}^k)\right] - F^* = \frac{1}{2}\mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(\hat{x}^k)\right] \leq \frac{\text{dist}^2_{\mathcal{X}}(x^0)}{2\delta\gamma_N\Sigma_k}.$$

Moreover, the average sequence $\hat{x}^k$ converges almost surely to a random point in the set $\mathcal{X}$, provided that exactness holds.

(ii) If additionally the linear regularity property (38) holds, then we have the following linear convergence rate for the last iterate x^k:

$$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^{k+1})\right] \leq \left(1 - \frac{\delta^2\gamma_N}{\kappa}\right)\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right],$$

or in terms of function values:

$$\mathbf{E}\left[F(x^k)\right] - F^* \leq \left(1 - \frac{\delta^2\gamma_N}{\kappa}\right)^k \frac{\gamma \text{dist}^2_{\mathcal{X}}(x^0)}{2}.$$

Proof: By taking expectation w.r.t. the entire history on both sides in (44) we get the following decrease in the distance to a point $x^* \in \mathcal{X}$:

$$\mathbf{E}\left[\|x^{k+1} - x^*\|^2\right] \leq \mathbf{E}\left[\|x^k - x^*\|^2\right] - 2(2\alpha_k - \alpha_k^2\gamma_N)\mathbf{E}\left[F(x^k)\right].$$

Further, denoting $r_k \overset{\text{def}}{=} \mathbf{E}\left[\|x^k - x^*\|^2\right]$ and noticing the lower bound $2 - \alpha_k\gamma_N \geq \delta\gamma_N$ for any stepsize satisfying $\delta \leq \alpha_k \leq \frac{2}{\gamma_N} - \delta$ for some $0 < \delta \leq \frac{1}{\gamma_N}$, we have:

$$2\delta\gamma_N\alpha_k\mathbf{E}\left[F(x^k)\right] \leq 2\alpha_k(2 - \alpha_k\gamma_N)\mathbf{E}\left[F(x^k)\right] \leq r_k - r_{k+1}.$$

If we add the entire history from $i = 0$ to $i = k-1$, we obtain:

$$2\delta\gamma_N\mathbf{E}\left[\sum_{i=0}^{k-1}\alpha_i F(x^i)\right] = \sum_{i=0}^{k-1} 2\delta\gamma_N\alpha_i\mathbf{E}\left[F(x^i)\right] \leq r_0 - r_k \leq r_0 = \|x^0 - x^*\|^2$$

for all $x^* \in \mathcal{X}$. If we choose $x^* = \Pi_{\mathcal{X}}(x^0)$ and use the convexity of function F, then we finally get:

$$2\Sigma_k\delta\gamma_N\mathbf{E}\left[F\left(\frac{1}{\Sigma_k}\sum_{i=0}^{k-1}\alpha_i x^i\right)\right] \leq 2\delta\gamma_N\mathbf{E}\left[\sum_{i=0}^{k-1}\alpha_i F(x^i)\right] \leq \text{dist}^2_{\mathcal{X}}(x^0).$$

This relation and $F^* = 0$ imply immediately the first part of our result. Moreover, by Theorem 8, x^k converges almost surely to a random point in the set $\mathcal{X}$. Therefore, the average sequence $\hat{x}^k = \frac{1}{k}\sum_{i=0}^{k-1} x^i$ also converges almost surely to the same random point in the set $\mathcal{X}$.
(ii) In order to prove linear convergence under linear regularity property (38) we use again inequality (44) and $F^* = 0$:

$$\begin{aligned}\mathbf{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] &\leq \|x^k - x^*\|^2 - 2(2\alpha_k - \alpha_k^2\gamma_N)\left(F(x^k) - F^*\right)\\ &\overset{(39)}{\leq} \|x^k - x^*\|^2 - \tfrac{2\alpha_k - \alpha_k^2\gamma_N}{\kappa}\mathrm{dist}^2_{\mathcal{X}}(x^k).\end{aligned}$$

Taking expectations w.r.t. the entire history, we obtain:

$$\mathbf{E}\left[\|x^{k+1} - x^*\|^2\right] \leq \mathbf{E}\left[\|x^k - x^*\|^2\right] - \frac{2\alpha_k - \alpha_k^2\gamma_N}{\kappa}\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^k)\right]. \tag{45}$$

Choosing $x^* = \Pi_{\mathcal{X}}(x^k)$, and using the inequality $\mathrm{dist}^2_{\mathcal{X}}(x^{k+1}) = \|x^{k+1} - \Pi_{\mathcal{X}}(x^{k+1})\|^2 \leq \|x^{k+1} - x^*\|^2$ together with (45), we finally get:

$$\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^{k+1})\right] \leq \left(1 - \frac{2\alpha_k - \alpha_k^2\gamma_N}{\kappa}\right)\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^k)\right].$$

Since for our choice of the stepsize $\delta \leq \alpha_k \leq \frac{2}{\gamma_N} - \delta$ for some $0 < \delta \leq \frac{1}{\gamma_N}$, we have $2\alpha_k - \alpha_k^2\gamma_N \geq \delta^2\gamma_N$, then the previous relation implies immediately:

$$\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^{k+1})\right] \leq \left(1 - \frac{\delta^2\gamma_N}{\kappa}\right)\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^k)\right].$$

which proves the second statement of the theorem. Finally, combining the convergence rate in distances with the right hand side inequality in (39) we get the convergence in expectation of value function. Q.E.D.
An immediate consequence of Theorem 9 is the following corollary:

COROLLARY 6. *Assume that the set $\mathcal{X}$ is nonempty and $\gamma_N = \frac{1}{N} + \left(1 - \frac{1}{N}\right)\gamma$. Let $\{x^k\}_{k\geq 0}$ be generated by algorithm **SPA** with the optimal constant stepsize $\alpha_k = 1/\gamma_N$. Then:*

(i) For the average point $\hat{x}^k = \frac{1}{k}\sum_{i=0}^{k-1} x^i$ we have the following sublinear convergence rate:

$$\mathbf{E}\left[F(\hat{x}^k)\right] - F^* = \frac{1}{2}\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}_S}(\hat{x}^k)\right] \leq \frac{\gamma_N \cdot \mathrm{dist}^2_{\mathcal{X}}(x^0)}{2k}.$$

(ii) If additionally the linear regularity property (38) holds, then we have the following linear convergence rate for the last iterate x^k:

$$\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^{k+1})\right] \leq \left(1 - \frac{1}{\gamma_N \cdot \kappa}\right)\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^k)\right], \tag{46}$$

or in terms of function values:

$$\mathbf{E}\left[F(x^k)\right] - F^* \leq \left(1 - \frac{1}{\gamma_N \cdot \kappa}\right)^k \frac{\gamma \mathrm{dist}^2_{\mathcal{X}}(x^0)}{2}.$$

Proof: By taking expectation w.r.t. the entire history on both sides in (44) we get the following decrease in the distance to a point $x^* \in \mathcal{X}$:

$$\mathbf{E}\left[\|x^{k+1}-x^*\|^2\right] \leq \mathbf{E}\left[\|x^k-x^*\|^2\right] - 2(2\alpha_k - \alpha_k^2\gamma_N)\mathbf{E}\left[F(x^k)\right].$$

Further, denoting $r_k \overset{\text{def}}{=} \mathbf{E}\left[\|x^k-x^*\|^2\right]$, we have:

$$2(2\alpha_k - \alpha_k^2\gamma_N)\mathbf{E}\left[F(x^k)\right] \leq r_k - r_{k+1}.$$

The fastest decrease is obtained maximizing $2\alpha_k - \alpha_k^2\gamma_N$ in α_k, which leads to the optimal stepsize $\alpha_k = 1/\gamma_N$. The rest of the proof follows exactly the same steps as in the proof of Theorem 9, observing that choosing $\delta = 1/\gamma_N$ we get $\alpha_k = 1/\gamma_N$. Q.E.D.

From Theorem 9 and Corollary 6 it follows that the convergence rates of algorithm **SPA** depend explicitly on the minibatch sample size N via the term γ_N. Moreover, we notice that the scheme **SPA** is very general and we can recover multiple existing projection algorithms from the literature. Further, we analyze some particular algorithms resulted from **SPA** and derive their convergence rates.

6.3. Average Projection algorithm: $N = m/\infty$ As $N \to \infty$, we have $\gamma_N \to \gamma$, and the linear rate in Corollary 6 converges to $1 - 1/(\kappa \cdot \gamma)$. This is also confirmed by the convergence rate given in Theorem 10 below. More precisely, when $N \to \infty$ the algorithm **SPA** becomes the deterministic gradient method for solving the smooth convex problem (6), which we call average projection algorithm:

Algorithm AvP

Choose $x^0 \in \mathbb{R}^n$ and positive stepsizes $\{\alpha_k\}_{k\geq 0}$. For $k \geq 0$ repeat:

1. Compute $x^{k+1} = x^k - \alpha_k \nabla F(x^k) \quad \left(\overset{\text{def}}{=} x^k - \alpha_k\left(x^k - \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x^k)\right]\right)\right)$

Under linear regularity condition (38) the sequence $\{x^k\}_{k\geq 0}$ generated by algorithm **AvP** is converging linearly:

THEOREM 10. *If the linear regularity property (38) holds, then we have the following linear convergence rate for the last iterate x^k generated by algorithm **AvP** with the optimal stepsize $\alpha_k = 1/\gamma$:*

$$\text{dist}^2_{\mathcal{X}}(x^{k+1}) \leq \left(1 - \frac{1}{\gamma \cdot \kappa}\right)\text{dist}^2_{\mathcal{X}}(x^k), \tag{47}$$

or in terms of function values:

$$F(x^k) - F^* \leq \left(1 - \frac{1}{\gamma \cdot \kappa}\right)^k \frac{\gamma \text{dist}^2_{\mathcal{X}}(x^0)}{2}.$$

Proof: Let x^* be any element of $\mathcal{X}$. Then, we have the following:

$$\begin{aligned}
\|x^{k+1}-x^*\|^2 &= \left\|x^k - x^* - \alpha_k\left(x^k - \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x^k)\right]\right)\right\|^2 \\
&= \|x^k-x^*\|^2 - 2\alpha_k\left\langle x^k - x^*, x^k - \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x^k)\right]\right\rangle + \alpha_k^2\left\|x^k - \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x^k)\right]\right\|^2 \\
&\overset{(37)}{\leq} \|x^k-x^*\|^2 - 2\alpha_k\left\langle x^k - x^*, x^k - \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x^k)\right]\right\rangle + \gamma\alpha_k^2\mathbf{E}\left[\left\|x^k - \Pi_{\mathcal{X}_S}(x^k)\right\|^2\right] \\
&= \|x^k-x^*\|^2 - 2\alpha_k\mathbf{E}\left[\left\langle x^k - \Pi_{\mathcal{X}_S}(x^k) + \Pi_{\mathcal{X}_S}(x^k) - x^*, x^k - \Pi_{\mathcal{X}_S}(x^k)\right\rangle\right] \\
&\qquad + \gamma\alpha_k^2\mathbf{E}\left[\left\|x^k - \Pi_{\mathcal{X}_S}(x^k)\right\|^2\right] \\
&= \|x^k-x^*\|^2 - 2\alpha_k\mathbf{E}\left[\left\|x^k - \Pi_{\mathcal{X}_S}(x^k)\right\|^2\right] + \gamma\alpha_k^2\mathbf{E}\left[\left\|x^k - \Pi_{\mathcal{X}_S}(x^k)\right\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&\quad - 2\alpha_k \mathbf{E}\left[\left\langle \Pi_{\mathcal{X}_S}(x^k) - x^*, x^k - \Pi_{\mathcal{X}_S}(x^k)\right\rangle\right] \\
&\leq \|x^k - x^*\|^2 - (2\alpha_k - \gamma\alpha_k^2)\mathbf{E}\left[\left\|x^k - \Pi_{\mathcal{X}_S}(x^k)\right\|^2\right] \\
&= \|x^k - x^*\|^2 - 2(2\alpha_k - \gamma\alpha_k^2)F(x^k). \qquad (48)
\end{aligned}
$$

where the second inequality follows from the optimality condition of the projection $\langle \Pi_{\mathcal{X}_S}(x^k) - x^*, x^k - \Pi_{\mathcal{X}_S}(x^k)\rangle \geq 0$ for all $x^* \in \mathcal{X} \subseteq \mathcal{X}_S$. From (48) we observe that the fastest decrease is obtained from maximizing $2\alpha_k - \gamma\alpha_k^2$, which leads to the optimal stepsize $\alpha_k = 1/\gamma$. For this choice of the stepsize, $\alpha_k = 1/\gamma$, and using $F^* = 0$ we obtain from (48):

$$
\begin{aligned}
\|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - \frac{2}{\gamma}(F(x^k) - F^*) \\
&\overset{(39)}{\leq} \|x^k - x^*\|^2 - \frac{1}{\gamma\kappa}\|x^k - x^*\|^2,
\end{aligned}
$$

which implies immediately the statement of the theorem. Q.E.D.

Note that from the proof of Theorem 10 it follows that we can achieve linear convergence for the last iterate generated by algorithm **AvP** with stepsizes satisfying $0 < \alpha_k < 2/\gamma$ Moreover, $\gamma\kappa$ represents the condition number of the convex feasibility problem (3) or of its stochastic reformulation (6) (see Theorem 3).

Let us consider finding a point in the finite intersection of convex sets $(\mathcal{X}_i)_{i\in[m]}$, that is $\mathcal{X} = \cap_{i=1}^m \mathcal{X}_i$. Further, we consider a uniform probability on $\Omega = [m]$ and we choose the minibach sample size $N = m$, then the average projection algorithm **AvP** becomes the barycentric method:

$$
\mathbf{AvP}(1/m): \quad x^{k+1} = x^k - \alpha_k \left(x^k - \frac{1}{m}\sum_{i=1}^m \Pi_{\mathcal{X}_i}(x^k) \right).
$$

The barycentric method was shown to converge asymptotically to a point in the intersection of the closed convex sets $(\mathcal{X}_i)_{i\in[m]}$, see e.g. [15]. Recall that we denoted $D = \mathrm{diag}(\|A_1\|^{-2}, \cdots, \|A_m\|^{-2})$. Let us derive convergence rates for the barycentric method $\mathbf{AvP}(1/m)$ for two particular cases of sets:

(i): Consider the problem of finding a solution to a linear system $Ax = b$, where A is an $m \times n$ matrix. In this case $\mathcal{X}_i = \{x : \ A_i^T x = b_i\}$. Then, from Theorem 10 the barycentric method $\mathbf{AvP}(1/m)$ with the optimal stepsize $\alpha_k = 1/\gamma$ converges linearly:

$$
\begin{aligned}
\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^k)\right] &\overset{(47)}{\leq} \left(1 - \frac{1}{\gamma\kappa}\right)^k \mathrm{dist}^2_{\mathcal{X}}(x^0) \\
&\overset{(14)+(28)}{=} \left(1 - \frac{\lambda^{\mathrm{nz}}_{\min}(A^T D A)}{\lambda_{\max}(A^T D A)}\right)^k \mathrm{dist}^2_{\mathcal{X}}(x^0).
\end{aligned}
$$

(ii): Consider now the more general problem of finding a solution to a system of linear inequalities $Ax \leq b$, where A is an $m \times n$ matrix. Then $\mathcal{X}_i = \{x : \ A_i^T x \leq b_i\}$. From Theorem 10 it follows that the barycentric method $\mathbf{AvP}(1/m)$ with the optimal stepsize $\alpha_k = 1/\gamma$ converges also linearly:

$$
\begin{aligned}
\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^k)\right] &\overset{(47)}{\leq} \left(1 - \frac{1}{\gamma\kappa}\right)^k \mathrm{dist}^2_{\mathcal{X}}(x^0) \\
&\overset{(15)+(29)}{=} \left(1 - \frac{1}{\max\limits_{i=1:m} \|A_i\|^2 \lambda_{\max}(A^T D A)\tilde{\kappa}}\right)^k \mathrm{dist}^2_{\mathcal{X}}(x^0).
\end{aligned}
$$

Note that from Theorem 10 it follows immediately that the basic barycentric method $x^{k+1} = \frac{1}{m}\sum_{i=1}^m \Pi_{\mathcal{X}_i}(x^k)$, i.e. stepsize $\alpha_k = 1$, converges linearly:

$$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(47)}{\leq} \left(1 - \frac{2-\gamma}{\kappa}\right)^k \text{dist}^2_{\mathcal{X}}(x^0).$$

However, our algorithmic framework leads to new schemes. For example, for a general probability distribution $(p_i)_{i\in[m]}$ on $\Omega = [m]$ and $N = m$, the average projection algorithm **AvP** has the iteration:

$$\mathbf{AvP}(p_i): \quad x^{k+1} = x^k - \alpha_k \left(x^k - \sum_{i=1}^m p_i \Pi_{\mathcal{X}_i}(x^k)\right).$$

If we choose the probabilities $p_i = \frac{\|A_i\|^2}{\|A\|_F^2}$, then this method has the following convergence rates for linear systems and linear inequalities:
(iii): For a linear system $Ax = b$, from Theorem 10 the $\mathbf{AvP}(\|A_i\|^2/\|A\|_F^2)$ method with the optimal stepsize $\alpha_k = 1/\gamma$ converges linearly:

$$\begin{aligned}\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] &\overset{(47)}{\leq} \left(1 - \frac{1}{\gamma\kappa}\right)^k \text{dist}^2_{\mathcal{X}}(x^0) \\ &\overset{(14)+(28)}{=} \left(1 - \frac{\lambda^{\text{nz}}_{\min}(A^TA)}{\lambda_{\max}(A^TA)}\right)^k \text{dist}^2_{\mathcal{X}}(x^0).\end{aligned}$$

(iv): For a system of linear inequalities $Ax \leq b$, from Theorem 10 the previous method with the optimal stepsize $\alpha_k = 1/\gamma$ converges also linearly:

$$\begin{aligned}\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] &\overset{(47)}{\leq} \left(1 - \frac{1}{\gamma\kappa}\right)^k \text{dist}^2_{\mathcal{X}}(x^0) \\ &\overset{(15)+(32)}{=} \left(1 - \frac{1}{\lambda_{\max}(A^TA)\tilde{\kappa}}\right)^k \text{dist}^2_{\mathcal{X}}(x^0).\end{aligned}$$

6.4. Stochastic Alternating Projection algorithm: $N = 1$ In this section we analyze in more detail a particular case of scheme **SPA** which uses a single projection for the updates. That is, in **SPA** we choose $N = 1$, which results in the Stochastic Alternating Projection (**SAP**) scheme:

Algorithm SAP
Choose $x^0 \in \mathbb{R}^n$ and positive stepsizes $\{\alpha_k\}_{k\geq 0}$
For $k \geq 0$ repeat:
1. Choose randomly a sample $S_k \sim \mathbf{P}$
2. Compute $x^{k+1} = x^k - \alpha_k \left(x^k - \Pi_{\mathcal{X}_{S_k}}(x^k)\right)$

Algorithm **SAP** can be viewed as a random implementation of the alternating projection method, which generates a sequence of iterates by projecting on the sets cyclically. The alternating projection algorithm has been proposed by Von Neumann [37] for the intersection problem of two subspaces in a Hilbert space, and it has many generalization and extensions [6, 16, 27]. A nice survey of the work in this area is given in [5]. The first convergence rate result for the alternating projection algorithm under the assumption that the intersection set has a nonempty interior has been given in [18]. Unlike the alternate projection method (which is deterministic), the algorithm **SAP** utilize random projections. The convergence rate of **SAP** for a finite intersection of simple

convex sets has been given recently in [27, 28]. From the convergence analysis of previous section it follows that the stepsize in **SAP** can be chosen as:

$$\delta \leq \alpha_k \leq 2-\delta,$$

since for $N=1$ we have $\gamma_N = 1$. Moreover, the optimal stepsize is $\alpha_k = 1$. However, it has been observed in practice that overrelaxations, that is $\alpha_k \in [1,\ 2]$, make **SAP** to perform better. Further note that for specific sets and probabilities we recover well known algorithms from literature:

(i): Consider the problem of finding a solution to a linear system $Ax = b$, where A is an $m \times n$ matrix. Further, assume $\Omega = \{e_1, \cdots, e_m\}$ and the probability distribution $\mathbf{P}(S = e_i) = \frac{\|A_i\|^2}{\|A\|_F^2}$. Then, **SAP** with $\alpha_k = 1$ is the randomized Kaczmarz algorithm from [36]:

$$x^{k+1} = x^k - \frac{A_i^T x^k - b_i}{\|A_i\|^2} A_i.$$

Moreover, for these choices of the probabilities and stepsize, our convergence analysis matches exactly the one in [36], that is **SAP** is converging linearly:

$$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(46)}{\leq} \left(1 - \frac{1}{\kappa}\right)^k \text{dist}^2_{\mathcal{X}}(x^0) \overset{(28)}{=} \left(1 - \frac{\lambda_{\min}^{\text{nz}}(A^TA)}{\|A\|_F^2}\right)^k \text{dist}^2_{\mathcal{X}}(x^0).$$

However, **SAP** generalizes the randomized Kaczmarz algorithm from [36], considering for a random matrix $S_k \in \mathbb{R}^{m\times q}$ the general iteration:

$$x^{k+1} = x^k - \alpha_k A^T S_k (S_k^T A A^T S_k)^\dagger S_k^T (Ax^k - b).$$

Notice that for constant stepsize $\alpha_k = 1$, the previous **SAP** scheme is equivalent with the randomized iterative method of [31]. For this choice of the stepsize, our convergence analysis matches exactly the one in [31]:

$$\begin{aligned}\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] &\overset{(46)}{\leq} \left(1 - \frac{1}{\kappa}\right)^k \text{dist}^2_{\mathcal{X}}(x^0) \\ &\overset{(24)}{=} \left(1 - \lambda_{\min}^{\text{nz}}(A^T \mathbf{E}\left[S(S^TAA^TS)^\dagger S^T\right] A)\right)^k \text{dist}^2_{\mathcal{X}}(x^0).\end{aligned}$$

(ii): Consider now the more general problem of finding a solution to a system of linear inequalities $Ax \leq b$, where A is an $m \times n$ matrix. Further, assume as above $\Omega = \{e_1, \cdots, e_m\}$ and the probability distribution $\mathbf{P}(S = e_i) = \frac{\|A_i\|^2}{\|A\|_F^2}$. Then, **SAP** with $\alpha_k = 1$ is the Algorithm 4.6 from [22]:

$$x^{k+1} = x^k - \frac{\Pi_+(A_i^T x^k - b_i)}{\|A_i\|^2} A_i.$$

For these choices of the probabilities and stepsize, our convergence analysis matches exactly the one in [22]:

$$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(46)}{\leq} \left(1 - \frac{1}{\kappa}\right)^k \text{dist}^2_{\mathcal{X}}(x^0) \overset{(32)}{=} \left(1 - \frac{1}{\tilde{\kappa}\|A\|_F^2}\right)^k \text{dist}^2_{\mathcal{X}}(x^0).$$

However, **SAP** generalizes the Algorithm 4.6 from [22], considering for a random vector $S_k \in \mathbb{R}^m_+$ the general iteration:

$$x^{k+1} = x^k - \alpha_k \frac{\Pi_+(S_k^T Ax - S_k^T b)}{\|A^T S_k\|^2} A^T S_k.$$

Under the settings of Theorem 4 we obtain:

$$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(46)}{\leq} \left(1-\frac{1}{\kappa}\right)^k \text{dist}^2_{\mathcal{X}}(x^0) \overset{(32)}{=} \left(1-\frac{1}{\tilde{\kappa}\sum_{S\in\Omega_r}\|A^TS\|^2}\right)^k \text{dist}^2_{\mathcal{X}}(x^0).$$

(iii): Finally, we can consider the convex feasibility problem where the intersection set has a nonempty interior. First, let us investigate when the sequence $\|x^k - x^*\|$ is decreasing. For $N=1$ and $\alpha_k \in [0,\ 2]$ it follows from (41) that:

$$\|x^{k+1}-x^*\|^2 \leq \|x^k-x^*\|^2 - (2\alpha_k-\alpha_k^2)\|x^k - \Pi_{\mathcal{X}_{S^k}}(x^k)\| \quad \forall k\geq 0,$$

that is the sequence $\|x^k - x^*\|$ is nonincreasing. Similarly, for $N\geq 1$ and $\alpha_k \in [0,\ 1]$ it follows that:

$$\begin{aligned}
\|x^{k+1}-x^*\|^2 &= \left\|(1-\alpha_k)(x^k-x^*) + \alpha_k\left(\frac{1}{N}\sum_{i=1}^N \Pi_{\mathcal{X}_{S_i^k}}(x^k) - x^*\right)\right\|^2 \\
&\leq (1-\alpha_k)\|x^k-x^*\|^2 + \alpha_k\left\|\frac{1}{N}\sum_{i=1}^N \Pi_{\mathcal{X}_{S_i^k}}(x^k) - x^*\right\|^2 \\
&\leq (1-\alpha_k)\|x^k-x^*\|^2 + \frac{\alpha_k}{N}\sum_{i=1}^N \|\Pi_{\mathcal{X}_{S_i^k}}(x^k) - x^*\|^2 \\
&= \|x^k-x^*\|^2 + \frac{\alpha_k}{N}\sum_{i=1}^N \left(\|\Pi_{\mathcal{X}_{S_i^k}}(x^k) - x^*\|^2 - \|x^k-x^*\|^2\right) \\
&\leq \|x^k-x^*\|^2 - \frac{\alpha_k}{N}\sum_{i=1}^N \|x^k - \Pi_{\mathcal{X}_{S_i^k}}(x^k)\|^2 \quad \forall k\geq 0.
\end{aligned}$$

The last inequality follows from the bound $\|x^k - \Pi_{\mathcal{X}_{S_i^k}}(x^k)\|^2 + \|\Pi_{\mathcal{X}_{S_i^k}}(x^k) - x^*\|^2 \leq \|x^k - x^*\|^2$ for all $x^* \in \mathcal{X} \subseteq \mathcal{X}_{S_i^k}$. Therefore, for $N\geq 1$ and $\alpha_k \in [0,\ 1]$ we also have a nonincreasing sequence $\|x^k - x^*\|$. In conclusion, for the two choices for N and α_k given above we have:

$$\|x^k - x^*\| \leq \|x^0 - x^*\| \quad \forall x^* \in \mathcal{X},\ k\geq 0.$$

An important application of the previous inequality is that when the set $\mathcal{X}$ contains a ball with radius δ centered in $\bar{x}$. By taking $x^* = \bar{x}$ in the previous relation, we have: $\|x^k - \bar{x}\| \leq \|x^0 - \bar{x}\|$ for all $k\geq 0$. This implies that under the settings of Theorem 7, one should choose the compact set $Q = \{x:\ \|x-\bar{x}\| \leq \|x^0 - \bar{x}\|\}$, such that the linear regularity constant given in (34) becomes:

$$\kappa = \frac{\|x^0-\bar{x}\|^2}{\delta^2 \min\limits_{S\in\Omega} p_S}, \tag{49}$$

since all the points of interest for which the linear regularity property has to hold are the iterates $\{x^k\}_{k\geq 0}$. Then, **SAP** with $\alpha_k = 1$ is the random projection algorithm from [27]. For this choice of the stepsize and under the setting of Theorem 7, the algorithm **SAP** attains the following linear rate:

$$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(46)}{\leq} \left(1-\frac{1}{\kappa}\right)^k \text{dist}^2_{\mathcal{X}}(x^0) \overset{(34)+(49)}{=} \left(1-\frac{p_{\min}\delta^2}{R^2}\right)^k \text{dist}^2_{\mathcal{X}}(x^0),$$

where $p_{\min} = \min_{S\in\Omega} p_S$ and $R = \|x^0 - \bar{x}\|$. A similar convergence rate has been derived in [27] for this particular scheme.

7. Conclusions We have proposed new stochastic reformulations of the classical convex feasibility problem and analyzed the problem conditioning parameters in relation with (linear) regularity assumptions on the individual convex sets. Then, we have introduced a general random projection algorithmic framework, which extends to the random settings many existing projection schemes, designed for the general convex feasibility problem. Based on the conditioning parameters, besides the asymptotic convergence results, we have also derived explicit sublinear and linear convergence rates for this general algorithm. The convergence rates show specific dependence on the number of projections averaged at each iteration. Our general random projection algorithm also allows to project simultaneously on several sets, thus providing great flexibility in matching the implementation of the algorithms on the parallel architecture at hand.

References

[1] A. Auslender, M. Teboulle, *A Log-Quadratic Projection Method for Convex Feasibility Problems*, Studies in Computational Mathematics, 8: 1–9, 2001.

[2] A. Beck and M. Teboulle, *A Linearly Convergent Algorithm for Solving a Class of Nonconvex/Affine Feasibility Problems*, In "Fixed-Point Algorithms for Inverse Problems in Science and Engineering", Springer Verlag series Optimization and Its Applications, 33–48, 2011.

[3] A. Beck and M. Teboulle, *Convergence rate analysis and error bounds for projection algorithms in convex feasibility problems* , Optimization Methods and Software, 18(4): 377–394, 2003.

[4] A. Beck and M. Teboulle, *A conditional gradient method with linear rate of convergence for solving convex linear systems*, Mathematical Methods of Operations Research 59(2): 235–247, 2004.

[5] H.H. Bauschke and J.M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Review 38(3): 367–426, 1996.

[6] H.H. Bauschke, D. Noll, *On cluster points of alternating projections*, Serdica Mathematical Journal, 39: 355–364, 2013.

[7] H.H. Bauschke, P.L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, Springer, New York, 2011.

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

[9] J.V. Burke and M.C. Ferris, *Weak sharp minima in mathematical programming*, SIAM Journal of Control and Optimization, 31(6): 1340–1359, 1993.

[10] C. Byrne and Y. Censor, *Proximity function minimization using multiple Bregman projections, with applications to split feasibility and Kullback-Leibler distance minimization*, Annals of Operations Research, 105(1): 77-98, 2001.

[11] Y. Censor, T. Elfving and G.T. Herman, *Averaging strings of sequential iterations for convex feasibility problems.* In: D. Butnariu, Y. Censor and S. Reich (editors), Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Elsevier Science Publishers, 101–114, 2001.

[12] 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.

[13] H. Choi, R.G. Baraniuk, *Multiple wavelet basis image denoising using Besov ball projections*, IEEE Signal Process. Lett. 11, 717 - 720, 2004.

[14] P.L. Combettes, *The convex feasibility problem in image recovery*, Advances in Imaging and Electron Physics, 95, 155–270, 1996.

[15] P.L. Combettes, *Hilbertian convex feasibility problem: convergence of projection methods*, Applied Mathematics & Optimization, 35: 311-330, 1997.

[16] F. Deutsch and H. Hundal, *The rate of convergence for the cyclic projections algorithm I: Angles between convex sets*, Journal of Approximation Theory, 142, 36-55, 2006

[17] J. Gu, H. Stark and Y. Yang, *Wide-band smart antenna design using vector space projection methods*, IEEE Trans. Antennas Propag., 52: 3228–3236, 2004.

[18] L.G. Gubin, B.T. Polyak and E.V. Raik, *The method of projections for finding the common point of convex sets*, USSR Computational Mathematics and Mathematical Physics, 7(6): 1–24, 1967.

[19] G.T. Herman, *Fundamentals of Computerized Tomography: Image Reconstruction from Projections*, Springer, New York, 2009.

[20] G.T. Herman, W. Chen, *A fast algorithm for solving a linear feasibility problem with application to intensity-modulated radiation therapy*, Linear Algebra Appl., 428, 1207–1217, 2008.

[21] S. Kaczmarz, *Angenaherte Auflosung von Systemen linearer Gleichungen*, Bull. Acad. Sci. Pologne, A35: 355–357, 1937.

[22] D. Leventhal and A. Lewis, *Randomized methods for linear constraints: convergence rates and conditioning*, Mathematics of Operations Research, 35(3): 641 - 654, 2010.

[23] A. Liew, H. Yan and N. Law, *POCS-based blocking artifacts suppression using a smoothness constraint set with explicit region modeling*, IEEE Trans. Circuits Syst. Video Technol. 15: 795–800, 2005.

[24] Y.M. Lu, M. Karzand and M. Vetterli, *Demosaicking by alternating projections: Theory and fast one-step implementation*, IEEE Trans. Image Process., 19(8): 2085–2098, 2010.

[25] T. Motzkin and I. Schoenberg, *The relaxation method for linear inequalities*, Canad. J. Math., 6: 393–404, 1954.

[26] I. Necoara, Yu. Nesterov, F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, submitted, 2015.

[27] A. Nedic, *Random Projection Algorithms for Convex Set Intersection Problems*, 49th IEEE Conference on Decision and Control, 7655–7660, 2010.

[28] A. Nedic, *Random Algorithms for Convex Minimization Problems*, Mathematical Programming, Series B, 129: 225–253, 2011.

[29] A. Patrascu and I. Necoara, *Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization*, Journal of Machine Learning Research, 2017 (to appear).

[30] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press; 3rd edition, 1996.

[31] R. M. Gower and P. Richtarik, *Randomized iterative methods for linear systems*, SIAM Journal on Matrix Analysis and Applications, 36(4): 1660–1690, 2015.

[32] A. Shapiro, D. Dentcheva and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory*, Siam, 2009.

[33] A.A. Samsonov, E.G. Kholmovski, D.L. Parker, C.R. Johnson, *POCSENSE: POCS- based reconstruction for sensitivity encoded magnetic resonance imaging*, Magn. Reson. Med., 52: 139–1406, 2004.

[34] G. Sharma, *Set theoretic estimation for problems in subtractive color*, Color Res. Appl., 25: 333–348, 2000.

[35] H. Stark and Y. Yang, *Vector Space Projections : A Numerical Approach to Signal and Image Processing*, Neural Nets and Optics, Wiley-Interscience, 1998.

[36] T. Strohmer and R. Vershynin, *A randomized Kaczmarz algorithm with exponential convergence*, Journal of Fourier Analysis and Applications, 15, 2009.

[37] J. von Neumann, *Functional operators*, Princeton University Press, 1950.

[38] M. Yukawa, I. Yamada, *Pairwise optimal weight realization: Acceleration technique for set-theoretic adaptive parallel subgradient projection algorithm*, IEEE Trans. Signal Process. 54: 4557–4571, 2006.