SIAM J. OPTIM.
Vol. 29, No. 4, pp. 2814–2852

RANDOMIZED PROJECTION METHODS FOR CONVEX FEASIBILITY: CONDITIONING AND CONVERGENCE RATES*

ION NECOARA†, PETER RICHTÁRIK‡, AND ANDREI PATRASCU†

Abstract. In this paper we develop a family of *randomized projection methods* (RPM) for solving the convex feasibility problem. Our approach is based on several new ideas and tools, including stochastic approximation of convex sets, stochastic reformulations, and conditioning of the convex feasibility problem. In particular, we propose four equivalent stochastic reformulations: stochastic smooth and nonsmooth optimization problems, the stochastic fixed point problem, and the stochastic feasibility problem. The last reformulation asks for a point which belongs to a certain random set with probability one. In this case, RPM can be interpreted as follows: we sample a batch of random sets in an independently and identically distributed fashion, perform projections of the current iterate on all sampled sets, and then combine these projections by averaging with a carefully designed extrapolation step. We prove that under stochastic linear regularity, RPM converges linearly, with a rate that has a natural interpretation as a condition number of the stochastic optimization reformulation and that depends explicitly on the number of sets sampled. In doing so, we extend the concept of condition number to general convex feasibility problems. This condition number depends on the linear regularity constant and an additional key constant which can be interpreted as a Lipschitz constant of the gradient of the stochastic optimization reformulation. Besides providing a general framework for the design and analysis of randomized projection schemes, our results resolve an open problem in the literature related to the theoretical understanding of observed practical efficiency of extrapolated parallel projection methods. In addition, we prove that our method converges sublinearly in the case when the stochastic linear regularity condition is not satisfied. Preliminary numerical results also show a better performance of our extrapolated step-size scheme over its constant step-size counterpart.

Key words. convex feasibility, stochastic reformulations, condition number, randomized projection methods, convergence rates

AMS subject classifications. 90C25, 90C15, 65K05

DOI. 10.1137/18M1167061

1. Introduction.

The *convex feasibility problem* seeks to find a point belonging to a nonempty closed convex set $\mathcal{X} \subset \mathbb{R}^n$:

$$\text{find} \quad x \in \mathcal{X}. \tag{1.1}$$

In many applications it is sufficient to relax this hard feasibility requirement and only aim to find a point which is not too far from $\mathcal{X}$. In particular, one chooses an error tolerance $\varepsilon > 0$ and aims to find a point x satisfying $\text{dist}^2_{\mathcal{X}}(x) \leq \varepsilon$, where $\text{dist}_{\mathcal{X}}(\cdot) \overset{\text{def}}{=} \min_{y \in \mathcal{X}} \|\cdot - y\|$ is the distance function and $\|\cdot\|$ is the standard Euclidean norm on $\mathbb{R}^n$. In the case when a randomized algorithm is used to find x, which renders x a random vector, one may further relax the soft feasibility requirement and replace

*Received by the editors January 25, 2018; accepted for publication (in revised form) June 19, 2019; published electronically November 7, 2019.
https://doi.org/10.1137/18M1167061
Funding: The work of the first and third authors was supported by the Executive Agency for Higher Education, Research and Innovation Funding (UEFISCDI), Romania, PNIII-P4-PCE-2016-0731, project ScaleFreeNet, 39/2017.

†Automatic Control and Systems Engineering Department, University Politehnica Bucharest, 060042 Bucharest, Romania (ion.necoara@acse.pub.ro, andrei.patrascu@acse.pub.ro).
‡King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia; University of Edinburgh, Edinburgh, United Kingdom; and Moscow Institute of Physics and Technology (MIPT), Dolgoprudny, Russia (peter.richtarik@kaust.edu.sa).

it with

$$\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x)\right] \leq \varepsilon, \tag{1.2}$$

where $\mathbf{E}[\cdot]$ denotes the expectation with respect to the randomness of the algorithm. Convex feasibility represents a modeling paradigm for solving many engineering and physics problems: radiation therapy planning [17], magnetic resonance imaging [28], wavelet denoising [11], color imaging [29], antenna design [14], sensor networks [6], data compression [20], neural networks [30], and machine learning and optimal control [26].

1.1. Projection methods. In applications, $\mathcal{X}$ is typically represented as the intersection of a large but finite number of simpler (closed convex) sets:

$$\mathcal{X} = \bigcap_{i=1}^m \mathcal{X}_i. \tag{1.3}$$

A popular class of methods for solving (1.1) with $\mathcal{X}$ represented in the form (1.3) are *projection methods*. They are very attractive in applications since they are able to handle problems of huge dimension and with a large number of sets in the intersection. Projection methods were first used for solving systems of linear equalities [18] or inequalities [21], and then extended to general convex feasibility problems, e.g., in [12]. The key subroutine of these methods is the Euclidean projection of the current iterate onto one or more of the sets $\mathcal{X}_i$ selected via a certain rule, e.g., cyclic, greedy, or random. For this approach to make sense, it is assumed that projections onto the individual sets can be computed efficiently (e.g., in closed form or via a fast method). The resulting projections are averaged to form the next iterate. This leads to the iterative process

$$x^{k+1} = \tfrac{1}{|J^k|} \textstyle\sum_{i\in J^k} \Pi_{\mathcal{X}_i}(x^k), \tag{1.4}$$

where $\emptyset \neq J^k \subseteq [m] \overset{\text{def}}{=} \{1,\ldots,m\}$ is the collection of sets active at iteration k, and $\Pi_{\mathcal{X}_i}$ is the projection operator onto the set $\mathcal{X}_i$. Variants performing projection onto a single set $\mathcal{X}_i$ in each iteration only (i.e., $|J^k| = 1$ for all k), and irrespectively of the particular set selection rule employed, are known as *alternating* projection methods [3, 4, 16, 23]. On the other hand, variants performing projections onto multiple sets in each iteration are typically referred to as *average* or *parallel* projection methods [10, 9]. The convergence properties and even the inherent limitations of projection methods have been intensely analyzed over the last decades, as can be seen in [2, 1, 3, 4, 10, 12, 16, 23, 24] and the references therein.

1.2. Extrapolation. It is well known that the practical performance of parallel projection methods can be enhanced, and often dramatically so, using *extrapolation*. This refers to the practice of moving *further* along the line connecting the last iterate and the average of the projections:

$$x^{k+1} = x^k + \alpha^k \left(\tfrac{1}{|J^k|} \textstyle\sum_{i\in J^k} \Pi_{\mathcal{X}_i}(x^k) - x^k \right), \tag{1.5}$$

where $\alpha^k \geq 1$ is the extrapolation parameter ($\alpha^k = 1$ means that no extrapolation is employed). There are several heuristic rules for deciding on the amount of extrapolation, some of which work very well in practice [5, 12, 9]. A particularly successful rule is to set α^k according to [12]

$$\alpha^k = \frac{\sum_{i\in J^k} w_i^k \left\|x^k - \Pi_{\mathcal{X}_i}(x^k)\right\|^2}{\left\|\sum_{i\in J^k} w_i^k (x^k - \Pi_{\mathcal{X}_i}(x^k))\right\|^2}, \tag{1.6}$$

where $\{w_i^k\}_{i\in J^k}$ are nonnegative weights adding up to one (typically one just sets $w_i^k = 1/|J^k|$ for all i). However, despite more than 80 years of research on projection methods, the empirical success of extrapolation schemes is not supported by theory. That is, to the best of our knowledge, there is no theory explaining why these methods require fewer iterations than their nonextrapolated variants.

1.3. Representation. Convex sets $\mathcal{X}$ can be represented in the form of an intersection of simple sets in multiple ways. For instance, given representation (1.3), one can define $\mathcal{X}_{ij} \overset{\text{def}}{=} \mathcal{X}_i \cap \mathcal{X}_j$ and write $\mathcal{X} = \bigcap_{ij} \mathcal{X}_{ij}$. One does not have to be limited to working with pairs of sets, but may introduce sets of the form $\mathcal{X}_S \overset{\text{def}}{=} \bigcap_{i\in S} \mathcal{X}_i$, where $S \subset [m]$. As long as the sets S cover $[m]$, we have $\mathcal{X} = \bigcap_S \mathcal{X}_S$. In many situations, such as in the case of linear feasibility (see Example 3.1), one can define an infinite number of reformulations. Assuming that multiple representations are available, which one is the best? *The current literature on convex feasibility does not have tools to ask such questions. It is almost entirely devoted to designing projection methods for a fixed representation which is assumed to be given as an input.* An exception to this are some recent works on linear feasibility [16, 33]. There are several difficulties with trying to decide among multiple representations. First, it is not possible to decouple the choice of a representation from the choice of a projection method as the latter cannot exist without the former. Indeed, while the choice of a representation will affect the cost of projections, the choice of a projection algorithm will dictate convergence speed, which in turn depends on the representation. The overall complexity of the representation-algorithm pair is given by the number of iterations needed to find x sufficiently close to $\mathcal{X}$ multiplied by the average projection cost.

1.4. Importance. While (1.3) appears as a "democratic" intersection of m sets, it is likely the case that some of these sets are more important than others. As an illustration, consider the scenario in which there exists $T \subset [m]$ such that $\mathcal{X} = \bigcap_{i\in T} \mathcal{X}_i$. Clearly, sets $\mathcal{X}_i$ for $i \in T$ are more important than sets $\mathcal{X}_i$ for $i \notin T$. This is an extreme scenario: if T is known, one should simply remove the nonimportant sets from the representation to begin with. However, even if none of the sets can be removed, it is often the case that some sets are more important than others in the sense that one should project on these sets more often. *We are not aware of any papers on convex feasibility that investigate the importance of sets.*

1.5. Outline. In section 2 we summarize selected key contributions of this paper. In section 3 we develop stochastic approximations and reformulations of the convex feasibility problem and analyze some key properties of these reformulations in section 4. Section 5 presents some concrete examples of stochastic approximations for a set, while in section 6 we present our main projection algorithms and their convergence properties. In section 7 we corroborate our theoretical results through numerical experiments.

1.6. Notation. For $x, y \in \mathbb{R}^n$, the standard inner product is denoted by $\langle x, y\rangle \overset{\text{def}}{=} x^\top y$, the standard Euclidean norm by $\|x\| \overset{\text{def}}{=} \sqrt{x^\top x}$, and the 0-"norm" (number of nonzeros of x) by $\|x\|_0$. By x_i we denote the ith component of vector x. For a positive integer m, let $[m] \overset{\text{def}}{=} \{1, 2, \ldots, m\}$. By e_i we denote the ith column of the identity matrix $I_m \in \mathbb{R}^{m\times m}$. Let Q be a matrix. By $\|Q\|_F$, Rank(Q), $Q_i^\top$, $\sigma_{\min}^+(Q)$, $\lambda_{\min}^+(Q)$, $\sigma_{\max}(Q)$, and $\lambda_{\max}(Q)$ we denote its Frobenius norm, rank, ith row, the smallest nonzero singular value, the smallest nonzero eigenvalue, the largest singular value, and the largest singular value, respectively. The projection operator onto a closed

convex set $\mathcal{X}$ is denoted by $\Pi_{\mathcal{X}}(\cdot)$, the distance of a point x from $\mathcal{X}$ is denoted by $\mathrm{dist}_{\mathcal{X}}(x) \stackrel{\mathrm{def}}{=} \min_z\{\|x-z\| : z \in \mathcal{X}\}$, and $\mathbb{I}_{\mathcal{X}}$ is the indicator function of $\mathcal{X}$.

2. Contributions. In this section we briefly review our key contributions and results, leaving details to the rest of the paper.

2.1. General framework. We *develop a unified framework for studying extrapolation, representation, and importance questions for general convex feasibility problems, together with randomized projection methods for solving such problems.* We do this via a *stochastic representation-algorithm codesign.* Given a closed convex set $\mathcal{X}$, we represent it in the form

$$\mathcal{X} = \bigcap_{S\in\Omega} \mathcal{X}_S, \tag{2.1}$$

where $\mathcal{X}_S$ are closed convex sets indexed by an arbitrary (and possibly countably or uncountably infinite) set Ω. For instance, $\{\mathcal{X}_S : S \in \Omega\}$ can be the collection of *all* closed convex sets containing $\mathcal{X}$ or a "very large" and "well-behaved" subcollection of such sets. The idea here is to *not commit* to any particular representation of $\mathcal{X}$, and to include as many sets as possible. We then extend Ω into a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ which plays several roles at the same time.

(i) Let a support of $\mathbf{P}$ be any $\Omega' \in \mathcal{F}$ for which $\mathbf{P}(\Omega') = 1$. If it is the case that $\mathcal{X} = \bigcap_{S\in\Omega'} \mathcal{X}_S$, then the probability measure effectively *selects* the representation of $\mathcal{X}$ as an intersection of a *smaller* collection of sets. By adjusting the support, we adjust the representation.

(ii) By sampling $S \sim \mathbf{P}$, we are *choosing* a set $\mathcal{X}_S$. In this way we achieve two goals at the same time. First, this defines a *random selection rule* which we shall use to design a *randomized projection method* (`RPM`) and an *expected projection method* (`EPM`), described in section 2.3. Second, the choice of probability measure is a natural way to assign *importance* to the sets $\mathcal{X}_S$.

The probability space and the sets $\mathcal{X}_S$ need to satisfy certain regularity properties for our framework to make sense; we leave such technicalities to subsequent sections (see section 3). Note that $\mathbf{P}$ is a *parameter* playing the dual role of controlling the representation of $\mathcal{X}$ as an intersection of sets, and defining the (importance) sampling procedure which in turn defines the algorithm.

2.2. Stochastic reformulations. Based on the stochastic representation of a convex set we derive several new equivalent stochastic reformulations of the convex feasibility problem that are governed by an arbitrary discrete/continuous random variable. In particular, our reformulation can be seen as a stochastic smooth/nonsmooth optimization problem, a stochastic fixed point problem, or a stochastic intersection problem (see (3.5)–(3.8)). We also prove sufficient conditions for the reformulations to be exact. Moreover, our stochastic reformulations allow us to deal easily and naturally with intersection of families of sets that may even be uncountable. By combining these reformulations with certain regularity assumptions on the individual convex sets we extend the concept of condition number from convex optimization to convex feasibility. More precisely, a key quantity in our main convergence analysis is a *stochastic* variant of linear regularity, which requires the existence of a constant $\mu > 0$ for which

$$\mu \cdot \mathrm{dist}^2_{\mathcal{X}}(x) \le \mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}_S}(x)\right] \quad \forall x \in \mathbb{R}^n. \tag{2.2}$$

Another important quantity is defined in terms of a Lipschitz-like constant $L > 0$ satisfying the following inequality:

$$\|\mathbf{E}\left[x - \Pi_{\mathcal{X}_S}(x)\right]\|^2 \le L \cdot \mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right] \quad \forall x \in \mathbb{R}^n. \tag{2.3}$$

TABLE 1

The key convergence results obtained in this paper for our two new methods: RPM and EPM. Here $L_N = \frac{1}{N} + (1 - \frac{1}{N})L$, *$L$ is the Lipschitz constant, μ is the (stochastic) linear regularity constant, and $\hat{x}$ is the average of the iterates $x^0, \ldots, x^{k-1}$.*

	RPM with any finite $N \geq 1$ (Theorem 6.4)	EPM ($=$ RPM with $N = \infty$) (Theorem 6.6)
$\mu = 0$	$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(\hat{x}^k)\right] \leq \frac{L_N}{k}\text{dist}^2_{\mathcal{X}}(x^0)$	$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(\hat{x}^k)\right] \leq \frac{L}{k}\text{dist}^2_{\mathcal{X}}(x^0)$
$\mu > 0$	$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] \leq (1 - \frac{\mu}{L_N})^k\text{dist}^2_{\mathcal{X}}(x^0)$	$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] \leq \left(1 - \frac{\mu}{L}\right)^k \text{dist}^2_{\mathcal{X}}(x^0)$

It follows from Jensen's inequality that (2.3) always holds for $L = 1$. It is also the case that $0 \leq \mu \leq L \leq 1$. The ratio $\texttt{cond} = \mu/L$ represents the condition number of the feasibility problem (2.1). For more details, see section 4 (Theorem 4.1). Note that these equivalent reformulations can also facilitate the development of new algorithmic schemes using domain specific insights.

2.3. Algorithms. We propose two projection algorithms (see section 6). The first is the *randomized projection method* (RPM):

1. draw N independent samples $S_1^k, \ldots, S_N^k \sim \mathbf{P}$,
2. update $x^{k+1} = x^k + \alpha^k \left(\frac{1}{N}\sum_{i=1}^N \Pi_{\mathcal{X}_{S_i^k}}(x^k) - x^k\right)$,

where $\Pi_{\mathcal{X}}(\cdot)$ denotes the projection operator onto $\mathcal{X}$, and $N \geq 1$ is a minibatch/ parallelism parameter. One important distinction between our algorithm and other existing ones is the use of an extrapolated step size $\alpha_k \sim 2/L_N$, depending on the conditioning parameter L and minibatch size N through $L_N = \frac{1}{N} + (1 - \frac{1}{N})L$, that, in general, is much larger than the constant step size $\alpha_k \equiv \alpha \in (0, 2)$ usually used in the literature. This extrapolation drastically accelerates the convergence of the method (see numerical results of section 7). Another feature of our algorithm is that it allows us to simultaneously project onto several sets ($N \geq 1$), thus providing flexibility in matching the implementation of the algorithm on the parallel architecture at hand. RPM can be equivalently interpreted as a *minibatch* stochastic gradient descent, a stochastic proximal point method, a stochastic fixed point method, and a stochastic projection method, with extrapolation, applied to the corresponding stochastic reformulations of the convex feasibility problem. Our second method is the *expected projection method* (EPM):

$$x^{k+1} = x^k + \alpha^k \left(\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x^k) \mid x^k\right] - x^k\right).$$

Here expectation is taken w.r.t. the random variable $S \sim \mathbf{P}$. For simplicity of notation, in what follows we omit this dependence. Note that EPM can be obtained as the limit case of RPM by letting $N \to +\infty$.

2.4. Convergence rates. Based on the conditioning parameters μ and L, as well as asymptotic convergence, we also derive explicit convergence rates for these two algorithms. Our convergence rates depend explicitly on the conditioning parameters and number of computed projections per iteration N (see Table 1).

2.4.1. Randomized projection method. We prove that the iterates of RPM with constant extrapolation parameter $0 < \alpha_k \equiv \alpha < \frac{2}{L_N}$ converge linearly to $\mathcal{X}$ with

iteration complexity (see Theorem 6.4)

$$O\left(\frac{1}{\alpha(2-\alpha L_N)\mu}\log\frac{1}{\varepsilon}\right), \tag{2.4}$$

where $L_N = \frac{1}{N} + (1-\frac{1}{N})L$, and L is defined in (2.3). Note that in the $L = 1$ case we get $L_N = 1$ for all N, and hence complexity does not improve with N. However, as long as $L < 1$ (and it can also be the case that $L \approx 0$), L_N decreases in N, which shows that extrapolation indeed works and improves complexity. *To the best of our knowledge, this is the first time a parallel projection method ($N > 1$) has been shown to have a better rate than its nonparallel variant ($N = 1$), whether extrapolation is used ($\alpha > 1$) or not ($\alpha = 1$).* Note that for any fixed N, the optimal constant extrapolation parameter is $\alpha_k \equiv \alpha = \frac{1}{L_N}$. As long as $N \geq 2$ and $L < 1$, we have $\alpha > 1$ and hence extrapolation improves complexity. *To the best of our knowledge, this is the first time extrapolated projection methods are shown to be better than their nonextrapolated variants. We have identified L as the key quantity determining whether extrapolation helps ($L < 1$) or not ($L = 1$), and how much (the smaller L, the more it helps).* Using optimal extrapolation we obtain the rate $O(\frac{L_N}{\mu}\log\frac{1}{\varepsilon})$ (see Table 1, bottom left). Note that $L_N > \frac{1}{N} = \frac{L_1}{N}$, which means that minibatching can never reach *perfect* linear speedup in N. However, as long as $L = O(\frac{1}{N})$, we have $L_N = O(\frac{1}{N})$, which means almost linear speedup in N. To the best of our knowledge, convergence rates of projection methods were only previously derived for extrapolation parameters belonging to the interval $(1, 2)$ and the existing convergence estimates do not show dependence on the number N of sets we project at each iteration (see, e.g., [1, 9, 12, 24, 27]). However, as discussed before, our convergence analysis allows us to derive convergence estimates for more general extrapolation rules that depend explicitly on N. Note that the optimal step size $\alpha = 1/L_N$ satisfies $\lim_{N\to+\infty} 1/L_N = 1/L$, and the adaptive step size (1.6) proposed in [12] can be seen as a practical online approximation of our extrapolated step size. Indeed, (1.6) is an empirical approximation of the ratio of the two expectations defining $1/L$ (see Remark 6.5).

2.4.2. Expected projection method. The complexity results for `EPM` are as one would expect: they are obtained by taking limits for $N \to +\infty$ in the corresponding complexity results for `RPM`. Again, the results presented in Table 1 are for the optimal extrapolation parameter, which in this case is equal to $\alpha^k = \lim_{N\to+\infty} 1/L_N = 1/L$. Some of the comments made in the above discussion are summarized in Table 2.

2.4.3. Results without linear regularity. Finally, we prove convergence of these two methods for the $\mu = 0$ case: complexity $O(\frac{L_N}{\varepsilon})$ for `RPM` and $O(\frac{L}{\varepsilon})$ for `EPM`, both in an ergodic sense. However, in this case it is not possible to show convergence to $\mathcal{X}$. Instead, we show that the expected squared distance from $\mathcal{X}_S$ approaches zero.

2.5. Related work. The papers most closely related to our work are [23, 24, 27]. Although these papers also consider general convex feasibility problems, they assume a fixed representation for the set $\mathcal{X}$, given either in terms of finite intersection of simple convex sets [23] or in terms of functional constraints [24, 27]. They also state that choices for the probabilities of sampling sets other than the uniform one may lead to better projection schemes, without any theoretical justification, however. For such fixed representations, [23, 24, 27] analyze only random projection algorithms that require projection onto a single set at each iteration, i.e., $N = 1$. Finally, they do not consider extrapolation $\alpha_k > 2$ in the convergence analysis, i.e., their convergence

TABLE 2

Convergence of RPM *($N = 1, 2, \dots$) and* EPM *($N = +\infty$) under stochastic linear regularity ($\mu > 0$) and constant extrapolation $\alpha^k \equiv \alpha$. In all cases, the iteration complexity is $O([\alpha(2 - \alpha L_N)\mu]^{-1} \log(1/\varepsilon))$; see (2.4). In this table, the logarithmic factor is suppressed and only the dominant term $(\alpha(2-\alpha L_N)\mu)^{-1}$ is displayed. Rows correspond to minibatch strategies (varying N), and columns correspond to unfavourable ($L = 1$) and favourable ($L < 1$) regime for minibatching, under no extrapolation ($\alpha = 1$) and optimal extrapolation ($\alpha = 1/L_N$). Observations: if $L = 1$, then neither minibatching nor extrapolation help, and the complexity is $1/\mu$. If $L < 1$ and no extrapolation is used, the rate improves at most by a factor of 2 as we move from $N = 1$ to $N = +\infty$. When extrapolation is used, the rate can improve by up to the factor $1/L$ as we increase the minibatch size. This can be huge: if $L = 10^{-3}$, say, this is an improvement by 3 orders of magnitude.*

	$L = 1$ (bad for minibatching)		$L < 1$ (good for minibatching)	
	No extrapolation $\alpha = 1$	Optim. extrapolation $\alpha = \frac{1}{L_N}$	No extrapolation $\alpha = 1$	Optim. extrapolation $\alpha = \frac{1}{L_N}$
$N = 1$ $(L_N = 1)$	$\frac{1}{\mu}$	$\frac{1}{\mu}$	$\frac{1}{\mu}$	$\frac{1}{\mu}$
$N \in \{1, 2, \dots\}$ $(L_N = \frac{1}{N} + (1 - \frac{1}{N})L)$	$\frac{1}{\mu}$	$\frac{1}{\mu}$	$\frac{1}{(2 - L_N)\mu}$	$\frac{L_N}{\mu}$
$N = +\infty$ $(L_N = L)$	$\frac{1}{\mu}$	$\frac{1}{\mu}$	$\frac{1}{(2 - L)\mu}$	$\frac{L}{\mu}$

results are derived for a diminishing step size, constant step size $\alpha \in (0, 2)$, or an ideal constant step size that depends on the radius of a ball contained in the feasible set $\mathcal{X}$. Note that our general convergence results for $N = 1$ recover the main convergence rates from [23, 24, 27]; see section 6.3 for a detailed comparison.

3. Stochastic approximations of sets. As we have seen in the introduction, projection methods are intimately tied with the representation of $\mathcal{X}$ as an intersection of sets. However, such representations are not unique. Let us illustrate this with an example.

Example 3.1 (linear feasibility). Let $\mathcal{X} = \{x : Ax = b\} \neq \emptyset$, where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

(i) *Trivial.* We can trivially represent $\mathcal{X}$ as itself (the "intersection" of a single set). This representation is not useful since a projection onto $\mathcal{X}$ is (at least) as hard as solving the original problem (1.1). Indeed, while (1.1) only requires us to find a point in $\mathcal{X}$, a projection step will lead to a point in $\mathcal{X}$ closest to the starting point.

(ii) *Natural.* The most natural representation of $\mathcal{X}$ is to write $\mathcal{X} = \bigcap_{i \in [m]} \mathcal{X}_i$, where $\mathcal{X}_i \overset{\text{def}}{=} \{x : A_i^\top x = b_i\}$ and $A_i^\top$ is the ith row of A. Projection onto the hyperplane $\mathcal{X}_i$ can be performed in closed form at a cost $O(n)$.

(iii) *Blocks.* More generally, we can choose sets of rows $\emptyset \neq S_1, S_2, \dots, S_t \subseteq [m]$ covering $[m]$, define $\mathcal{X}_{S_i} \overset{\text{def}}{=} \bigcap_{j \in S_i} \mathcal{X}_j$ and represent $\mathcal{X}$ as $\mathcal{X} = \bigcap_{i \in [t]} \mathcal{X}_{S_i}$. Projection onto $\mathcal{X}_{S_i}$ involves forming a $|S_i| \times |S_i|$ matrix at a cost $O(n \times |S_i|^2)$, and solving a linear system involving this matrix, which costs $O(|S_i|^3)$ if a direct method is used. Note that we recover the natural representation from (ii) as a special case by setting $t = m$ and $S_i = \{i\}$ for all i. On the other hand, the trivial representation from (i) arises as a special case by setting $t = 1$ and $S_1 = [m]$.

(iv) *Sketching.* Finally, for $S \in \mathbb{R}^{m\times q}$ define $\mathcal{X}_S \stackrel{\text{def}}{=} \{x : S^\top Ax = S^\top b\}$, and let $\Omega \stackrel{\text{def}}{=} \bigcup_{q=1}^{m} \mathbb{R}^{m\times q}$. Clearly, $\mathcal{X} \subset \mathcal{X}_S$ for all S and $\mathcal{X} = \bigcap_{S\in\Omega} \mathcal{X}_S$. However, there are many different subsets $\Omega' \subseteq \Omega$ for which $\mathcal{X} = \bigcap_{S\in\Omega'} \mathcal{X}_S$. The trivial approximation from (i) corresponds to choosing $\Omega' = \{I_m\} \in \mathbb{R}^{m\times m}$ (the $m \times m$ identity matrix). The natural representation from (ii) corresponds to choosing $\Omega' = \{e_1, e_2, \ldots, e_m\} \subset \mathbb{R}^{m\times 1}$, where e_i is the ith unit basis vector in $\mathbb{R}^m$. The block representation from (iii) corresponds to choosing $\Omega' = \{I_{:S_i} : i \in [t]\} \in \mathbb{R}^{m\times|S_i|}$, where $I_{:S_i}$ is the column submatrix of I_m corresponding to columns belonging to S_i.

Given a multitude of possible representations to choose from, how do we choose? Clearly, the answer depends on the combined effect of the projection cost and the number of iterations needed for the method to converge. While one only needs a single step of a projection method to solve the problem using the trivial representation, the cost of such a step will typically be prohibitive. On the other hand, while projecting onto $\mathcal{X}_i$ is cheap, many iterations will be needed for a projection method to converge. As we have seen, these two examples correspond to the extreme choices $\Omega' = I_m$ and $\Omega' = \{e_1, e_2, \ldots, e_m\}$, respectively. However, we can choose *any* other $\Omega' \subset \Omega$ for which $\mathcal{X} = \bigcap_{S\in\Omega'} \mathcal{X}_S$. Which one is the best? This is a very difficult question, further complicated by the fact that it is not possible to fully decouple the choice of the representation from the choice of the projection algorithm as these two choices are intimately intertwined.

3.1. Stochastic representation. In this work we construct a systematic approach to asking such representation questions. In particular, we fix some ground set Ω and with each $S \in \Omega$ associate a closed convex set $\mathcal{X}_S$ containing $\mathcal{X}$. The way that this can be done depends on the *input* representation and structure of $\mathcal{X}$; the case of linear feasibility is illustrated in Example 3.1. Other examples are given in section 5. Since $\mathcal{X} \subset \mathcal{X}_S$ for all $S \in \Omega$, we have a first relaxation

$$\mathcal{X} \subseteq \mathcal{X}' \stackrel{\text{def}}{=} \bigcap_{S\in\Omega} \mathcal{X}_S. \tag{3.1}$$

While we will have particular interest in situations when $\mathcal{X} = \mathcal{X}'$, let us postpone this requirement until later. Since we study projection methods which need to identify a set or sets $\mathcal{X}_S$ to project on in each iteration, we need to have a way of *selecting* $S \in \Omega$. If Ω is finite or countable, there are many selection rules that can in principle be used (e.g., cyclic, greedy, random). However, since we wish to allow for Ω to be infinite and even uncountable, in this paper we shall focus on *random selection rules*, which in turn lead to stochastic projection methods. In particular, we define a probability measure $\mathbf{P}$ on Ω and consider a further relaxation of the set $\mathcal{X}'$, which we refer to as the *probabilistic intersection*:

$$\mathcal{X}'' \stackrel{\text{def}}{=} \bigcap_{S\sim\mathbf{P}} \mathcal{X}_S \stackrel{\text{def}}{=} \{x : \mathbf{P}(x \in \mathcal{X}_S) = 1\}. \tag{3.2}$$

Clearly, from (3.1) and (3.2) we have the inclusions

$$\mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{X}''. \tag{3.3}$$

If $\mathcal{X} = \mathcal{X}''$ (a condition we call *exactness*), we can think of $\mathcal{X}''$ as a new representation of $\mathcal{X}$ in the form of an intersection of sets $\mathcal{X}_S$ over $S \in \Omega$ having measure (probability) 1. In this way we have arrived at a parametric family of representations of $\mathcal{X}$ depending on the choice of the ground set Ω and on the choice of the probability

measure $\mathbf{P}$. The support of $\mathbf{P}$ can be interpreted as Ω'. Hence, by choosing the probability measure $\mathbf{P}$, we are also choosing the representation of $\mathcal{X}$. Under exactness (we shall propose a sufficient condition for exactness in Lemma 3.5), we have managed to re-represent $\mathcal{X}$ in the form of the probabilistic intersection $\mathcal{X}''$ (see (3.2)), and can study the following natural stochastic projection method (and its extensions that include multiple projections and step sizes/relaxation parameters): sample $S_k \sim \mathbf{P}$ and set $x^{k+1} = \Pi_{\mathcal{X}_{S_k}}(x^k)$. Now, we are ready to introduce the concept of a *stochastic approximation* of $\mathcal{X}$.

DEFINITION 3.2. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and consider a random set-valued mapping from Ω to $2^{\mathbb{R}^n}$ which maps $S \in \Omega$ to $\mathcal{X}_S \subseteq \mathbb{R}^n$. If* (i) *$\mathcal{X}_S$ is a closed and convex subset of $\mathbb{R}^n$ for all $S \in \Omega$,* (ii) *$\mathcal{X} \subseteq \mathcal{X}_S$ for all $S \in \Omega$, and* (iii) *$S \mapsto \operatorname{dist}^2_{\mathcal{X}_S}(x)$ is measurable for all $x \in \mathbb{R}^n$, then we say that this mapping is a stochastic approximation of $\mathcal{X}$. If, moreover, the expectation operator satisfies $\mathbf{E}[\operatorname{dist}^2_{\mathcal{X}_S}(x)] = 0 \Leftrightarrow x \in \mathcal{X}$, then we say that the stochastic approximation is exact.*

3.2. Stochastic reformulations. From our definition it follows that either the set $\mathcal{X}$ is represented as an exact countable/uncountable intersection of the stochastic approximation sets $\mathcal{X}_S$, that is, $\mathcal{X} = \bigcap_{S\in\Omega} \mathcal{X}_S$, or approximated by this intersection, that is, $\mathcal{X} \subseteq \bigcap_{S\in\Omega} \mathcal{X}_S (:= \mathcal{X}')$. Thus, we consider the following convex feasibility problem, which may be a relaxation of the potentially difficult original problem (1.1):

$$\text{find} \quad x \in \mathcal{X}' \stackrel{\text{def}}{=} \bigcap_{S\in\Omega} \mathcal{X}_S. \tag{3.4}$$

We propose below four stochastic reformulations of the convex feasibility problem (3.4).

1. *Stochastic fixed point problem.*

$$\text{Find a fixed point of the expectation operator} \quad \Pi(x) \stackrel{\text{def}}{=} \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]. \tag{3.5}$$

2. *Stochastic nonsmooth optimization problem.*

$$\text{Minimize} \quad f(x) \stackrel{\text{def}}{=} \mathbf{E}\left[\mathbb{I}_{\mathcal{X}_S}(x)\right] \quad \text{subject to} \quad x \in \mathbb{R}^n. \tag{3.6}$$

3. *Stochastic smooth optimization problem.*

$$\text{Minimize} \quad F(x) \stackrel{\text{def}}{=} \tfrac{1}{2}\mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right] \quad \text{subject to} \quad x \in \mathbb{R}^n. \tag{3.7}$$

4. *Stochastic feasibility problem.*

$$\text{Find} \quad x \in \mathcal{X}'' \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : \mathbf{P}(x \in \mathcal{X}_S) = 1\}. \tag{3.8}$$

Thus, our reformulation can be seen as a stochastic smooth/nonsmooth optimization problem, a stochastic fixed point problem, or a stochastic feasibility problem. Note that the stochastic smooth optimization formulation (3.7) has also been considered in [23, 27] for solving the convex feasibility problem (3.4). However, the other three stochastic reformulations from above seem to be new. These four reformulations allow for researchers from various communities to leverage their domain-specific insights for the development of new algorithms. The equivalence of the above stochastic reformulations is captured by the following theorem.

THEOREM 3.3 (equivalence of reformulations). *The stochastic reformulations* (3.5)–(3.8) *of the convex feasibility problem* (3.4) *are equivalent. That is, the set of fixed points of $x \mapsto \mathbf{E}[\Pi_{\mathcal{X}_S}(x)]$ is equal to the set of minimizers of the objective functions f or F, and to the set $\mathcal{X}''$.*

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

(3.6)$\Leftrightarrow$(3.8) The $\mathbf{P}$-measurable function $\mathbb{I}_{\mathcal{X}_S}(x)$ is nonnegative and thus the set of minimizers in (3.6) are those x for which $\mathbf{E}[\mathbb{I}_{\mathcal{X}_S}(x)] = 0$, 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 $\mathbf{P}(x \in \mathcal{X}_S) = 1$.

(3.7)$\Leftrightarrow$(3.8) The function $F_S(x) = \frac{1}{2}\|x - \Pi_{\mathcal{X}_S}(x)\|^2$ is nonnegative and thus the set of minimizers in (3.7) are those x for which $\mathbf{E}[\|x - \Pi_{\mathcal{X}_S}(x)\|^2] = 0$, 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 $\mathbf{P}(x \in \mathcal{X}_S) = 1$.

(3.7)$\Rightarrow$(3.5) Since $\|\mathbf{E}[x - \Pi_{\mathcal{X}_S}(x)]\|^2 \leq \mathbf{E}[\|x - \Pi_{\mathcal{X}_S}(x)\|^2]$, it follows that the set of minimizers of (3.7) is included in the set of fixed points of the average projection operator $\Pi(x) = \mathbf{E}[\Pi_{\mathcal{X}_S}(x)]$ from (3.5). It remains to prove the other inclusion (3.5)$\Rightarrow$(3.7). Let x be a fixed point of the average projection operator, that is, $x = \mathbf{E}[\Pi_{\mathcal{X}_S}(x)]$. Then, for any $z \in \bigcap_{S\in\Omega} \mathcal{X}_S$, 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

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

for all $z \in \bigcap_{S\in\Omega} \mathcal{X}_S$. Thus, the sum of the two nonnegative terms is zero, implying that each term is zero, that is, $\mathbf{E}[\|x - \Pi_{\mathcal{X}_S}(x)\|^2] = 0$ and therefore the set of fixed points of (3.5) are included into the set of minimizers of (3.7). ∎

Discussion. There is an interesting connection between reformulations (3.6) and (3.7). Indeed, for any nonempty closed convex set $\mathcal{X}_S$, the convex function

$$F_S(x) = \tfrac{1}{2}\|x - \Pi_{\mathcal{X}_S}(x)\|^2 = \min_{z\in\mathbb{R}^n} \mathbb{I}_{\mathcal{X}_S}(z) + \tfrac{1}{2}\|z - x\|^2 \tag{3.9}$$

is known as the Moreau approximation of the nonsmooth indicator function $\mathbb{I}_{\mathcal{X}_S}$. Thus, F_S has Lipschitz continuous gradient with constant $L_{F_S} = 1$ [23]. This implies that the function F has Lipschitz continuous gradient with constant $L_F = 1$. For the convex feasibility problem (3.4), with Ω finite, the following basic alternating projection algorithm has been extensively studied in the literature [1, 3, 23, 15]:

$$\texttt{B-AP}: \quad \text{choose } S_k \text{ cyclic/random and update } x^{k+1} = x^k - \alpha(x^k - \Pi_{\mathcal{X}_{S_k}}(x^k))$$

with the constant step size $\alpha \in (0, 2)$. First, based on our new reformulations (3.5)–(3.8), B-AP can be interpreted as stochastic gradient descent, a stochastic proximal point method, a stochastic fixed point method, or a stochastic projection method. For example, when solving the stochastic fixed point problem (3.5), we use the stochastic projection map $x \mapsto \Pi_{\mathcal{X}_S}(x)$ for a sample $S \sim \mathbf{P}$ followed by a relaxation step $(1 - \alpha)x + \alpha\Pi_{\mathcal{X}_S}(x)$, which leads to a random variant of B-AP. Second, since $f_S = \mathbb{I}_{\mathcal{X}_S}$ is nonsmooth, we apply a stochastic proximal point method to (3.6), i.e., $\Pi_{\mathcal{X}_S}(x) = \arg\min_z \mathbb{I}_{\mathcal{X}_S}(z) + \frac{1}{2}\|z - x\|^2$, followed by the previous relaxation step, thus obtaining again B-AP. Third, when solving the smooth stochastic optimization problem (3.7), $\min_x F(x) = \min_x \mathbf{E}[F_S(x)]$, we do not have access to the gradient of F, $\nabla F(x) = \mathbf{E}[\nabla F_S(x)] = \mathbf{E}[x - \Pi_{\mathcal{X}_S}(x)]$, and thus we repeatedly sample $S \sim \mathbf{P}$ and use an unbiased gradient sample $\nabla F_S(x) = x - \Pi_{\mathcal{X}_S}(x)$ in the stochastic gradient method with step size $\alpha \in (0, 2)$, leading again to B-AP. Fourth, when solving the stochastic

feasibility problem (3.8), which can be written equivalently as find $x \in \bigcap_{S\sim\mathbf{P}} \mathcal{X}_S$, we typically do not have explicit access to full stochastic intersection. Then, we sample $S \sim \mathbf{P}$ and perform a projection onto $\mathcal{X}_S$ followed by a relaxation step, leading to B-AP. However, in section 6 we propose a general algorithmic framework RPM for solving the four equivalent problems, which uses a minibatch of sampled sets instead of a single one and larger step size (extrapolation) than $\alpha \in (0, 2)$, leading, in general, to a faster scheme than B-AP.

3.3. Exactness. As Theorem 3.3 shows, the four equivalent stochastic reformulations are relaxations of the convex feasibility problem (3.4), that is, $\mathcal{X}' \subseteq \mathcal{X}''$. Thus, for any family of stochastic approximation sets over a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, denoted hereafter by $(\mathcal{X}_S)_{S\sim\mathbf{P}}$, we have the inclusions (3.3): $\mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{X}''$. Therefore, it is natural to investigate when these inclusions hold with equality. Note that when the original set $\mathcal{X}$ has the representation (3.4), i.e., $\mathcal{X} = \bigcap_{S\in\Omega} \mathcal{X}_S$, with Ω finite or countable, and the probability of choosing any $S \in \Omega$ is strictly positive, i.e., $p_S > 0$ for all $S \in \Omega$, then we have exactness, i.e., $\mathcal{X} = \mathcal{X}' = \mathcal{X}''$. Moreover, if $\mathcal{X} = \mathcal{X}''$, then the stochastic reformulations (3.5), (3.6), (3.7), and (3.8) are exact (i.e., equivalent to the convex feasibility problems (1.1) and (3.4)). However, this need not be the case, not without additional assumptions. To see this, consider $\mathcal{X} = \bigcap_{i=1}^m \mathcal{X}_i$, $\Omega = [m]$, $S = i$, and the probabilities $p_1 = 1$ and $p_2 = \cdots = p_m = 0$. Then, $(\mathcal{X}_i)_{i\sim\mathbf{P}}$ with the probability distribution $\mathbf{P}$ defined before constitutes a stochastic approximation of $\mathcal{X}$ as defined in Definition 3.2. However, $\mathcal{X}'' = \mathcal{X}_1$, which is not necessarily equal to $\mathcal{X}$. Therefore, no solution of the stochastic reformulations can guarantee bounds on the infeasibility of the original problem (1.1) without some additional assumptions. In view of the above discussion, we need to enforce a regularity assumption, which we call *exactness*.

Assumption 3.4 (exactness). Stochastic reformulations (3.5), (3.6), (3.7), and (3.8) of the convex feasibility problems (1.1) and (3.4) are exact. That is, $\mathcal{X} = \mathcal{X}' = \mathcal{X}''$.

In the next lemma we give a sufficient condition for exactness.

LEMMA 3.5. *Exactness holds, i.e., $\mathcal{X} = \mathcal{X}' = \mathcal{X}''$, provided that there exists $\mu > 0$ such that the following inequality (we call it the "stochastic linear regularity property") holds:*

$$\mu \cdot \operatorname{dist}^2_{\mathcal{X}}(x) \leq \mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right] \quad \forall x \in \mathbb{R}^n. \tag{3.10}$$

Proof. The set $\mathcal{X}''$ of optimal points of the stochastic smooth optimization problem (3.7) satisfies $F(x) = 0$ for all $x \in \mathcal{X}''$. Moreover, $F(x) = \mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right] = \mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right]$ holds. Therefore, for any $x \in \mathcal{X}''$ we have $\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right] = 0$. From (3.10) we conclude that $\operatorname{dist}^2_{\mathcal{X}}(x) = 0$, which means that $x \in \mathcal{X}$. Combined with (3.3), this implies that $\mathcal{X} = \mathcal{X}''$ holds. ∎

Since $\operatorname{dist}_{\mathcal{X}_S}(x) \leq \operatorname{dist}_{\mathcal{X}}(x)$ it follows immediately from (3.10) that $\mu \leq 1$. The feasibility problem is ill-conditioned when μ is very small. Moreover, linear regularity is related to Slater's condition, as can be seen from Example 3.6 below (see also the discussion from [2]). Linear regularity has been used frequently in the convex feasibility literature for proving linear convergence of projection methods [7, 2, 13, 23] or for dealing with uncountable intersection of convex sets [26, 27]. Notice that linear regularity can be a conservative condition for exactness and it does not hold for any collection of convex sets as the following example shows.

Example 3.6. Let $\mathcal{X}_1 = \{x : |x_1|^p \le x_2\}$ 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 $\text{dist}^2_{\mathcal{X}}(x) = x_1^2 + x_2^2$ and $\text{dist}^2_{\mathcal{X}_1}(x) + \text{dist}^2_{\mathcal{X}_2}(x) = x_2^2$. Clearly there is no $\mu > 0$ such that $\mu(x_1^2 + x_2^2) \le x_2^2$ for all $|x_1|^p = x_2,\ x_1 \ge 0$, since by replacing x_2 we obtain

$$\mu(x_1^2 + x_1^{2p}) \le x_1^{2p} \quad \Rightarrow \quad \mu\left(\frac{1}{x_1^{2p-2}} + 1\right) \le 1,$$

and we can take x_1 close to 0.

In the next section we derive some key properties of the stochastic reformulations (3.5)–(3.8).

4. Properties of F and Π. In convex optimization, the condition number of the problem, usually defined in terms of the Lipschitz and strong convexity constants, plays a key role in the development of first-order algorithms and in their convergence analysis [25]. In what follows, we combine the previous stochastic reformulations with certain regularity assumptions on the individual sets so that we extend the concept of condition number from convex optimization to convex feasibility. The conditioning parameters will play an important role in the development of a new projection method and in its convergence analysis.

4.1. Properties of objective function F. First, we concentrate on the reformulation of the feasibility problems (1.1) and (3.4) as the smooth stochastic optimization problem (3.7). Clearly, the objective function F of (3.7) is not strongly convex. However, the linear regularity condition (3.10) defined in terms of the sets $\mathcal{X}_S$ and $\mathcal{X}$ is equivalent to the quadratic functional growth condition on F introduced in [22], which was introduced as a relaxation of the strong convexity condition. Indeed, under exactness, we have $\mathcal{X} = \mathcal{X}'' = \arg\min_x F(x)$, the optimal value $F^* = 0$, and $F(x) = \mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right]$. Then, condition (3.10) can be rewritten equivalently as

$$F(x) - F^* \ge \frac{\mu}{2}\|x - \Pi_{\mathcal{X}}(x)\|^2 \quad \forall x \in \mathbb{R}^n, \tag{4.1}$$

which coincides with the definition of quadratic functional growth for F defined in [22]. Note that for proving linear convergence of first-order methods for smooth convex optimization we usually require strong convexity of the objective function, an assumption which does not hold for many applications, including problem (3.7). In [22] it was proved that several first-order methods can achieve linear convergence on nonstrongly convex optimization problems involving an objective function F with a Lipschitz continuous gradient and satisfying the relaxed strong convexity condition (4.1). This paper extends these results to the stochastic framework (see section 6), i.e., we prove in Theorem 6.3 that a stochastic projection scheme for (3.4) (or equivalently a mini-batch stochastic gradient descent for (3.7)) also achieves linear convergence under similar settings. In conclusion, μ from (3.10) plays the role of a relaxation of strong convexity constant for the objective function F. It remains to also identify an appropriate Lipschitz constant for the gradient of F. Clearly, F has Lipschitz continuous gradient with global constant $L_F = 1$. However, below we introduce a smaller Lipschitz constant for ∇F. Indeed, let $L \ge 0$ be the smallest constant satisfying

$$\|\mathbf{E}\left[x - \Pi_{\mathcal{X}_S}(x)\right]\|^2 \le L \cdot \mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right] \quad \forall x \in \mathbb{R}^n. \tag{4.2}$$

By Jensen's inequality, $L \le L_F = 1$. However, for specific sets and probability distributions $\mathbf{P}$, it is possible for L to be strictly smaller than 1; see the examples

below. The next theorem clearly shows that the linear regularity inequality (3.10) and the smooth regularity inequality (4.2) on the sets $(\mathcal{X}_S)_{S\sim\mathbf{P}}$ play a similar role to the strong convexity and Lipschitz gradient conditions for an objective function, respectively.

THEOREM 4.1. *Let the linear regularity condition* (3.10) *and the smooth condition* (4.2) *hold. Then, the following primal bounds are valid for the smooth objective function* F *with* $0 < \mu \leq L \leq 1$*:*

$$\frac{\mu}{2}\|x - \Pi_{\mathcal{X}}(x)\|^2 \leq F(x) - F^* \leq \frac{L}{2}\|x - \Pi_{\mathcal{X}}(x)\|^2 \quad \forall x \in \mathbb{R}^n. \tag{4.3}$$

Similarly, the following dual bounds $\frac{1}{2L}\|\nabla F(x)\|^2 \leq F(x) - F^* \leq \frac{1}{2\mu}\|\nabla F(x)\|^2$ *hold for all* $x \in \mathbb{R}^n$.

Proof. Under the linear regularity condition (3.10) we have (4.1), which represents the left-hand side inequality in (4.3). For proving the right-hand side in (4.3) we use a property of the projection [24]:

$$\|x - \Pi_{\mathcal{X}_S}(x)\|^2 \leq \|x - z\|^2 - \|\Pi_{\mathcal{X}_S}(x) - z\|^2 \quad \forall z \in \mathcal{X}_S. \tag{4.4}$$

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

$$\begin{aligned}
&\mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right] \\
&= \|x - \Pi_{\mathcal{X}}(x)\|^2 + \mathbf{E}\left[\|\Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}_S}(x)\|^2\right] + 2\langle x - \Pi_{\mathcal{X}}(x), \mathbf{E}\left[\Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}_S}(x)\right]\rangle \\
&\overset{(4.4)}{\leq} 2\|x - \Pi_{\mathcal{X}}(x)\|^2 - \mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right] + 2\langle x - \Pi_{\mathcal{X}}(x), \mathbf{E}\left[\Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}_S}(x)\right]\rangle \\
&= -\mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right] + 2\langle x - \Pi_{\mathcal{X}}(x), \mathbf{E}\left[x - \Pi_{\mathcal{X}_S}(x)\right]\rangle,
\end{aligned}$$

In conclusion, we get

$$\mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right] \leq \langle x - \Pi_{\mathcal{X}}(x), \mathbf{E}\left[x - \Pi_{\mathcal{X}_S}(x)\right]\rangle. \tag{4.5}$$

Furthermore, using the Cauchy–Schwartz inequality and (4.2) in (4.5) we get

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

which leads to $F(x) - F^* \leq \frac{L}{2}\|x - \Pi_{\mathcal{X}}(x)\|^2$, that is, the right-hand side inequality in (4.3) holds. This proves the first statement of the theorem, i.e., (4.3). For proving the second statement, i.e., the dual bounds, we first notice that since the Jensen-type inequality (4.2) always holds for some $L \leq 1$ and using the expression of F and that $F^* = 0$, we can easily find the left-hand side inequality for the dual bounds:

$$\frac{1}{2}\|\nabla F(x)\|^2 \leq L(F(x) - F^*) \quad \forall x \in \mathbb{R}^n. \tag{4.6}$$

Then, combining (3.10) and (4.5) and using the Cauchy–Schwartz inequality, we get

$$\begin{aligned}
\mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right] &\leq \|x - \Pi_{\mathcal{X}}(x)\|\|\mathbf{E}\left[x - \Pi_{\mathcal{X}_S}(x)\right]\| \\
&\leq \sqrt{\mu^{-1}\mathbf{E}\left[\|x - \Pi_{\mathcal{X}_S}(x)\|^2\right]}\|\mathbf{E}\left[x - \Pi_{\mathcal{X}_S}(x)\right]\|,
\end{aligned}$$

which leads to $\mu(F(x) - F^*) \leq \frac{1}{2}\|\nabla F(x)\|^2$. Combining the previous inequality with (4.6) we obtain the second statement of the theorem, i.e., the dual bounds. ∎

Theorem 4.1 states that F is strongly convex with constant μ and has Lipschitz continuous gradient with constant L when restricted along any segment $[x, \Pi_{\mathcal{X}}(x)]$. Indeed, since $\nabla F(\Pi_{\mathcal{X}}(x)) = 0$, from (4.3) we obtain for all $x \in \mathbb{R}^n$ that

$$\frac{\mu}{2}\|x - \Pi_{\mathcal{X}}(x)\|^2 + \langle \nabla F(\Pi_{\mathcal{X}}(x)), x - \Pi_{\mathcal{X}}(x)\rangle + F^* \\ \leq F(x) \leq \frac{L}{2}\|x - \Pi_{\mathcal{X}}(x)\|^2 + \langle \nabla F(\Pi_{\mathcal{X}}(x)), x - \Pi_{\mathcal{X}}(x)\rangle + F^*,$$

which are exactly the strong convexity condition on F and the Lipschitz continuity condition on ∇F, respectively, along any segment $[x, \Pi_{\mathcal{X}}(x)]$; see [22, 25] for more details. It follows that $0 \leq \mu \leq L \leq 1$ and the ratio $\texttt{cond} = \mu/L$ represents the condition number of the stochastic optimization problem (3.7), or of the convex feasibility problem (3.4). *To the best of our knowledge, Theorem* 4.1 *is the first result extending the concept of condition number from convex optimization to the convex feasibility problem* (3.4). Next, we provide some nontrivial examples for which our new Lipschitz constant L satisfies $L < L_F = 1$ (the global Lipschitz constant of the gradient ∇F).

Linear equalities. Let us 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 approximation sets $\mathcal{X}_S = \{x : S^T Ax = S^T b\}$ taking any matrix $S \in \mathbb{R}^{m\times q}$ (for any S we have $\mathcal{X} \subseteq \mathcal{X}_S$). Then, we get the following characterization for L.

THEOREM 4.2. *Let us consider the linear system $\mathcal{X} = \{x : Ax = b\}$, where $A \in \mathbb{R}^{m\times n}$. Further, take 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,* (4.2) *holds with*

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

Proof. Clearly, for x satisfying $Ax = b$ the inequality (4.2) holds for any $L \leq 1$. It remains to prove for x satisfying $Ax - b \neq 0$. However, since $\mathcal{X}_S = \{x : S^T Ax = S^T b\}$, the projection of x onto $\mathcal{X}_S$ can be computed explicitly via the pseudoinverse: $\Pi_{\mathcal{X}_S}(x) = x - A^T S(S^T AA^T S)^\dagger S^T(Ax - b)$ and the relation we need to prove becomes

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

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

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

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

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

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

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

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 obtain

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

which proves the statement of the theorem. ∎

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

COROLLARY 4.3. *Let us consider the linear system* $\mathcal{X} = \{x : \ Ax = b\}$, *where* $A \in \mathbb{R}^{m\times n}$ *has* $\operatorname{rank}(A) \geq 2$ *and* $D = \operatorname{diag}(\|A_1\|^{-2}, \ldots, \|A_m\|^{-2})$. *Further, let us consider* $\Omega = \{e_1, \ldots, e_m\}$, *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* (4.2) *holds with*

$$1 > L = \begin{cases} \frac{\lambda_{\max}(A^T A)}{\|A\|_F^2} & \text{if} \quad \mathbf{P}(S = e_i) = \frac{\|A_i\|^2}{\|A\|_F^2}, \\ \frac{\lambda_{\max}(A^T DA)}{m} & \text{if} \quad \mathbf{P}(S = e_i) = \frac{1}{m}. \end{cases} \tag{4.7}$$

Proof. In this case we have $S(S^T AA^T S)^\dagger S^T = e_i(e_i^T AA^T e_i)^\dagger e_i^T = \|A_i\|^{-2} e_i e_i^T$. Then, from Theorem 4.2 we get for the probability $\mathbf{P}(S = e_i) = \|A_i\|^2/\|A\|_F^2$,

$$\begin{aligned} L = \lambda_{\max}\left(A^T \mathbf{E}\left[\frac{1}{\|A_i\|^2} e_i e_i^T\right] A\right) &= \lambda_{\max}\left(A^T \sum_{i=1}^m \frac{\|A_i\|^2}{\|A\|_F^2} \frac{1}{\|A_i\|^2} e_i e_i^T A\right) \\ &= \lambda_{\max}\left(\frac{AA^T}{\|A\|_F^2}\right). \end{aligned}$$

In the last equality we used that the maximum eigenvalues of $A^T A$ and AA^T coincide. Moreover, the trace of the matrix $\frac{AA^T}{\|A\|_F^2}$ is equal to 1 and thus

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

Therefore, if $\operatorname{rank}(A) \geq 2$, then $L = \lambda_{\max}(\frac{AA^T}{\|A\|_F^2}) < 1$.

Similarly, from Theorem 4.2 we obtain for the uniform probability $\mathbf{P}(S = e_i) = \frac{1}{m}$,

$$L = \lambda_{\max}\left(A^T \sum_{i=1}^m \frac{1}{m} \frac{1}{\|A_i\|^2} e_i e_i^T A\right) = \lambda_{\max}\left(\frac{A^T DA}{m}\right) = \lambda_{\max}\left(\frac{AA^T D}{m}\right),$$

where we used 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, $U = A^T D$ and $V = A$. Moreover, the trace of the matrix $\frac{AA^T D}{m}$ is equal to 1 and thus

$$\sum_{i=1}^m \lambda_i\left(\frac{AA^T D}{m}\right) = \operatorname{Trace}\left(\frac{AA^T D}{m}\right) = 1.$$

If $\operatorname{rank}(A) \geq 2$, then $L = \lambda_{\max}(\frac{AA^T D}{m}) < 1$ also for the uniform probability distribution. ∎

Linear inequalities. 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 approximation sets $\mathcal{X}_S = \{x : \ S^T Ax \leq S^T b\}$, 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 $\mathcal{X} \subseteq \mathcal{X}_S$. Moreover, $\mathcal{X}_S = \mathbb{R}^n$ provided that $S^T A = 0$ (when $A^T S = 0$ we use the convention $A^T S/\|A^T S\|^2 = 0$). Then, we have the following characterization for L.

THEOREM 4.4. *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 $\mathcal{X}_S = \{x : \ S^T Ax \leq S^T b\}$, where $S \in \Omega = \mathbb{R}^m_+$, and a probability distribution $\mathbf{P}$ on Ω. Then,* (4.2) *holds with*

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

Proof. Clearly, for x satisfying $Ax \leq b$ the inequality (4.2) holds for any $L \leq 1$. It remains to prove for x satisfying $Ax \not\leq b$. However, since $\mathcal{X}_S = \{x : \ S^T Ax \leq S^T b\}$, the projection of x onto $\mathcal{X}_S$ can be computed explicitly as

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

where $\Pi_+(v) = \max(0, v)$, and the relation we need to prove becomes

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

or equivalently

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

Further, if we define the event $\mathcal{I}(x) = \{S \in \Omega : \ S^T(Ax-b) > 0\}$, and $E(x) = \int_{\mathcal{I}(x)} S(S^T AA^T S)^{-1} S^T d\mathbf{P}$ and $E = \int_\Omega S(S^T AA^T S)^{-1} S^T d\mathbf{P}$, then the previous relation can be written as follows:

$$\left\|A^T E(x)(Ax-b)\right\|^2 \leq L(Ax-b)^T E(x)(Ax-b).$$

Note that both matrices $E(x)$ and E are positive semidefinite and $E(x) \preceq E$ for all x such that $Ax \not\leq b$. It follows that L is an upper bound on the following function:

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

However, it is easy to find an upper bound for this function $\mathcal{R}(x)$ for each fixed x, namely, $\mathcal{R}(x) \leq \lambda_{\max}(A^T E(x)A)$ for all $x : \ Ax \not\leq b$. Since $E(x) \preceq E$, we have $A^T E(x)A \preceq A^T EA$ and consequently $\lambda_{\max}(A^T E(x)A) \leq \lambda_{\max}(A^T EA)$. Moreover, there exists x such that $\mathcal{I}(x) = \Omega$. Thus, we have

$$L = \lambda_{\max}(A^T EA) = \lambda_{\max}\left(A^T\mathbf{E}\left[S(S^T AA^T S)^{-1} S^T\right] A\right).$$

Using now Jensen's inequality for the convex function $\lambda_{\max}(\cdot)$, 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 $P_S = A^TS(S^TAA^TS)^{-1}S^TA$ satisfies $P_S^2 = P_S$, so that all the eigenvalues of P_S are either 0 or 1. Then, we get

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

which proves the statement of the theorem. ∎

From the previous theorem it follows that for particular choices of the probability $\mathbf{P}$ we have $L < 1$; see, e.g., the next corollary, whose proof is similar to that of Corollary 4.3.

COROLLARY 4.5. *Let us consider the system of linear inequalities*

$$\mathcal{X} = \{x : \ Ax \leq b\},$$

where $A \in \mathbb{R}^{m\times n}$ *has* $\operatorname{rank}(A) \geq 2$ *and* $D = \operatorname{diag}(\|A_1\|^{-2}, \ldots, \|A_m\|^{-2})$. *Further, let us consider* $\Omega = \{e_1, \ldots, e_m\}$, *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, (4.2) *holds with*

$$1 > L = \begin{cases} \frac{\lambda_{\max}\left(A^TA\right)}{\|A\|_F^2} & \text{if} \quad \mathbf{P}(S = e_i) = \frac{\|A_i\|^2}{\|A\|_F^2}, \\ \frac{\lambda_{\max}\left(A^TDA\right)}{m} & \text{if} \quad \mathbf{P}(S = e_i) = \frac{1}{m}. \end{cases} \tag{4.8}$$

The reader may find other examples of convex feasibility problems with $L < 1$; we believe that this paper opens a window of opportunity for theoretical research and applications related to convex feasibility.

4.2. Properties of operator $\mathbf{\Pi = E\left[\Pi_{\mathcal{X}_S}\right]}$. It is well known that the projection operator is firmly nonexpansive [24]:

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

Then, the average projection operator $\Pi(x) = \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]$ is also firmly nonexpansive:

$$\begin{aligned}\langle \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right] - \mathbf{E}\left[\Pi_{\mathcal{X}_S}(y)\right], x - y\rangle &\geq \mathbf{E}\left[\|\Pi_{\mathcal{X}_S}(x) - \Pi_{\mathcal{X}_S}(y)\|^2\right] \\ &\geq \|\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right] - \mathbf{E}\left[\Pi_{\mathcal{X}_S}(y)\right]\|^2 \quad \forall x, y \in \mathbb{R}^n.\end{aligned}$$

Similar to Theorem 4.1, we can also derive contraction inequalities for the average projection operator Π.

THEOREM 4.6. *Under the linear regularity condition* (3.10) *and the smooth condition* (4.2) *the following bounds hold for the average projection operator* $\Pi(x) = \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right]$ *with fixed point* $x^* = \Pi_{\mathcal{X}}(x)$*:*

$$(1 - L)\|x - x^*\|^2 \leq \langle \Pi(x) - \Pi(x^*), x - x^*\rangle \leq (1 - \mu)\|x - x^*\|^2 \quad \forall x \in \mathbb{R}^n. \tag{4.9}$$

Proof. In order to prove the right-hand side inequality, we choose in the nonexpansive inequality of the operator Π 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 (3.10) leads to

$$\langle \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right] - \Pi_{\mathcal{X}}(x), x - \Pi_{\mathcal{X}}(x)\rangle \;\leq\; \|x - \Pi_{\mathcal{X}}(x)\|^2 - \mathbf{E}\left[\|\Pi_{\mathcal{X}_S}(x) - x\|^2\right] \overset{(3.10)}{\leq} (1-\mu)\,\|x - \Pi_{\mathcal{X}}(x)\|^2.$$

For the left-hand side inequality we proceed as follows:

$$\begin{aligned}
&\langle \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right] - \Pi_{\mathcal{X}}(x), x - \Pi_{\mathcal{X}}(x)\rangle = \|x - \Pi_{\mathcal{X}}(x)\|^2 + \langle \mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right] - x, x - \Pi_{\mathcal{X}}(x)\rangle \\
&\geq\; \|x - \Pi_{\mathcal{X}}(x)\|^2 - \|x - \Pi_{\mathcal{X}}(x)\|\|\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x)\right] - x\| \\
&\geq\; \|x - \Pi_{\mathcal{X}}(x)\|^2 - \|x - \Pi_{\mathcal{X}}(x)\|\sqrt{L\mathbf{E}\left[\|\Pi_{\mathcal{X}_S}(x) - x\|^2\right]} \\
&=\; \|x - \Pi_{\mathcal{X}}(x)\|^2 - \|x - \Pi_{\mathcal{X}}(x)\|\sqrt{2L(F(x) - F^*)} \\
&\overset{(4.3)}{\geq} \|x - \Pi_{\mathcal{X}}(x)\|^2 - \|x - \Pi_{\mathcal{X}}(x)\|L\sqrt{\|x - \Pi_{\mathcal{X}}(x)\|^2} = (1-L)\|x - \Pi_{\mathcal{X}}(x)\|^2,
\end{aligned}$$

where in the first inequality we used the Cauchy–Schwartz inequality, in the second inequality we used (4.2), and in the third inequality we used (4.3). ∎

Thus, from an operator theory perspective Theorem 4.6 shows that the average projection operator Π is a contraction with contraction constant $1 - \mu < 1$ when restricted along any segment $[x, \Pi_{\mathcal{X}}(x)]$.

5. Examples of stochastic approximations. In this section we provide concrete examples of stochastic approximations of sets, considering both scenarios: finite intersection and infinite intersection.

5.1. Examples: Finite intersection. We consider the set $\mathcal{X}$ represented as the intersection of a finite family of convex sets:

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

where $\mathcal{X}_i$ are nonempty closed convex sets. We also assume that $\mathcal{X} \neq \emptyset$. In several papers, such as [23, 24, 3, 2], a *linear regularity property* for $\mathcal{X} = \bigcap_{i=1}^m \mathcal{X}_i$ was introduced stating that there exists $\mu_{\max} > 0$ such that

$$\mu_{\max} \cdot \operatorname{dist}^2_{\mathcal{X}}(x) \leq \max_{i\in[m]} \operatorname{dist}^2_{\mathcal{X}_i}(x) \quad \forall x \in \mathbb{R}^n. \tag{5.1}$$

Based on this condition, a linear convergence rate, depending on the constant $\mu_{\max}$, has been derived for the basic alternating projection algorithm `B-AP` with constant step size $\alpha \in (0, 2)$; see [23, 24, 3, 2]. Note that our definition of linear regularity (3.10) extends the one given in (5.1) for finite intersection to the more general convex feasibility problem (3.4). More precisely, in order to show linear convergence for our general algorithmic framework introduced in this paper, we require the linear regularity property (3.10) for the set $\mathcal{X} = \bigcap_{S\in\Omega} \mathcal{X}_S$. For a uniform probability over the set $\Omega = [m] \overset{\text{def}}{=} \{1, 2, \ldots, m\}$ we have

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

This shows that our constant μ satisfies $\mu \leq \mu_{\max}/m$. Further, we analyze (3.10) and estimate the constant μ for several representative cases of stochastic approximation sets for $\mathcal{X}$.

5.1.1. Standard. Let $\mathcal{X}_S = \mathcal{X}_i$ for all $i \in \Omega = [m]$, endowed with some probability $p_i \geq 0$. Since $\bigcap_{i=1}^m \mathcal{X}_i = \mathcal{X} \subseteq \mathcal{X}_S$, we get a stochastic approximation of $\mathcal{X}$ as in Definition 3.2 and the optimal set is

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

Hence, a sufficient condition for exactness, i.e., $\mathcal{X} = \mathcal{X}''$, is to require $p_i > 0$ for all $i \in [m]$. Moreover, under this condition and (5.1), the linear regularity (3.10) holds with $\mu = \mu_{\max} p_{\min}$, where $p_{\min} = \min_{i \in [m]} p_i$. Indeed, it follows from

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

5.1.2. 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 $\mathcal{X}_S = \bigcap_{i \in S} \mathcal{X}_i$ with probability p_S. Since $\mathcal{X} \subseteq \mathcal{X}_S$, we get a stochastic approximation. Moreover,

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

A sufficient condition for the last set to be equal to $\mathcal{X}$, i.e., 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 the standard case. Moreover, under this condition and (5.1) it follows that linear regularity (3.10) holds with $\mu = \mu_{\max} p_{\min}$, where $p_{\min} = \min_{S: p_S > 0} p_S$. This is due to the fact that $\max_{i \in [m]} \operatorname{dist}^2_{\mathcal{X}_i}(x) \leq \sum_{i=1}^m \operatorname{dist}^2_{\mathcal{X}_i}(x)$, that $\operatorname{dist}^2_{\mathcal{X}_i}(x) \leq \operatorname{dist}^2_{\bigcap_{j \in S} \mathcal{X}_j}(x) = \operatorname{dist}^2_{\mathcal{X}_S}(x)$ for all $i \in S$, and assuming a collection of subsets S covering $[m]$.

5.1.3. Convex combination. Let $\|x\|_0$ denote the 0-norm of vector x (its number of nonzeros). Fix $r \in [m]$, and let us consider a countable subset Ω_r defined as

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

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

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

This is clearly a stochastic approximation, that is, $\mathcal{X} \subseteq \mathcal{X}_S$, since $\sum_{i=1}^m S_i = 1$ and $S \geq 0$ (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 5.1.1. If, additionally, we assume that Ω_r contains the basic vectors, i.e., $\{e_1, \ldots, e_m\} \subseteq \Omega_r$, and $\mathcal{X}_S$ is defined as above, then exactness holds when $p_i = \mathbf{P}(S = e_i) > 0$ for all $i \in [m]$. Indeed, if $x \in \mathcal{X}''$, then

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

which implies $x \in \mathcal{X}_i$, provided that $p_i > 0$, for all $i \in [m]$. Moreover, under this condition and (5.1) it follows that linear regularity (3.10) holds with $\mu = \mu_{\max} p_{\min}$, 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

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

5.1.4. 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 $A \in \mathbb{R}^{m\times n}$. Let $q \le m$, $\Omega \subseteq \mathbb{R}^{m\times q}$, and let $\mathbf{P}$ be a probability distribution 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 $\bigcap_{S\in\Omega} \mathcal{X}_S = \{x : \ SAx = Sb \ \forall S \in \Omega\}$. If we can find m linearly independent columns in the family of matrices $(S)_{S\in\Omega}$, then $\mathcal{X} = \bigcap_{S\in\Omega} \mathcal{X}_S$. Next we derive sufficient conditions for exactness and an estimate for μ.

THEOREM 5.1. *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* (3.10) *holds with constant*

$$\mu = \lambda^{\mathrm{nz}}_{\min}(A^T \mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right] A) > 0. \tag{5.2}$$

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

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

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

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

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 the expectation in both sides of $\operatorname{dist}_{\mathcal{X}_S}$ and also using the previous property of P_S, we further obtain

$$\begin{aligned} \mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right] &= \mathbf{E}\left[\|P_S(x - \Pi_{\mathcal{X}}(x))\|^2\right] \mathbf{E}\left[(x - \Pi_{\mathcal{X}}(x))^T P_S (x - \Pi_{\mathcal{X}}(x))\right] \\ &= (x - \Pi_{\mathcal{X}}(x))^T \mathbf{E}\left[P_S\right](x - \Pi_{\mathcal{X}}(x)). \end{aligned} \tag{5.3}$$

On the other hand, it is well known from the Courant–Fischer theorem that for any $C \in \mathbb{R}^{m\times n}$ we have $\|Cz\| \ge \sigma^{\mathrm{nz}}_{\min}(C)\|z\|$ for all $z \in \operatorname{Im}(C^T)$, where we recall that $\sigma^{\mathrm{nz}}_{\min}$ 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

$$\|E^{1/2}Az\| \ge \sigma^{\mathrm{nz}}_{\min}(E^{1/2}A)\|z\| \quad \forall z \in \operatorname{Im}(A^T E^{1/2}). \tag{5.4}$$

Further, since we assume $E = \mathbf{E}\left[S(S^T AA^T S)^\dagger S^T\right] \succ 0$, we have $E^{1/2} \succ 0$ and $\operatorname{Im}(A^T) = \operatorname{Im}(A^T E^{1/2})$. Moreover, we have the fact that $x - \Pi_{\mathcal{X}}(x) \in \operatorname{Im}(A^T)$.

Therefore, by applying the relation (5.4) for $z = x - \Pi_{\mathcal{X}}(x)$, observing that $\mathbf{E}[P_S] = A^T E A$, and by combining relations (5.3) and (5.4), we have

$$\begin{aligned}\mathbf{E}\left[\text{dist}^2_{\mathcal{X}_S}(x)\right] &= \|E^{1/2}A(x - \Pi_{\mathcal{X}}(x))\|^2 \overset{(5.4)}{\geq} \left(\sigma^{\text{nz}}_{\min}(E^{1/2}A)\right)^2 \text{dist}^2_{\mathcal{X}}(x) \\ &= \lambda^{\text{nz}}_{\min}(A^T E A)\text{dist}^2_{\mathcal{X}}(x) = \lambda^{\text{nz}}_{\min}(\mathbf{E}[P_S])\text{dist}^2_{\mathcal{X}}(x) \\ &= \lambda^{\text{nz}}_{\min}(\mathbf{E}\left[A^T S(S^T A A^T S)^\dagger S^T A\right])\text{dist}^2_{\mathcal{X}}(x) \quad \forall x \in \mathbb{R}^n.\end{aligned}$$

This final relation implies our statement. ∎

In [16] it has been proved that, when we consider discrete samplings, such as $S \in \Omega = \{e^1, \ldots, e^m\}$, and full row rank matrices A, the matrix $\mathbf{E}\left[S(S^T A A^T S)^\dagger S^T\right]$ is positive definite, that is, it satisfies our assumption considered in the previous theorem. A simple consequence of the previous theorem is the following corollary.

COROLLARY 5.2. *If $\Omega = \{e_1, \ldots, e_m\}$ and $D = \text{diag}(\|A_1\|^{-2}, \ldots, \|A_m\|^{-2})$, then for two choices of the probability $\mathbf{P}$ on Ω the linear regularity constant takes the form*

$$(0,1] \ni \mu = \begin{cases} \frac{\lambda^{\text{nz}}_{\min}(A^T A)}{\|A\|_F^2} & \text{if} \quad \mathbf{P}(S = e_i) = \frac{\|A_i\|^2}{\|A\|_F^2}, \\ \frac{\lambda^{\text{nz}}_{\min}(A^T D A)}{m} & \text{if} \quad \mathbf{P}(S = e_i) = \frac{1}{m}. \end{cases} \tag{5.5}$$

Proof. If $\Omega = \{e_1, \ldots, e_m\}$ and the probability satisfies $\mathbf{P}(S = e_i) = \|A_i\|^2/\|A\|_F^2$, 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 (5.2) becomes

$$\begin{aligned}&\lambda^{\text{nz}}_{\min}(A^T \mathbf{E}\left[S(S^T A A^T S)^\dagger S^T\right] A) = \lambda^{\text{nz}}_{\min}\left(A^T \mathbf{E}\left[\frac{e_i e_i^T}{\|A_i\|^2}\right] A\right) \\ &= \lambda^{\text{nz}}_{\min}\left(A^T \sum_{i=1}^m \frac{\|A_i\|^2}{\|A\|_F^2} \frac{e_i e_i^T}{\|A_i\|^2} A\right) = \lambda^{\text{nz}}_{\min}\left(A^T \frac{I_m}{\|A\|_F^2} A\right) = \frac{\lambda^{\text{nz}}_{\min}(A^T A)}{\|A\|_F^2}.\end{aligned}$$

For the uniform probability $\mathbf{P}(S = e_i) = 1/m$, the expression in (5.2) becomes

$$\lambda^{\text{nz}}_{\min}(A^T \mathbf{E}\left[S(S^T A A^T S)^\dagger S^T\right] A) = \lambda^{\text{nz}}_{\min}\left(A^T \sum_{i=1}^m \frac{1}{m} \frac{e_i e_i^T}{\|A_i\|^2} A\right) = \frac{\lambda^{\text{nz}}_{\min}(A^T D A)}{m},$$

where $D = \text{diag}(\|A_1\|^{-2}, \ldots, \|A_m\|^{-2})$. These prove our statements. ∎

5.1.5. Inequality constraints. Let $q \leq m$, $\Omega \subseteq \mathbb{R}_+^{m \times q}$, where $\mathbb{R}_+^{m \times q}$ is 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 let $\mathbf{P}$ be a probability distribution 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, $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_m)$. 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 $\bigcap_{S \in \Omega} \mathcal{X}_S = \{x : \ S^T \mathcal{F}(x) \leq 0 \ \forall S \in \Omega\}$. If there exist m linearly independent columns in the family of matrices $(S)_{S \in \Omega}$, then $\mathcal{X} = \bigcap_S \mathcal{X}_S$. Next, we provide estimates for the linear regularity constant μ for some particular sets. First, we consider finding a point in the intersection of a finite number of half-spaces $\mathcal{X} = \{x \in \mathbb{R}^n : Ax \leq b\}$.

THEOREM 5.3. *Let $\mathcal{X} = \{x \in \mathbb{R}^n : \ Ax \leq b\}$ and consider stochastic approximation half-spaces $\mathcal{X}_S = \{x \in \mathbb{R}^n : \ S^T Ax \leq S^T b\}$, 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 $\mathcal{X}$ by $\tilde{\mu}$ [7, 22]. Then, under exactness the linear regularity property* (3.10) *holds with constant*

$$\mu = \tilde{\mu} \left(\max_{S \in \Omega_r} \|A^T S\|^2 \right)^{-1} \left(\min_{S \in \Omega_r} p_S \right). \tag{5.6}$$

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

$$\operatorname{dist}_{\mathcal{X}_S}(x) = \frac{\Pi_+(S^T Ax - S^T b)}{\|A^T S\|} \geq \frac{\Pi_+(S^T Ax - S^T b)}{\max\limits_{S \in \Omega_r} \|A^T S\|}. \tag{5.7}$$

From the Markov inequality we have

$$\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right] \left(\max_{S \in \Omega_r} \operatorname{dist}^2_{\mathcal{X}_S}(x) \right)^{-1} \geq \mathbf{P}\left(\operatorname{dist}^2_{\mathcal{X}_S}(x) \geq \max_{S \in \Omega_r} \operatorname{dist}^2_{\mathcal{X}_S}(x) \right).$$

Combining the previous inequality with (5.7), we obtain

$$\begin{aligned} \mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right] &\geq \mathbf{P}\left(\operatorname{dist}^2_{\mathcal{X}_S}(x) \geq \max_{S \in \Omega_r} \operatorname{dist}^2_{\mathcal{X}_S}(x) \right) \cdot \max_{S \in \Omega_r} \operatorname{dist}^2_{\mathcal{X}_S}(x) \\ &\geq \min_{S \in \Omega_r} \ p_S \cdot \max_{S \in \Omega_r} \operatorname{dist}^2_{\mathcal{X}_S}(x) \overset{(5.7)}{\geq} \min_{S \in \Omega_r} \ p_S \cdot \frac{\max\limits_{S \in \Omega_r} \Pi^2_+(S^T Ax - S^T b)}{\max\limits_{S \in \Omega_r} \|A^T S\|^2}. \end{aligned} \tag{5.8}$$

On the other hand, for a polyhedral set the Hoffman inequality is valid (see [7, 22]). Since we assume exactness and Ω_r has a finite number of elements, there exists some positive Hoffman constant $\tilde{\mu} > 0$ such that the Hoffman inequality holds:

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

Using the previous Hoffman inequality in (5.8) leads to our statement. ∎

However, for a specific choice of the probability we can get a better estimate for μ.

COROLLARY 5.4. *Under the same assumptions as in Theorem* 5.3 *but with the particular probability distribution $\mathbf{P} = (p_S)_{S \in \Omega_r}$ given by*

$$p_S = \|A^T S\|^2 / \sum_{S \in \Omega_r} \|A^T S\|^2,$$

the linear regularity property (3.10) *holds with constant*

$$\mu = \tilde{\mu} \left(\sum_{S \in \Omega_r} \|A^T S\|^2 \right)^{-1}. \tag{5.9}$$

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

$$\operatorname{dist}_{\mathcal{X}_S}(x) = \Pi_+(S^TAx - S^Tb)\|A^TS\|^{-1}.$$

Using the expressions for the distance and probability, we have

$$\begin{aligned}\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right] &= \sum_{S\in\Omega_r} p_S \operatorname{dist}^2_{\mathcal{X}_S}(x) = \sum_{S\in\Omega_r} \frac{\|A^TS\|^2}{\sum_{S\in\Omega_r}\|A^TS\|^2}\cdot\frac{\Pi^2_+(S^TAx - S^Tb)}{\|A^TS\|^2} \\ &= \frac{1}{\sum_{S\in\Omega_r}\|A^TS\|^2}\sum_{S\in\Omega_r}\Pi^2_+(S^TAx - S^Tb). \qquad (5.10)\end{aligned}$$

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

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

Using the previous Hoffman inequality in (5.10) leads to our statement. ∎

Finally, following similar ideas to those in [23, 15], 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

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

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 $\mathbf{P} = (p_S)_{S\in\Omega}$ (e.g., a subset of matrices with nonnegative entries). Then, under exactness the linear regularity property (3.10) holds over any compact set Q with constant [23]

$$\mu = \delta^2\Big(\min_{S\in\Omega} p_S\Big)\Big(\max_{x\in Q}\|x - \bar{x}\|^2\Big)^{-1} \quad \forall x\in Q. \qquad (5.11)$$

5.2. Examples: Infinite intersection. Assume now that $\mathcal{X} = \bigcap_{S\in\Omega}\mathcal{X}_S$ for some (possibly infinite) index set Ω and the sets $\mathcal{X}_S \subseteq \mathbb{R}^n$. Many interesting applications can be modeled as the intersection of an infinite (countable/uncountable) number of simple convex sets; see, e.g., [26, 27] for some control and machine learning applications. Let $\mathbf{P}$ be a probability measure on Ω. Then, if we choose $S \sim \mathbf{P}$, $(\mathcal{X}_S)_{S\sim\mathbf{P}}$ is a stochastic approximation of $\mathcal{X}$. Recall that $\mathcal{X}'' = \{x :\ \mathbf{P}(x\in\mathcal{X}_S) = 1\}$.

5.2.1. 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 $\langle g, z - S\rangle \le 0$ for all $z\in\mathcal{X}$. If we let

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

then clearly $\mathcal{X}\subseteq\mathcal{X}_S$ for all $S\in\mathbb{R}^n$. Given any probability $\mathbf{P}$ over $\Omega = \mathbb{R}^n$, we get a stochastic approximation of $\mathcal{X}$. In this case we can guarantee $\mathcal{X}\subseteq\bigcap_{S\in\mathbb{R}^n}\mathcal{X}_S$.

5.2.2. Supporting half-spaces. A particular case of the convex feasibility problem is the so-called split feasibility problem [8]:

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

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 half-space

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

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 half-space $\mathcal{X}_S = \{x : \ c_S^T x \le b_S\}$ can be written equivalently as

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

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, $\mathcal{X} \subseteq \mathcal{X}_S$ for all $S \in \mathbb{R}^n$. Therefore, given any distribution $\mathbf{P}$ over $\mathbb{R}^n$, this construction forms a stochastic approximation of $\mathcal{X}$. In fact, in this case we have $\mathcal{X} = \bigcap_{S\in\mathbb{R}^n} \mathcal{X}_S$. Indeed, it is straightforward that we have $\mathcal{X} \subseteq \bigcap_{S\in\mathbb{R}^n} \mathcal{X}_S$. For the other inclusion, let us take any $x \in \bigcap_{S\in\mathbb{R}^n} \mathcal{X}_S$. 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 $\langle Ax - \Pi_{\mathcal{Z}}(Ax), Ax - \Pi_{\mathcal{Z}}(Ax)\rangle \le 0$, which holds if and only if $Ax = \Pi_{\mathcal{Z}}(Ax)$, that is, $x \in \mathcal{X}$.

5.2.3. 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 $\bar{x} \in \Omega$:

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

where $\mathcal{X}_S = \{x : \ (x - \bar{x})^T(S - \bar{x}) \le 0\}$. If $\mathbf{P}$ is a probability distribution over Ω, and we sample $S \sim \mathbf{P}$, then we get a stochastic approximation of $\mathcal{X}$ according to our Definition 3.2. Moreover, we have the relation $\mathcal{X} = \bigcap_{S\in\Omega} \mathcal{X}_S$.

For feasibility problems with an infinite (uncountable) intersection of sets it is difficult to derive conditions for exactness. We leave this issue for future work. However, motivated by these finite/infinite convex feasibility examples, we introduce next a random projection algorithmic framework and provide a detailed convergence analysis for it.

6. Randomized projection method. In this section we propose the following general randomized minibatch projection algorithm with an extrapolated step size.

Algorithm `RPM` **(general case).**

Choose $x^0 \in \mathbb{R}^n$, minibatch size $N \ge 1$, and extrapolated step sizes $\{\alpha_k\}_{k\ge 0}$.
For $k \ge 0$ repeat:

1. Draw N independent samples, $S_1^k, \ldots, S_N^k \sim \mathbf{P}$.
2. Compute $x^{k+1} = x^k + \alpha_k \left(\dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \Pi_{\mathcal{X}_{S_i^k}}(x^k) - x^k \right)$.

Based on our new reformulations (3.5)–(3.8), RPM can be interpreted as a minibatch stochastic proximal point, a minibatch stochastic gradient descent, a minibatch stochastic fixed point method, or a minibatch stochastic projection method followed by a relaxation (extrapolated) step. Indeed, we usually do not have explicit access to operators or functions involving expectation $\mathbf{E}[\cdot]$ or stochastic intersection $x \in \bigcap_{S\sim\mathbf{P}} \mathcal{X}_S$ (i.e., $\mathbf{P}(x \in \mathcal{X}_S) = 1$). Therefore, once we encounter one of these situations, instead of evaluation of the expectation or full stochastic intersection, we repeatedly draw at each iteration a minibatch of samples $S_1, \ldots, S_N \sim \mathbf{P}$ and use the corresponding stochastic map with respect to the chosen samples followed by a relaxation step with relaxation parameter α. For example, first, when solving the stochastic fixed point problem (3.5), we use the minibatch stochastic projection map $x \mapsto 1/N \sum_{i=1}^N \Pi_{\mathcal{X}_{S_i}}(x)$ for a minibatch of samples $S_1, \ldots, S_N$ and then we perform a relaxation step $(1-\alpha)x + \alpha/N \sum_{i=1}^N \Pi_{\mathcal{X}_{S_i}}(x)$, which leads to RPM. Second, since $f_S = \mathbb{I}_{\mathcal{X}_S}$ is nonsmooth, we apply the minibatch stochastic proximal point method on the stochastic optimization problem (3.6), i.e., $\Pi_{\mathcal{X}_{S_i}}(x) = \arg\min_z \mathbb{I}_{\mathcal{X}_{S_i}}(z) + \frac{1}{2}\|z-x\|^2$, followed by a linear combination between the previous iterate and the average of minibatch projections, thus obtaining again RPM. Third, when solving the smooth stochastic optimization problem (3.7), $\min_x F(x) = \mathbf{E}[F_S(x)]$, we do not have access to the gradient of F, $\nabla F(x) = \mathbf{E}[\nabla F_S(x)] = \mathbf{E}[x - \Pi_{\mathcal{X}_S}(x)]$, and thus we draw a minibatch $S_1, \ldots, S_N$ to form a sample average function $\hat{F}_N = 1/N \sum_{i=1}^N F_{S_i}$ of F, use the minibatch gradient $\nabla \hat{F}_N(x) = 1/N \sum_{i=1}^N (x - \Pi_{\mathcal{X}_{S_i}}(x))$, and perform a gradient update with step size α, leading again to RPM. Note that we have $\arg\min_x f(x) = \arg\min_x F(x)$ (see Theorem 3.3). This property and the relation $F(x) = \frac{1}{2}\mathbf{E}\left[\|\nabla F_S(x)\|^2\right]$, where $\nabla F_S(x) = x - \Pi_{\mathcal{X}_S}(x)$, will be essential in the next sections for proving linear convergence for the stochastic gradient method on this particular problem. Fourth, when solving the stochastic intersection problem (3.8), $x \in \bigcap_{S\sim\mathbf{P}} \mathcal{X}_S$, we typically do not have explicit access to $\bigcap_{S\sim\mathbf{P}} \mathcal{X}_S$. We take a minibatch of samples $S_1, \ldots, S_N$, perform projection onto each $\mathcal{X}_{S_i}$ followed by a relaxation step, leading again to RPM. One important distinction between our algorithm and other existing ones is the use of an extrapolated (also referred to as overrelaxed) step size α_k, depending on the conditioning problem parameter L (see Theorem 6.2), that, in general, is much larger than the constant step size $\alpha \in (0, 2)$ usually used in the literature. Another feature of our algorithm is that it allows one to project simultaneously onto several sets ($N \geq 1$), thus providing flexibility in matching the implementation of the algorithm on the parallel architecture at hand.

6.1. Convergence analysis for RPM. The stochastic reformulations and the conditioning problem parameters play an essential role in the derivation of the extrapolated step size of our new projection scheme and in its convergence analysis. More precisely, our convergence is based on two important properties of the family of the stochastic approximation $(\mathcal{X}_S)_{S\sim\mathbf{P}}$. For simplicity, we recall them here again. First, for stochastic approximation $(\mathcal{X}_S)_{S\sim\mathbf{P}}$ there exists $L \leq 1$ satisfying the smooth regularity inequality (4.2):

$$\|\mathbf{E}\left[x - \Pi_{\mathcal{X}_S}(x)\right]\|^2 \leq L \cdot \mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right] \quad \forall x \in \mathbb{R}^n. \tag{6.1}$$

However, for specific sets and distributions $\mathbf{P}$, we proved in section 4 that L can be much smaller than 1. Second, for the stochastic approximation $(\mathcal{X}_S)_{S\sim\mathbf{P}}$ there exists $\mu \geq 0$ satisfying the stochastic linear regularity property (3.10):

$$\mu \cdot \operatorname{dist}^2_{\mathcal{X}}(x) \leq \mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right] \quad \forall x \in \mathbb{R}^n. \tag{6.2}$$

However, we have proved in section 5 that for specific sets and distributions $\mathbf{P}$, the constant μ can be nonzero. Based on properties (6.1) and (6.2), the smooth objective function F of the stochastic optimization problem (3.7) satisfies Theorem 4.1, and in particular we have $0 \leq \mu \leq L \leq 1$ and the inequalities

$$\frac{\mu}{2}\|x - \Pi_{\mathcal{X}}(x)\|^2 \leq F(x) - F^* \leq \frac{L}{2}\|x - \Pi_{\mathcal{X}}(x)\|^2 \quad \forall x \in \mathbb{R}^n. \tag{6.3}$$

In section 4 we proved that inequality (6.3) expresses that the objective function F is strongly convex with constant μ and has Lipschitz continuous gradient with constant L when restricted on any segment $[x, \Pi_{\mathcal{X}}(x)]$. Thus, the ratio $\texttt{cond} = \mu/L$ represents the condition number of the convex feasibility problem (3.4) or its stochastic reformulation (3.7). In [22] it has been proved that gradient-based methods converge linearly if and only if the inequalities (6.3) hold. We show below that these conditions are also sufficient for the minibatch stochastic gradient method `RPM` to converge linearly on problem (3.7) and that the rates depend explicitly on the conditioning parameters L and μ and the size of the minibatch N. We start with a basic result from probability theory [24].

LEMMA 6.1 (supermartingale convergence lemma). *Let v^k and u^k be sequences of nonnegative random variables such that $\mathbf{E}\left[v^{k+1}|\mathcal{F}_k\right] \leq v^k - u^k$ a.s. for all $k \geq 0$, where $\mathcal{F}_k$ denotes the collection $\{v^0, \ldots, v^k, u^0, \ldots, u^k\}$. Then, v^k converges to a random variable v a.s. and $\sum_{k=0}^{\infty} u^k < \infty$ a.s.*

Let x^* be any element of $\mathcal{X}$. Now, we are ready to derive the following asymptotic convergence result.

THEOREM 6.2. *Assume that the set $\mathcal{X}$ is nonempty and define*

$$L_N = \frac{1}{N} + \left(1 - \frac{1}{N}\right) L \leq 1.$$

Let $\{x^k\}_{k\geq 0}$ be generated by algorithm `RPM` with extrapolated step sizes $0 < \alpha_k < \frac{2}{L_N}$. Then, we have the following average decrease in square distance:

$$\mathbf{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] \leq \|x^k - x^*\|^2 - 2(2\alpha_k - \alpha_k^2 L_N)F(x^k) \quad \forall k \geq 0,\ x^* \in \mathcal{X}. \tag{6.4}$$

Moreover, the fastest decrease is given by the constant step size $\alpha_k = 1/L_N$. If, additionally, exactness holds and the step size satisfies $\delta \leq \alpha_k \leq \frac{2}{L_N} - \delta$ for some $0 < \delta \leq \frac{1}{L_N}$, then at least one subsequence of x^k converges a.s. to a random point in the set $\mathcal{X}$ and $\lim_{k\to\infty} F(x^k) = 0$ a.s.

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

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - \frac{2\alpha_k}{N}\sum_{i=1}^{N}\left\langle x^k - x^*, x^k - \Pi_i^k\right\rangle + \frac{\alpha_k^2}{N^2}\left\|\sum_{i=1}^{N}(x^k - \Pi_i^k)\right\|^2 \\ &\leq \|x^k - x^*\|^2 - \frac{2\alpha_k}{N}\sum_{i=1}^{N}\left\|x^k - \Pi_i^k\right\|^2 + \frac{\alpha_k^2}{N^2}\left\|\sum_{i=1}^{N}(x^k - \Pi_i^k)\right\|^2 \end{aligned} \tag{6.5}$$

$$= \|x^k - x^*\|^2 - \frac{2\alpha_k}{N}\sum_{i=1}^N \left\|x^k - \Pi_i^k\right\|^2 + \frac{\alpha_k^2}{N^2}\left(\sum_{i=1}^N \|x^k - \Pi_i^k\|^2 + \sum_{i\neq j;i,j=1}^N \langle x^k - \Pi_i^k, x^k - \Pi_j^k\rangle\right),$$

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, using the definition of F,

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

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

$$\begin{aligned}
&\mathbf{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] \\
&\quad\leq \|x^k - x^*\|^2 - 4\alpha_k F(x^k) \\
&\qquad + \frac{\alpha_k^2}{N^2}\left(2NF(x^k) + \sum_{i\neq j}\left\langle \mathbf{E}\left[x^k - \Pi_i^k \mid x^k\right], \mathbf{E}\left[x^k - \Pi_j^k \mid x^k\right]\right\rangle\right) \\
&\quad= \|x^k - x^*\|^2 + \left(\frac{2\alpha_k^2}{N} - 4\alpha_k\right)F(x^k) + \frac{\alpha_k^2(N^2-N)}{N^2}\left\|\mathbf{E}\left[x^k - \Pi_{\mathcal{X}_S}(x^k) \mid x^k\right]\right\|^2 \\
&\quad\overset{(6.1)}{\leq} \|x^k - x^*\|^2 + \left(\frac{2\alpha_k^2}{N} - 4\alpha_k\right)F(x^k) + \frac{\alpha_k^2(N-1)}{N}L\mathbf{E}\left[\|x^k - \Pi_{\mathcal{X}_S}(x^k)\|^2 \mid x^k\right] \\
&\quad= \|x^k - x^*\|^2 - 2(2\alpha_k - \alpha_k^2 L_N)F(x^k).
\end{aligned}$$

Clearly, the fastest decrease is obtained by maximizing $2\alpha_k - \alpha_k^2 L_N$ in α_k, that is, the maximum is obtained for constant step size $\alpha_k = 1/L_N$. Further, for the step sizes satisfying $0 < \delta \leq \alpha_k \leq \frac{2}{L_N} - \delta$ we have $2\alpha_k - \alpha_k^2 L_N \geq \delta^2 L_N > 0$. Then, from the 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, by continuity of F, $F(\tilde{x}^*) = 0$ a.s. This leads to $\tilde{x}^* \in \mathcal{X}''$ a.s. When exactness holds, that is, $\mathcal{X} = \mathcal{X}''$, then at least one subsequence of x^k converges a.s. to a random point $\tilde{x}^*$ from the set $\mathcal{X}$. ∎

The next theorem provides rates of convergence for the sequence x^k generated by `RPM` that depend explicitly on the conditioning parameters L and μ and the number of samples N.

THEOREM 6.3. *Assume that the set $\mathcal{X}$ is nonempty and define*

$$L_N = \frac{1}{N} + \left(1 - \frac{1}{N}\right)L.$$

Let $\{x^k\}_{k\geq 0}$ be generated by algorithm `RPM` with step size satisfying $\delta \leq \alpha_k \leq \frac{2}{L_N} - \delta$ for some $0 < \delta \leq \frac{1}{L_N}$. Then we have the following.

(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 in terms of function values:*

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

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

(ii) *If additionally the linear regularity property* (6.2) *holds, then we have the following linear convergence rate for the last iterate* x^k*:*

$$\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^{k+1})\right] \leq \left(1-\delta^2 L_N \mu\right)\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^k)\right],$$
$$\mathbf{E}\left[F(x^k)\right] - F^* \leq \left(1-\delta^2 L_N\mu\right)^k \frac{L\mathrm{dist}^2_{\mathcal{X}}(x^0)}{2}.$$

Proof. (i) By taking expectation w.r.t. the entire history on both sides in (6.4) we get the following decrease in the distance to a point $x^* \in \mathcal{X}$:

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

Further, writing $r_k = \mathbf{E}\left[\|x^k - x^*\|^2\right]$ and noticing the lower bound $2-\alpha_k L_N \geq \delta L_N$ for any step size satisfying $\delta \leq \alpha_k \leq \frac{2}{L_N} - \delta$ for some $0 < \delta \leq \frac{1}{L_N}$, we have

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

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

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

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

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

This relation and $F^* = 0$ imply the first statement. Moreover, without loss of generality, from Theorem 6.2 we can assume that x^k converges almost surely to a random point in the set $\mathcal{X}$. Therefore, the average sequence $\hat{x}^k$ also converges almost surely to the same random point.

(ii) To prove linear convergence under linear regularity (6.2) we again use inequality (6.4) and $F^* = 0$:

$$\begin{aligned}\mathbf{E}\left[\|x^{k+1}-x^*\|^2 \mid x^k\right] &\leq \|x^k - x^*\|^2 - 2(2\alpha_k - \alpha_k^2 L_N)\left(F(x^k) - F^*\right)\\ &\overset{(6.3)}{\leq} \|x^k - x^*\|^2 - (2\alpha_k - \alpha_k^2 L_N)\,\mu\,\mathrm{dist}^2_{\mathcal{X}}(x^k).\end{aligned}$$

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

$$\mathbf{E}\left[\|x^{k+1}-x^*\|^2\right] \leq \mathbf{E}\left[\|x^k - x^*\|^2\right] - (2\alpha_k - \alpha_k^2 L_N)\mu\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^k)\right]. \tag{6.6}$$

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 (6.6), we finally get

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

Since for our choice of the step size $\delta \leq \alpha_k \leq \frac{2}{L_N}-\delta$, for some $0<\delta\leq\frac{1}{L_N}$, we have $2\alpha_k-\alpha_k^2 L_N \geq \delta^2 L_N$, and so the previous relation implies immediately

$$\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^{k+1})\right] \leq \left(1-\delta^2 L_N\mu\right)\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^k)\right],$$

which proves the second statement of the theorem. Finally, combining the convergence rate in distances with the right-hand side inequality in (6.3) we get the convergence rate in expectation of value function. ∎

An immediate consequence of Theorem 6.3 is the following result.

THEOREM 6.4. *Assume that $\mathcal{X}\neq\emptyset$ and $L_N=\frac{1}{N}+\left(1-\frac{1}{N}\right)L$. Let $\{x^k\}_{k\geq 0}$ be generated by* RPM *with constant step size $0<\alpha_k\equiv\alpha<\frac{2}{L_N}$ or optimal constant step size $\alpha_k\equiv\alpha=\frac{1}{L_N}$. Then we have the following.*

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

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

(ii) *If additionally the linear regularity property* (6.2) *holds, then we have the following linear convergence rate for the last iterate x^k:*

$$\mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^k)\right] \leq \left(1-\alpha(2-\alpha L_N)\mu\right)^k \mathrm{dist}^2_{\mathcal{X}}(x^0) \tag{6.7}$$
$$\xrightarrow{\alpha=\frac{1}{L_N}} \mathbf{E}\left[\mathrm{dist}^2_{\mathcal{X}}(x^k)\right] \leq \left(1-\frac{\mu}{L_N}\right)^k \mathrm{dist}^2_{\mathcal{X}}(x^0).$$

Proof. By taking expectation w.r.t. the entire history on both sides in (6.4) we get the following decrease in the distance to a point $x^*\in\mathcal{X}$:

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

The fastest decrease is obtained maximizing $2\alpha_k-\alpha_k^2 L_N$ in α_k, which leads to the optimal step size $\alpha_k=1/L_N$. The rest of the proof follows the same steps as those of the proof of Theorem 6.3. ∎

Remark 6.5. Theorem 6.2 shows that in order to have a decrease in average distance (6.4), the step size α_k has to satisfy $0<\alpha_k<\frac{2}{L_N}$, where $L_N=\frac{1}{N}+(1-\frac{1}{N})L$ and L satisfies the smooth regularity condition (4.2). Note that if $L<1$, then $L_N<1$ and consequently $\frac{2}{L_N}>2$. Thus, our method uses an extrapolated step size $\alpha\sim\frac{2}{L_N}$, depending on the conditioning problem parameter L, that, in general, is much larger than the constant step size $\alpha\sim 2$ usually used in the literature (see, e.g., the B-AP method from section 2.2) [5]. To the best of our knowledge, the convergence rates from Theorems 6.3 and 6.4 are the first showing an explicit dependence on the conditioning parameters L and μ of the general convex feasibility problem, and on the minibatch sample size N via the term L_N. For example, we have obtained for RPM

a linear rate given by $1-\frac{N\mu}{1+(N-1)L}$ (see Theorem 6.4) which is definitely better for $L<1$ and $N>1$ than the linear rate of B-AP given by $1-\mu$ for $N=1$ (see, e.g., [23, 24, 27]).

Moreover, our convergence results are the first explaining when extrapolation combined with minibatching works, that is, we can accelerate a random *minibatch* projection algorithm provided that there exists a parameter $L<1$ such that the sets describing the stochastic approximation $(\mathcal{X}_S)_{S\sim\mathbf{P}}$ satisfy the smooth regularity condition (4.2), work with a minibatch size $N>1$, and use an *extrapolated* step size $\alpha\sim 2/L_N$. Note also that for several important sets we can estimate the Lipschitz constant L and consequently L_N; see section 4. As is usual in optimization, when the Lipschitz constant L cannot be determined, search procedures [25] or adaptive procedures [12] may be used. More specifically, when it is difficult to compute L, then, inspired by [12], we can use the following online approximation of it:

$$L=\max_{k\geq 0}\frac{\|\mathbf{E}\left[x^k-\Pi_{\mathcal{X}_S}(x^k)\right]\|^2}{\mathbf{E}\left[\|x^k-\Pi_{\mathcal{X}_S}(x^k)\|^2\right]}\simeq L^{(k)}=\frac{\|\sum_{i=1}^N w_i^k(x^k-\Pi_{\mathcal{X}_{S_i^k}}(x^k))\|^2}{\sum_{i=1}^N w_i^k\|x^k-\Pi_{\mathcal{X}_{S_i^k}}(x^k)\|^2}, \tag{6.8}$$

where the weights w_i^k satisfy $\sum_{i=1}^N w_i^k=1$ and $w_i^k\geq 0$. The asymptotic convergence and also the effectiveness of projection algorithms based on an adaptive extrapolated step size $0<\alpha_k<\frac{2}{L^{(k)}}$ has been shown, e.g., in [5, 10, 12]. Note that $\lim_{N\to+\infty}L_N=L$ and consequently for large minibatch size N the adaptive extrapolation rule $0<\alpha_k<\frac{2}{L^{(k)}}$ is interpreted by our theory as an online numerical approximation of our fixed extrapolation rule $0<\alpha_k<\frac{2}{L_N}$.

6.2. Expected projection method. As $N\to\infty$, RPM becomes the gradient method for solving the smooth problem (3.7), which we call the *expected projection method*.

Algorithm EPM.
Choose $x^0\in\mathbb{R}^n$ and positive step sizes $\{\alpha_k\}_{k\geq 0}$. For $k\geq 0$ repeat:
Compute $x^{k+1}=x^k-\alpha_k\nabla F(x^k)\quad\left(:=x^k-\alpha_k\left(x^k-\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x^k)|x^k\right]\right)\right)$.

For EPM we have the following convergence results.

THEOREM 6.6. *Assume that the set $\mathcal{X}\neq\emptyset$. Let $\{x^k\}_{k\geq 0}$ be generated by EPM with constant step size $0<\alpha_k\equiv\alpha<\frac{2}{L}$ or optimal constant step size $\alpha_k\equiv\alpha=\frac{1}{L}$. Then we have the following.*

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

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

(ii) *If additionally the linear regularity property* (6.2) *holds, then we have the following linear convergence rate for the last iterate x^k:*

$$\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^k)\right]\leq(1-\alpha(2-\alpha L)\mu)^k\operatorname{dist}^2_{\mathcal{X}}(x^0) \tag{6.9}$$
$$\xrightarrow{\alpha=\frac{1}{L}}\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^k)\right]\leq\left(1-\frac{\mu}{L}\right)^k\operatorname{dist}^2_{\mathcal{X}}(x^0).$$

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

$$\begin{aligned}
\|x^{k+1}-x^*\|^2 &= \|x^k-x^*\|^2 - 2\alpha_k\left\langle x^k-x^*, x^k-\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x^k)\right]\right\rangle \\
&\quad + \alpha_k^2\left\|x^k-\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x^k)\right]\right\|^2 \\
&\overset{(6.1)}{\leq} \|x^k-x^*\|^2 - 2\alpha_k\left\langle x^k-x^*, x^k-\mathbf{E}\left[\Pi_{\mathcal{X}_S}(x^k)\right]\right\rangle \\
&\quad + L\alpha_k^2\mathbf{E}\left[\left\|x^k-\Pi_{\mathcal{X}_S}(x^k)\right\|^2\right] \\
&= \|x^k-x^*\|^2 - 2\alpha_k\mathbf{E}\left[\left\|x^k-\Pi_{\mathcal{X}_S}(x^k)\right\|^2\right] + L\alpha_k^2\mathbf{E}\left[\left\|x^k-\Pi_{\mathcal{X}_S}(x^k)\right\|^2\right] \\
&\quad - 2\alpha_k\mathbf{E}\left[\left\langle \Pi_{\mathcal{X}_S}(x^k)-x^*, x^k-\Pi_{\mathcal{X}_S}(x^k)\right\rangle\right] \\
&\leq \|x^k-x^*\|^2 - (2\alpha_k - L\alpha_k^2)\mathbf{E}\left[\left\|x^k-\Pi_{\mathcal{X}_S}(x^k)\right\|^2\right] \\
&= \|x^k-x^*\|^2 - 2(2\alpha_k - L\alpha_k^2)F(x^k), \qquad (6.10)
\end{aligned}$$

where the second inequality follows from $\langle \Pi_{\mathcal{X}_S}(x^k)-x^*, x^k-\Pi_{\mathcal{X}_S}(x^k)\rangle \geq 0$ for all $x^* \in \mathcal{X} \subseteq \mathcal{X}_S$. From (6.10) we observe that the fastest decrease is obtained from maximizing $2\alpha_k - L\alpha_k^2$, which leads to the optimal step size $\alpha_k = 1/L$. The rest of the proof follows the same steps as in the proof of Theorem 6.3. ∎

From Theorem 6.6 it follows that we can achieve linear convergence given in terms of $1-\mu/L$ for EPM with step size satisfying $0 < \alpha_k < 2/L$. Recall that $\texttt{cond} = \mu/L$ represents the condition number of the convex feasibility problem (3.4) or of its stochastic reformulation (3.7) (see Theorem 4.1). Let us now derive convergence rates for some particular sets. For example, let us consider finding a point in the finite intersection of convex sets $(\mathcal{X}_i)_{i\in[m]}$, that is, $\mathcal{X} = \bigcap_{i=1}^m \mathcal{X}_i$. Further, we consider a uniform probability on $\Omega = [m]$. Then EPM is similar to barycentric method in [9, 12]:

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

EPM$(1/m)$ 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., [9, 12]. Our convergence analysis (Theorem 6.6) allows one to easily derive convergence rates for EPM$(1/m)$, as shown by the following examples.

(i) Consider 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 6.6 the sequence generated by EPM$(1/m)$ with constant optimal step size $\alpha = 1/L$ converges linearly:

$$\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(6.9)}{\leq} \left(1-\frac{\mu}{L}\right)^k \text{dist}^2_{\mathcal{X}}(x^0) \overset{(4.7)+(5.5)}{=} \left(1-\frac{\lambda^{\text{nz}}_{\min}\left(A^TDA\right)}{\lambda_{\max}(A^TDA)}\right)^k \text{dist}^2_{\mathcal{X}}(x^0).$$

(ii) Consider now the more general problem of finding a solution to a system of linear inequalities $Ax \leq b$, where $A \in \mathbb{R}^{m\times n}$ and $\mathcal{X}_i = \{x : A_i^T x \leq b_i\}$. From Theorem 6.6, the sequence generated by EPM$(1/m)$ with constant optimal step size $\alpha = 1/L$ also converges linearly:

$$\begin{aligned}
\mathbf{E}\left[\text{dist}^2_{\mathcal{X}}(x^k)\right] &\overset{(6.9)}{\leq} \left(1-\frac{\mu}{L}\right)^k \text{dist}^2_{\mathcal{X}}(x^0) \\
&\overset{(4.8)+(5.6)}{=} \left(1-\frac{\tilde{\mu}}{\max\limits_{i=1:m}\|A_i\|^2\lambda_{\max}(A^TDA)}\right)^k \text{dist}^2_{\mathcal{X}}(x^0).
\end{aligned}$$

Note that from Theorem 6.6 it follows that the basic EPM method

$$x^{k+1} = \frac{1}{m}\sum_{i=1}^{m} \Pi_{\mathcal{X}_i}(x^k),$$

i.e., step size $\alpha = 1$, also converges linearly:

$$\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(6.10)}{\leq} (1-(2-L)\mu)^k \operatorname{dist}^2_{\mathcal{X}}(x^0).$$

(iii) However, our algorithmic framework leads to new schemes with possibly better rates. For example, for a general probability distribution $(p_i)_{i\in[m]}$ on $\Omega = [m]$ and $N = m$, we get a new EPM with iteration

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

For systems of linear equalities and inequalities, choosing the probabilities $p_i = \frac{\|A_i\|^2}{\|A\|_F^2}$ the sequence generated by this method has the following convergence rates.

(iii′) For a linear system $Ax = b$, the sequence generated by $\texttt{EPM}(\frac{\|A_i\|^2}{\|A\|_F^2})$ with the constant optimal step size $\alpha = 1/L$ converges linearly (Theorem 6.6):

$$\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(6.9)}{\leq} \left(1-\frac{\mu}{L}\right)^k \operatorname{dist}^2_{\mathcal{X}}(x^0) \overset{(4.7)+(5.5)}{=} \left(1-\frac{\lambda^{\text{nz}}_{\min}\left(A^TA\right)}{\lambda_{\max}(A^TA)}\right)^k \operatorname{dist}^2_{\mathcal{X}}(x^0).$$

(iii″) For a system of linear inequalities $Ax \leq b$, the sequence generated by $\texttt{EPM}(\frac{\|A_i\|^2}{\|A\|_F^2})$ with optimal step size $\alpha = 1/L$ also converges linearly:

$$\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(6.9)}{\leq} \left(1-\frac{\mu}{L}\right)^k \operatorname{dist}^2_{\mathcal{X}}(x^0) \overset{(4.8)+(5.9)}{=} \left(1-\frac{\tilde{\mu}}{\lambda_{\max}(A^TA)}\right)^k \operatorname{dist}^2_{\mathcal{X}}(x^0).$$

In general, $\lambda_{\max}(A^TA) \leq \max_i \|A_i\|^2 \lambda_{\max}(A^TDA)$ and thus $\texttt{EPM}(\|A_i\|^2/\|A\|_F^2)$ is usually better than $\texttt{EPM}(1/m)$ on linear inequalities. The reader can choose other probabilities that may lead to better rates.

6.3. Random alternating projection method. Note that RPM is general and for $N = 1$ we recover multiple existing projection algorithms from the literature. In this case we use a single projection per iteration, i.e., $N = 1$, which results in the *basic* randomized alternating projection method B-AP with step size α_k:

$$\texttt{B-AP}: \quad \text{choose random } S_k \sim \mathbf{P} \text{ and update } x^{k+1} = x^k - \alpha_k(x^k - \Pi_{\mathcal{X}_{S_k}}(x^k)).$$

This can be viewed as a random implementation of the cyclic alternating projection method on the convex feasibility problem (3.4) or as a stochastic gradient method on the equivalent stochastic reformulation (3.7). The cyclic variant has been proposed by Von Neumann [32] for the intersection of two subspaces in a Hilbert space, and it has many generalization and extensions [4, 13, 23, 24]. A detailed survey of the work in this area is given in [3]. The first convergence rate result for the B-AP algorithm, under the assumption that the intersection set has a nonempty interior, was given in [15]. The convergence rates of B-AP for $\alpha_k = 1$ and a finite intersection of convex sets have been given recently in [23]. From our convergence analysis it follows that we

can work with an infinite intersection of sets and the step size in B-AP can be chosen as $0 < \alpha_k < 2$, since for $N = 1$ we have $L_N = 1$. Moreover, the optimal step size is $\alpha = 1$, for which the linear convergence rate, under linear regularity, becomes (see Theorem 6.3)

$$\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^k)\right] \leq (1-\mu)^k \, \mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^0)\right].$$

The same convergence rate was derived, e.g., in [24]. However, it was observed in practice that extrapolation, that is, $\alpha_k \in (1,2)$, makes B-AP perform better. Note that from our general convergence analysis, for particular choices of sets and probabilities, we recover well-known algorithms from literature.

(i) Consider finding a solution to a linear system $Ax = b$, where A is an $m \times n$ matrix. Further, assume that $\Omega = \{e_1, \ldots, e_m\}$ and take the probability distribution $\mathbf{P}(S = e_i) = \|A_i\|^2/\|A\|_F^2$. Then, B-AP with $\alpha_k = 1$ is the randomized Kaczmarz algorithm [31]:

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

Moreover, for these choices of the probabilities and step size, our convergence analysis matches exactly the one in [31], that is,

$$\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(6.7)}{\leq} (1-\mu)^k \operatorname{dist}^2_{\mathcal{X}}(x^0) \overset{(5.5)}{=} \left(1 - \frac{\lambda_{\min}^{\mathrm{nz}}\left(A^T A\right)}{\|A\|_F^2}\right)^k \operatorname{dist}^2_{\mathcal{X}}(x^0).$$

The next example shows the possibly slow behavior of B-AP (Kaczmarz) for an ill-conditioned linear system.

Example 6.7. Given $a > 0$, consider two hyperplanes $\mathcal{X}_1 = \{x \in \mathbb{R}^2 : \ x_2 = 0\}$ and $\mathcal{X}_2 = \{x \in \mathbb{R}^2 : \ x_1 + a x_2 = 0\}$ with the intersection $\mathcal{X} = \{0\}$. The linear regularity constant is $\mu = 1/(a^2 + 2 + \sqrt{(a^2+2)^2 - 4}) \geq 1/(2a^2+4)$. Then, by choosing $x^0 = [1 \ \ 0]$, B-AP generates $\{x^k\}_{k \geq 0}$ such that

$$\left(1 - 1/(1+a^2)\right)^k \leq \|x^k - x^*\|^2 \leq \left(1 - 1/(2a^2+4)\right)^k \quad \forall k \geq 0.$$

Notice that for large a, the linear rate can be arbitrarily slow.

However, our B-AP generalizes the randomized Kaczmarz algorithm from [31], 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 step size $\alpha_k = 1$, the previous iteration is equivalent to the randomized iterative method of [16]. For this choice of the step size, our convergence analysis exactly matches the one in [16]:

$$\begin{aligned}\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^k)\right] &\overset{(6.7)}{\leq} (1-\mu)^k \operatorname{dist}^2_{\mathcal{X}}(x^0) \\ &\overset{(5.2)}{=} \left(1 - \lambda_{\min}^{\mathrm{nz}}(A^T \mathbf{E}\left[S(S^T A A^T S)^\dagger S^T\right] A)\right)^k \operatorname{dist}^2_{\mathcal{X}}(x^0).\end{aligned}$$

Clearly, for some choices of random matrix S better rates may be obtained than for basic randomized Kaczmarz [31], as proved in [33] (see also our numerical section below). Moreover, our convergence analysis allows one to choose an extrapolated step size $\alpha_k \in (1,2)$, while in [16] convergence is proved only for $\alpha_k \leq 1$.

(ii) Consider now the more general problem of finding a solution to a system of linear inequalities $Ax \leq b$, where $A \in \mathbb{R}^{m\times n}$. Further, assume as above $\Omega = \{e_1, \ldots, e_m\}$ and take the probability distribution $\mathbf{P}(S = e_i) = \|A_i\|^2/\|A\|_F^2$. Then, B-AP with $\alpha_k = 1$ is Algorithm 4.6 from [19]:

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

For these choices of the probabilities and step size, our convergence analysis exactly matches the one in [19]:

$$\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(6.7)}{\leq} (1-\mu)^k \operatorname{dist}^2_{\mathcal{X}}(x^0) \overset{(5.9)}{=} \left(1 - \frac{\tilde{\mu}}{\|A\|_F^2}\right)^k \operatorname{dist}^2_{\mathcal{X}}(x^0).$$

However, B-AP generalizes Algorithm 4.6 from [19], considering for a random vector $S_k \in \mathbb{R}^m_+$ the general iteration

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

Thus, we obtained a new algorithm for solving linear inequalities. Following a similar reasoning to that in [33] (where systems of linear equalities were considered), we believe that this update may lead to better convergence rates for certain choices of the random vector S than the one corresponding to the simple choice $S = e_i$. However, this analysis is beyond the scope of this paper. In particular, from Corollary 5.4, we get

$$\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(6.7)}{\leq} (1-\mu)^k \operatorname{dist}^2_{\mathcal{X}}(x^0) \overset{(5.9)}{=} \left(1 - \frac{\tilde{\mu}}{\sum_{S\in\Omega_r}\|A^T S\|^2}\right)^k \operatorname{dist}^2_{\mathcal{X}}(x^0).$$

(iii) Finally, we can consider the convex feasibility problem where the intersection set has a nonempty interior. First, let us investigate the case when the sequence $\|x^k - x^*\|$ is decreasing. For $N = 1$ and $\alpha_k \in (0, 2)$ it follows from (6.5) 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. For $N \geq 1$ and $\alpha_k \in (0, 1]$ we have for all $k \geq 0$,

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

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 of N and α_k given above we have $\|x^k - x^*\| \leq \|x^0 - x^*\|$ for all $x^* \in \mathcal{X}$ and $k \geq 0$. An important application of the previous inequality is that when the set $\mathcal{X}$ contains a ball with radius δ centered on $\bar{x}$. 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 (5.11), one should choose the compact set $Q = \{x : \ \|x - \bar{x}\| \leq \|x^0 - \bar{x}\|\}$ such that the linear regularity constant given in (5.11) becomes

$$\mu = \delta^2 \left(\min_{S \in \Omega} \ p_S \right) \|x^0 - \bar{x}\|^{-2} \tag{6.11}$$

since all the points of interest for which the linear regularity property has to hold are the iterates $\{x^k\}_{k \geq 0}$. Hence, for step size $\alpha_k = 1$ the sequence generated by B-AP attains the following linear rate:

$$\mathbf{E}\left[\operatorname{dist}^2_{\mathcal{X}}(x^k)\right] \overset{(6.7)}{\leq} (1-\mu)^k \operatorname{dist}^2_{\mathcal{X}}(x^0) \overset{(5.11)+(6.11)}{=} \left(1 - \frac{p_{\min}\delta^2}{R^2}\right)^k \operatorname{dist}^2_{\mathcal{X}}(x^0),$$

where $p_{\min} = \min_{S \in \Omega} p_S$ and $R = \|x^0 - \bar{x}\|$. Under the same settings, a similar rate was derived in [23].

7. Numerical simulations. The practical performance of projection algorithms has been investigated in many papers; see, e.g., [6, 8, 10, 11, 14, 17, 20, 28, 29, 30]. In this section we present some preliminary numerical results for our extrapolated randomized projection algorithms RPM and EPM. In all tests we consider the accuracy $\epsilon = 0.01$ and for data generated randomly the entries of the matrix $A \in \mathbb{R}^{m \times n}$ and vector $b \in \mathbb{R}^m$ are taken in $[-2,\ 2]$. For all the implementations we consider *full* iterations for EPM and for RPM ($1 \leq N < m$) we plot only trajectory points at kN/m. Moreover, we use either extrapolated step size $\alpha = 1.9/L_N$ or constant step size $\alpha = 1.9 < 2$.

In the first numerical test we consider random linear systems of equalities $Ax = b$ with $X_{e_i} = \{x : \ A_i^T x = b_i\}$. In Figure 1, the left plot presents the growth of our extrapolated (also referred to as *overrelaxed*) step size $\alpha = 2/L_N$, where recall that $L_N = 1/N + (1 - 1/N)L$ and $L = \lambda_{\max}(A^T A)/\|A\|_F^2$, for several randomly generated matrices A of dimension $m \in \{20, 40, 60, 80, 100\}$ and $n \in \{30, 60, 90, 120, 150\}$ over the values $N \in \{1, \ldots, m\}$.[1] Since for these matrices $\operatorname{rank}(A) \geq 2$, from Corollary 4.3 it follows that $L = \lambda_{\max}(A^T A)/\|A\|_F^2 < 1$ and consequently $L_N < 1$. The red line indicates the constant value $\alpha = 2$ usually considered in the literature. We clearly see that our extrapolated step size $\alpha = 2/L_N$ is much larger than the step size $\alpha = 2$, as is also predicted by our Corollary 4.3. This extrapolation leads to an acceleration of our algorithm RPM, which is confirmed in numerical tests; see the middle and the right-hand plots of Figure 1. More precisely, the plot in the middle compares the *full* iterations of our RPM algorithm with the overrelaxed step size $\alpha = 1.9/L_N$ versus constant step size $\alpha = 1.9$ implementation for different values of N. We clearly see that RPM with $\alpha = 1.9/L_N$ is even 10 times faster than its corresponding constant counterpart $\alpha = 1.9$. The right-hand plot displays the behavior of the extrapolated RPM and EPM for different values of N, including $N = 1$, which corresponds to $L_N = 1$ and consequently $\alpha = 1.9$. We observe that it is beneficial for the RPM scheme to

[1] See online version for color figures.

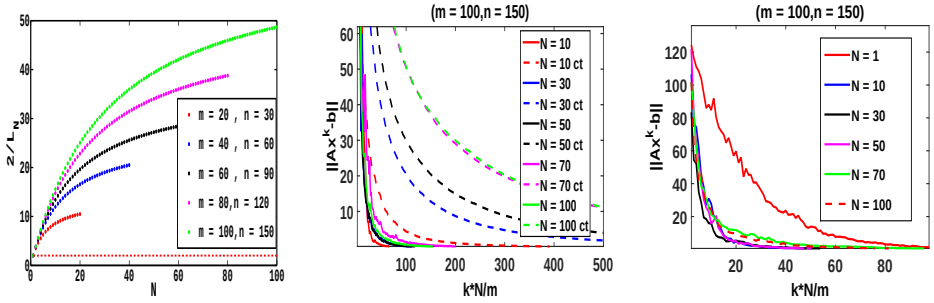

FIG. 1. *Random linear systems. Left: growth of step size $\alpha = 2/L_N$. Middle: behavior of extrapolated (overrelaxed) RPM vs. constant RPM for various N. Right: behavior of extrapolated RPM for various N. The $N = m$ case corresponds to EPM.*

perform more than one projection at each iteration ($N > 1$), and EPM is not necessarily the most efficient one—we found that usually an $N \sim \mathcal{O}(m/2)$ produces the best performance in terms of the total computational cost.

In the second test we analyze the behavior of extrapolated RPM and EPM on solving linear systems $Ax = b$ with either an ill-conditioned matrix $A = \alpha I_n + \beta ee^T$ for some choice of the parameters α and β as in [33], or using a real data minibatch CIFAR-10 (from an image classification dataset; see https://www.cs.toronto.edu/~kriz/cifar.html) and consider a kernel ridge regression task using Gaussian kernel. That is, given the training set $\{z_i, y_i\}_{i=1}^m$ we form the Gaussian kernel matrix $K \in \mathbb{R}^{m\times m}$, described by $K_{ij} = e^{-\zeta\|z_i - z_j\|^2}$ for some $\zeta > 0$, and the labels vector $y \in \mathbb{R}^m$, and then the regression task requires the solution of a linear system with $A = \lambda I_n + K$ and $b = y$ (in our tests we used $\lambda = 0.01$ and $\zeta = 1$). In Figure 2, the left- and right-hand plots present the behavior of extrapolated RPM and EPM (corresponding to $N = m$) for the matrix $A = \alpha I_n + \beta ee^T$ having the condition number 50 and the matrix $A = \lambda I_n + K$ using CIFAR-10, for various choices of N. For both cases we found that usually RPM with $N \sim \mathcal{O}(m/2)$ produces the best performance in terms of the total computational cost. Finally, the plot in the middle presents the behavior of RPM when the matrix $A = \alpha I_n + \beta ee^T$ has a large condition number (10^3). In this case we consider two representations for sets: either $X_{e_i} = \{x : A_i^T x = b_i\}$ or $X_S = \{x : S^T Ax = S^T b\}$ with S chosen with uniform probability from the simplex $\{S \in \mathbb{R}^m : S \geq 0, \sum_{i=1}^m S_i = 1\}$. For ill-conditioned systems, we observe that RPM using random sets X_S produces better results than using deterministic sets X_{e_i}; see also results in [33].

Finally, in the third set of experiments we consider solving either linear inequalities $Ax \leq b$ or linear programs (LPs) $\min_x c^T x \mid Ax = b,\ x \geq 0$. For LPs we consider the equivalent reformulation of finding a point in the intersection of two sets $X_1 = \{x \in \mathbb{R}^{2n+m} : Ax = b\}$ and $X_2 = \mathbb{R}^n_+ \times \mathbb{R}^m \times \mathbb{R}^n_+$, where

$$A = \begin{bmatrix} 0 & A^T & -I_n \\ A & 0 & 0 \\ c^T & b^T & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} -c \\ b \\ 0 \end{bmatrix}, \qquad x = \begin{bmatrix} x \\ \lambda \\ \mu \end{bmatrix}.$$

In Figure 3, the plot on the left-hand side presents the performance of extrapolated RPM/EPM when the matrix A describing the linear inequalities is random and $X_{e_i} = \{x : A_i^T x \leq b_i\}$ for various N. As for linear systems, we observe that by increasing N up to a certain threshold accelerates the RPM algorithm, and for $N = 1$

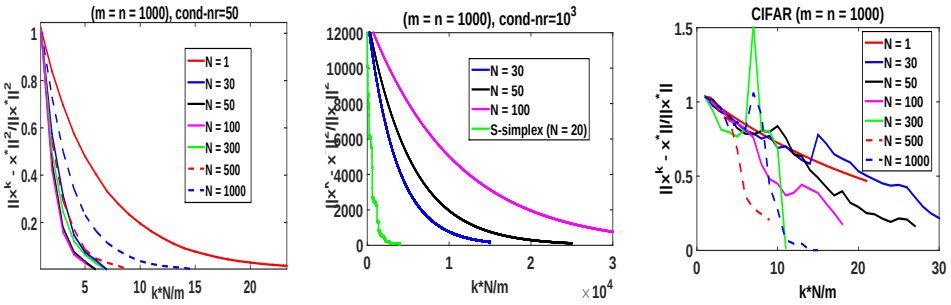

FIG. 2. *Performance of extrapolated (overrelaxed)* RPM *on linear systems with* $A = \alpha I_n + \beta ee^T$. *Left: condition number* 50. *Middle: condition number* 10^3. *Right: kernel ridge regression using CIFAR*-10. *The* $N = m$ *case corresponds to* EPM.

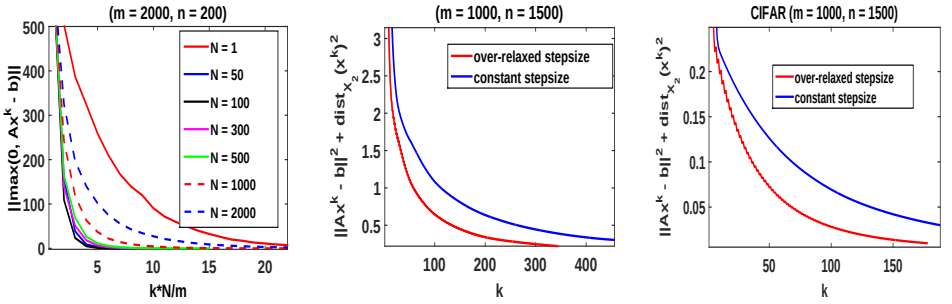

FIG. 3. *Performance of extrapolated (overrelaxed)* RPM *and* EPM *on linear inequalities and LPs. Left: randomly generated linear inequalities* $Ax \leq b$. *Middle: random LP. Right: LP with* A *taken from kernel* K.

and consequently step size $\alpha = 1.9$ we get the worst performance. The $N = m$ case corresponds to extrapolated EPM. Finally, we consider linear programs where matrix A is generated randomly (middle plot) or a block taken from the kernel K of CIFAR-10 (right-hand plot). For finding a point in the intersection of the sets X_1 and X_2 described above we implement EPM with constant step size $\alpha = 1.9$ and adaptive extrapolated step size $\alpha_k = 1.9/(1/2 + L^{(k)}/2)$, where

$$L^{(k)} = \|(x^k - \Pi_{X_1}(x^k)) + (x^k - \Pi_{X_2}(x^k))\|^2 / (2\|x^k - \Pi_{X_1}(x^k)\|^2 + 2\|x^k - \Pi_{X_2}(x^k)\|^2)$$

(see [5, 10, 12] and Remark 6.5). Note that EPM with adaptive step size has a better performance than its constant step-size counterpart. In conclusion, from the numerical results we see a better performance of the extrapolated ($\alpha \sim 2/L_N$) RPM over the constant step size ($\alpha \sim 2$) implementation.

REFERENCES

[1] 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 Optim. Appl. 49, Springer, New York, 2011, pp. 33–48.

[2] A. BECK AND M. TEBOULLE, *Convergence rate analysis and error bounds for projection algorithms in convex feasibility problems*, Optim. Methods Softw., 18 (2003), pp. 377–394.

[3] H. BAUSCHKE AND J. BORWEIN, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.

[4] H. Bauschke and D. Noll, *On cluster points of alternating projections*, Serdica Math. J., 39 (2013), pp. 355–364.
[5] H. Bauschke, P. Combettes, and S. Kruk, *Extrapolation algorithm for affine-convex feasibility problems*, Numer. Algorithms, 41 (2006), pp. 239–274.
[6] D. Blatt and A. Hero, *Energy based sensor network source localization via projection onto convex sets*, IEEE Trans. Signal Process., 54 (2006), pp. 3614–3619.
[7] J. Burke and M. Ferris, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.
[8] C. Byrne and Y. Censor, *Proximity function minimization using multiple Bregman projections, with applications to split feasibility and Kullback–Leibler distance minimization*, Ann. Oper. Res., 105 (2001), pp. 77–98.
[9] Y. Censor, T. Elfving, and G. T. Herman, *Averaging strings of sequential iterations for convex feasibility problems*, in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Stud. Comput. Math. 8, D. Butnariu, S. Reich, and Y. Censor, eds., Elsevier, Amsterdam, 2001, pp. 101–114.
[10] 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*, Comput. Optim. Appl., 51 (2012), pp. 1065–1088.
[11] H. Choi and R. Baraniuk, *Multiple wavelet basis image denoising using Besov ball projections*, IEEE Signal Process. Lett., 11 (2004), pp. 717–720.
[12] P. L. Combettes, *Hilbertian convex feasibility problem: Convergence of projection methods*, Appl. Math. Optim., 35 (1997), pp. 311–330.
[13] F. Deutsch and H. Hundal, *The rate of convergence for the cyclic projections algorithm* I: *Angles between convex sets*, J. Approx. Theory, 142 (2006), pp. 36–55.
[14] J. Gu, H. Stark, and Y. Yang, *Wide-band smart antenna design using vector space projection methods*, IEEE Trans. Antennas Propag., 52 (2004), pp. 3228–3236.
[15] L. Gubin, B. Polyak, and E. Raik, *The method of projections for finding the common point of convex sets*, Comput. Math. Math. Phys., 7 (1967), pp. 1–24.
[16] R. M. Gower and P. Richtarik, *Randomized iterative methods for linear systems*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1660–1690.
[17] G. Herman and W. Chen, *A fast algorithm for solving a linear feasibility problem with application to intensity-modulated radiation therapy*, Linear Algebra Appl., 428 (2008), pp. 1207–1217.
[18] S. Kaczmarz, *Angenaherte Auflosung von Systemen linearer Gleichungen*, Bull. Intern. Acad. Polonaise Sci. Lett., 35 (1937), pp. 355–357 (in Polish); *Approximate solution of systems of linear equations*, Internat. J. Control, 57 (1993), pp. 1269–1271 (in English).
[19] D. Leventhal and A. Lewis, *Randomized methods for linear constraints: Convergence rates and conditioning*, Math. Oper. Res., 35 (2010), pp. 641–654.
[20] 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 (2005), pp. 795–800.
[21] T. Motzkin and I. Schoenberg, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 393–404.
[22] I. Necoara, Y. Nesterov, and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Math. Program., 175 (2019), pp. 69–107.
[23] A. Nedic, *Random projection algorithms for convex set intersection problems*, in Proceedings of the 49th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2010, pp. 7655–7660.
[24] A. Nedic, *Random algorithms for convex minimization problems*, Math. Program., 129 (2011), pp. 225–253.
[25] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.
[26] A. Patrascu and I. Necoara, *Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization*, J. Mach. Learn. Res., 18 (2018), pp. 1–42.
[27] B. Polyak, *Random algorithms for solving convex inequalities*, in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Stud. Comput. Math. 8, D. Butnariu, S. Reich, and Y. Censor, eds., Elsevier, Amsterdam, 2001, pp. 409–422.
[28] A. Samsonov, E. Kholmovski, D. Parker, and C. Johnson, *POCS-based reconstruction for sensitivity encoded magnetic resonance imaging*, Magn. Reson. Med., 52 (2004), pp. 1397–1406.
[29] G. Sharma, *Set theoretic estimation for problems in subtractive color*, Color Res. Appl., 25 (2000), pp. 333–348.

[30] H. Stark and Y. Yang, *Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics*, Wiley Ser. Telecommun. Signal Process., Wiley Interscience, New York, 1998.
[31] T. Strohmer and R. Vershynin, *A randomized Kaczmarz algorithm with exponential convergence*, J. Fourier Anal. Appl., 15 (2009), Article 262.
[32] J. von Neumann, *Functional Operators*, Princeton University Press, Princeton, NJ, 1950.
[33] S. Tu, S. Venkataraman, A. Wilson, A. Gittens, M. Jordan, and B. Recht, *Breaking Locality Accelerates Block Gauss–Seidel*, preprint, https://arxiv.org/abs/1701.03863, 2017.