Randomized Iterative Methods for Linear Systems

Robert M. Gower* Peter Richtárik†

June 16, 2015

Abstract

We develop a novel, fundamental and surprisingly simple *randomized iterative method* for solving consistent linear systems. Our method has five different but equivalent interpretations: sketch-and-project, constrain-and-approximate, random intersect, random linear solve and random fixed point. By varying its two parameters—a positive definite matrix (defining geometry), and a random matrix (sampled in an i.i.d. fashion in each iteration)—we recover a comprehensive array of well known algorithms as special cases, including the randomized Kaczmarz method, randomized Newton method, randomized coordinate descent method and random Gaussian pursuit. We naturally also obtain variants of all these methods using blocks and importance sampling. However, our method allows for a much wider selection of these two parameters, which leads to a number of new specific methods. We prove exponential convergence of the *expected norm of the error* in a single theorem, from which existing complexity results for known variants can be obtained. However, we also give an exact formula for the evolution of the expected iterates, which allows us to give *lower bounds* on the convergence rate.

Keywords: linear systems, stochastic methods, iterative methods, randomized Kaczmarz, randomized Newton, randomized coordinate descent, random pursuit, randomized fixed point.

1 Introduction

The need to solve linear systems of equations is ubiquitous in essentially all quantitative areas of human endeavour, including industry and science. Linear systems are a central problem in numerical linear algebra, and play an important role in computer science, mathematical computing, optimization, signal processing, engineering, numerical analysis, computer vision, machine learning, and many other fields.

For instance, in the field of large scale optimization, there is a growing interest in inexact and approximate Newton-type methods for [5, 9, 1, 33, 32, 10], which can benefit from fast subroutines for calculating approximate solutions of linear systems. In machine learning, applications arise for the problem of finding optimal configurations in Gaussian Markov Random Fields [26], in graph-based semi-supervised learning and other graph-Laplacian problems [2], least-squares SVMs, Gaussian processes and more.

In a large scale setting, direct methods are generally not competitive when compared to iterative approaches. While classical iterative methods are deterministic, recent breakthroughs suggest that

*School of Mathematics, The University of Edinburgh, United Kingdom (e-mail: gowerrobert@gmail.com)
†School of Mathematics, The University of Edinburgh, United Kingdom (e-mail: peter.richtarik@ed.ac.uk)

randomization can play a powerful role in the design and analysis of efficient algorithms [31, 15, 18, 7, 34, 14, 17, 24] which are in many situations competitive or better than existing deterministic methods.

1.1 Contributions

Given a real matrix $A \in \mathbb{R}^{m\times n}$ and a real vector $b \in \mathbb{R}^m$, in this paper we consider the linear system

$$Ax = b. \tag{1}$$

We shall assume throughout that the system is consistent: there exists x^* for which $Ax^* = b$.

We now comment on the main contribution of this work:

1. **New method.** We develop a novel, fundamental, and surprisingly simple *randomized iterative method* for solving (1).

2. **Five equivalent formulations.** Our method allows for several seemingly different but nevertheless equivalent formulations. First, it can be seen as a *sketch-and-project* method, in which the system (1) is replaced by its *random sketch*, and then the current iterate is projected onto the solution space of the sketched system. We can also view it as a *constrain-and-approximate* method, where we constrain the next iterate to live in a particular random affine space passing through the current iterate, and then pick the point from this subspace which best approximates the optimal solution. We can also interpret the method as the iterative solution of a sequence of random (and simpler) linear equations, or as a *randomized fixed point method*. Finally, the method also allows for a simple geometrical interpretation: the new iterate is defined as the unique intersection of two random affine spaces which are orthogonal complements.

3. **Special cases.** These multiple viewpoints enrich our interpretation of the method, and enable us to draw previously unknown links between several existing algorithms. Our algorithm has two parameters, an $n \times n$ positive definite matrix B defining geometry of the space, and a random matrix S. Through combinations of these two parameters, in special cases our method recovers several well known algorithms. For instance, we recover the randomized Kaczmarz method of Strohmer and Vershyinin [31], randomized coordinate descent method of Leventhal and Lewis [15], random pursuit [21, 30, 29, 28] (with exact line search), and the stochastic Newton method recently proposed by Qu et al [24]. However, our method is more general, and leads to i) various generalizations and improvements of the aforementioned methods (e.g., block setup, importance sampling), and ii) completely new methods. Randomness enters our framework in a very general form, which allows us to obtain a *Gaussian Kaczmarz method*, *Gaussian descent*, and more.

4. **Complexity: general results.** When A has full column rank, our framework allows us to determine the complexity of these methods using a single analysis. Our main results are summarized in Table 1, where $\{x^k\}$ are the iterates of our method, Z is a random matrix dependent on the data matrix A, parameter matrix B and random parameter matrix S, defined as

$$Z \stackrel{\text{def}}{=} A^T S(S^T AB^{-1}A^T S)^{-1}S^T A, \tag{2}$$

$\mathbf{E}\left[x^{k+1}-x^*\right] = \left(I - B^{-1}\mathbf{E}\left[Z\right]\right)\mathbf{E}\left[x^k - x^*\right]$	Theorem 4.1
$\left\|\mathbf{E}\left[x^{k+1}-x^*\right]\right\|_B^2 \leq \rho^2 \ \cdot \ \left\|\mathbf{E}\left[x^k - x^*\right]\right\|_B^2$	Theorem 4.1
$\mathbf{E}\left[\left\|x^{k+1}-x^*\right\|_B^2\right] \leq \rho \ \cdot \ \mathbf{E}\left[\left\|x^k - x^*\right\|_B^2\right]$	Theorem 4.2

Table 1: Our main complexity results. The convergence rate is: $\rho = 1 - \lambda_{\min}(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2})$.

and $\|x\|_B \overset{\text{def}}{=} \sqrt{\langle x, x\rangle_B}$, where $\langle x, y\rangle_B \overset{\text{def}}{=} x^T B y$, for all $x, y \in \mathbb{R}^n$. We show that the convergence rate ρ is always bounded between zero and one. We also show that as soon as $\mathbf{E}\left[Z\right]$ is invertible (which can only happen if A has full column rank, which then implies that x^* is unique), we have $\rho < 1$, and the method converges. Besides establishing a bound involving the *expected norm of the error* (see last line of Table 1), we also obtain bounds involving the *norm of the expected error* (second line of Table 1). Studying the expected sequence of iterates directly is very fruitful, as it allows us to establish an *exact characterization* of the evolution of the expected iterates (see first line of Table 1) through a *linear fixed point iteration*.

Both of these theorems on the convergence of the error can be recast as iteration complexity bounds. For instance, using standard arguments, from Theorem 4.1 in Table 1 we observe that for a given $\epsilon > 0$ we have that

$$k \geq \frac{1}{1-\rho}\log\left(\frac{1}{\epsilon}\right) \quad \Rightarrow \quad \left\|\mathbf{E}\left[x^k - x^*\right]\right\|_B \leq \epsilon \left\|x^0 - x^*\right\|_B. \tag{3}$$

5. **Complexity: special cases.** Besides these generic results, which hold without any major restriction on the sampling matrix S (in particular, it can be either discrete or continuous), we give a specialized result applicable to discrete sampling matrices S (see Theorem 5.1). In the special cases for which rates are known, our analysis recovers the existing rates.

6. **Extensions.** Our approach opens up many avenues for further development and research. For instance, it is possible to extend the results to the case when A is not necessarily of full column rank. Furthermore, as our results hold for a wide range of distributions, new and efficient variants of the general method can be designed for problems of specific structure by fine-tuning the stochasticity to the structure. Similar ideas can be applied to design randomized iterative algorithms for finding the inverse of a very large matrix.

1.2 Background and Related Work

The Kaczmarz method dates back to the 30's [13]. Research into the Kaczmarz method was reignited by Strohmer and Vershynin [31], who proved that a randomized variant of Kaczmarz enjoys an exponential error decay (aka "linear convergence"). This has trigged much research into developing and analyzing randomized linear solvers.

Leventhal and Lewis [15] develop similar bounds for randomized coordinate descent type methods (closely related to Gauss-Seidel methods in the linear algebra literature) for solving positive

definite or least squares problem. They also extend the use of Randomized Kaczmarz (RK) for solving inequalities [15].

The RK method and its analysis have been extended to the least-squares problem [18, 34] and to a block variant [19, 20]. In [17] the authors extend the randomized coordinate descent and the RK methods for solving underdetermined systems. The authors of [17, 25] analyze side-by-side the randomized coordinate descent and RK method, for least-squares, using a convenient notation in order to point out their similarities. Our work takes the next step, by analyzing these, and many other methods, through a genuinely single analysis. Also in the spirit of unifying the analysis of different methods, in [22] the authors provide a unified analysis of iterative Schwarz methods and Kaczmarz methods.

The use of random Gaussian directions as search directions in zero-order (derivative-free) minimization algorithm was recently suggested [21]. More recently, Gaussian directions have been combined with exact and inexact line-search into a single *random pursuit* framework [28], and further utilized within a randomized variable metric method [29, 30].

2 One Algorithm in Five Disguises

Our method has *two parameters*: i) an $n \times n$ positive definite matrix B which is used to define the B-inner product and the induced B-norm by

$$\langle x, y\rangle_B \overset{\text{def}}{=} \langle Bx, y\rangle, \qquad \|x\|_B \overset{\text{def}}{=} \sqrt{\langle x, x\rangle_B}, \tag{4}$$

where $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product, and ii) a random matrix $S \in \mathbb{R}^{m\times q}$, to be drawn in an i.i.d. fashion at each iteration. As we shall see later, we will often consider setting $B = I$, $B = A$ (if A is positive definite) or $B = A^TA$ (if A is of full column rank). We stress that we do not restrict the number of columns of S; indeed, we even allow q to vary (and hence, q is a random variable).

2.1 Four Viewpoints

Starting from $x^k \in \mathbb{R}^n$, our method draws a random matrix S and uses it to generate a new point $x^{k+1} \in \mathbb{R}^n$. This iteration can be formulated in *five seemingly different but equivalent ways:*

1. Sketching Viewpoint: Sketch-and-Project. x^{k+1} is the nearest point to x^k which solves a *sketched* version of the original linear system:

$$\boxed{x^{k+1} \quad = \quad \arg\min_{x\in\mathbb{R}^n} \left\|x - x^k\right\|_B^2 \quad \text{subject to} \quad S^TAx = S^Tb} \tag{5}$$

This viewpoint arises very naturally. Indeed, since the original system (1) is assumed to be complicated, we replace it by a simpler system—a *random sketch* of the original system—whose solution set $S^TAx = S^Tb$ contains all solutions of the original system. However, this system will typically have many solutions, so in order to define a method, we need a way to select one of them. The idea is to try to preserve as much of the information learned so far as possible, as condensed in the current point x^k. Hence, we pick the solution which is closest to x^k.

2. Optimization Viewpoint: Constrain-and-Approximate. x^{k+1} is the best approximation of x^* in a random space passing through x^k:

$$\boxed{x^{k+1} \quad = \quad \arg\min_{x\in\mathbb{R}^n} \left\|x \ - x^*\right\|_B^2 \quad \text{subject to} \quad x = x^k + B^{-1}A^TSy, \quad y \text{ is free}} \tag{6}$$

The above step has the following interpretation[1]. We choose a random affine space containing x^k, and constrain our method to choose the next iterate from this space. We then do as well as we can on this space; that is, we pick x^{k+1} as the point which best approximates x^*. Note that x^{k+1} does not depend on which solution x^* is used in (6) (this can be best seen by considering the geometric viewpoint, discussed next).

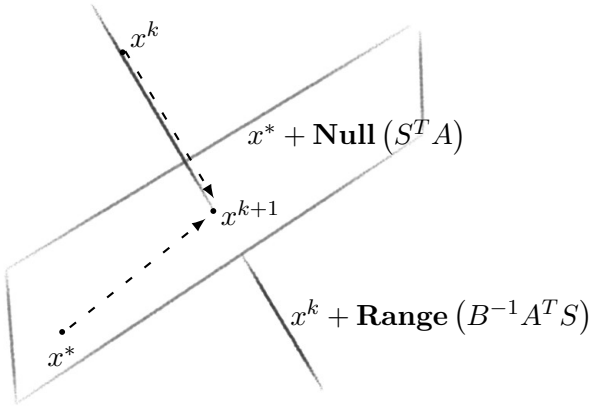


Figure 1: The geometry of our algorithm. The next iterate, x^{k+1}, arises as the intersection of two random affine spaces: $x^k + \mathbf{Range}\left(B^{-1}A^TS\right)$ and $x^* + \mathbf{Null}\left(S^TA\right)$ (see (7)). The spaces are orthogonal complements of each other with respect to the B-inner product, and hence x^{k+1} can equivalently be written as the projection, in the B-norm, of x^k onto $x^* + \mathbf{Null}\left(S^TA\right)$ (see (5)), or the projection of x^* onto $x^k + \mathbf{Range}\left(B^{-1}A^TS\right)$ (see (6)). The intersection x^{k+1} can also be expressed as the solution of a system of linear equations (see (8)). Finally, the new error $x^{k+1} - x^*$ is the projection, with respect to the B-inner product, of the current error $x^k - x^*$ onto $\mathbf{Null}\left(S^TA\right)$. This gives rise to a random fixed point formulation (see (12)).

[1]Formulation (6) is similar to the framework often used to describe Krylov methods [16, Chapter 1], which is

$$x^{k+1} \stackrel{\text{def}}{=} \arg\min_{x\in\mathbb{R}^n} \|x - x^*\|_B^2 \quad \text{s. t.} \quad x \in x^0 + \mathcal{K}_{k+1},$$

where $\mathcal{K}_{k+1} \subset \mathbb{R}^n$ is a $(k+1)$–dimensional subspace. Note that the constraint $x \in x^0 + \mathcal{K}_{k+1}$ is an affine space that contains x^0, as opposed to x^k in our formulation (6). The objective $\|x - x^*\|_B^2$ is a generalization of the residual, where $B = A^TA$ is used to characterize minimal residual methods [23, 27] and $B = A$ is used to describe the Conjugate Gradients method [12]. Progress from iteration to the next is guaranteed by using expanding nested search spaces at each iteration, that is, $\mathcal{K}_k \subset \mathcal{K}_{k+1}$. In our setting, progress is enforced by using x^k as the displacement term instead of x^0. This also allows for a simple recurrence for updating x^k to arrive at x^{k+1}, which facilitates the analyses of the method. In the Krylov setting, to arrive at an explicit recurrence, one needs to carefully select a basis for the nested spaces that allows for short recurrence.

3. Geometric viewpoint: Random Intersect. x^{k+1} is the (unique) intersection of two affine spaces:

$$\boxed{\{x^{k+1}\} \quad = \quad \left(x^* + \mathbf{Null}\left(S^T A\right)\right) \quad \bigcap \quad \left(x^k + \mathbf{Range}\left(B^{-1}A^T S\right)\right)} \tag{7}$$

First, note that the first affine space above does not depend on the choice of x^* from the set of optimal solutions of (1). A basic result of linear algebra says that the nullspace of an arbitrary matrix is the orthogonal complement of the range space of its transpose. Hence, whenever we have $h \in \mathbf{Null}\left(S^T A\right)$ and $y \in \mathbb{R}^q$, where q is the number of rows of S, then $\langle h, A^T S y\rangle = 0$. It follows that the two spaces in (7) are orthogonal complements with respect to the B-inner product and as such, they intersect at a unique point (see Figure 1).

4. Algebraic viewpoint: Random Linear Solve. x^{k+1} is the (unique) solution (in x) of a linear system (with variables x and y):

$$\boxed{x^{k+1} \quad = \quad \text{solution of the linear system} \quad S^T A x = S^T b, \quad x = x^k + B^{-1}A^T S y} \tag{8}$$

Note that this system is clearly equivalent to (7), and can alternatively be written as:

$$\begin{pmatrix} S^T A & 0 \\ B & -A^T S \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} S^T b \\ B x^k \end{pmatrix}. \tag{9}$$

Hence, our method reduces the solution of the (complicated) linear system (1) into a sequence of (hopefully simpler) random systems of the form (9).

The equivalence between these four viewpoints is formally captured in the next statement.

Theorem 2.1. *The four viewpoints are equivalent: they all produce the same (unique) point x^{k+1}.*

Proof. The proof is simple, and follows directly from the above discussion. In particular, see the caption of Figure 1. □

2.2 The Fifth Viewpoint

In order to arrive at the fifth viewpoint, we shall make the following additional assumption.

Assumption 2.1. *With probability 1, $S^T A$ has full row rank.*

Recalling that S is a $m \times q$ matrix (with q possibly being random), this assumption implies that $q = \mathbf{Rank}\left(S^T A\right) \leq n$. Moreover, note that

$$\dim\left(\mathbf{Range}\left(B^{-1}A^T S\right)\right) = q, \qquad \dim\left(\mathbf{Null}\left(S^T A\right)\right) = n - q. \tag{10}$$

For instance, Assumption 2.1 holds if S is a random column vector which with probability 1 stays away from the null space of A^T. If this assumption holds, then the matrix $S^T A B^{-1} A^T S$ is invertible with probability 1. If this is the case, then we can write x^{k+1} explicitly in closed form, which will be useful in the convergence analysis.

Theorem 2.2. *Under Assumption 2.1, our algorithm takes, with probability 1, the explicit form:*

$$\boxed{x^{k+1} \quad = \quad x^k - B^{-1}A^T S(S^T A B^{-1} A^T S)^{-1} S^T (A x^k - b)} \tag{11}$$

The above statement can be verified by a direct examination of any of the four equivalent formulations (5), (6), (7) and (8) (it is easiest to start with (8)).

We are now ready to describe the last interpretation of our algorithm.

5. Algebraic viewpoint: Random Fixed Point. Note that iteration (11) can be written as

$$\boxed{x^{k+1} - x^* \quad = \quad (I - B^{-1}Z)(x^k - x^*)} \tag{12}$$

where we used $Ax^* = b$ and

$$Z \stackrel{\text{def}}{=} A^T S(S^T A B^{-1} A^T S)^{-1} S^T A. \tag{13}$$

Matrix Z plays a central role in our analysis, and can be used to construct explicit projection matrices of the two projections depicted in Figure 1. We formalize this as Lemma 2.1.

Lemma 2.1. *With respect to the geometry induced by the B-inner product, we have that*

(i) $B^{-1}Z$ projects orthogonally onto the q–dimensional subspace $\mathbf{Range}\left(B^{-1}A^TS\right)$

(ii) $(I - B^{-1}Z)$ projects orthogonally onto $(n-q)$–dimensional subspace $\mathbf{Null}\left(S^TA\right)$.

Proof. By verifying that

$$(B^{-1}Z)^2 = B^{-1}Z, \tag{14}$$

we see that both $B^{-1}Z$ and $I - B^{-1}Z$ are projection matrices. Furthermore,

$$B^{-1}Z(B^{-1}A^TS) = B^{-1}A^TS, \qquad \text{and} \qquad B^{-1}Zy = 0, \quad \forall y \in \mathbf{Null}\left(S^TA\right),$$

shows that $B^{-1}Z$, and consequently $I - B^{-1}Z$, are orthogonal projections with respect to the B-inner product. □

This lemma also shows that $I - B^{-1}Z$ is a contraction with respect to the B-norm, so (12) can be seen as a *randomized fixed point method*. While $I - B^{-1}Z$ is not a strict contraction, under some assumptions on S, it will be a strict contraction in expectation. This ensures convergence.

3 Special Cases: Examples

Below we briefly mention how by selecting the parameters S and B of our method we recover several existing methods. The list is by no means comprehensive and merely serves the purpose of an illustration of the flexibility of our algorithm. All the associated complexity results we present in this section, can be recovered from Theorem 5.1, presented later in Section 5.

3.1 The One Step Method

When S is an $m \times m$ invertible matrix with probability one, then the system $S^TAx = S^Tb$ is equivalent to solving $Ax = b$, thus the solution to (5) must be $x^{k+1} = x^*$, independently of matrix B. Our convergence theorems also predict this one step behaviour, since $\rho = 0$ (see Table 1).

3.2 Randomized Kaczmarz

If we choose $S = e^i$ (unit coordinate vector in $\mathbb{R}^m$) and $B = I$ (the identity matrix), in view of (5) we obtain the method:

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\|x - x^k\right\|_2^2 \quad \text{subject to} \quad A_{i:}x = b_i. \tag{15}$$

Using (11), these iterations can be calculated with

$$\boxed{x^{k+1} = x^k - \frac{A_{i:}x^k - b_i}{\|A_{i:}\|_2^2}(A_{i:})^T} \tag{16}$$

Complexity. When i is selected at random, this is the randomized Kaczmarz (RK) method [31]. A specific non-uniform probability distribution for S can yield simple and easily interpretable (but not necessarily optimal) complexity bound. In particular, by selecting i with probability proportional to the magnitude of row i of A, that is $p_i = \|A_{i:}\|_2^2 / \|A\|_F^2$, it follows from Theorem 5.1 that RK enjoys the following complexity bound:

$$\mathbf{E}\left[\left\|x^k - x^*\right\|_2^2\right] \leq \left(1 - \frac{\lambda_{\min}\left(A^TA\right)}{\|A\|_F^2}\right)^k \left\|x^0 - x^*\right\|_2^2. \tag{17}$$

This result was first established by Strohmer and Vershynin [31]. We also provide new convergence results in Theorem 4.1, based on the convergence of the norm of the expected error. Theorem 4.1 applied to the RK method gives

$$\left\|\mathbf{E}\left[x^k - x^*\right]\right\|_2^2 \leq \left(1 - \frac{\lambda_{\min}\left(A^TA\right)}{\|A\|_F^2}\right)^{2k} \left\|x^0 - x^*\right\|_2^2. \tag{18}$$

Now the convergence rate appears squared, which is a better rate, though, the expectation has moved inside the norm, which is a weaker form of convergence.

Analogous results for the convergence of the norm of the expected error holds for all the methods we present, though we only illustrate this with the RK method.

Re-interpretation as SGD with exact line search. Using the Constrain-and-Approximate formulation (6), the randomized Kaczmarz method can also be written as

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \|x - x^*\|_2^2 \quad \text{subject to} \quad x = x^k + t(A_{i:})^T, \quad t \in \mathbb{R},$$

with probability p_i. Writing the least squares function $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ as

$$f(x) = \sum_{i=1}^m p_i f_i(x), \qquad f_i(x) = \frac{1}{2p_i}(A_{i:}x - b_i)^2,$$

we see that the random vector $\nabla f_i(x) = \frac{1}{p_i}(A_{i:}x - b_i)(A_{i:})^T$ is an unbiased estimator of the gradient of f at x. That is, $\mathbf{E}\left[\nabla f_i(x)\right] = \nabla f(x)$. Notice that RK takes a step in the direction $-\nabla f_i(x)$. This is true even when $A_{i:}x - b_i = 0$, in which case, the RK does not take any step. Hence, RK takes a step in the direction of the negative stochastic gradient. This means that it is equivalent to the Stochastic Gradient Descent (SGD) method. However, the stepsize choice is very special: RK chooses the stepsize which leads to the point which is closest to x^* in the Euclidean norm.

3.3 Randomized Coordinate Descent: positive definite case

If A is positive definite, then we can choose $B = A$ and $S = e^i$ in (5), which results in

$$x^{k+1} \stackrel{\text{def}}{=} \arg\min_{x \in \mathbb{R}^n} \left\|x - x^k\right\|_A^2 \quad \text{subject to} \quad (A_{i:})^T x = b_i, \tag{19}$$

where we used the symmetry of A to get $(e^i)^T A = A_{i:} = (A_{:i})^T$. The solution to the above, given by (11), is

$$\boxed{x^{k+1} = x^k - \frac{(A_{i:})^T x^k - b_i}{A_{ii}} e^i} \tag{20}$$

Complexity. When i is chosen randomly, this is the *Randomized CD* method (CD-pd). Applying Theorem 5.1, we see the probability distribution $p_i = A_{ii}/\mathbf{Tr}\,(A)$ results in a convergence with

$$\mathbf{E}\left[\left\|x^k - x^*\right\|_A^2\right] \leq \left(1 - \frac{\lambda_{\min}\,(A)}{\mathbf{Tr}\,(A)}\right)^k \left\|x^0 - x^*\right\|_A^2. \tag{21}$$

This result was first established by Leventhal and Lewis [15].

Interpretation. Using the Constrain-and-Approximate formulation (6), this method can be interpreted as

$$x^{k+1} = \arg\min \|x - x^*\|_A^2 \quad \text{subject to} \quad x = x^k + te^i, \quad t \in \mathbb{R}, \tag{22}$$

with probability p_i. It is easy to check that the function $f(x) = \frac{1}{2}x^T Ax + b^T x$ satisfies: $\|x - x^*\|_A^2 = 2f(x) + b^T x^*$. Therefore, (22) is equivalent to

$$x^{k+1} = \arg\min f(x) \quad \text{subject to} \quad x = x^k + te^i, \quad t \in \mathbb{R}. \tag{23}$$

The iterates (20) can also be written as

$$x^{k+1} = x^k - \frac{1}{L_i}\nabla_i f(x^k) e^i,$$

where $L_i = A_{ii}$ is the Lipschitz constant of the gradient of f corresponding to coordinate i and $\nabla_i f(x^k)$ is the ith partial derivative of f at x^k.

3.4 Randomized Block Kaczmarz

Our framework also extends to new block formulations of the randomized Kaczmarz method. Let R be a random subset of $[m]$ and let $S = I_{:R}$ be a column concatenation of the columns of the $m \times m$ identity matrix I indexed by R. Further, let $B = I$. Then (5) specializes to

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\|x - x^k\right\|_2^2 \quad \text{subject to} \quad A_{R:}x = b_R.$$

In view of (11), this can be equivalently written as

$$\boxed{x^{k+1} = x^k - (A_{R:})^T (A_{R:}(A_{R:})^T)^{-1}(A_{R:}x^k - b_R)}$$

For this to be well defined, we need to ensure that $A_{R:}(A_{R:})^T$ is invertible with probability 1; this is ensured by Assumption 2.1.

Currently, only block Kaczmarz methods with R defining a partition of $[m]$ have been analysed using a row paving of A, see [19, 20]. With our framework, we can analyse the convergence of the iterates for a large set of possible random subsets R, including partitions.

Complexity. From Theorem 4.2 we obtain the following new complexity result:

$$\mathbf{E}\left[\|x^k - x^*\|_2^2\right] \leq \left(1 - \lambda_{\min}\left(\mathbf{E}\left[(A_{R:})^T(A_{R:}(A_{R:})^T)^{-1}A_{R:}\right]\right)\right)^k \|x^0 - x^*\|_2^2.$$

3.5 Randomized Newton: positive definite case

If A is symmetric positive definite, then we can choose $B = A$ and $S = I_{:C}$, a column concatenation of the columns of I indexed by C, which is a random subset of $[n]$. In view of (5), this results in

$$x^{k+1} \overset{\text{def}}{=} \arg\min_{x \in \mathbb{R}^n} \left\|x - x^k\right\|_A^2 \quad \text{subject to} \quad (A_{:C})^T x = b_C. \tag{24}$$

In view of (11), we can equivalently write the method as

$$\boxed{x^{k+1} \quad = \quad x^k - I_{:C}((I_{:C})^T A I_{:C})^{-1}(I_{:C})^T(Ax^k - b)} \tag{25}$$

Complexity. Clearly, iteration (25) is well defined as long as C is nonempty with probability 1. Such C is in [24] referred to by the name "non-vacuous" sampling. From Theorem 4.2 we obtain the following convergence rate:

$$\begin{aligned}\mathbf{E}\left[\left\|x^k - x^*\right\|_A^2\right] &\leq \rho^k \|x^0 - x^*\|_A^2 \\ &= \left(1 - \lambda_{\min}\left(\mathbf{E}\left[I_{:C}((I_{:C})^T A I_{:C})^{-1}(I_{:C})^T A\right]\right)\right)^k \left\|x^0 - x^*\right\|_A^2. \end{aligned} \tag{26}$$

The convergence rate of this particular method was first established in [24]. Moreover, it was shown in [24] that $\rho < 1$ if one additionally assumes that the probability that $i \in C$ is positive for each column $i \in [n]$, i.e., that C is a "proper" sampling.

Interpretation. Using formulation (6), and in view of the equivalence between $f(x)$ and $\|x - x^*\|_A^2$ discussed in Section 3.3, the Randomized Newton method can be equivalently written as

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad x = x^k + I_{:C}\, y, \quad y \in \mathbb{R}^{|C|}.$$

The next iterate is determined by advancing from the previous iterate over a subset of coordinates such that f is minimized. Hence, an exact line search is performed in a $|C|$ dimensional subspace.

Method (25) was fist studied by Qu et al [24], and referred therein as "Method 1", or *Randomized Newton Method*. The name comes from the observation that the method inverts random principal submatrices of A and that in the special case when $C = [n]$ with probability 1, it specializes to the Newton method (which in this case converges in a single step).

3.6 Randomized Coordinate Descent: least-squares version

By choosing $S = Ae^i =: A_{:i}$ as the ith column of A and $B = A^TA$, the resulting iterates (6) are given by

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \quad \text{subject to} \quad x = x^k + t\, e^i, \quad t \in \mathbb{R}. \tag{27}$$

When i is selected at random, this is the Randomized Coordinate Descent method (*CD-LS*) applied to the least-squares problem: $\min_x \|Ax - b\|_2^2$. Using (11), these iterations can be calculated with

$$\boxed{x^{k+1} = x^k - \frac{(A_{:i})^T(Ax^k - b)}{\|A_{:i}\|_2^2} e^i} \tag{28}$$

Complexity. Applying Theorem 5.1, we see that by selecting i with probability proportional to magnitude of column i of A, that is $p_i = \|A_{:i}\|_2^2 / \|A\|_F^2$, results in a convergence with

$$\mathbf{E}\left[\left\|x^k - x^*\right\|_{A^TA}^2\right] \leq \rho^k \|x^0 - x^*\|_{A^TA}^2 = \left(1 - \frac{\lambda_{\min}\left(A^TA\right)}{\|A\|_F^2}\right)^k \left\|x^0 - x^*\right\|_{A^TA}^2. \tag{29}$$

This result was first established by Leventhal and Lewis [15].

Interpretation. Using the Constrain-and-Approximate formulation (6), the CD-LS method can be interpreted as

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \|x - x^*\|_{A^TA}^2 \quad \text{subject to} \quad x = x^k + te^i, \quad t \in \mathbb{R}. \tag{30}$$

The CD-LS method selects a coordinate to advance from the previous iterate x^k, then performs an exact minimization of the least squares function over this line. This is equivalent to applying coordinate descent to the least squares problem $\min_{x \in \mathbb{R}^n} f(x) \stackrel{\text{def}}{=} \frac{1}{2}\|Ax - b\|_2^2$. The iterates (27) can be written as

$$x^{k+1} = x^k - \frac{1}{L_i} \nabla_i f(x^k) e^i,$$

where $L_i \stackrel{\text{def}}{=} \|A_{:i}\|_2^2$ is the Lipschitz constant of the gradient corresponding to coordinate i and $\nabla_i f(x^k)$ is the ith partial derivative of f at x^k.

4 Convergence: General Theory

We shall present two complexity theorems: we first study the convergence of $\left\|\mathbf{E}\left[x^k - x^*\right]\right\|$, and then move on to analysing the convergence of $\mathbf{E}\left[\left\|x^k - x^*\right\|\right]$.

4.1 Two types of convergence

The following lemma explains the relationship between the convergence of the norm of the expected error and the expected norm of the error.

Lemma 4.1. *Let $x \in \mathbb{R}^n$ be a random vector, $\|\cdot\|$ a norm induced by an inner product and fix $x^* \in \mathbb{R}^n$. Then*

$$\left\|\mathbf{E}\left[x - x^*\right]\right\|^2 = \mathbf{E}\left[\|x - x^*\|^2\right] - \mathbf{E}\left[\|x - \mathbf{E}\left[x\right]\|^2\right].$$

Proof. Note that $\mathbf{E}\left[\|x - \mathbf{E}\left[x\right]\|^2\right] = \mathbf{E}\left[\|x\|^2\right] - \|\mathbf{E}\left[x\right]\|^2$. Adding and subtracting $\|x^*\|^2 - 2\left\langle \mathbf{E}\left[x\right], x^*\right\rangle$ from the right hand side and grouping the appropriate terms yields the desired result. □

To interpret this lemma, note that $\mathbf{E}\left[\|x - \mathbf{E}\left[x\right]\|^2\right] = \sum_{i=1}^n \mathbf{E}\left[(x_i - \mathbf{E}\left[x_i\right])^2\right] = \sum_{i=1}^n \mathbf{Var}(x_i)$, where x_i denotes the ith element of x. This lemma shows that the convergence of x to x^* under the expected norm of the error is a stronger form of convergence than the convergence of the norm of the expected error, as the former also guarantees that the variance of x_i converges to zero, for $i = 1, \ldots, n$.

4.2 The Rate of Convergence

All of our convergence theorems (see Table 1) depend on the convergence rate

$$\rho \overset{\text{def}}{=} 1 - \lambda_{\min}(B^{-1}\mathbf{E}\left[Z\right]) = 1 - \lambda_{\min}(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}). \tag{31}$$

To show that the rate is meaningful, in Lemma 4.2 we prove that $0 \leq \rho \leq 1$. We also provide a meaningful lower bound for ρ.

Lemma 4.2. *Let Assumption 2.1 hold. The quantity ρ defined in* (31) *satisfies:*

$$0 \leq 1 - \frac{\mathbf{E}\left[q\right]}{n} \leq \rho \leq 1.$$

Proof. Since the mapping $A \mapsto \lambda_{\max}(A)$ is convex, by Jensen's inequality we get

$$\lambda_{\max}(\mathbf{E}\left[B^{-1}Z\right]) = \lambda_{\max}(B^{-1}\mathbf{E}\left[Z\right]) \leq \mathbf{E}\left[\lambda_{\max}(B^{-1}Z)\right]. \tag{32}$$

Recalling from Lemma 2.1 that $B^{-1}Z$ is a projection, the spectrum of $B^{-1}Z$ is contained in $\{0, 1\}$. Thus $\lambda_{\max}(B^{-1}Z) \leq 1$, and from (32) we conclude that $\lambda_{\max}(B^{-1}\mathbf{E}\left[Z\right]) \leq 1$. The inequality $\lambda_{\min}(B^{-1}\mathbf{E}\left[Z\right]) \geq 0$ can be shown analogously using convexity of the mapping $A \mapsto -\lambda_{\min}(A)$. Thus

$$\lambda_{\min}(B^{-1}\mathbf{E}\left[Z\right]) = \lambda_{\min}(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}) \in [0, 1]$$

and consequentially $0 \leq \rho \leq 1$. As the trace of a matrix is equal to the sum of its eigenvalues, we have

$$\mathbf{E}\left[\mathbf{Tr}\left(B^{-1}Z\right)\right] = \mathbf{Tr}\left(\mathbf{E}\left[B^{-1}Z\right]\right) \geq n\,\lambda_{\min}(\mathbf{E}\left[B^{-1}Z\right]). \tag{33}$$

As $B^{-1}Z$ projects onto a q–dimensional subspace (Lemma 2.1) we have $\mathbf{Tr}\left(B^{-1}Z\right) = q$. Thus rewriting (33) gives $1 - \mathbf{E}\left[q\right]/n \leq \rho$. □

The lower bound bound on ρ in item 1 has a natural interpretation which makes intuitive sense. We shall present it from the perspective of the Constrain-and-Approximate formulation (6). As the dimension (q) of the search space $B^{-1}A^TS$ increases (see (10)), the lower bound on ρ decreases, and a faster convergence is possible. For instance, when S is restricted to being a random column vector, as it is in the RK (16), CD-LS (28) and CD-pd (21) methods, the convergence rate is

bounded with $1 - 1/n \leq \rho$. Using (3), this translates into the simple iteration complexity bound of $k \geq n \log(1/\epsilon)$. On the other extreme, when the search space is large, then the lower bound is close to zero, allowing room for the method to be faster.

We now characterize circumstances under which ρ is strictly smaller than one.

Lemma 4.3. *Let Assumption 2.1 hold. If $\mathbf{E}[Z]$ is invertible, then $\rho < 1$, A has full column rank and x^* is unique.*

Proof. Assume that $\mathbf{E}[Z]$ is invertible. First, this means that $B^{-1/2}\mathbf{E}[Z]B^{-1/2}$ is positive definite, which in view of (31) means that $\rho < 1$. If A did not have full column rank, then there would be $0 \neq x \in \mathbb{R}^n$ such that $Ax = 0$. However, we then have $Zx = 0$ and also $\mathbf{E}[Z]x = 0$, contradicting the assumption that $\mathbf{E}[Z]$ is invertible. Finally, since A has full column rank, x^* must be unique (recall that we assume throughout the paper that the system $Ax = b$ is consistent). □

4.3 Exact Characterization and Norm of Expectation

We now state a theorem which exactly characterizes the evolution of the expected iterates through a linear fixed point iteration. As a consequence, we obtain a convergence result for the norm of the expected error. While we do not highlight this in the text, this theorem can be applied to all the particular instances of our general method we detail throughout this paper.

For any $M \in \mathbb{R}^{n\times n}$ let us define

$$\|M\|_B \overset{\text{def}}{=} \max_{\|x\|_B=1} \|Mx\|_B. \tag{34}$$

Theorem 4.1 (Norm of expectation)**.** *If Assumption 2.1 holds, then for every $x^* \in \mathbb{R}^n$ satisfying $Ax = b$ we have*

$$\mathbf{E}\left[x^{k+1} - x^*\right] = \left(I - B^{-1}\mathbf{E}[Z]\right)\mathbf{E}\left[x^k - x^*\right]. \tag{35}$$

Moreover, the spectral radius and the induced B-norm of the iteration matrix $I - B^{-1}\mathbf{E}[Z]$ are both equal to ρ:

$$\lambda_{\max}(I - B^{-1}\mathbf{E}[Z]) = \|I - B^{-1}\mathbf{E}[Z]\|_B = \rho.$$

Therefore,

$$\left\|\mathbf{E}\left[x^k - x^*\right]\right\|_B \leq \rho^k \left\|x^0 - x^*\right\|_B. \tag{36}$$

Proof. Taking expectations conditioned on x^k in (12), we get

$$\mathbf{E}\left[x^{k+1} - x^* \mid x^k\right] = (I - B^{-1}\mathbf{E}[Z])(x^k - x^*). \tag{37}$$

Taking expectation again gives

$$\begin{aligned}
\mathbf{E}\left[x^{k+1} - x^*\right] &= \mathbf{E}\left[\mathbf{E}\left[x^{k+1} - x^* \mid x^k\right]\right] \\
&\overset{(37)}{=} \mathbf{E}\left[(I - B^{-1}\mathbf{E}[Z])(x^k - x^*)\right] \\
&= (I - B^{-1}\mathbf{E}[Z])\mathbf{E}\left[x^k - x^*\right].
\end{aligned}$$

Applying the norms to both sides we obtain the estimate

$$\left\|\mathbf{E}\left[x^{k+1}-x^*\right]\right\|_B \leq \left\|I-B^{-1}\mathbf{E}\left[Z\right]\right\|_B \left\|\mathbf{E}\left[x^{k}-x^*\right]\right\|_B.$$

It remains to prove that $\rho = \left\|I-B^{-1}\mathbf{E}\left[Z\right]\right\|_B$ and then unroll the recurrence. According to the definition of operator norm (34), we have

$$\left\|I-B^{-1}\mathbf{E}\left[Z\right]\right\|_B^2 = \max_{\left\|B^{1/2}x\right\|_2=1} \left\|B^{1/2}(I-B^{-1}\mathbf{E}\left[Z\right])x\right\|_2^2.$$

Substituting $B^{1/2}x = y$ in the above gives

$$\begin{aligned}
\left\|I-B^{-1}\mathbf{E}\left[Z\right]\right\|_B^2 &= \max_{\|y\|_2=1} \left\|B^{1/2}(I-B^{-1}\mathbf{E}\left[Z\right])B^{-1/2}y\right\|_2^2 \\
&= \max_{\|y\|_2=1} \left\|(I-B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2})y\right\|_2^2 \\
&= \lambda_{\max}^2(I-B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}) \\
&= \left(1-\lambda_{\min}(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2})\right)^2 \\
&= \rho^2,
\end{aligned}$$

where in the third equality we used the symmetry of $(I-B^{-1}\mathbf{E}\left[Z\right]B^{-1})$ when passing from the operator norm to the spectral radius. Note that the symmetry of $\mathbf{E}\left[Z\right]$ derives from the symmetry of Z. □

4.4 Expectation of Norm

We now turn to analysing the convergence of the expected norm of the error, for which we need the following technical lemma.

Lemma 4.4. *If $\mathbf{E}\left[Z\right]$ is positive definite, then*

$$\langle \mathbf{E}\left[Z\right]y, y\rangle \geq (1-\rho)\left\|y\right\|_B^2, \quad \forall y \in \mathbb{R}^n. \tag{38}$$

Proof. As $\mathbf{E}\left[Z\right]$ and B are positive definite, we get

$$\begin{aligned}
1-\rho = \lambda_{\min}(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}) &= \max_t \left\{t \quad | \quad B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2} - t\cdot I \succeq 0\right\} \\
&= \max_t \left\{t \quad | \quad \mathbf{E}\left[Z\right] - t\cdot B \succeq 0\right\}.
\end{aligned}$$

Therefore, $\mathbf{E}\left[Z\right] \succeq (1-\rho)B$, and the result follows. □

Theorem 4.2 (Expectation of norm)**.** *Let Assumption 2.1 hold and furthermore suppose that $\mathbf{E}\left[Z\right]$ is positive definite, where Z is defined in* (13)*. Then*

$$\mathbf{E}\left[\left\|x^k-x^*\right\|_B^2\right] \leq \rho^k \left\|x^0-x^*\right\|_B^2, \tag{39}$$

where $\rho < 1$ is given in (31)*.*

Proof. Let $r^k = x^k - x^*$. Taking the expectation of (12) conditioned on r^k we get

$$\begin{aligned}\mathbf{E}\left[\|r^{k+1}\|_B^2 \,|\, r^k\right] &\overset{(12)}{=} \mathbf{E}\left[\|(I - B^{-1}Z)r^k\|_B^2 \,|\, r^k\right] \\ &\overset{(14)}{=} \mathbf{E}\left[\left\langle (B-Z)r^k, r^k\right\rangle \,|\, r^k\right] \\ &= \|r^k\|_B^2 - \left\langle \mathbf{E}\left[Z\right] r^k, r^k\right\rangle \overset{\text{(Lemma (4.4))}}{\leq} \rho \cdot \|r^k\|_B^2.\end{aligned}$$

Taking expectation again and unrolling the recurrence gives the result. □

The convergence rate ρ of the expected norm of the error is "worse" than the ρ^2 rate of convergence of the norm of the expected error in Theorem 4.1 This should not be misconstrued as Theorem 4.1 offering a "better" convergence rate than Theorem 4.2, because, as explained in Lemma 4.1, convergence of the expected norm of the error is a stronger type of convergence.

5 Methods Based on Discrete Sampling

When S has a discrete distribution, we can establish under reasonable assumptions when $\mathbf{E}\left[Z\right]$ is positive definite (Proposition 5.1), we can optimize the convergence rate in terms of the chosen probability distribution, and finally, determine a probability distribution for which the convergence rate is expressed in terms of the scaled condition number (Theorem 5.1).

Assumption 5.1 (Complete Discrete Sampling)**.** *The random matrix S has a discrete distribution. In particular, $S = S_i \in \mathbb{R}^{m\times q_i}$ with probability $p_i > 0$, where $S_i^T A$ has full row rank and $q_i \in \mathbb{N}$, for $i = 1, \ldots, r$. Furthermore $\mathbf{S} \overset{def}{=} [S_1, \ldots, S_r] \in \mathbb{R}^{n\times\sum_{i=1}^r q_i}$ is such that $A^T\mathbf{S}$ has full row rank.*

The choice of S in all the methods we describe in Section 3 satisfy this assumption. In particular, if A has full column rank, $S = e^i$ with probability $p_i = 1/n$, for $i = 1, \ldots, n$, then $\mathbf{S} = I$ and S is a complete discrete sampling. From any basis of $\mathbb{R}^n$ we can construct a complete discrete sampling in an analogous way.

Using a complete discrete sampling guarantees convergence of the resulting method.

Proposition 5.1. *Let S be a complete discrete sampling, then $\mathbf{E}\left[Z\right]$ is positive definite.*

Proof. Let

$$D \overset{\text{def}}{=} \operatorname{diag}\left(\sqrt{p_1}((S_1)^T AB^{-1}A^T S_1)^{-1/2}, \ldots, \sqrt{p_r}((S_r)^T AB^{-1}A^T S_r)^{-1/2}\right) \tag{40}$$

which is a block diagonal matrix, and is well defined and invertible as $S_i^T A$ has full row rank for $i = 1, \ldots, r$. Taking the expectation of Z (13) gives

$$\begin{aligned}\mathbf{E}\left[Z\right] &= \sum_{i=1}^r A^T S_i (S_i^T AB^{-1}A^T S_i)^{-1} S_i^T A p_i \\ &= A^T \left(\sum_{i=1}^r S_i \sqrt{p_i}(S_i^T AB^{-1}A^T S_i)^{-1/2}(S_i^T AB^{-1}A^T S_i)^{-1/2}\sqrt{p_i} S_i^T\right) A \\ &= \left(A^T\mathbf{S}D\right)\left(D\mathbf{S}^T A\right), \end{aligned} \tag{41}$$

which is positive definite because $A^T\mathbf{S}$ has full row rank and D is invertible. □

With $\mathbf{E}[Z]$ positive definite, we can apply the convergence Theorem 4.1 and 4.2, and the resulting method converges.

5.1 Optimal Probabilities

We can choose the discrete probability distribution that optimizes the convergence rate. For this, according to Theorems 4.2 and 4.1 we need to find $p = (p_1, \ldots, p_r)$ that maximizes the minimal eigenvalue of $B^{-1/2}\mathbf{E}[Z]B^{-1/2}$. Let S be a complete discrete sampling and fix the sample matrices $S_1, \ldots, S_r$. Let us denote $Z = Z(p)$ as a function of $p = (p_1, \ldots, p_r)$. Then we can also think of the spectral radius as a function of p where

$$\rho(p) = 1 - \lambda_{\min}(B^{-1/2}\mathbf{E}[Z(p)]B^{-1/2}).$$

Letting

$$\Delta_r \stackrel{\text{def}}{=} \left\{ p = (p_1, \ldots, p_r) \in \mathbb{R}^r \ : \ \sum_{i=1}^r p_i = 1, \ p \geq 0 \right\},$$

the problem of minimizing the spectral radius (i.e., optimizing the convergence rate) can be written as

$$\rho^* \quad \stackrel{\text{def}}{=} \quad \min_{p \in \Delta_r} \rho(p) \quad = \quad 1 - \max_{p \in \Delta_r} \lambda_{\min}(B^{-1/2}\mathbf{E}[Z(p)]B^{-1/2}).$$

This can be cast as a convex optimization problem, by first re-writing

$$\begin{aligned} B^{-1/2}\mathbf{E}[Z(p)]B^{-1/2} &= \sum_{i=1}^r p_i \left(B^{-1/2}A^T S_i (S_i^T A B^{-1} A^T S_i)^{-1} S_i^T A B^{-1/2} \right) \\ &= \sum_{i=1}^r p_i \left(V_i (V_i^T V_i)^{-1} V_i^T \right), \end{aligned}$$

where $V_i = B^{-1/2}A^T S_i$. Thus

$$\rho^* \quad = \quad 1 - \max_{p \in \Delta_r} \lambda_{\min} \left(\sum_{i=1}^r p_i V_i (V_i^T V_i)^{-1} V_i^T \right). \tag{42}$$

To obtain p that maximizes the smallest eigenvalue, we solve

$$\begin{aligned} \max_{p,t} \quad & t \\ \text{subject to} \quad & \sum_{i=1}^r p_i \left(V_i (V_i^T V_i)^{-1} V_i^T \right) \succeq t \cdot I, \\ & p \in \Delta_r. \end{aligned} \tag{43}$$

Despite (43) being a convex semi-definite program, which is apparently a harder problem than solving the original linear system, investing the time into solving (43) can pay off, as we show in Section 7.5. Though for a practical method based on this, we would need to develop an approximate solution to (43) which can be efficiently calculated.

5.2 Convenient Probabilities

Next we develop a choice of probability distribution that yields a convergence rate that is easy to interpret. This result is new and covers a wide range of methods, including randomized Kaczmarz, randomized coordinate descent, as well as their block variants. However, it is more general, and covers many other possible particular algorithms, which arise by choosing a particular set of sample matrices S_i, for $i = 1, \ldots, r$.

Theorem 5.1. *Let S be a complete discrete sampling such that $S = S_i \in \mathbb{R}^m$ with probability*

$$p_i \;=\; \frac{\mathbf{Tr}\left(S_i^T A B^{-1} A^T S_i\right)}{\left\|B^{-1/2} A^T \mathbf{S}\right\|_F^2}, \quad \textit{for} \quad i = 1, \ldots, r. \tag{44}$$

Then the iterates (11) *satisfy*

$$\mathbf{E}\left[\left\|x^k - x^*\right\|_B^2\right] \leq \rho_c^k \left\|x^0 - x^*\right\|_B^2, \tag{45}$$

where

$$\rho_c = 1 - \frac{\lambda_{\min}\left(\mathbf{S}^T A B^{-1} A^T \mathbf{S}\right)}{\left\|B^{-1/2} A^T \mathbf{S}\right\|_F^2}. \tag{46}$$

Proof. Let $t_i = \mathbf{Tr}\left((S^i)^T A B^{-1} A^T S^i\right)$, and with (44) in (40) we have

$$D^2 = \frac{1}{\left\|B^{-1/2} A^T \mathbf{S}\right\|_F^2} \mathrm{diag}\left(t_1((S^1)^T A B^{-1} A^T S^1)^{-1}, \ldots, t_r((S^r)^T A B^{-1} A^T S^r)^{-1}\right),$$

thus

$$\lambda_{\min}(D^2) = \frac{1}{\left\|B^{-1/2} A^T \mathbf{S}\right\|_F^2} \min_i \left\{\frac{t_i}{\lambda_{\max}((S^i)^T A B^{-1} A^T S^i)}\right\} \geq \frac{1}{\left\|B^{-1/2} A^T \mathbf{S}\right\|_F^2}. \tag{47}$$

Applying the above in (41) gives

$$\begin{aligned}
\lambda_{\min}\left(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}\right) &= \lambda_{\min}\left(B^{-1/2} A^T \mathbf{S} D^2 \mathbf{S}^T A B^{-1/2}\right) \\
&= \lambda_{\min}\left(\mathbf{S}^T A B^{-1} A^T \mathbf{S} D^2\right) \\
&\geq \lambda_{\min}\left(\mathbf{S}^T A B^{-1} A^T \mathbf{S}\right) \lambda_{\min}(D^2) \\
&\geq \frac{\lambda_{\min}\left(\mathbf{S}^T A B^{-1} A^T \mathbf{S}\right)}{\left\|B^{-1/2} A^T \mathbf{S}\right\|_F^2},
\end{aligned} \tag{48}$$

where we used that if $B, C \in \mathbb{R}^{n\times n}$ are positive definite $\lambda_{\min}(BC) \geq \lambda_{\min}(B)\lambda_{\min}(C)$. Finally

$$1 - \lambda_{\min}\left(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}\right) \leq 1 - \frac{\lambda_{\min}\left(\mathbf{S}^T A B^{-1} A^T \mathbf{S}\right)}{\left\|B^{-1/2} A^T \mathbf{S}\right\|_F^2}. \tag{49}$$

The result (45) follows by applying Theorem 4.2. □

The convergence rate $\lambda_{\min}\left(\mathbf{S}^T AB^{-1}A^T\mathbf{S}\right) / \left\|B^{-1/2}A^T\mathbf{S}\right\|_F^2$ is known as the scaled condition number, and naturally appears in other numerical schemes, such as matrix inversion [8, 6]. When $S^i = s^i \in \mathbb{R}^n$ is a column vector then

$$p_i \;=\; \left((s^i)^T AB^{-1}A^T s^i\right) / \left\|B^{-1/2}A^T\mathbf{S}\right\|_F^2,$$

for $i = 1, \dots r$. In this case, the bound (47) is an equality and D^2 is a scaled identity, so (48) and consequently (49) are equalities. For block methods, it is different story, and there is much more slack in the inequality (49). So much so, the convergence rate (46) does not indicate any advantage of using a block method (contrary to numerical experiments). To see the advantage of a block method, we need to use the exact expression for $\lambda_{\min}(D^2)$ given in (47). Though this results in a somewhat harder to interpret convergence rate, a matrix paving could be used explore this block convergence rate, as was done for the block Kaczmarz method [20, 19].

By appropriately choosing B and S, this theorem applied to RK method (15), the CD-LS method (27) and the CD-pd method (19), yields the convergence results (17), (29) and (21), respectively, for single column sampling or block methods alike.

This theorem also suggests a preconditioning strategy, in that, a faster convergence rate will be attained if $\mathbf{S}$ is an approximate inverse of $B^{-1/2}A^T$. For instance, in the RK method where $B = I$, this suggests that an accelerated convergence can be attained if S is a random sampling of the rows of a preconditioner (approximate inverse) of A.

6 Methods Based on Gaussian Sampling

In this section we shall describe variants of our method in the case when S is a Gaussian vector with mean $0 \in \mathbb{R}^m$ and a positive definite covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$. That is, $S = \zeta \sim N(0, \Sigma)$. This applied to (11) results in iterations of the form

$$\boxed{x^{k+1} = x^k - \frac{\zeta^T(Ax^k - b)}{\zeta^T AB^{-1}A^T\zeta} B^{-1}A^T\zeta} \tag{50}$$

Due to the symmetry of the multivariate normal distribution, there is a zero probability that $\zeta \in \mathbf{Null}\left(A^T\right)$ for any nonzero matrix A. Thus Assumption 2.1 holds for A nonzero, and (50) is well defined with probability 1.

Unlike the discrete methods in Section 3, to calculate an iteration of (50) we need to compute the product of a matrix with a dense vector ζ. This significantly raises the cost of an iteration. Though in our numeric tests in Section 7, the faster convergence of the Gaussian method often pays off for their high iteration cost.

To analyze the complexity of the resulting method let $\xi \overset{\text{def}}{=} B^{-1/2}A^T S$, which is also Gaussian, distributed as $\xi \sim N(0, \Omega)$, where $\Omega \overset{\text{def}}{=} B^{-1/2}A^T\Sigma AB^{-1/2}$. In this section we assume A *has full column rank*, so that Ω is always positive definite. The complexity of the method can be established through a simple computation:

$$\rho = 1 - \lambda_{\min}(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}) = 1 - \lambda_{\min}\left(\mathbf{E}\left[B^{-1/2}ZB^{-1/2}\right]\right) = 1 - \lambda_{\min}\left(\mathbf{E}\left[M_\xi\right]\right),$$

where by (13) and $\xi = B^{-1/2}A^T S$ we have $M_\xi \overset{\text{def}}{=} \xi\xi^T / \left\|\xi\right\|_2^2$. Thus the convergence rate of any method where S is Gaussian depends on the spectral properties of $\mathbf{E}\left[M_\xi\right]$. This can be revealing.

From (**??**) we obtain the lower bound $\rho \geq 1 - 1/n$. Furthermore, we prove in Lemma 4.1 in the supplementary material that $\mathbf{E}[M_\xi]$ is always positive definite. Thus Theorem 4.2 guarantees that the expected norm of the error of all Gaussian methods converges exponentially to zero. When $n = 2$, then in Lemma 8.2 of the Appendix we prove that

$$\mathbf{E}[M_\xi] = \frac{\Omega^{1/2}}{\mathbf{Tr}\left(\Omega^{1/2}\right)},$$

which yields a very favourable convergence rate. This expression does not hold for $n > 2$, and instead, we conjecture that

$$\mathbf{E}[M_\xi] \succeq \frac{\Omega}{\mathbf{Tr}(\Omega)},$$

for all n and perform numeric tests in Section 7.2 to support this.

6.1 Gaussian Kaczmarz

Let $B = I$ and choose $\Sigma = I$ so that $S = \eta \sim N(0, I)$. Then (50) has the form

$$\boxed{x^{k+1} = x^k - \frac{\eta^T(Ax^k - b)}{\|A^T\eta\|_2^2} A^T\eta} \tag{51}$$

which we call the *Gaussian Kaczmarz* (GK) method, for it is the analogous method to the Randomized Karcmarz method in the discrete setting. Using the formulation (6), for instance, the GK method can be interpreted as

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \|x - x^*\|^2 \quad \text{subject to} \quad x = x^k + A^T\eta\lambda, \quad \lambda \in \mathbb{R}.$$

Thus at each iteration, a random normal Gaussian vector η is drawn and a search direction is formed by $A^T\eta$. Then, starting from the previous iterate x^k, an exact line search is performed over this search direction so that the euclidean distance from the optimal is minimized.

6.2 Gaussian Least-Squares

Let $B = A^TA$ and choose $S \sim N(0, \Sigma)$ with $\Sigma = AA^T$. It will be convenient to write $S = A\eta$, where $\eta \sim N(0, I)$. Then method (50) then has the form

$$\boxed{x^{k+1} = x^k - \frac{\eta^TA^T(Ax^k - b)}{\|A\eta\|_2^2}\eta} \tag{52}$$

which we call the *Gauss-LS* method. This method has a natural interpretation through formulation (6) as

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 \quad \text{subject to} \quad x = x^k + \eta\lambda, \quad \lambda \in \mathbb{R}.$$

That is, starting from x^k, we take a step in a random (Gaussian) direction, then perform an exact line search over this direction that minimizes the least squares error. Thus the Gauss-LS method is the same as applying the Random Pursuit method [29] with exact line search to the Least-squares function.

6.3 Gaussian Positive Definite

When A is positive definite, we achieve an accelerated Gaussian method. Let $B = A$ and choose $S = \eta \sim N(0, I)$. Method (50) then has the form

$$\boxed{x^{k+1} = x^k - \frac{\eta^T(Ax^k - b)}{\|\eta\|_A^2}\eta} \tag{53}$$

which we call the *Gauss-pd* method.

Using formulation (6), the method can be interpreted as

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} f(x) \overset{\text{def}}{=} \tfrac{1}{2}x^T Ax - b^T x \quad \text{subject to} \quad x = x^k + \eta\lambda, \quad \lambda \in \mathbb{R}.$$

That is, starting from x^k, we take a step in a random (Gaussian) direction, then perform an exact line search over this direction. Thus the Gauss-pd method is equivalent to applying the Random Pursuit method [29] with exact line search to $f(x)$.

7 Numerical Experiments

We perform some preliminary numeric tests. Everything was coded and run in MATLAB R2014b. Let $\kappa_2 = 1/\|A\|\left\|A^\dagger\right\|$ be the 2−norm condition number, where $A^\dagger$ is a pseudo-inverse of A. In comparing different methods for solving overdetermined systems, we use the relative error measure $\left\|Ax^k - b\right\|_2 / \|b\|_2$, while for positive definite systems we use $\left\|x^k - x^*\right\|_A / \|x^*\|_A$ as a relative error measure. We run each method until the relative error is below 10^{-4} or until 300 seconds in time is exceeded. We use $x_0 = 0 \in \mathbb{R}^n$ as an initial point. In each figure we plot the relative error in percentage, thus starting with 100%.

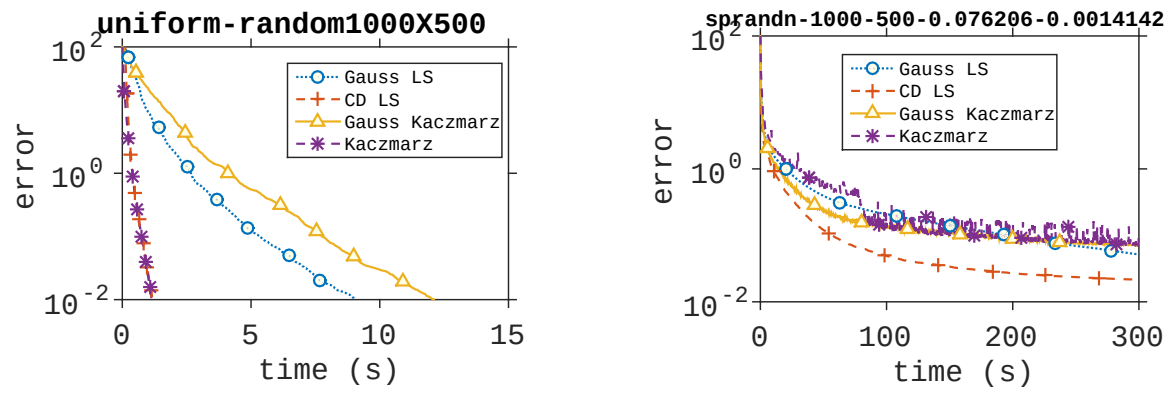
In implementing the methods we used the convenient probability distributions (44).

7.1 Overdetermined linear systems

First we compare the methods Gauss-LS, CD-LS, Gauss-Kaczmarz and RK methods on synthetic linear systems generated with the matrix functions `rand` and `sprandn`, see Figure 2. The high iteration cost of the Gaussian methods resulted in poor performance on the dense problem generated using `rand` in Figure 2a. In Figure 2b we compare the methods on a sparse linear system generated using the MATLAB sparse random matrix function `sprandn`$(m, n,$`density,rc`$)$, where `density` is the percentage of nonzero entries and `rc` is the reciprocal of the condition number. On this sparse problem the Gaussian methods are more efficient, and converge at a similar rate to the discrete sampling methods.

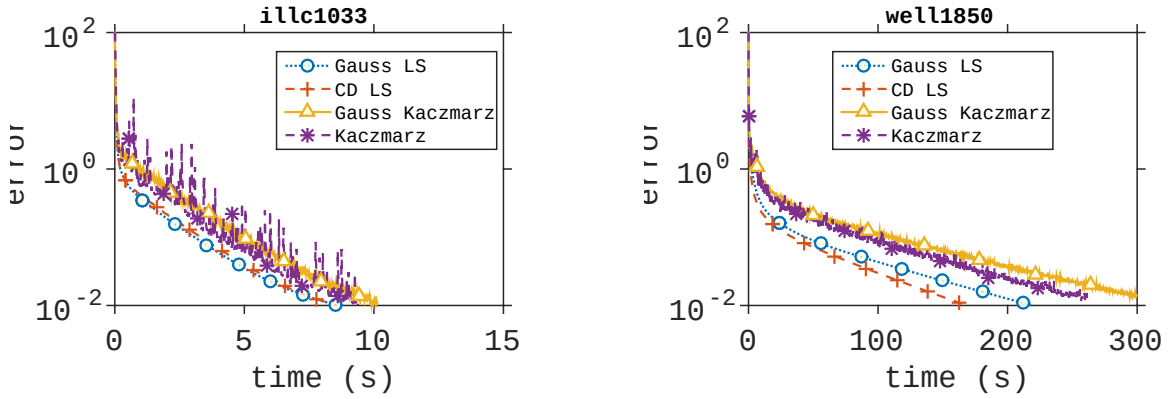
In Figure 3 we test two overdetermined linear systems taken from the the Matrix Market collection [3]. The collection also provides the right-hand side of the linear system. Both of these systems are very well conditioned, but do not have full column rank, thus Theorem 4.2 does not apply. The four methods have a similar performance on Figure 3a, while the Gauss-LS and CD-LS method converge faster on 3b as compared to the Gauss-Kaczmarz and Kaczmarz methods.

Finally, we test two problems, the `SUSY` problem and the `covtype.binary` problem, from the library of support vector machine problems LIBSVM [4]. These problems do not form consistent linear systems, thus only the Guass-LS and CD-LS methods are applicable, see Figure 4. This is



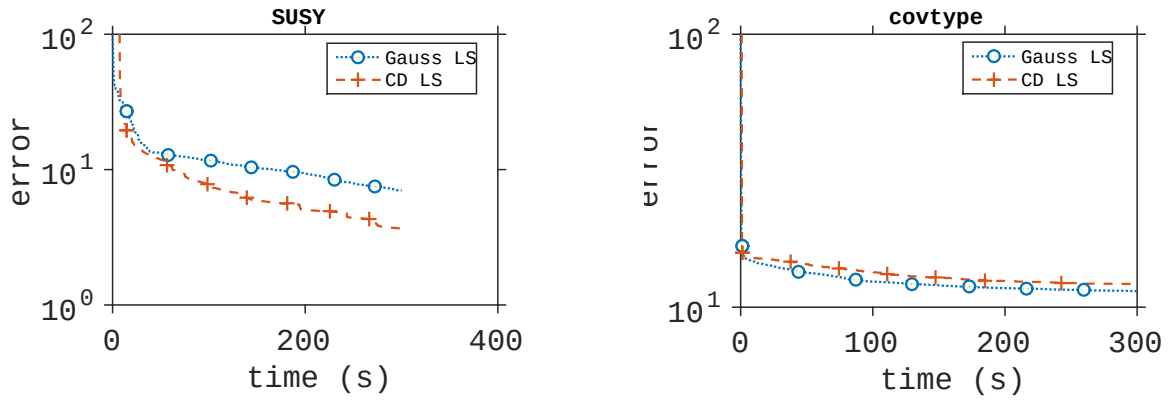
(a) `rand` (b) `sprandn`

Figure 2: The performance of the Gauss-LS, CD-LS, Gauss-Kaczmarz and RK methods on synthetic MATLAB generated problems (a) `rand`(n, m) with $(m; n) = (1000, 500)$ (b) `sprandn`$(m, n$,`density`,`rc`$)$ with $(m; n) = (1000, 500)$, `density`$= 1/\log(nm)$ and `rc`$= 1/\sqrt{mn}$. In both experiments dense solutions were generated with $x^* =$`rand`$(n, 1)$ and $b = Ax^*$.



(a) `illc1033` (b) `well1033`

Figure 3: The performance of the Gauss-LS, CD-LS, Gauss-Kaczmarz and RK methods on linear systems (a) `well1033` where $(m; n) = (1850, 750)$, $nnz = 8758$ and $\kappa_2 = 1.8$ (b) `illc1033` where $(m; n) = (1033; 320)$, $nnz = 4732$ and $\kappa_2 = 2.1$, from the Matrix Market [3].



(a) SUSY (b) covtype.binary

Figure 4: The performance of Gauss LS and CD LS methods on two LIBSVM test problems: (a) SUSY: $(m; n) = (5 \times 10^6; 18)$ (b) covtype.binary: $(m; n) = (581,012; 54)$.

equivalent to applying the Guass-pd and CD-pd to the least squares system $A^T Ax = A^T b$, which is always consistent.

Despite the higher iteration cost of the Gaussian methods, their performance in these tests is comparable to the discrete methods. This suggests that the convergence rate ρ of the Gaussian methods is at least as good as their discrete counterparts.

7.2 Bound for Gaussian convergence

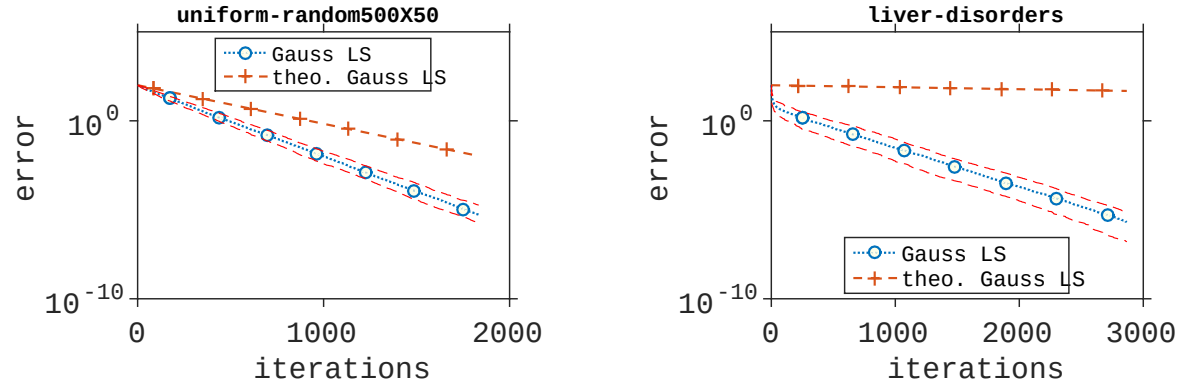
For $\xi \sim N(0, \Omega)$, we **conjecture** that

$$1 - \lambda_{\min}\left(\mathbf{E}\left[\frac{\xi\xi^T}{\|\xi\|_2^2}\right]\right) \leq 1 - \lambda_{\min}\left(\frac{\Omega}{\mathbf{Tr}\,(\Omega)}\right). \tag{54}$$

In numeric tests, this bound holds. In particular, in Figures 5a and 5b we plot the evolution of the error over the number iterations of Gauss-LS and the conjectured convergence rate (54) on a random Gaussian matrix and the `liver-disorders` problem [4]. Furthermore, we ran the Gauss-LS method 100 times and plot as dashed lines the 95% and 5% quantiles. These tests indicate that the convergence of the error is well within the conjectured bound (54). If (54) holds, then the convergence rate of the Gauss-LS method is the same as CD-LS, which is $1 - \lambda_{\min}(A^T A)/\|A\|_F^2$.

7.3 Positive Definite

First we compare the two methods Gauss-pd and CD-pd on synthetic data in Figure 6. Using the MATLAB function `hilbert`, we can generate positive definite matrices with very high condition number, see Figure 6(LEFT). Both methods converge slowly and, despite the full density, the Gauss-pd method has a similar performance to CD-pd. In Figure (6)(RIGHT) we compare the two methods on a system generated by the MATLAB function `sprandsym` (m, n, `density`, `rc`, `type`), where `density` is the percentage of nonzero entries, `rc` is the reciprocal of the condition number and `type=1` returns a positive definite matrix. The Gauss-pd method is more efficient at bringing



(a) `rand`(n, m) (b) `liver-disorders`

Figure 5: A comparison between the Gauss-LS method and the conjectured rate of convergence $\rho_{conj} = 1 - \lambda_{\min}(A^T A)/\|A\|_F^2$ on (a) `rand`(n, m) with $(m; n) = (500, 50)$ and a dense solution generated with x^* =`rand`$(n, 1)$ (b) `liver-disorders` with $(m; n) = (345, 6)$

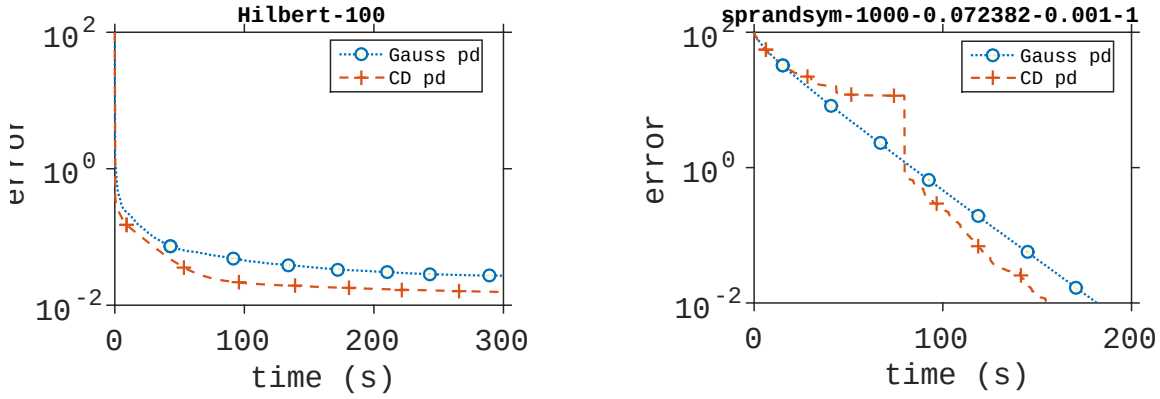
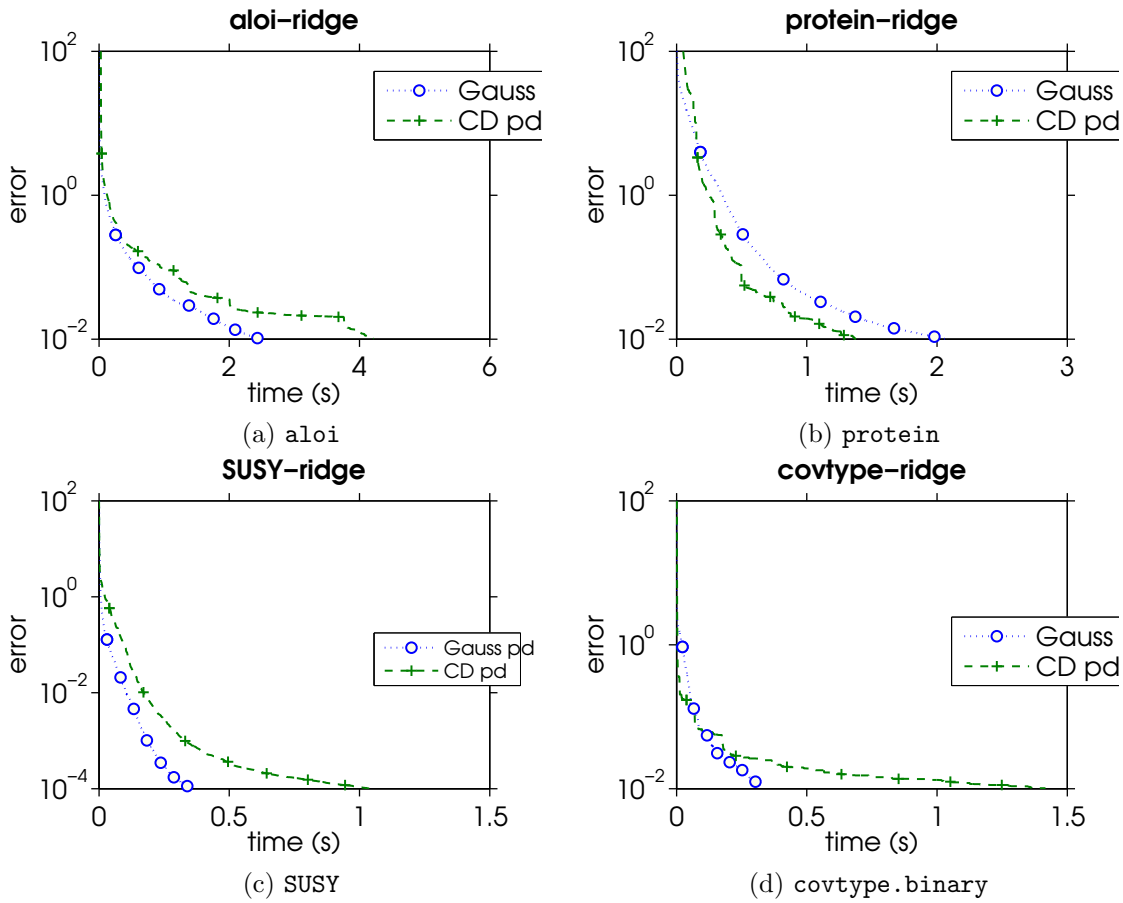


Figure 6: Synthetic MATLAB generated problem. The Gaussian methods are more efficient on sparse matrices. LEFT: The `Hilbert Matrix` with $n = 100$ and condition number $\|A\| \left\|A^{-1}\right\| = 6.5953 \times 10^{19}$. RIGHT: Sparse random matrix $A =$ `sprandsym` (n, `density`, `rc`, `type`) with $n = 1000$, `density`$= 1/\log(n^2)$ and `rc` $= 1/n = 0.001$. Dense solution generated with x^* =`rand`$(n, 1)$.



(a) `aloi` (b) `protein`

(c) `SUSY` (d) `covtype.binary`

Figure 7: The performance of Gaussian and Coordinate Descent pd methods on four ridge regression problems: (a) `aloi`: $(m;n) = (108,000;128)$ (b) `protein`: $(m;n) = (17,766;357)$ (c) `SUSY`: $(m;n) = (5 \times 10^6;18)$ (d) `covtype.binary`: $(m;n) = (581,012;54)$.

the error below 1%, and the CD-pd method is more efficient at bringing the error below 0.1%, on this sparse problem.

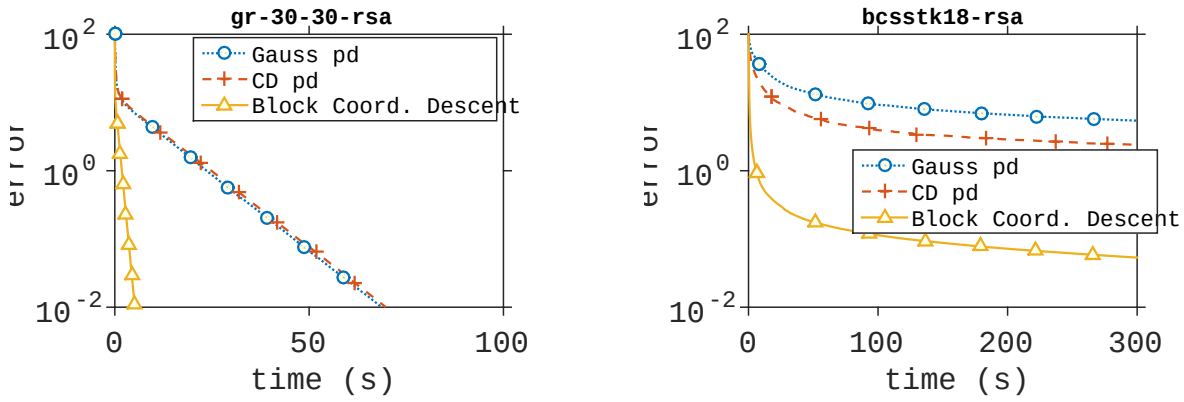
Next we test the Newton system $\nabla^2 f(x_0)d = -\nabla f(x_0)$ from four ridge-regression problems (55) using data from LIBSVM [4] where

$$\min_{x\in\mathbb{R}^n} f(x) \stackrel{\text{def}}{=} \|Ax-b\|_2^2 + \lambda\|x\|_2^2, \qquad \nabla f(x_0) = A^T b, \qquad \nabla^2 f(x) = A^T A + \lambda I. \tag{55}$$

We use $\lambda = 1$ as the regularization parameter. In reaching a low precision solution with 1% error, the CD-pd method and Gauss-pd method have a comparable performance, see Figure 7. Though, in bringing the error below 1%, the Gauss-pd method was more efficient, with the exception of the `protein` problem, where the CD-pd method was more efficient.

7.4 Block methods

To appraise the performance gain in using block variants, we performs tests with the the Randomized Newton method for positive definite matrices, which we will now refer to as the Block CD-pd



(a) `gr_30_30-rsa` (b) `bcsstk18`

Figure 8: The performance of the Gauss-pd, CD-pd and the Block CD-pd methods on two linear systems from the MatrixMarket (a) `gr_30_30-rsa` with $n = 900$, $nnz = 4322$ (`density`$= 0.53\%$) and $\kappa_2 = 12$. (b) `bcsstk18` with $n = 11948$, $nnz = 80519$ (`density`$= 0.1\%$) and $\kappa_2 = 4.3 \cdot 10^{10}$.

method. We compare the Gauss-pd, CD-pd and Block CD-pd methods on two positive definite matrices from the Matrix Market collection [3], see Figure 8. The right-hand side was generated using `rand(n,1)`. The size of blocks q in the Block CD-pd method was set to $q = \sqrt{n}$. To solve the $q \times q$ system required in the Block CD-pd, we use MATLAB's built-in direct solver, sometimes referred to as "back-slash". The Block CD-pd method converged much faster on both problems. The lower condition number ($\kappa_2 = 12$) of the `gr_30_30-rsa` problem resulted in fast convergence of all methods, see Figure 8a. While the high condition number ($\kappa_2 = 4.3 \cdot 10^4$) of the `bcsstk18` problem, resulted in a slow convergence for all methods, see Figure 8b.

Despite the clear advantage of using the block variant of the CD-pd method in Figure 8, applying a block method that uses a direct solver can be infeasible on very ill-conditioned problems. As an example, applying the Block CD-pd to the Hilbert system, and using MATLAB back-slash solver to solve the inner $q \times q$ systems, resulted in large numerical inaccuracies, and ultimately, prevented the method from converging. This occurred because the submatrices of the Hilbert matrix are also very ill-conditioned.

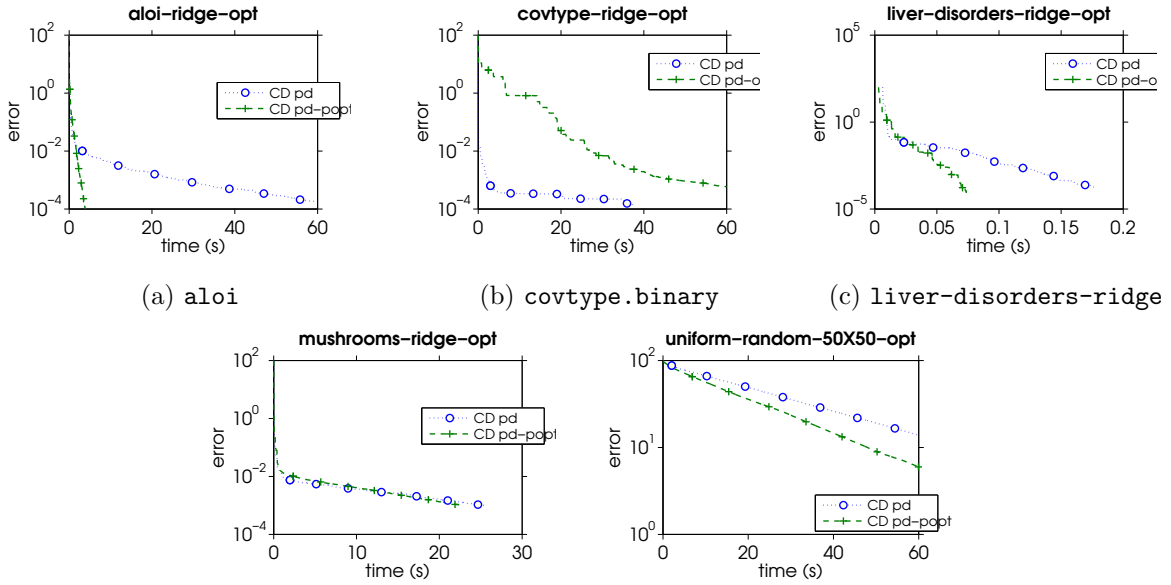
7.5 Comparison between Optimized and Convenient

We compare the practical performance of using the convenient probabilities (44) against using the optimized probabilities by solving (43).

In Table 2 we compare the different convergence rates for the CD-pd method, where ρ_c is the convenient convergence rate, ρ^* the optimized convergence rate, $1/n$ is the lower bound, and in the final "optimized time(s)" column the time taken to compute ρ^*. In Figure 9, we compare the empirical convergence of the CD-pd method when using the convenient probabilities (44) and CD-pd-opt, the CD-pd method with the optimized probabilities, on four ridge regression problems and a uniform random matrix. In most cases using the optimized probabilities results in a faster convergence, see Figures 9a, 9c and 9e. In particular, the 9.457 second spent calculating the optimal probabilities for `aloi` paid off with a convergence that was 55 seconds faster. The `mushrooms`

name	ρ_c	ρ^*	$1-1/n$	optimized time(s)
`rand`(50,50)	$1-2\cdot 10^{-6}$	$1-3.05\cdot 10^{-6}$	$1-2.10^{-2}$	1.076
`mushrooms-ridge`	$1-5.86\cdot 10^{-6}$	$1-7.15\cdot 10^{-6}$	$1-8.93\cdot 10^{-3}$	5.777
`aloi-ridge`	$1-2.17\cdot 10^{-7}$	$1-1.26\cdot 10^{-4}$	$1-7.81\cdot 10^{-3}$	9.457
`liver-disorders-ridge`	$1-5.16\cdot 10^{-4}$	$1-8.25\cdot 10^{-3}$	$1-1.67\cdot 10^{-1}$	0.413
`covtype-ridge`	$1-7.57\cdot 10^{-14}$	$1-1.48\cdot 10^{-6}$	$1-1.85\cdot 10^{-2}$	1.449

Table 2: Optimizing the rate for CD-pd



(a) `aloi` (b) `covtype.binary` (c) `liver-disorders-ridge`

(d) `mushrooms-ridge-opt` (e) `uniform-random-50X50-opt`

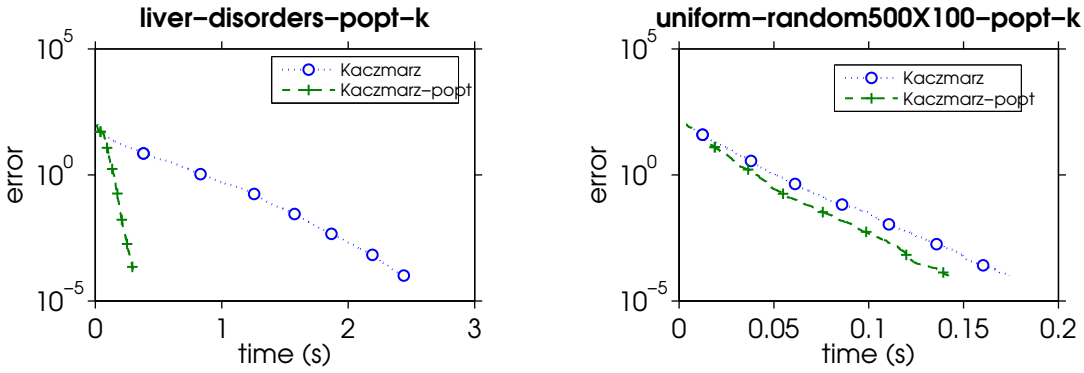
Figure 9: The performance of CD-pd and optimized CD-pd methods on (a) `aloi`: $(m;n) = (108,000;128)$ (b) `covtype.binary`: $(m;n) = (581,012;54)$ (c) `liver-disorders`: $(m;n) = (345,6)$ (c)`mushrooms`: $(m;n) = (8124,112)$ (d) `uniform-random-50X50`

problem was insensitive to the choice of probabilities 9d. Finally despite ρ^* being much less than ρ_c on `covtype`, see Table 2, using optimized probabilities resulted in a much slower method, see Figure 9b. This goes as warning, that optimizing an upper bound on the rate of convergence, does not guarantee that the method will be faster in practice.

In Table 3 we compare the different convergence rates for the RK method. In Figure 10, we then compare the empirical convergence of the RK method when using the convenient probabilities (44) and RK-opt, the RK method with the optimized probabilities by solving (43). The rates ρ^* and ρ_c for the `rand`(500,100) problem are similar, and accordingly, both the convenient and optimized variant converge at a similar rate in practice, see Figure 10b. While the difference in the rates ρ^* and ρ_c for the `liver-disorders` is more pronounced, and in this case, the 1.762 seconds invested in obtaining the optimized probability distribution paid off in practice, as the optimized method converged 2.135 seconds before the RK method with the convenient probability distribution, see Figure 10a.

name	ρ_c	ρ^*	$1-1/n$	optimized time(s)
`rand`(500,100)	$1-3.37\cdot10^{-3}$	$1-4.27\cdot10^{-3}$	$1-1.10^{-2}$	57.643
`liver-disorders`	$1-5.16\cdot10^{-4}$	$1-4.04\cdot10^{-3}$	$1-1.67\cdot10^{-1}$	1.762

Table 3: Optimizing the rate for Kaczmarz



(a) `liver-disorders-popt-k` (b) `rand`(500,100)

Figure 10: The performance of Kaczmarz and optimized Kaczmarz methods on (a) `liver-disorders`: $(m;n)=(345,6)$ (b) `rand`(500,100)

We conclude from these tests that the choice of probability distribution can greatly affect the performance of the method. Thus it is worthwhile to develop approximate solutions to (42).

8 Conclusion

We present a unifying framework for the randomized Kaczmarz method, randomized Newton method, randomized coordinate descent method and random Gaussian pursuit. Not only can we recover these methods by selecting appropriately the parameters S and B, but also, we can analyse them and their block variants through a single Theorem 4.2. Furthermore, we obtain a new lower bound for all these methods in Theorem 4.1, and in the discrete case, recover all known convergence rates expressed in terms of the scaled condition number in Theorem 5.1.

The Theorem 5.1 also suggests a preconditioning strategy. Developing preconditioning methods are important for reaching a higher precision solution on ill-conditioned problems. For as we have seen in the numerical experiments, the randomized methods struggle to bring the solution within 10^{-2} relative error when the matrix is ill-conditioned.

This is also a framework on which randomized methods for linear systems can be designed. As an example, we have designed new RK block variant and a new Gaussian Kaczmarz method. Furthermore, the flexibility of our framework and the general convergence Theorems 4.2 and 4.1 allows one to tailor the probability distribution of S to a particular problem class. For instance, other continuous distributions such uniform, or other discrete distributions such Poisson might be more suited to a particular class of problems.

Numeric tests reveal that the new Guassian methods designed for overdetermined systems are competitive on sparse problems, as compared to the Karczmarz and CD-LS methods. The Gauss-

pd also proved competitive as compared to CD-pd on all tests. Though, when applicable, the combined efficiency of using a direct solver and an iterative procedure, such as in Block CD-pd method, proved the most efficient.

The work opens up many possible future venues of research. Including investigating accelerated convergence rates through preconditioning strategies based on Theorem 5.1 or by obtaining approximate optimized probability distributions (43).

Acknowledgments

The authors would like to thank Prof. Sandy Davie for useful discussions relating to Lemma 8.2.

References

[1] S Bellavia. "An Inexact Interior Point Method". *Journal of Optimization Theory and Applications* 96.1 (1998), pp. 109–121.

[2] Y. Bengio, O. Delalleau, and N. Le Roux. "Label Propagation and Quadratic Criterion". In: *Semi-Supervised Learning*. Ed. by O. Chapelle, B. Schölkopf, and A. Zien. MIT Press, 2006, pp. 193–216.

[3] R. F. Boisvert et al. "Matrix Market : A Web Resource for Test Matrix Collections". In: *The Quality of Numerical Software: Assessment and Enhancement*. Ed. by R. Boisvert. London: Chapman & Hall, 1997, pp. 125–137.

[4] C.-C. Chang and C.-J. Lin. "LIBSVM: A Library for Support Vector Machines". *ACM Transactions on Intelligent Systems and Technology* 2.3 (Apr. 2011), pp. 1–27.

[5] R. S. Dembo, S. C. Eisenstat, and T. Steihaug. "Inexact Newton Methods". *SIAM Journal on Numerical Analysis* 19.2 (1982), pp. 400–408.

[6] J. W. Demmel. "The Probability that a Numerical Analysis Problem is Difficult". *Mathematics of Computation* 50.182 (1988), pp. 449–449.

[7] P. Drineas et al. "Faster Least Squares Approximation". *Numerische Mathematik* 117.2 (2011), pp. 219–249.

[8] A. Edelman. "On the Distribution of a Scaled Condition Number". *Mathematics of Computation* 58.197 (1992), pp. 185–185.

[9] S. C. Eisenstat and H. F. Walker. "Choosing the Forcing Terms in an Inexact Newton Method". *SIAM Journal on Scientific Computing* 17 (1994), pp. 16–32.

[10] J. Gondzio. "Convergence Analysis of an Inexact Feasible Interior Point Method for Convex Quadratic Programming". *SIAM Journal on Optimization* 23.3 (2013), pp. 1510–1527.

[11] R. Heijmans. "When Does the Expectation of a Ratio Equal the Ratio of Expectations?" *Statistical Papers* 40 (1999), pp. 107–115.

[12] M. R. Hestenes and E. Stiefel. "Methods of Conjugate Gradients for Solving Linear Systems". *Journal of research of the National Bureau of Standards* 49.6 (1952).

[13] M. S. Karczmarz. "Angenaherte Auflosung von Systemen linearer Gleichungen". *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles. Série A, Sciences Mathématiques* 35 (1937), pp. 355–357.

[14] Y. T. Lee and A. Sidford. "Efficient Accelerated Coordinate Descent Methods and Faster Algorithms for Solving Linear Systems". *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS* (2013), pp. 147–156.

[15] D. Leventhal and A. S. Lewis. "Randomized Methods for Linear Constraints: Convergence Rates and Conditioning". *Mathematics of Operations Research* 35.3 (2010), p. 22.

[16] J. Liesen and Z. Strakos. *Krylov Subspace Methods : Principles and Analysis*. Oxford: Oxford University Press, 2014, pp. 1–50.

[17] A. Ma et al. "Convergence Properties of the Randomized Extended Gauss-Seidel and Kaczmarz methods". *arXiv:1503.08235* (2015), pp. 1–16.

[18] D. Needell. "Randomized Kaczmarz solver for noisy linear systems". *BIT* 50.2 (2010), pp. 395–403.

[19] D. Needell and J. A. Tropp. "Paved with Good Intentions: Analysis of a Randomized Block Kaczmarz Method". *Linear Algebra and Its Applications* 441.August (2012), pp. 199–221.

[20] D. Needell, R. Zhao, and A. Zouzias. "Randomized Block Kaczmarz Method with Projection for Solving Least Squares". *arXiv:1403.4192* (2014).

[21] Y. Nesterov. *Random Gradient-Free Minimization of Convex Functions*. Tech. rep. Louvain: ECORE Universite catholique de Louvain, 2011, pp. 1–34.

[22] P. Oswald and W. Zhou. "Convergence analysis for Kaczmarz-type methods in a Hilbert space framework". *Linear Algebra and its Applications* 478 (2015), pp. 131–161.

[23] C. C. Paige and M. A. Saunders. "Solution of Sparse Indefinite Systems of Linear Equations". *SIAM J. Numer. Anal.* 12.4 (1975), pp. 617–629.

[24] Z. Qu et al. "SDNA: Stochastic Dual Newton Ascent for Empirical Risk Minimization". *arXiv:1502.02268v1* (2015).

[25] A. Ramdas. "Rows vs Columns for Linear Systems of Equations - Randomized Kaczmarz or Coordinate Descent ?" *arXiv:1406.5295* (2014).

[26] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Vol. 104. Monographs on Statistics and Applied Probability. London: Chapman & Hall, 2005.

[27] Y. Saad and M. H. Schultz. "GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems". *SIAM Journal on Scientific and Statistical Computing* 7.3 (1986), pp. 856–869.

[28] S. U. Stich, C. L. Müller, and B. Gärtner. "Optimization of Convex Functions with Random Pursuit". *SIAM Journal on Optimization* 23.2 (2014), pp. 1284–1309.

[29] S. U. Stich. "Variable Metric Random Pursuit". *arXiv:1210.5114v3* (2014).

[30] S. U. Stich. "Convex Optimization with Random Pursuit". PhD thesis. ETH Zurich, 2014.

[31] T. Strohmer and R. Vershynin. "A Randomized Kaczmarz Algorithm with Exponential Convergence". *Journal of Fourier Analysis and Applications* 15.2 (2009), pp. 262–278.

[32] C. Wang and A. Xu. "An Inexact Accelerated Proximal Gradient Method and a Dual Newton-CG Method for the Maximal Entropy Problem". *Journal of Optimization Theory and Applications* 157.2 (2013), pp. 436–450.

[33] X.-Y. Zhao, D. Sun, and K.-C. Toh. "A Newton-CG Augmented Lagrangian Method for Semidefinite Programming". *SIAM Journal on Optimization* 20.4 (2010), pp. 1737–1765.

[34] A. Zouzias and N. M. Freris. "Randomized Extended Kaczmarz for Solving Least-Squares". *SIAM Journal on Matrix Analysis and Applications* 34.2 (2013), pp. 773–793.

Appendix

8.1 The Expected Gaussian Projection Matrix is Positive Definite

Lemma 8.1. *Let $\xi \sim N(0,\Omega)$ and $\Omega \in \mathbb{R}^{n\times n}$ be a positive definite matrix then $\mathbf{E}\left[\frac{\xi\xi^T}{\xi^T\xi}\right]$ is positive definite and satisfies the bounds*

$$\Omega\frac{1}{n\cdot\lambda_{\max}^2(\Omega)} \preceq \mathbf{E}\left[\frac{\xi\xi^T}{\|\xi\|_2^2}\right] \preceq \Omega\frac{1}{n\cdot\lambda_{\min}^2(\Omega)}. \tag{56}$$

Proof. Let $\xi = \Omega^{1/2}\eta$ where $\eta \sim N(0,I)$. First we collect two results. Note that from the extremal characterization of eigenvalues

$$\lambda_{\max}(\Omega) = \max_{\eta\in\mathbb{R}^n}\frac{\|\eta\|_\Omega}{\|v\|} \quad \text{and} \quad \lambda_{\min}(\Omega) = \min_{\eta\in\mathbb{R}^n}\frac{\|\eta\|_\Omega}{\|\eta\|}$$

we have

$$\frac{1}{\lambda_{\max}^2(\Omega)} \leq \frac{\|\eta\|_2^2}{\|\eta\|_\Omega^2} \leq \frac{1}{\lambda_{\min}^2(\Omega)}. \tag{57}$$

Furthermore, using the independence of $\eta\eta^T/\|\eta\|_2^2$ and $\|\eta\|_2^2$ we have that $\operatorname{cov}\left(\eta\eta^T/\|\eta\|_2^2, \|\eta\|_2^2\right) = 0$, thus according to [11] we have

$$\mathbf{E}\left[\frac{\eta\eta^T}{\|\eta\|_2^2}\right] = \frac{\mathbf{E}\left[\eta\eta^T\right]}{\mathbf{E}\left[\|\eta\|_2^2\right]} = \frac{1}{n}I.$$

Now using $\xi\xi^T/\|\xi\|_2^2 = \Omega^{1/2}\eta\eta^T\Omega^{1/2}/\|\eta\|_\Omega^2$, and taking expectation and using (57) we have

$$\mathbf{E}\left[\frac{\xi\xi^T}{\|\xi\|_2^2}\right] \preceq \frac{1}{\lambda_{\min}^2(\Omega)}\Omega^{1/2}\mathbf{E}\left[\frac{\eta\eta^T}{\|\eta\|_2^2}\right]\Omega^{1/2} = \frac{1}{n\cdot\lambda_{\min}^2(\Omega)}\Omega,$$

where we used $\mathbf{E}\left[\eta\eta^T/\|\eta\|_2^2\right] = \frac{1}{n}I$. The left hand side of (56) follows using analogous arguments. □

8.2 Gaussian 2D Expected Projection

Lemma 8.2. *Let $\xi \sim N(0,\Omega)$ and $\Omega \in \mathbb{R}^{2\times 2}$ be a positive definite matrix, then*

$$\mathbf{E}\left[\frac{\xi\xi^T}{\xi^T\xi}\right] = \frac{\Omega^{1/2}}{\mathbf{Tr}\left(\Omega^{1/2}\right)}. \tag{58}$$

Proof. To prove this, first we reduce to the problem to determining (58) for uncorrelated Gaussian random variables. This part of the proof is valid for Gaussian vectors of any dimension.

Let us write $S(\xi)$ for the random vector $\xi/\|\xi\|_2$ (if $\xi = 0$, we set $S(\xi) = 0$). Using this notation, we can write

$$\mathbf{E}\left[\xi(\xi^T\xi)^{-1}\xi^T\right] = \mathbf{E}\left[S(\xi)(S(\xi))^T\right] = \mathbf{Cov}\left[S(\xi)\right],$$

where the last identity follows since $\mathbf{E}\left[S(\xi)\right] = 0$, which in turn holds as the Gaussian distribution is centrally symmetric.

Using the spectral decomposition $\Omega = UDU^T$, where U is an orthogonal matrix and D is a diagonal matrix containing the eigenvalues, then $\xi = Uu$ where $u \sim N(0, D)$. Moreover, note that

$$S(U^T\xi) = \frac{U^T\xi}{\|U^T\xi\|_2} = \frac{U^T\xi}{\|\xi\|_2} = U^T S(\xi).$$

Multiplying both sides by U we obtain $US(U^T\xi) = S(\xi)$, from which we conclude that

$$\mathbf{Cov}\left[S(\xi)\right] = U\mathbf{Cov}\left[S(U^T\xi)\right]U^T = U\mathbf{Cov}\left[S(u)\right]U^T. \tag{59}$$

Now based on Lemma 8.3 we have

$$\mathbf{Cov}\left[S(u)\right] = \frac{D^{1/2}}{\mathbf{Tr}\left(D^{1/2}\right)}. \tag{60}$$

Plugging this into (59), we get

$$\mathbf{Cov}\left[S(\xi)\right] = \frac{UD^{1/2}U^T}{\mathbf{Tr}\left(D^{1/2}\right)} = \frac{\Omega^{1/2}}{\mathbf{Tr}\left(\Omega^{1/2}\right)},$$

as desired. □

Lemma 8.3. *Let $\sim N(0, D)$ and $D \in \mathbb{R}^{2\times 2}$ be a diagonal positive definite matrix, then*

$$\mathbf{E}\left[\frac{\xi\xi^T}{\xi^T\xi}\right] = \frac{D^{1/2}}{\mathbf{Tr}\left(D^{1/2}\right)}. \tag{61}$$

Proof. Let σ_x^2 and σ_y^2 be the two diagonal elements of D. First, suppose that $\sigma_x = \sigma_y$. Then $\xi = \sigma_x\eta$ where $\eta \sim N(0, I)$ and

$$\mathbf{E}\left[\frac{\xi\xi^T}{\xi^T\xi}\right] = \frac{\sigma_x^2}{\sigma_x^2}\mathbf{E}\left[\frac{\eta\eta^T}{\eta^T\eta}\right] = \frac{1}{n}I = \frac{D^{1/2}}{\mathbf{Tr}\left(D^{1/2}\right)}.$$

Now suppose that $\sigma_x \neq \sigma_y$.

Off-diagonal elements: To calculate the off-diagonal term in (60) we integrate

$$\mathbf{E}\left[\frac{\xi_1\xi_2}{\xi_1^2+\xi_2^2}\right] = \frac{1}{2\pi\sigma_x\sigma_y}\int_{\mathbb{R}^2}\frac{xy}{x^2+y^2}e^{-\frac{1}{2}\left(x^2/\sigma_x^2+y^2/\sigma_y^2\right)}dxdy \overset{\text{def}}{=} \int_{\mathbb{R}^2} h(x,y)dxdy.$$

As $-h(x,y) = h(-x,y)$ and $-h(x,y) = h(x,-y)$, we have that $\int_{\mathbb{R}^2} h(x,y)dxdy = 0$.

Diagonal elements: If σ_x and σ_y were integers then $\frac{\xi_1^2}{\xi_1^2+\xi_2^2} \sim B(\sigma_y/2, \sigma_x/2)$, where $B(\sigma_x, \sigma_y)$ is the Beta distribution. The expected value of which is known to be $\sigma_x/(\sigma_x+\sigma_y)$. Unfortunately as σ_x and σ_y are not necessarily integer, we must calculate the diagonal terms of the covariance matrix by integrating

$$\mathbf{E}\left[\frac{\xi_1^2}{\xi_1^2+\xi_2^2}\right] = \frac{1}{2\pi\sigma_x\sigma_y}\int_{\mathbb{R}^2}\frac{x^2}{x^2+y^2}e^{-\frac{1}{2}\left(x^2/\sigma_x^2+y^2/\sigma_y^2\right)}dxdy.$$

Using polar coordinates $x = R\cos(\theta)$ and $y = R\sin(\theta)$ we have

$$\int_{\mathbb{R}^2}\frac{x^2}{x^2+y^2}e^{-\frac{1}{2}\left(x^2/\sigma_x^2+y^2/\sigma_y^2\right)}dxdy = \int_0^{2\pi}\int_0^{\infty} R\cos^2(\theta)e^{-\frac{R^2}{2}\left(\cos(\theta)^2/\sigma_x^2+\sin(\theta)^2/\sigma_y^2\right)}dRd\theta. \tag{62}$$

Let $C(\theta) = \left(\cos(\theta)^2/\sigma_x^2 + \sin(\theta)^2/\sigma_y^2\right)$. Note that

$$\int_0^{\infty} Re^{-\frac{C(\theta)R^2}{2}}dR = -\frac{1}{C(\theta)}e^{-\frac{C(\theta)R^2}{2}}\Big|_0^{\infty} = \frac{1}{C(\theta)}. \tag{63}$$

This applied in (62) gives

$$\mathbf{E}\left[\frac{\xi_1^2}{\xi_1^2+\xi_2^2}\right] = \frac{1}{2\pi\sigma_x\sigma_y}\int_0^{2\pi}\frac{\cos^2(\theta)}{\cos(\theta)^2/\sigma_x^2+\sin(\theta)^2/\sigma_y^2}d\theta = \frac{b}{\pi}\int_0^{\pi}\frac{\cos^2(\theta)}{\cos^2(\theta)+b^2\sin^2(\theta)}d\theta,$$

where $b = \sigma_x/\sigma_y$. Multiplying the numerator and denominator of the integrand by $\sec^4(x)$ gives the integral

$$\mathbf{E}\left[\frac{\xi_1^2}{\xi_1^2+\xi_2^2}\right] = \frac{b}{\pi}\int_0^{\pi}\frac{\sec^2(\theta)}{\sec(\theta)^2\left(1+b^2\tan^2(\theta)\right)}d\theta.$$

Substituting $u = \tan(\theta)$ so that $u^2+1 = \sec^2(\theta)$ and $du = \sec^2(\theta)d\theta$ and using the partial fractions

$$\frac{1}{(u^2+1)(1+b^2u^2)} = \frac{1}{1-b^2}\left(\frac{1}{u^2+1} - \frac{b^2}{b^2u^2+1}\right),$$

gives the integral

$$\begin{aligned}\int\frac{du}{(u^2+1)(1+b^2u^2)} &= \frac{1}{1-b^2}\left(\arctan(u) - b\arctan(bu)\right)\\ &= \frac{1}{1-b^2}\left(\theta - b\arctan(b\tan(\theta))\right).\end{aligned}$$

To apply the limits of integration, we must take care because of the singularity at $\theta = \pi/2$. For this, consider the limits

$$\lim_{\theta\to(\pi/2)^-}\arctan(b\tan(\theta)) = \frac{\pi}{2}, \qquad \lim_{\theta\to(\pi/2)^+}\arctan(b\tan(\theta)) = -\frac{\pi}{2}.$$

Applying this to

$$\lim_{t\to(\pi/2)^-}\frac{1}{1-b^2}\left(\theta - b\arctan(b\tan(\theta))\right)\Big|_0^t = \frac{1}{1-b^2}\frac{\pi}{2}(1-b) = \frac{\pi}{2(1+b)}.$$

Applying a similar argument for calculating the limits from $\pi/2^+$ to π, we find

$$\mathbf{E}\left[\frac{\xi_1^2}{\xi_1^2+\xi_2^2}\right] = \frac{2b}{\pi}\frac{\pi}{2(1+b)} = \frac{\sigma_x}{\sigma_y+\sigma_x}.$$

Repeating the same steps with x swapped for y we obtain the other diagonal element. □