MATHEMATICS OF OPERATIONS RESEARCH
Vol. 35, No. 3, August 2010, pp. 641–654
ISSN 0364-765X | EISSN 1526-5471 | 10 | 3503 | 0641

DOI 10.1287/moor.1100.0456

Randomized Methods for Linear Constraints: Convergence Rates and Conditioning

D. Leventhal, A. S. Lewis
School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853
{leventhal@orie.cornell.edu, aslewis@orie.cornell.edu, http://people.orie.cornell.edu/~aslewis}

We study randomized variants of two classical algorithms: coordinate descent for systems of linear equations and iterated projections for systems of linear inequalities. Expanding on a recent randomized iterated projection algorithm of Strohmer and Vershynin (Strohmer, T., R. Vershynin. 2009. A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.* **15** 262–278) for systems of linear equations, we show that, under appropriate probability distributions, the linear rates of convergence (in expectation) can be bounded in terms of natural linear-algebraic condition numbers for the problems. We relate these condition measures to distances to ill-posedness and discuss generalizations to convex systems under metric regularity assumptions.

Key words: coordinate descent; linear constraint; condition number; randomization; error bound; iterated projections; averaged projections; distance to ill-posedness; metric regularity
MSC2000 subject classification: Primary: 15A12, 15A39, 65F10, 90C25
OR/MS subject classification: Primary: programming, nonlinear, unconstrained
History: Received August 8, 2008; revised April 12, 2010, and May 25, 2010. Published online in *Articles in Advance* August 4, 2010.

1. Introduction. The condition number of a problem measures the sensitivity of a solution to small perturbations in its input data. For many problems that arise in numerical analysis, there is often a simple relationship between the condition number of a problem instance and the distance to the set of ill-posed problems—those problem instances whose condition numbers are infinite (Demmel [10]). For example, with respect to the problem of inverting a matrix A, it is known (see Horn and Johnson [21], for example) that if A is perturbed to $A+E$ for a sufficiently small matrix E, then

$$\frac{\|(A+E)^{-1}-A^{-1}\|}{\|A^{-1}\|} \le \|A^{-1}\|\|E\| + O(\|E\|^2).$$

Thus a condition measure for this problem may be taken as $\|A^{-1}\|$. The classical Eckart-Young theorem (Eckart and Young [15]) relates this condition measure to the distance to ill-posedness.

THEOREM 1.1 (ECKART-YOUNG). *For any nonsingular matrix, A,*

$$\min_E\{\|E\|\colon A+E \text{ is singular}\} = \frac{1}{\|A^{-1}\|}.$$

We are typically concerned with relative condition numbers as introduced by Demmel [10]. For example, with respect to the problem of matrix inversion, the relative condition number is $k(A) := \|A\|\|A^{-1}\|$, the commonly used condition measure.

Condition numbers are also important from an algorithmic perspective. In the example of matrix inversion, the sensitivity of a problem under perturbations could be relevant due to errors in either the initial problem data or accumulated rounding error. Hence it is natural that condition numbers affect algorithm speed. For example, in the context of linear programming, Renegar defined a condition measure based on the distance to ill-posedness (Renegar [40])—similar to the Eckart-Young result—and showed its effect on the convergence rate of interior point methods (Renegar [41]).

As another example, consider the problem of finding a solution to the system $Ax=b$, where A is a positive-definite matrix. It was shown by Akaike [1] that the steepest descent method is linearly convergent with rate $((k(A)-1)/(k(A)+1))^2$ and that this bound is asymptotically tight for almost all choices of initial iterates. Similarly, it is well known (see Golub and van Loan [16]) that the conjugate gradient method applied to the same problem is also linearly convergent with rate $(\sqrt{k(A)}-1)/(\sqrt{k(A)}+1)$.

From a computational perspective, a related and important area of study is that of error bounds. Given a subset of a Euclidean space, an error bound is an inequality that bounds the distance from a test vector to the specified subset in terms of some residual function that is typically easy to compute. An error bound can thus be used both as part of a stopping rule during implementation of an algorithm as well as an aide in proving algorithmic

convergence. A comprehensive survey of error bounds for a variety of problems arising in optimization can be found in Pang [35].

With regard to the problem of solving a nonsingular linear system $Ax = b$, one connection between condition measures and error bounds is immediate. Let x^* be a solution to the system and let x be any other vector. Then

$$\|x - x^*\| = \|A^{-1}A(x - x^*)\| = \|A^{-1}(Ax - b)\| \leq \|A^{-1}\|\|Ax - b\|,$$

so the distance to the solution set is bounded by a constant multiple of the residual vector, $\|Ax - b\|$, and this constant is the same one that appears in the context of conditioning and distance to infeasibility. As we discuss later, these relationships are not necessarily confined to systems of linear equations.

Error bounds often make a prominent appearance in algorithmic convergence proofs. In particular, a comprehensive series of papers by Luo and Tseng in the early 1990s showed the unifying power of the error bound idea in demonstrating linear convergence for a wide variety of algorithms applied to a rich class of problems. Particularly relevant for this current work is Luo and Tseng's elegant and novel proof of linear convergence of the coordinate descent method for smooth convex minimization (see Luo and Tseng [26]), along with their discussion of duality-based interpretations of this convergence, one example of which is the iterated projection algorithm for linear equations. This approach generalizes broadly, for example, to gradient projection and reduced gradient schemes and matrix splitting algorithms (see Luo and Tseng [28, 29]), to nonconvex minimization (see Luo and Tseng [31]), to composite and linearly constrained convex minimization (see Luo and Tseng [27, 30]), and to variational inequalities (see Tseng [44]).

In the current work, as in Luo and Tseng's development, we are interested in the linear convergence of some basic algorithms and make fundamental use of the error bound idea. Our work differs from that development in two ways: first, we aim to quantify the convergence rate explicitly in terms of *natural linear-algebraic condition numbers*, and secondly, our (very simple) derivations of these explicit rates involve *randomized* versions of the algorithms. These randomized methods are motivated by a recent iterated projection scheme for systems of linear equations due to Strohmer and Vershynin [43].

The rest of the paper is organized as follows. In §2, we define some notation used throughout the rest of this paper. In §3, we consider the problem of solving a linear system $Ax = b$ and show that a randomized coordinate descent scheme, implemented according to a specific probability distribution, is linearly convergent with a rate expressible in terms of traditional conditioning measures. In §4, we build upon the work of Strohmer and Vershynin [43] by considering randomized iterated projection algorithms for linear inequality systems. In particular, we show how randomization can provide convergence rates in terms of the traditional Hoffman error bound (Hoffman [20]) as well as in terms of Renegar's distance to infeasibility (Renegar [39]). We remark, as observed by Luo and Tseng [26], that since iterated projection algorithms can be interpreted, via duality, as coordinate descent schemes, the Strohmer-Vershynin randomized projection method for linear equations is simply a version of the randomized coordinate descent scheme of §3, but since the convergence proofs are so simple and intuitive, we give both. In §5, we consider randomized iterated projection algorithms for general convex sets and, under appropriate metric regularity assumptions, obtain local convergence rates in terms of the modulus of regularity.

Classical deterministic versions of the simple algorithms we consider here have been widely studied, in part due to the extreme simplicity of each iteration. Especially after Luo and Tseng's work, the fact of their linear convergence is well known. However, as remarked on linear systems of equations in Strohmer and Vershynin [43], randomized versions are interesting for several reasons. The randomized iterated projection method for linear equations from which this work originated may have some practical promise, even compared with conjugate gradients; see, for example, Strohmer and Vershynin [43]. Furthermore, from a theoretical perspective (our emphasis here) randomization provides a framework for simplifying the convergence analysis, allowing easy bounds on the rates of linear convergence in terms of natural linear-algebraic condition measures, such as relative condition numbers, Hoffman constants, and the modulus of metric regularity.

2. Notation. On the Euclidean space $\mathbf{R}^n$, we denote the Euclidean norm by $\|\cdot\|$. Let e_i denote the column vector with a one in the ith position and zeros elsewhere.

We consider m-by-n real matrices A. We denote the set of rows of A by $\{a_1^T, \ldots, a_m^T\}$ and the set of columns by $\{A_1, \ldots, A_n\}$. The *spectral norm* of A is the quantity $\|A\|_2 := \max_{\|x\|=1} \|Ax\|$, and the *Frobenius norm* is $\|A\|_F := \sum_{i,j} a_{ij}^2$. Additionally, these norms satisfy

$$\|A\|_F \leq \sqrt{n}\|A\|_2. \tag{1}$$

For an arbitrary matrix, A, let $\|A^{-1}\|_2$ be the smallest constant M such that $\|Ax\|_2 \geq 1/M\|x\|_2$ for all vectors x. In the case $m \geq n$, if A has singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, then M can also be expressed as the reciprocal of the minimum singular value σ_n, and, if A is invertible, this quantity equals the spectral norm of A^{-1}.

The *relative condition number* of A is the quantity $k(A) := \|A\|_2\|A^{-1}\|_2$; related to this is the *scaled condition number* introduced by Demmel [11], defined by $\kappa(A) := \|A\|_F\|A^{-1}\|_2$. From this, it is easy to verify (using the singular value decomposition, for example) the following relationship between condition numbers:

$$1 \leq \frac{\kappa(A)}{\sqrt{n}} \leq k(A). \tag{2}$$

Now suppose the matrix A is n-by-n symmetric and positive-definite. The *energy norm* (or A-norm), denoted $\|\cdot\|_A$, is defined by $\|x\|_A := \sqrt{x^T Ax}$. The inequality

$$\|x\|_A^2 \leq \|A^{-1}\|_2 \cdot \|Ax\|^2 \quad \text{for all } x \in \mathbf{R}^n \tag{3}$$

is useful later. Furthermore, if A is simply positive-semidefinite, we can generalize inequality (3) as

$$x^T Ax \leq \frac{1}{\underline{\lambda}(A)}\|Ax\|^2, \tag{4}$$

where $\underline{\lambda}(A)$ is the smallest nonzero eigenvalue of A. We denote the trace of A by $\mathrm{tr}A$: It satisfies the inequality

$$\|A\|_F \geq \frac{\mathrm{tr}A}{\sqrt{n}}. \tag{5}$$

Given a nonempty closed convex set S, let $P_S(x)$ be the projection of x onto S: that is, $P_S(x)$ is the vector y that is the optimal solution to $\min_{z\in S}\|x-z\|_2$. Additionally, define the distance from x to a set S by

$$d(x,S) = \min_{z\in S}\|x-z\|_2 = \|x - P_S(x)\|. \tag{6}$$

The following useful inequality is standard:

$$\|y-x\|^2 - \|P_S(y)-x\|^2 \geq \|y-P_S(y)\|^2 \quad \text{for all } x \in S,\ y \in \mathbf{R}^n. \tag{7}$$

3. Randomized coordinate descent. Let A be an n-by-n symmetric positive-definite matrix. We consider a linear system of the form $Ax = b$, with solution $x^* = A^{-1}b$. We consider the equivalent problem of minimizing the strictly convex quadratic function

$$f(x) = \tfrac{1}{2}x^T Ax - b^T x,$$

and we note the standard relationship

$$f(x) - f(x^*) = \tfrac{1}{2}\|x-x^*\|_A^2. \tag{8}$$

Suppose our current iterate is x and we obtain a new iterate x_+ by performing an exact line search in the nonzero direction d: that is, x_+ is the solution to $\min_{x+Rd} f$. This gives us

$$x_+ = x + \frac{(b-Ax)^T d}{d^T Ad}d$$

and

$$f(x_+) - f(x^*) = \frac{1}{2}\|x_+ - x^*\|_A^2 = \frac{1}{2}\|x-x^*\|_A^2 - \frac{((Ax-b)^T d)^2}{2d^T Ad}. \tag{9}$$

One natural choice of a set of easily computable search directions is to choose d from the set of canonical unit vectors, $\{e_1, \ldots, e_n\}$. Note that, when using search direction e_i, we can compute the new point

$$x_+ = x + \frac{b_i - a_i^T x}{a_{ii}}e_i,$$

using only $2n+2$ arithmetic operations. If the search direction is chosen at each iteration by successively cycling through the set of coordinate directions, then the algorithm is known to be linearly convergent but with a rate not easily expressible in terms of typical matrix quantities (see Golub and van Loan [16] or Quarteroni et al. [38]). However, by choosing a coordinate direction as a search direction randomly according to an appropriate probability distribution, we can obtain a convergence rate in terms of the scaled or relative condition numbers. In considering the following algorithm, we will weaken our assumptions and merely require the matrix A to be positive-*semi*definite.

Algorithm: Randomized coordinate descent. Consider a system $Ax = b$ for an n-by-n nonzero symmetric positive-semidefinite matrix A, and let $x_0 \in \mathbf{R}^n$ be an arbitrary starting point. For $j = 0, 1, 2, \ldots,$ compute

$$x_{j+1} = x_j + \frac{b_i - a_i^T x_j}{a_{ii}} e_i,$$

where e_i is the ith canonical unit vector, and at each iteration j, the index i is chosen independently at random from the set $\{1, \ldots, n\}$, with distribution

$$P\{i = k\} = \frac{a_{kk}}{\mathrm{tr}A}.$$

Notice in the algorithm that the matrix A may be singular but that, nonetheless, $a_{ii} > 0$ almost surely. If A is merely positive-semidefinite, solutions of the system $Ax = b$ coincide with minimizers of the function f, and consistency of the system is equivalent to f being bounded below. We now have the following result.

THEOREM 3.1. *Consider a consistent system $Ax = b$ for an n-by-n nonzero symmetric positive-semidefinite matrix A, and define the corresponding objective and error by*

$$f(x) = \tfrac{1}{2} x^T A x - b^T x,$$
$$\delta(x) = f(x) - \min f.$$

Then the randomized coordinate descent algorithm is linearly convergent in expectation: indeed, for each iteration $j = 0, 1, 2, \ldots,$

$$\mathbf{E}[\delta(x_{j+1}) \mid x_j] \le \left(1 - \frac{\underline{\lambda}(A)}{\mathrm{tr}A}\right)\delta(x_j).$$

In particular, if A is positive-definite and $x^* = A^{-1}b$, we have the equivalent property

$$\mathbf{E}\left[\|x_{j+1} - x^*\|_A^2 \mid x_j\right] \le \left(1 - \frac{1}{\|A^{-1}\|_2 \mathrm{tr}A}\right)\|x_j - x^*\|_A^2.$$

Hence, the expected reduction in the squared error $\|x_j - x^*\|_A^2$ is at least a factor

$$1 - \frac{1}{\sqrt{n}\kappa(A)} \le 1 - \frac{1}{nk(A)}$$

at each iteration.

PROOF. We make basic use of Equation (8). Note that if coordinate direction e_i is chosen during iteration j, then Equation (9) shows

$$f(x_{j+1}) = f(x_j) - \frac{(b_i - a_i^T x_j)^2}{2a_{ii}}.$$

Hence, using

$$\mathbf{E}[f(x_{j+1}) \mid x_j] = f(x_j) - \sum_{i=1}^{n} \frac{a_{ii}}{\mathrm{tr}(A)} \frac{(b_i - a_i^T x_j)^2}{2a_{ii}},$$

we deduce

$$\mathbf{E}[f(x_{j+1}) \mid x_j] = f(x_j) - \frac{1}{2\mathrm{tr}A}\|Ax_j - b\|^2. \tag{10}$$

Using inequality (4) (with the vector x replaced by the vector $x_j - x^*$ for any solution x^*) and Equation (8), we easily verify

$$\tfrac{1}{2}\|Ax_j - b\|^2 \ge \underline{\lambda}(A)\delta(x_j),$$

and the first result follows. Applying Equation (8) provides the second result. The final result comes from applying inequalities (1), (2), and (5). □

Consider for a moment the case when the system $Ax = b$ is inconsistent. In that case, the quantity $\|Ax - b\|$ is bounded below by some strictly positive constant. Equation (10), therefore, implies the existence of a constant $\epsilon > 0$ such that

$$\mathbf{E}[f(x_{j+1}) \mid x_j] \le f(x_j) - \epsilon, \quad \text{for all } j.$$

We know $f(x_{j+1}) \le f(x_j)$. The description of the algorithm implies that, at each iteration, the probability that we observe $f(x_{j+1}) \le f(x_j) - \epsilon$ is at least some fixed positive constant. Hence $f(x_j) \downarrow -\infty$ almost surely.

The simple idea behind the proof of Theorem 3.1 is the main engine driving the remaining results in this paper. Fundamentally, the idea is to choose a probability distribution so that the expected distance to the solution from the new iterate is the distance to the solution from the old iterate minus some multiple of a residual. Then, using some type of error bound to bound the distance to a solution in terms of the residual, we obtain expected linear convergence of the algorithm.

Before we expand upon the previous result, it is worth considering the probabilistic consequences of linear convergence in expectation. Specifically, as demonstrated in the following theorem, this implies the iterates converge almost surely to the solution set in Theorem 3.1.

PROPOSITION 3.1. *Consider a constant $r \in [0, 1)$ and a sequence of nonnegative random variables $\{Y_j\}_{j\geq 0}$ satisfying*

$$Y_{j+1} \leq Y_j,$$
$$\mathbf{E}[Y_{j+1} \mid Y_j] \leq rY_j,$$

and consider that Y_0 is bounded above almost surely. Then $\lim_{j\to\infty} Y_j = 0$ almost surely.

PROOF. By assumption, Y_j is nonnegative and monotonically decreasing, implying that Y_j converges to some nonnegative random variable Y almost surely (see, for example, Billingsley [8]). Furthermore, we know that the conditional expectation of Y_j given Y_0 satisfies the following inequality:

$$\mathbf{E}[Y_{j+1} \mid Y_0] = \mathbf{E}[\mathbf{E}[Y_{j+1} \mid Y_j] \mid Y_0] \leq \mathbf{E}[rY_j \mid Y_0] = r\mathbf{E}[Y_j \mid Y_0]$$

by assumption. By induction, it follows that

$$\mathbf{E}[Y_j \mid Y_0] \leq r^j \ Y_0.$$

Finally, applying the dominated convergence theorem, it follows that

$$\mathbf{E}[Y \mid Y_0] = \mathbf{E}[\lim_j Y_j \mid Y_0] = \lim_j \mathbf{E}[Y_j \mid Y_0] \leq \lim_j r^j Y_0 = 0.$$

From $\mathbf{E}[Y \mid Y_0] = 0$ and $Y \geq 0$ almost surely, we can conclude that $Y = 0$ almost surely. □

Specifically, choosing S to be the set of minimizers of $f(x)$, as defined in Theorem 3.1, and letting $Y_j = \min_{s\in S}(x_j - s)^T A(x_j - s)$, we obtain that the iterates $\{x_j\}_{j\geq 0}$ generated by the randomized coordinate descent algorithm satisfy $Ax_j \to b$ almost surely.

Now let us consider the more general problem of finding a solution to a linear system $Ax = b$ where A is an $m \times n$ matrix. More generally, since the system might be inconsistent, we seek a "least squares solution" by minimizing the function $\|Ax - b\|^2$. The minimizers are exactly the solutions of the positive-semidefinite system $A^T Ax = A^T b$, to which we could easily apply the previous algorithm; however, as usual, we wish to avoid computing the new matrix $A^T A$ explicitly. Instead, we can proceed as follows.

Algorithm: Nonsymmetric randomized coordinate descent. Consider a linear system $Ax = b$ for a nonzero m-by-n matrix A. Let $x_0 \in \mathbf{R}^n$ be an arbitrary initial point and let $r_0 = b - Ax_0$ be the initial residual. For each $j = 0, 1, \ldots,$ compute

$$\alpha_j = \frac{A_i^T r_j}{\|A_i\|^2},$$
$$x_{j+1} = x_j + \alpha_j e_i,$$
$$r_{j+1} = r_j - \alpha_j A_i,$$

where, at each iteration j, the index i is chosen independently at random from the set $\{1, \ldots, n\}$, with distribution

$$P\{i = k\} = \frac{\|A_k\|^2}{\|A\|_F^2} \quad (k = 1, 2, \ldots, n).$$

(In the formula for α_j, notice by assumption that $A_i \neq 0$ almost surely.)

Note that the step size at each iteration can be obtained by directly minimizing the residual in the respective coordinate direction. However, the algorithm can also be viewed as the application of the algorithm for positive definite systems on the system of normal equations, $A^T Ax = A^T b$, without actually having to compute the matrix $A^T A$. Given the motivation of directly minimizing the residual, we would expect that the nonsymmetric randomized coordinate descent algorithm would converge to a least-squares solution, even in the case where the underlying system is inconsistent. The next result shows that this is, in fact, the case.

THEOREM 3.2. *Consider any linear system $Ax = b$, where the matrix A is nonzero. Define the least-squares residual and the error by*

$$f(x) = \tfrac{1}{2}\|Ax - b\|^2,$$
$$\delta(x) = f(x) - \min f.$$

Then the nonsymmetric randomized coordinate descent algorithm is linearly convergent in expectation to a least squares solution for the system: for each iteration $j = 0, 1, 2, \ldots,$

$$\mathbf{E}[\delta(x_{j+1}) \mid x_j] \le \left(1 - \frac{\lambda(A^TA)}{\|A\|_F^2}\right)\delta(x_j).$$

In particular, if A has full column rank, we have the equivalent property

$$\mathbf{E}[\|x_{j+1} - \hat{x}\|_{A^TA}^2 \mid x_j] \le \left(1 - \frac{1}{\kappa(A)^2}\right)\|x_j - \hat{x}\|_{A^TA}^2,$$

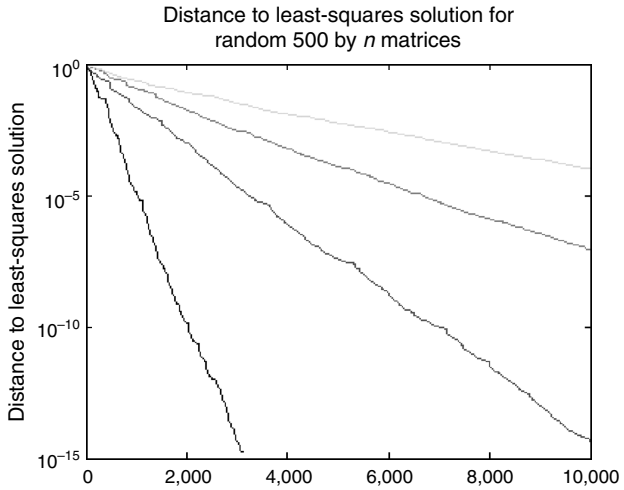
where $\hat{x} = (A^TA)^{-1}A^Tb$ is the unique least-squares solution.

PROOF. It is easy to verify, by induction on j, that the iterates x_j are exactly the same as the iterates generated by the randomized coordinate descent algorithm, when applied to the positive-semidefinite system $A^TAx = A^Tb$, and furthermore, that the residuals satisfy $r_j = b - Ax_j$ for all $j = 0, 1, 2, \ldots$. Hence, the results follow directly from Theorem 3.1. □

By the coordinate descent nature of this algorithm, once we have computed the initial residual r_0 and column norms $\{\|A_i\|^2\}_{i=1}^n$, we can perform each iteration in $O(n)$ time, just as in the positive-definite case. Specifically, this new iteration takes $4n + 1$ arithmetic operations, compared with $2n + 2$ for the positive-definite case.

For a computational example, we apply the nonsymmetric randomized coordinate descent algorithm to random $500 \times n$ matrices, where each element of A and b is an independent Gaussian random variable and we let n take values 50, 100, 150, and 200; see Figure 1.

Note that in the above examples, the theoretical bound provided by Theorem 3.2 predicts the actual behavior of the algorithm reasonably well.



n	Theoretical rate	Observed rate
50	0.98995	0.98911
100	0.99681	0.99670
150	0.99855	0.99837
200	0.99928	0.99908

FIGURE 1. Nonsymmetric randomized coordinate descent.

Note. Negative slopes decrease with n.

4. Randomized iterated projections. Iterated projection algorithms share important characteristics with coordinate descent algorithms and indeed may in some sense be considered dual to each other (Luo and Tseng [26]). Much convergence theory exists; a comprehensive overview on iterated projections can be found in Deutsch [12]. Randomized iterated projection methods have also been considered by many authors in a variety of mathematical settings. Convergence results for very general frameworks can be found in Amemiya and Ando [3], Bruck [9], Dye et al. [14], and Bauschke [4], among others. Results on randomized algorithms for convex feasibility problems in $\mathbf{R}^n$ have been further developed by Polyak [37] and Amaldi et al. [2], for example, including convergence theory for infeasible systems (see also Luo [25] and Luo and Tseng [32] for connections with the incremental gradient method). However, even for linear systems of equations, standard developments do not provide bounds on convergence rates in terms of natural linear-algebraic condition measures. By contrast, a recent paper of Strohmer and Vershynin [43] obtained a natural convergence rate via the following randomized iterated projection algorithm, which also provided the motivation for our work in the previous section.

Algorithm: Randomized iterated projections. Consider a linear system $Ax = b$ for a nonzero $m \times n$ matrix A. Let $x_0 \in \mathbf{R}^n$ be an arbitrary initial point. For each $j = 0, 1, \ldots,$ compute

$$x_{j+1} = x_j - \frac{a_i^T x_j - b_i}{\|a_i\|^2} a_i,$$

where, at each iteration j, the index i is chosen independently at random from the set $\{1, \ldots, m\}$, with distribution

$$P\{i = k\} = \frac{\|a_k\|^2}{\|A\|_F^2} \quad (k = 1, 2, \ldots, m).$$

Notice that the new iterate x_{j+1} is simply the orthogonal projection of the old iterate x_j onto the hyperplane $\{x\colon a_i^T x = b_i\}$. At first sight, the choice of probability distribution may seem curious, since we could rescale the equations arbitrarily without having any impact on the projection operations. However, following Strohmer and Vershynin [43], we emphasize that the aim is to understand linear convergence rates in terms of *linear-algebraic* condition measures associated with the original system, rather than in terms of *geometric* notions associated with the hyperplanes. This randomized algorithm has the following behavior.

Theorem 4.1 (Strohmer-Vershynin [43]). *Given any matrix A with full column rank, suppose the linear system $Ax = b$ has solution x^*. Then the randomized iterated projections algorithm converges linearly in expectation: for each iteration $j = 0, 1, 2, \ldots,$*

$$\mathbf{E}[\|x_{j+1} - x^*\|_2^2 \mid x_j] \le \left(1 - \frac{1}{\kappa(A)^2}\right) \|x_j - x^*\|_2^2.$$

As we remarked earlier, following observations of Luo and Tseng, iterated projections and coordinate descent are related via duality. Consider the problem of finding the least-squares solution to the system $Ax = b$ above. The dual of the corresponding optimization problem

$$\min_x \left\{ \tfrac{1}{2}\|x\|^2 \colon Ax = b \right\}$$

is, after a change of sign, the strictly convex quadratic minimization problem

$$\min_y \tfrac{1}{2}\|A^T y\|^2 - b^T y.$$

If the randomized coordinate descent scheme of the previous section generates the sequence of points $\{y_j\}$, then it is not hard to check that the points $x_j = A^T y_j$ comprise exactly the sequence generated by the Strohmer-Vershynin method, the randomized iterated projections algorithm, giving a proof of Theorem 4.1. However, since a direct proof is so simple and intuitive, we essentially reproduce it here, as a special case of a more general result.

We seek to generalize the above algorithm and convergence result to systems of linear inequalities of the form

$$\begin{cases} a_i^T x \le b_i & (i \in I_{\le}), \\ a_i^T x = b_i & (i \in I_{=}), \end{cases} \tag{11}$$

where the disjoint index sets $I_{\le}$ and $I_{=}$ partition the set $\{1, 2, \ldots, m\}$. To do so, staying with the techniques of the previous section, we need a corresponding error bound for a system of linear inequalities. First, given a vector $x \in \mathbf{R}^n$, define the vector x^+ by $(x^+)_i = \max\{x_i, 0\}$. Then a starting point for this subject is a result by Hoffman [20].

THEOREM 4.2 (HOFFMAN [20]). *For any right-hand side vector $b \in \mathbf{R}^m$, let S_b be the set of feasible solutions of the linear system (11). Then there exists a constant L, independent of b, with the following property*:

$$x \in \mathbf{R}^n \qquad \text{and} \qquad S_b \neq \varnothing \;\Rightarrow\; d(x, S_b) \leq L\|e(Ax-b)\|, \tag{12}$$

where the function e: $\mathbf{R}^m \to \mathbf{R}^m$ is defined by

$$e(y)_i = \begin{cases} y_i^+ & (i \in I_{\leq}), \\ y_i & (i \in I_{=}). \end{cases}$$

In the above result, each component of the vector $e(Ax-b)$ indicates the error in the corresponding inequality or equation. In particular $e(Ax-b)=0$ if and only if $x \in S_b$. Thus Hoffman's result provides a linear bound for the distance from a trial point x to the feasible region in terms of the size of the a posteriori error associated with x.

We call the minimum constant L such that property (12) holds, the *Hoffman constant* for the system (11). Several authors give geometric or algebraic meaning to this constant, or exact expressions for it, including Güler et al. [18], Ng and Zheng [34], Li [24], Ho and Tunçel [19], Zhang [45], and the survey of Pang [35]. In the case of linear equations (that is, $I_{\leq} = \varnothing$), an easy calculation using the singular value decomposition shows that the Hoffman constant is just the reciprocal of the smallest nonzero singular value of the matrix A, and hence equals $\|A^{-1}\|_2$ when A has full column rank.

For the problem of finding a solution to a system of linear inequalities, we consider a randomized algorithm generalizing the randomized iterated projections algorithm.

Algorithm: Randomized iterated projections for inequalities. Consider the system of inequalities (11). Let x_0 be an arbitrary initial point. For each $j = 0, 1, \ldots,$ compute

$$\beta_j = \begin{cases} (a_i^T x_j - b_i)^+ & (i \in I_{\leq}), \\ a_i^T x_j - b_i & (i \in I_{=}), \end{cases}$$

$$x_{j+1} = x_j - \frac{\beta_j}{\|a_i\|^2} a_i,$$

where, at each iteration j, the index i is chosen independently at random from the set $\{1, \ldots, m\}$, with distribution

$$P\{i = k\} = \frac{\|a_k\|^2}{\|A\|_F^2} \quad (k = 1, 2, \ldots, m).$$

In the above algorithm, notice that $\beta_j = e(Ax_j - b)_i$ and that x_{j+1} is just the orthogonal projection onto the halfspace or hyperplane defined by the constraint with index i. We can now generalize Theorem 4.1 as follows.

THEOREM 4.3. *Suppose the system (11) has nonempty feasible region S. Then the randomized iterated projections for inequalities algorithm converges linearly in expectation: for each iteration $j = 0, 1, 2, \ldots,$*

$$\mathbf{E}[d(x_{j+1}, S)^2 \mid x_j] \leq \left(1 - \frac{1}{L^2\|A\|_F^2}\right) d(x_j, S)^2,$$

where L is the Hoffman constant.

PROOF. Note that if the index i is chosen during iteration j, then it follows that

$$\begin{aligned} \|x_{j+1} - P_S(x_{j+1})\|_2^2 \leq \|x_{j+1} - P_S(x_j)\|_2^2 &= \left\| x_j - \frac{e(Ax_j - b)_i}{\|a_i\|^2} a_i - P_S(x_j) \right\|_2^2 \\ &= \|x_j - P_S(x_j)\|_2^2 + \frac{e(Ax_j - b)_i^2}{\|a_i\|^2} - 2\frac{e(Ax_j - b)_i}{\|a_i\|^2} a_i^T (x_j - P_S(x_j)). \end{aligned}$$

Note that $P_S(x_j) \in S$. Hence, if $i \in I_{\leq}$, then $a_i^T P_S(x_j) \leq b_i$, and $e(Ax_j - b)_i \geq 0$, so

$$e(Ax_j - b)_i a_i^T (x_j - P_S(x_j)) \geq e(Ax_j - b)_i (a_i^T x_j - b_i) = e(Ax_j - b)_i^2.$$

On the other hand, if $i \in I_=$, then $a_i^T P_S(x_j) = b_i$, so

$$e(Ax_j - b)_i a_i^T (x_j - P_S(x_j)) = e(Ax_j - b)_i (a_i^T x_j - b_i) = e(Ax_j - b)_i^2.$$

Putting these two cases together with the previous inequality shows

$$d(x_{j+1}, S)^2 \le d(x_j, S)^2 - \frac{e(Ax_j - b)_i^2}{\|a_i\|^2}.$$

Taking the expectation with respect to the specified probability distribution, it follows that

$$\mathbf{E}[d(x_{j+1}, S)^2 \mid x_j] \le d(x_j, S)^2 - \frac{\|e(Ax_j - b)\|^2}{\|A\|_F^2},$$

and the result now follows by the Hoffman bound. □

Since Hoffman's bound is not independent of the scaling of the matrix A, it is not surprising that a normalizing constant such as the $\|A\|_F^2$ term appears in the result.

For a computational example, we consider linear inequality systems $Ax \le b$, where the elements of A are independent standard Gaussian random variables and b is chosen so that the resulting system has a nonempty interior. We consider matrices A that are $500 \times n$, letting n take values 50, 100, 150, and 200. We then apply the randomized iterated projections for inequalities algorithm to these problems and observe the computational results in Figure 2.

Another natural conditioning measure for linear inequality systems is the distance to infeasibility, defined by Renegar [39] and shown (Renegar [41]) to govern the convergence rate of interior point methods for linear programming. It is interesting, therefore, from a theoretical perspective, to obtain a linear convergence rate for iterated projection algorithms in terms of this condition measure as well. For simplicity, we concentrate on the inequality case. To begin, let us recall the following results.

The *distance to infeasibility* (Renegar [39]) for the system $Ax \le b$ is the number

$$\mu = \inf\{\max\{\|\Delta A\|_2, \|\Delta b\|\}\colon (A + \Delta A)x \le b + \Delta b \text{ is infeasible}\}.$$

THEOREM 4.4 (RENEGAR [39, THEOREM 1.1]). *Consider the system $Ax \le b$. Suppose the distance to infeasibility $\mu > 0$. Then there exists a point $\hat{x}$ in the feasible region S satisfying $\|\hat{x}\| \le \|b\|/\mu$. Furthermore, any point $x \in \mathbf{R}^n$ satisfies the inequality*

$$d(x, S) \le \frac{\max\{1, \|x\|\}}{\mu} \|(Ax - b)^+\|.$$

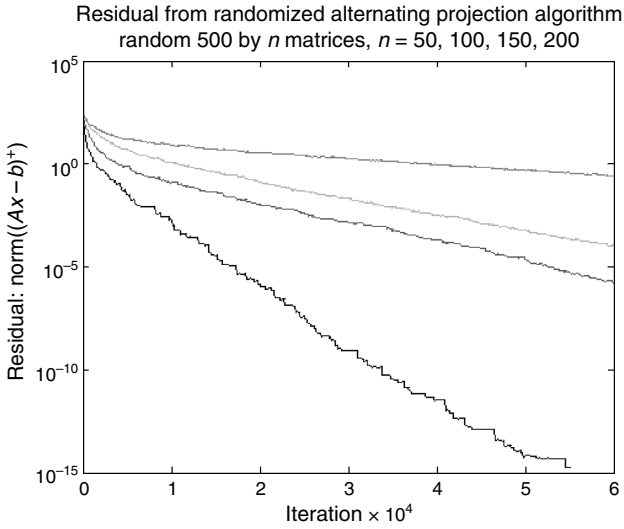


FIGURE 2. Randomized iterated projections for inequalities.

Note. Negative slopes decrease with n.

Using this, we can bound the linear convergence rate for the randomized iterated projections for inequalities algorithm in terms of the distance to infeasibility, as follows. As before, let $S = \{x\colon Ax \le b\}$. Suppose we start the algorithm at the initial point $x_0 = 0$ and notice that $\|x_j - \hat{x}\|$ is nonincreasing in j by inequality (7). Applying Theorem 4.4, we see that for all $j = 1, 2, \ldots,$

$$\|x_j\| \le \|\hat{x}\| + \|x_j - \hat{x}\| \le \|\hat{x}\| + \|x_0 - \hat{x}\| \le \frac{2\|b\|}{\mu},$$

so

$$d(x_j, S) \le \max\left\{\frac{1}{\mu}, \frac{2\|b\|}{\mu^2}\right\}\|(Ax_j - b)^+\|.$$

Using this inequality in place of Hoffman's bound in the proof of Theorem 4.3 gives

$$\mathbf{E}[d(x_{j+1}, S)^2 \mid x_j] \le \left[1 - \frac{1}{\|A\|_F^2(\max\{\frac{1}{\mu}, \frac{2\|b\|}{\mu^2}\})^2}\right] d(x_j, S)^2.$$

Although this bound may not be the best possible (and, in fact, it may not be as good as the bound provided in Theorem 4.3), this result simply emphasizes a relationship between algorithm speed and conditioning measures that appears naturally in other contexts. When equality constraints are present, an analogous argument applies, with the residual $(Ax - b)^+$ replaced by $e(Ax - b)$, as in Hoffman's theorem (Theorem 4.2.). In the next section, we proceed with these ideas in a more general framework.

5. Metric regularity and local convergence. The previous section concerned global rates of linear convergence. If, instead, we are interested in *local* rates, we can re-examine a generalization of our problem through an alternative perspective of set-valued mappings. Consider a set-valued mapping $\Phi\colon \mathbf{R}^n \rightrightarrows \mathbf{R}^m$ and the problem of solving the associated constraint system of the form $b \in \Phi(x)$ for the unknown vector x. For example, finding a feasible solution to $Ax \le b$ is equivalent to finding an x such that

$$b \in Ax + \mathbf{R}^m_+. \tag{13}$$

Related to this is the idea of metric regularity of set-valued mappings. We say the set-valued mapping Φ is *metrically regular* at $\bar{x}$ for $\bar{b} \in \Phi(\bar{x})$ if there exists $\gamma > 0$ such that

$$d(x, \Phi^{-1}(b)) \le \gamma d(b, \Phi(x)) \quad \text{for all } (x, b) \text{ near } (\bar{x}, \bar{b}), \tag{14}$$

where $\Phi^{-1}(b) = \{x\colon b \in \Phi(x)\}$. Furthermore, the *modulus of regularity* is the infimum of all constants γ such that equation (14) holds. Metric regularity is strongly connected with a variety of ideas from variational analysis; a good background reference is Rockafellar and Wets [42].

Metric regularity generalizes the error bounds discussed in previous sections at the expense of only guaranteeing a bound in local terms. For example, if Φ is a single-valued linear map, then the modulus of regularity (at any $\bar{x}$ for any $\bar{b}$) corresponds to the typical conditioning measure $\|\Phi^{-1}\|$ (with $\|\Phi^{-1}\| = \infty$ implying the map is not metrically regular), and if Φ is a smooth single-valued mapping, then the modulus of regularity is the reciprocal of the minimum singular value of the Jacobian, $\nabla\Phi(x)$. From an alternative perspective, metric regularity provides a framework for generalizing the Eckart-Young result on the distance to ill-posedness of linear mappings cited in Theorem 1.1. Specifically, if we define the *radius of metric regularity* at $\bar{x}$ for $\bar{b}$ for a set-valued mapping Φ between finite dimensional spaces by

$$\mathrm{rad}\Phi(\bar{x} \mid \bar{b}) = \inf\{\|E\|\colon \Phi + E \text{ not metrically regular at } \bar{x} \text{ for } \bar{b} + E(\bar{x})\},$$

where the infimum is over all linear mappings E, then one obtains the strikingly simple relationship (Dontchev et al. [13])

$$\text{modulus of regularity of } \Phi \text{ at } \bar{x} \text{ for } \bar{b} = \frac{1}{\mathrm{rad}\Phi(\bar{x} \mid \bar{b})}, \tag{15}$$

assuming only that Φ has a closed graph.

We will not be using the above result directly. Here, we simply use the fundamental idea of metric regularity that says that the distance from a point to the solution set, $d(x, \Phi^{-1}(b))$, is locally bounded by some constant times a "residual." For example, in the case where Φ corresponds to the linear inequality system (13), it follows that $d(b, \Phi(x)) = \|(Ax - b)^+\|$, implying that the modulus of regularity is in fact a global bound and equal to

the Hoffman bound. More generally, we wish to emphasize that metric regularity ties together several of the ideas from previous sections at the expense of those results now, only holding locally instead of globally.

In what follows, assume that all distances are Euclidean distances. We wish to consider how the modulus of regularity of Φ affects the convergence rate of iterated projection algorithms. We remark that linear convergence for iterated projection methods on convex sets has been very widely studied. For example, for two closed, convex sets, regularity conditions for linear convergence were proved by Bauschke and Borwein [6], generalizing results found in Gubin et al. [17]. Broad surveys of the topic for multiple sets can be found in Deutsch [12] and Bauschke and Borwein [7]. Our aim here is to observe, by analogy with previous sections, how randomization makes the linear convergence rate easy to interpret in terms of metric regularity.

Let $S_1, S_2, \ldots, S_m$ be closed convex sets in a Euclidean space $\mathbf{E}$ such that $\bigcap_i S_i \neq \varnothing$. Then, in a manner similar to Lewis et al. [23], we can endow the product space $\mathbf{E}^m$ with the inner product

$$\langle (u_1, u_2, \ldots, u_m), (v_1, v_2, \ldots, v_m) \rangle = \sum_{i=1}^{m} \langle u_i, v_i \rangle$$

and consider the set-valued mapping $\Phi\colon \mathbf{E} \to \mathbf{E}^m$ given by

$$\Phi(x) = (S_1 - x, S_2 - x, \ldots, S_m - x). \tag{16}$$

Then it clearly follows that $\bar{x} \in \bigcap_i S_i \Leftrightarrow 0 \in \Phi(\bar{x})$. Under appropriate regularity assumptions, we obtain the following local convergence result.

THEOREM 5.1. *Suppose the set-valued mapping Φ given by Equation (16) is metrically regular at $\bar{x}$ for 0 with regularity modulus γ. Let $\bar{\gamma}$ be any constant strictly larger than γ and let x_0 be any initial point sufficiently close to $\bar{x}$. Furthermore, suppose that $x_{j+1} = P_{S_i}(x_j)$ with probability $1/m$ for $i = 1, \ldots, m$. Then*

$$\mathbf{E}[d(x_{j+1}, S)^2 \mid x_j] \leq \left(1 - \frac{1}{m\bar{\gamma}^2}\right) d(x_j, S)^2.$$

PROOF. First, note that by inequality (7), the distance $\|x_j - \bar{x}\|$ is nonincreasing in j. Hence, if x_0 is sufficiently close to $\bar{x}$, then x_j is as well for all $j \geq 0$. Then, again using inequality (7) (applied to the set S_i), we have, for all points $x \in S \subset S_i$,

$$\|x_j - x\|^2 - \|x_j - P_{S_i}(x_j)\|^2 \geq \|P_{S_i}(x_j) - x\|^2.$$

Taking the minimum over $x \in S$, we deduce

$$d(x_j, S)^2 - \|x_j - P_{S_i}(x_j)\|^2 \geq d(P_{S_i}(x_j), S)^2.$$

Hence

$$\begin{aligned}
\mathbf{E}[d(x_{j+1}, S)^2 \mid x_j] &= \frac{1}{m} \sum_{i=1}^{m} d(P_{S_i}(x_j), S)^2 \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \left[d(x_j, S)^2 - d(x_j, S_i)^2 \right] \\
&= d(x_j, S)^2 - \frac{1}{m} \sum_{i=1}^{m} d(x_j, S_i)^2 \\
&= d(x_j, S)^2 - \frac{1}{m} d(0, \Phi(x_j))^2 \\
&\leq \left(1 - \frac{1}{m\bar{\gamma}^2}\right) d(x_j, S)^2,
\end{aligned}$$

using the definition of metric regularity. □

It is well known that metric regularity at $\bar{x}$ for 0 is a stronger assumption than necessary for the linear convergence of iterated projection schemes. For global convergence, "bounded linear regularity" suffices (Bauschke and Borwein [7, Theorem 5.7]). The above local convergence result is valid as long as Equation (14) holds for all points x near $\bar{x}$ with $\bar{b} = 0$ fixed, as opposed to the metric regularity assumption, which requires it to hold

for all b near $\bar{b}$ as well. For more on this distinction, see the comparison of metric regularity and bounded linear regularity in Bauschke [5, Remark 3.8] and the discussion of metric subregularity in Leventhal [22]. Nonetheless, for the purposes of the current discussion, we maintain our focus on metric regularity, which is a more robust general-purpose variational-analytic tool, quantifiable by calculus-like coderivative formulae (Rockafellar and Wets [42, Theorem 9.40]), and generalizing traditional condition measures such as the distance to infeasibility via the radius formula (15).

For a moment, let $m = 2$ and consider the sequence of iterates $\{x_j\}_{j\geq 0}$ generated by the randomized iterated projection algorithm. By idempotency of the projection operator, there is no benefit to projecting onto the same set in two consecutive iterations, so the subsequence consisting of different iterates corresponds exactly to that of the nonrandomized iterated projection algorithm. In particular, if $x_j \in S_1$, then

$$d(P_{S_2}(x_j), S)^2 \leq d(x_j, S)^2 - d(x_j, S_2)^2 = d(x_j, S)^2 - [d(x_j, S_2)^2 + d(x_j, S_1)^2],$$

since $d(x_j, S_1) = 0$. This gives us the following corollary, which also follows through more standard deterministic arguments.

COROLLARY 5.1. *If Φ is metrically regular at $\bar{x}$ for 0 with regularity modulus γ and $\bar{\gamma}$ is larger than γ, then for x_0 sufficiently close to $\bar{x}$, the 2-set iterated projection algorithm is linearly convergent and*

$$d(x_{j+1}, S)^2 \leq \left(1 - \frac{1}{\bar{\gamma}^2}\right) d(x_j, S)^2.$$

Note that this is very similar to a result of Bauschke and Borwein [6] under a slightly different regularity assumption. Furthermore, consider the following refined version of the m-set randomized algorithm. Suppose $x_0 \in S_1$ and $i_0 = 1$. Then for $j = 1, 2, \ldots,$ let i_j be chosen uniformly at random from $\{1, \ldots, m\}\backslash\{i_{j-1}\}$ and $x_{j+1} = P_{S_{i_j}}(x_j)$. Then we obtain the following similar result.

COROLLARY 5.2. *If Φ is metrically regular at $\bar{x}$ for 0 with regularity modulus γ and $\bar{\gamma}$ is larger than γ, then for x_0 sufficiently close to $\bar{x}$, the refined m-set randomized iterated projection algorithm is linearly convergent in expectation and*

$$\mathbf{E}[d(x_{j+1}, S)^2 \mid x_j,\ i_{j-1}] \leq \left(1 - \frac{1}{(m-1)\bar{\gamma}^2}\right) d(x_j, S)^2.$$

A simple but effective product space formulation by Pierra [36] has the benefit of reducing the problem of finding a point in the intersection of finitely many sets to the problem of finding a point in the intersection of two sets. Using the notation above, we consider the closed set in the product space given by

$$T = S_1 \times S_2 \times \cdots \times S_m$$

and the subspace

$$L = \{Ax\colon x \in E\},$$

where the linear mapping $A\colon \mathbf{E} \to \mathbf{E}^m$ is defined by $Ax = (x, x, \ldots, x)$. Again, notice that $\bar{x} \in \bigcap_i S_i \Leftrightarrow (\bar{x}, \ldots, \bar{x}) \in T \cap L$. One interesting aspect of this formulation is that projections in the product space $\mathbf{E}^m$ relate back to projections in the original space $\mathbf{E}$ by

$$\begin{aligned} (z_1, \ldots, z_m) \in P_T(Ax) \quad &\Leftrightarrow \quad z_i \in P_{S_i}(x) \quad (i = 1, 2, \ldots, m), \\ (P_L(z_1, \ldots, z_m))_i \quad &= \quad \frac{1}{m}(z_1 + z_2 + \cdots + z_m) \quad (i = 1, \ldots, m). \end{aligned}$$

This formulation provides a nice analytical framework. We can use the above equivalence of projections to consider the *method of averaged projections* directly, defined as follows.

Algorithm: Averaged projections. Let $S_1, \ldots, S_m \subseteq E$ be nonempty closed convex sets. Let x_0 be an initial point. For $j = 1, 2, \ldots,$ let

$$x_{j+1} = \frac{1}{m} \sum_{i=1}^{m} P_{S_i}(x_j).$$

Simply put, at each iteration, the algorithm projects the current iterate onto each set individually and takes the average of those projections as the next iterate. In the product space formulation, this is equivalent to $x_{j+1} = P_L(P_T(x_j))$. Expanding on the work of Pierra [36], additional convergence theory for this algorithm has been examined by Bauschke and Borwein [6]. Under appropriate regularity conditions, the general idea is that convergence of the iterated projection algorithm for two sets implies convergence of the averaged projection algorithm for m sets. In a similar sense, we prove the following result in terms of randomized projections.

THEOREM 5.2. *Suppose $S = \bigcap_{i=1}^m S_i$ is nonempty. If the randomized projection algorithm of Theorem 5.1 is linearly convergent in expectation with rate α, then so is the averaged projections algorithm.*

PROOF. Let x_j be the current iterate, let x_{j+1}^{AP} be the new iterate in the method of averaged projections, and let x_{j+1}^{RP} be the new iterate in the method of uniformly randomized projections. Then note that

$$x_{j+1}^{\mathrm{AP}} = \frac{1}{m}\sum_{i=1}^m P_{S_i}(x_j) = \mathbf{E}[x_{j+1}^{\mathrm{RP}}].$$

By convexity of the S_i's, it follows that

$$d(x_{j+1}^{\mathrm{AP}}, S) = d(\mathbf{E}[x_{j+1}^{\mathrm{RP}} \mid x_j], S) \le \mathbf{E}[d(x_{j+1}^{\mathrm{RP}}, S) \mid x_j] \le \alpha d(x_j, S)$$

by Jensen's inequality. □

Hence, the method of averaged projections converges no more slowly than the method of uniformly random projections. In particular, under the assumptions of Theorem 5.1, the method of averaged projections converges with rate no larger than $1 - 1/m\bar{\gamma}^2$.

Acknowledgments. After the initial submission of this work, the authors learned of a very recent and related randomized coordinate descent scheme in Nesterov [33]. The authors thank two anonyomous referees, who substantially improved the presentation of this work, in particular pointing out the relevance of the work of Luo and Tseng [26] on error bounds.

References

[1] Akaike, H. 1959. On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. *Ann. Inst. Statist. Math.* **11** 1–16.
[2] Amaldi, E., P. Belotti, R. Hauser. 2005. Randomized relaxation methods for the maximum feasible subsystem problem. M. Jünger, V. Kaibel, eds. *Integer Programming and Combinatorial Optimization*, *Lecture Notes in Computer Science*, Vol. 3509. Springer-Verlag, Berlin, 249–264.
[3] Amemiya, I., T. Ando. 1965. Convergence of random products of contractions in Hilbert space. *Acta. Sci. Math.* **26** 239–244.
[4] Bauschke, H. H. 1995. A norm convergence result on random products of relaxed projections in Hilbert space. *Trans. Amer. Math. Soc.* **347** 1365–1373.
[5] Bauschke, H. H. 2001. Projection algorithms: Results and open problems. D. Butnariu, Y. Censor, S. Reich, eds. *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, Elsevier, Amsterdam, 11–22.
[6] Bauschke, H. H., J. M. Borwein. 1993. On the convergence of von Neumann's alternating projection algorithm for two sets. *Set-Valued Anal.* **1** 185–212.
[7] Bauschke, H. H., J. M. Borwein. 1996. On projection algorithms for solving convex feasibility problems. *SIAM Rev.* **38** 367–426.
[8] Billingsley, P. 1986. *Probability and Measure*. John Wiley & Sons, New York.
[9] Bruck, R. E. 1982. Random products of contractions in metric and Banach spaces. *J. Math. Anal. Appl.* **88** 319–332.
[10] Demmel, J. W. 1987. On condition numbers and the distance to the nearest ill-posed problem. *Numerische Mathematik* **51** 251–289.
[11] Demmel, J. W. 1988. The probability that a numerical analysis problem is difficult. *Math. Comput.* **50** 449–480.
[12] Deutsch, F. 2001. *Best Approximation in Inner Product Spaces*. Springer-Verlag, New York.
[13] Dontchev, A. L., A. S. Lewis, R. T. Rockafellar. 2003. The radius of metric regularity. *Trans. Amer. Math. Soc.* **355** 493–517.
[14] Dye, J., M. A. Khamsi, S. Reich. 1991. Random products of contractions in Banach spaces. *Trans. Amer. Math. Soc.* **325** 87–99.
[15] Eckart, C., G. Young. 1936. The approximation of one matrix by another of low rank. *Psychometrika* **1** 211–218.
[16] Golub, G., C. van Loan. 1996. *Matrix Computations*. Johns Hopkins University Press, Baltimore.
[17] Gubin, L. G., B. T. Polyak, E. V. Raik. 1967. The method of projections for finding the common point of convex sets. *U.S.S.R. Comput. Math. Math. Phys.* **7** 1–24.
[18] Güler, O., A. J. Hoffman, U. Rothblum. 1995. Approximations to solutions to systems of linear inequalities. *SIAM J. Matrix Anal. Appl.* **16** 688–696.
[19] Ho, J. C. K., L. Tunçel. 2002. Reconciliation of various complexity and condition measures for linear programming problems and a generalization of Tardos' theorem. *Foundations Comput. Math.: Proc. Smalefest* 2000, World Scientific, Singapore, 93–148.
[20] Hoffman, A. J. 1952. On approximate solutions of systems of linear inequalities. *J. Res. National Bureau Standards* **49** 263–265.
[21] Horn, R., C. Johnson. 1999. *Matrix Analysis*. Cambridge University Press, Cambridge, UK.
[22] Leventhal, D. 2009. Metric subregularity and the proximal point method. *J. Math. Anal. Appl.* **360** 681–688.
[23] Lewis, A. S., D. R. Luke, J. Malick. 2009. Local convergence for alternating and averaged nonconvex projections. *Foundations Comput. Math.* **9** 485–513.
[24] Li, W. 1993. The sharp Lipschitz constants for feasible and optimal solutions of a perturbed linear program. *Linear Algebra Appl.* **187** 15–40.
[25] Luo, Z.-Q. 1991. On the convergence of the LMS algorithm with adaptive learning rate for linear feedback networks. *Neural Comput.* **3** 226–245.
[26] Luo, Z.-Q., P. Tseng. 1992a. On the convergence of the coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.* **72** 7–35.

[27] Luo, Z.-Q., P. Tseng. 1992b. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM J. Control Optim.* **30** 408–425.
[28] Luo, Z.-Q., P. Tseng. 1992c. Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem. *SIAM J. Optim.* **2** 43–54.
[29] Luo, Z.-Q., P. Tseng. 1993a. Error bound and reduced gradient projection algorithms for convex minimization over a polyhedral set. *SIAM J. Optim.* **3** 43–59.
[30] Luo, Z.-Q., P. Tseng. 1993b. On the convergence rate of dual ascent methods for strictly convex minimization. *Math. Oper. Res.* **18** 846–867.
[31] Luo, Z.-Q., P. Tseng. 1993c. Error bounds and convergence analysis of feasible descent methods: A general approach. *Ann. Oper. Res.* **46** 157–178.
[32] Luo, Z.-Q., P. Tseng. 1994. Analysis of an approximate gradient projection method with applications to the back propagation algorithm. *Optim. Methods Software* **4** 85–101.
[33] Nesterov, Y. 2010. Efficiency of coordinate descent methods on huge-scale optimization problems. CORE Discussion Paper 2010/2, Université Catholique de Louvain, Center for Operations Research and Economics, Louvain-la-Neuve, Belgium.
[34] Ng, K. F., X. Y. Zheng. 2004. Hoffman's least error bounds for linear inequalities. *J. Global Optim.* **30** 391–403.
[35] Pang, J.-S. 1997. Error bounds in mathematical programming. *Math. Programming* **79** 299–332.
[36] Pierra, G. 1984. Decomposition through formalization in a product space. *Math. Programming* **28** 96–115.
[37] Polyak, B. T. 2001. Random algorithms for solving convex inequalities. D. Butnariu, Y. Censor, S. Reich, eds. *Inherently Parallel Algorithms in Feasibility and Other Applications*. Elsevier, Amsterdam.
[38] Quarteroni, A., R. Sacco, F. Saleri. 2007. *Numerical Mathematics*. Springer-Verlag, New York.
[39] Renegar, J. 1994. Perturbation theory for linear programming. *Math. Programming* **65** 73–91.
[40] Renegar, J. 1995a. Incorporating conditions measures into the complexity theory of linear programming. *SIAM J. Optim.* **5** 506–524.
[41] Renegar, J. 1995b. Linear programming, complexity theory and elementary functional analysis. *Math. Programming* **70** 279–351.
[42] Rockafellar, R. T., R. J.-B. Wets. 1998. *Variational Analysis*. Springer-Verlag, Berlin.
[43] Strohmer, T., R. Vershynin. 2009. A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.* **15** 262–278.
[44] Tseng, P. 1995. On linear convergence of iterative methods for the variational inequality problem. *J. Comput. Appl. Math.* **60** 237–252.
[45] Zhang, S. 2000. Global error bounds for convex conic problems. *SIAM J. Optim.* **10** 836–851.