Convergence Properties of the Randomized Extended Gauss-Seidel and Kaczmarz Methods

Anna Ma
Claremont Graduate University

Deanna Needell
Claremont McKenna College

Aaditya Ramdas
University of California - Berkeley

CONVERGENCE PROPERTIES OF THE RANDOMIZED EXTENDED GAUSS-SEIDEL AND KACZMARZ METHODS

ANNA MA, DEANNA NEEDELL, AADITYA RAMDAS

Abstract. The Kaczmarz and Gauss-Seidel methods both solve a linear system $\boldsymbol{X\beta} = \boldsymbol{y}$ by iteratively refining the solution estimate. Recent interest in these methods has been sparked by a proof of Strohmer and Vershynin which shows the *randomized* Kaczmarz method converges linearly in expectation to the solution. Lewis and Leventhal then proved a similar result for the randomized Gauss-Seidel algorithm. However, the behavior of both methods depends heavily on whether the system is under or overdetermined, and whether it is consistent or not. Here we provide a unified theory of both methods, their variants for these different settings, and draw connections between both approaches. In doing so, we also provide a proof that an extended version of randomized Gauss-Seidel converges linearly to the least norm solution in the underdetermined case (where the usual randomized Gauss Seidel fails to converge). We detail analytically and empirically the convergence properties of both methods and their extended variants in all possible system settings. With this result, a complete and rigorous theory of both methods is furnished.

1. Introduction. We consider solving a linear system of equations

$$\boldsymbol{X\beta} = \boldsymbol{y}, \tag{1.1}$$

for a (real or complex) $m \times n$ matrix $\boldsymbol{X}$, in various problem settings. Recent interest in the topic was reignited when Strohmer and Vershynin [26] proved the linear[1] convergence rate of the Randomized Kaczmarz (RK) algorithm that works on the rows of $\boldsymbol{X}$ (data points). Following that, Leventhal and Lewis [15] proved the linear convergence of a Randomized Gauss-Seidel (RGS), i.e. Randomized Coordinate Descent, algorithm that works on the columns of $\boldsymbol{X}$ (features).

When the system of equations is inconsistent (i.e. has no exact solution), as is typically the case when $m > n$ in real-world overconstrained systems, RK is known to not converge to the ordinary least squares solution

$$\boldsymbol{\beta_{LS}} := \arg\min_{\boldsymbol{\beta}} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 \tag{1.2}$$

as studied by Needell [18]. Zouzias and Freris [30] extended the RK method with the modified Randomized Extended Kaczmarz (REK) algorithm, which linearly converges to $\boldsymbol{\beta_{LS}}$. Interestingly, in this setting, we will argue in Section 3.3 that RGS does converge to $\boldsymbol{\beta_{LS}}$ without any special extensions.

1.1. Motivation and contribution. The above introduction represents only half the story. When $m < n$, there are fewer constraints than variables, and the system has infinitely many solutions. In this case, especially if we have no prior reason to believe any additional sparsity in the signal structure, we are often interested in finding the least Euclidean norm solution:

$$\boldsymbol{\beta_{LN}} := \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{X\beta}. \tag{1.3}$$

While RGS converges to $\boldsymbol{\beta_{LS}}$ in the overcomplete setting, we shall argue in Section 3.3 that in the undercomplete setting it does not converge to $\boldsymbol{\beta_{LN}}$. We will also argue that RK does converge to $\boldsymbol{\beta_{LN}}$ without any extensions in this setting.

---

[1] Mathematicians often refer to linear convergence as exponential convergence.

The main contribution of our paper is to provide a unified theory of these related iterative methods. We will also construct an extension to RGS that parallels REK, which unlike RGS does converge to $\boldsymbol{\beta_{LN}}$ (just as REK, unlike RK, converges to $\boldsymbol{\beta_{LS}}$). Some desired properties for this algorithm include that it should *also* converge linearly, not require much extra computation, and work well in simulations. We shall see that our Randomized Extended Gauss-Seidel (REGS) method does indeed possess these desired properties. A summary of this unified theory is provided in Table 1.1.

Method	Overconstrained, consistent : convergence to $\boldsymbol{\beta^\star}$?	Overconstrained, inconsistent : convergence to $\boldsymbol{\beta_{LS}}$?	Underconstrained : convergence to $\boldsymbol{\beta_{LN}}$?
RK	Yes [26]	No [18, Thm. 2.1]	Yes (Sec. 3.3)
REK	Yes [30]	Yes [30]	Yes (Sec. 3.3)
RGS	Yes [15]	Yes [15]	No (Sec. 3.3)
REGS	Yes (Remark 1)	Yes (Sec. 4.3)	Yes (Thm. 4.1)

<div align="center">TABLE 1.1</div>

Summary of convergence properties for the overdetermined and consistent setting, overdetermined and inconsistent setting, and underdetermined settings. We write $\boldsymbol{\beta^\star}$ to denote the solution to (1.1) in the overdetermined consistent setting, with $\boldsymbol{\beta_{LS}}$ and $\boldsymbol{\beta_{LN}}$ being defined in (1.2) and (1.3) for the other two settings.

1.2. Paper Outline. In Section 2 we recap the three main existing algorithms mentioned in the introduction (RK, RGS, REK). We discuss the performance of these algorithms in the three natural settings described in Table 1.1 in Section 3. Section 4 introduces our proposed algorithm (REGS) and proves its linear convergence to the least norm solution, completing the theoretical framework. Lastly, we end with some simulation experiments in Section 5 to demonstrate the tightness and usefulness of our theory, and conclude in Section 6.

2. Existing Algorithms and Related Work. In this section, we will summarize the algorithms mentioned in the introduction, i.e. RK, RGS and REK. We will describe their iterative update rules and mention their convergence guarantees, leaving the details of convergence to the next section. Throughout the paper we will use the notation $\boldsymbol{X}^i$ to represent the ith row of $\boldsymbol{X}$ (or ith entry in the case of a vector) and $\boldsymbol{X}_{(j)}$ to denote the jth column of a matrix $\boldsymbol{X}$. We will write the estimation $\boldsymbol{\beta}$ as a column vector. We write vectors and matrices in boldface, and constants in standard font.

2.1. Randomized Kaczmarz (RK). The Kaczmarz method was first introduced in the notable work of Kaczmarz [14]. It has gained recent interest in tomography research where it is known as the *Algebraic Reconstruction Technique* (ART) [8, 17, 1, 13]. Although in its original form the method selects rows in a deterministic fashion (often simply cyclically), it has been well observed that a random selection scheme reduces the possibility of a poor choice of row ordering [9, 12]. Earlier convergence analysis of the randomized variant were obtained (e.g. [29]), but yielded bounds with expressions that were difficult to evaluate. Strohmer and Vershynin [26] showed that the RK method described above has an expected linear convergence rate to the solution $\boldsymbol{\beta^\star}$ of (1.1), and are the first to provide an explicit convergence rate in expectation which depends only on the geometric properties of the system. This work was extended by Needell [18] to the inconsistent case, analyzed almost surely

<div align="center">2</div>

by Chen and Powell [3], accelerated in several ways [7, 6, 24, 21, 20], and extended to more general settings [15, 25, 19].

We describe here the randomized variant of the Kaczmarz method put forth by Strohmer and Vershynin [26]. Taking $\boldsymbol{X}, \boldsymbol{y}$ as input and starting from an arbitrary initial estimate for $\boldsymbol{\beta}$ (for example $\boldsymbol{\beta}_0 = \boldsymbol{0}$), RK repeats the following in each iteration. First, a random row $i \in \{1, ..., m\}$ is selected with probability proportional to its Euclidean norm, i.e.

$$\Pr(\text{row} = i) = \frac{\|\boldsymbol{X}^i\|_2^2}{\|\boldsymbol{X}\|_F^2},$$

where $\|\boldsymbol{X}\|_F$ denotes the Frobenius norm of $\boldsymbol{X}$. Then, project the current iterate onto that row, i.e.

$$\boldsymbol{\beta}_{t+1} := \boldsymbol{\beta}_t + \frac{(y^i - \boldsymbol{X}^i \boldsymbol{\beta}_t)}{\|\boldsymbol{X}^i\|_2^2}(\boldsymbol{X}^i)^*, \tag{2.1}$$

where here and throughout $\boldsymbol{X}^*$ denotes the (conjugate) transpose of $\boldsymbol{X}$.

Intuitively, this update can be seen as greedily satisfying the ith equation in the linear system. Indeed, it is easy to see that after the update,

$$\boldsymbol{X}^i \boldsymbol{\beta}_{t+1} = y^i. \tag{2.2}$$

Referring to (1.2) and defining

$$L(\boldsymbol{\beta}) = \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \tfrac{1}{2}\sum_{i=1}^{m}(y^i - \boldsymbol{X}^i \boldsymbol{\beta})^2,$$

we can alternatively interpret this update as stochastic gradient descent (choosing a random data-point on which to update), where the step size is the inverse Lipschitz constant of the stochastic gradient

$$\nabla^2 \tfrac{1}{2}(y^i - \boldsymbol{X}^i \boldsymbol{\beta})^2 = \|\boldsymbol{X}^i\|_2^2.$$

2.2. Randomized Extended Kaczmarz (REK). For inconsistent systems, the RK method does not converge to the least-squares solution as one might desire. This fact is clear since the method at each iteration projects completely onto a selected solution space, being unable to break the so-called *convergence horizon*. One approach to overcome this is to use relaxation parameters, so that the estimates are not projected completely onto the subspace at each iteration [28, 27, 2, 10]. Recently, Zouzias and Freris [30] proposed a variant of the RK method motivated by the work of Popa [23] which instead includes a random projection to iteratively reduce the component of $\boldsymbol{y}$ which is orthogonal to the range of $\boldsymbol{X}$. This method, named Randomized Extended Kaczmarz (REK) can be described by the following iteration updates, which can be initialized with $\boldsymbol{\beta}_0 = \boldsymbol{0}$ and $\boldsymbol{z}_0 = \boldsymbol{y}$:

$$\boldsymbol{\beta}_{t+1} := \boldsymbol{\beta}_t + \frac{(y^i - z_t^i - \boldsymbol{X}^i \boldsymbol{\beta}_t)}{\|\boldsymbol{X}^i\|_2^2}(\boldsymbol{X}^i)^*, \quad \boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \frac{\langle \boldsymbol{X}_{(j)}, \boldsymbol{z}_t \rangle}{\|\boldsymbol{X}_{(j)}\|_2^2}\boldsymbol{X}_{(j)}. \tag{2.3}$$

Here, a column $j \in \{1, ..., n\}$ is also selected at random with probability proportional to its Euclidean norm:

$$\Pr(\text{column} = j) = \frac{\|\boldsymbol{X}_{(j)}\|_2^2}{\|\boldsymbol{X}\|_F^2}, \tag{2.4}$$

3

and again $\boldsymbol{X}_{(j)}$ denotes the jth column of $\boldsymbol{X}$. Here, $\boldsymbol{z}_t$ approximates the component of $\boldsymbol{y}$ which is orthogonal to the range of $\boldsymbol{X}$, allowing for the iterates $\boldsymbol{\beta}_t$ to converge to the true least-squares solution of the system. Zouzias and Freris [30] prove that REK converges linearly in expectation to this solution $\boldsymbol{\beta_{LS}}$.

2.3. Randomized Gauss-Seidel (RGS). Again taking $\boldsymbol{X}, \boldsymbol{y}$ as input and starting from an arbitrary $\boldsymbol{\beta}_0$, the Randomized Gauss-Seidel (RGS) method (or the Randomized Coordinate Descent method) repeats the following in each iteration. First, a random column $j \in \{1, ..., n\}$ is selected as in (2.4). We then minimize the objective $L(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ with respect to this coordinate to get

$$\boldsymbol{\beta}_{t+1} := \boldsymbol{\beta}_t + \frac{\boldsymbol{X}_{(j)}^*(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_t)}{\|\boldsymbol{X}_{(j)}\|_2^2}\boldsymbol{e}_{(j)} \tag{2.5}$$

where $\boldsymbol{e}_{(j)}$ is the jth coordinate basis column vector (all zeros with a 1 in the jth position). It can be seen as greedily minimizing the objective with respect to the jth coordinate. Indeed, letting $\boldsymbol{X}_{(-j)}, \boldsymbol{\beta}^{-j}$ represent $\boldsymbol{X}$ without its jth column and $\boldsymbol{\beta}$ without its jth coordinate,

$$\frac{\partial L}{\partial \boldsymbol{\beta}^j} = -\boldsymbol{X}_{(j)}^*(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = -\boldsymbol{X}_{(j)}^*(\boldsymbol{y} - \boldsymbol{X}_{(-j)}\boldsymbol{\beta}^{-j} - \boldsymbol{X}_{(j)}\boldsymbol{\beta}^j). \tag{2.6}$$

Setting this equal to zero for the coordinate-wise minimization, we get the aforementioned update (2.5) for $\boldsymbol{\beta}^j$. Alternatively, since $[\nabla L(\boldsymbol{\beta})]^j = -\boldsymbol{X}_{(j)}^*(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, the above update can intuitively be seen as a univariate descent step where the step size is the inverse Lipschitz constant of the gradient along the jth coordinate, since the (j, j) entry of the Hessian is

$$[\nabla^2 L(\boldsymbol{\beta})]_{j,j} = (\boldsymbol{X}^*\boldsymbol{X})_{j,j} = \|\boldsymbol{X}_{(j)}\|_2^2.$$

Leventhal and Lewis [15] showed that this algorithm has an expected linear convergence rate. We will detail the convergence properties of this algorithm and the others in the next section.

3. Problem Variations. We first examine the differences in behavior of the two algorithms RGS and RK in three distinct but related settings. This will highlight the opposite behaviors of these two similar algorithms.

When the system of equations (1.1) has a unique solution, we represent this by $\boldsymbol{\beta}^\star$. This happens when $m \geq n$, and the system is consistent. Assuming that $\boldsymbol{X}$ has full column rank,

$$\boldsymbol{\beta}^\star = (\boldsymbol{X}^*\boldsymbol{X})^{-1}\boldsymbol{X}^*\boldsymbol{y}, \tag{3.1}$$

and then $\boldsymbol{X}\boldsymbol{\beta}^\star = \boldsymbol{y}$.

When (1.1) does not have any consistent solution, we refer to the least-squares solution of (1.2) as $\boldsymbol{\beta_{LS}}$. This could happen in the overconstrained case, when $m > n$. Again, assuming that $\boldsymbol{X}$ has full column rank, we have

$$\boldsymbol{\beta_{LS}} = (\boldsymbol{X}^*\boldsymbol{X})^{-1}\boldsymbol{X}^*\boldsymbol{y}, \tag{3.2}$$

and we can write $\boldsymbol{r} := \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta_{LS}}$ as the residual vector.

When (1.1) has infinitely many solutions, we call the minimum Euclidean norm solution given by (1.3), $\boldsymbol{\beta_{LN}}$. This could happen in the underconstrained case, when $m < n$. Assuming that $\boldsymbol{X}$ has full row rank, we have

$$\boldsymbol{\beta_{LN}} = \boldsymbol{X}^*(\boldsymbol{XX}^*)^{-1}\boldsymbol{y}. \tag{3.3}$$

In the above notation, the LS stands for Least Squares and LN for Least Norm. We shall return to each of these three situations in that order in future sections.

One of our main contributions is to achieve a unified understanding of the behavior of RK and RGS in these different situations. The literature for RK deals mainly with the first two settings only (see [26], [18], [30]). In the third setting, one readily obtains convergence to an *arbitrary* solution (see e.g. (3) of [16]), but the convergence to the least norm solution is not often studied (likely for practical reasons). The literature for RGS typically focuses on more general setups than our specific quadratic least squares loss function $L(\beta)$ (see Nesterov [22] or Richtárik and Takáč [25]). However, for both the purposes of completeness, and for a more thorough understanding of the relationship between RK and RGS, it turns out to be crucial to analyze all three settings (for equations (1.1)-(1.3)).

1. When $\boldsymbol{\beta}^\star$ is a unique consistent solution, we present proofs of the linear convergence of both algorithms - the results are known from papers by [26] and [15] but are presented here in a novel manner so that their relationship becomes clearer and direct comparison is easily possible.
2. When $\boldsymbol{\beta_{LS}}$ is the (inconsistent) least squares solution, we show why RGS iterates converge linearly to $\boldsymbol{\beta_{LS}}$, but RK iterates do not - making RGS preferable. These facts are not hard to see, but we make it more intuitively and mathematically clear why this should be the case.
3. When $\boldsymbol{\beta_{LN}}$ is the minimum norm consistent solution, we explain why RK converges linearly to it, but RGS iterates do not (both so far seemingly undocumented observations) - making RK preferable.

Together, the above three points complete the picture (with solid accompanying intuition) of the opposing behavior of RK and RGS. Later, we will present our variant of the RGS method, the Randomized Extended Gauss-Seidel (REGS), and compare with the corresponding variant of RK (REK). This new analysis will complete the unified framework for these methods.

3.1. Overconstrained System, Consistent. Here we will assume that $m > n$, $\boldsymbol{X}$ has full column rank, and the system is consistent, so $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^\star$. First, let us write the updates used by both algorithms in a revealing fashion. If RK and RGS select row i and column j at step $t + 1$, and $\boldsymbol{e}^i$ (resp. $\boldsymbol{e}_{(j)}$) is the ith coordinate basis row (resp. column) vector, then the updates can be rewritten as:

$$(\text{RK}) \qquad \boldsymbol{\beta}_{t+1} := \boldsymbol{\beta}_t + \frac{\boldsymbol{e}^i \boldsymbol{r}_t}{\|\boldsymbol{X}^i\|_2^2}(\boldsymbol{X}^i)^* \tag{3.4}$$

$$(\text{RGS}) \qquad \boldsymbol{\beta}_{t+1} := \boldsymbol{\beta}_t + \frac{\boldsymbol{X}_{(j)}^* \boldsymbol{r}_t}{\|\boldsymbol{X}_{(j)}\|_2^2}\boldsymbol{e}_j \tag{3.5}$$

where $\boldsymbol{r}_t = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_t = \boldsymbol{X}\boldsymbol{\beta}^\star - \boldsymbol{X}\boldsymbol{\beta}_t$ is the residual vector at iteration t. Then multiplying both equations by $\boldsymbol{X}$ gives

$$\text{(RK)} \qquad \boldsymbol{X}\boldsymbol{\beta}_{t+1} := \boldsymbol{X}\boldsymbol{\beta}_t + \frac{\boldsymbol{X}^i(\boldsymbol{\beta}^\star - \boldsymbol{\beta}_t)}{\|\boldsymbol{X}^i\|_2^2}\boldsymbol{X}(\boldsymbol{X}^i)^* \qquad (3.6)$$

$$\text{(RGS)} \qquad \boldsymbol{X}\boldsymbol{\beta}_{t+1} := \boldsymbol{X}\boldsymbol{\beta}_t + \frac{\boldsymbol{X}^*_{(j)}\boldsymbol{X}(\boldsymbol{\beta}^\star - \boldsymbol{\beta}_t)}{\|\boldsymbol{X}_{(j)}\|_2^2}\boldsymbol{X}_{(j)}. \qquad (3.7)$$

We now come to an important difference, which is the key update equation for RK and RGS.

First, from the update (3.4) for RK, we have that $\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t$ is parallel to $\boldsymbol{X}^i$. Also, $\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^\star$ is orthogonal to $\boldsymbol{X}^i$ (since $\boldsymbol{X}^i(\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^\star) = y^i - y^i = 0$). Then by the Pythagorean theorem,

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^\star\|_2^2 = \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\|_2^2 - \|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|_2^2. \qquad (3.8)$$

Note that from the update (3.7), we have that $\boldsymbol{X}\boldsymbol{\beta}_{t+1} - \boldsymbol{X}\boldsymbol{\beta}_t$ is parallel to $\boldsymbol{X}_{(j)}$. Also, $\boldsymbol{X}\boldsymbol{\beta}_{t+1} - \boldsymbol{X}\boldsymbol{\beta}^\star$ is orthogonal to $\boldsymbol{X}_{(j)}$ (since $\boldsymbol{X}^*_{(j)}(\boldsymbol{X}\boldsymbol{\beta}_{t+1} - \boldsymbol{X}\boldsymbol{\beta}^\star) = \boldsymbol{X}^*_{(j)}(\boldsymbol{X}\boldsymbol{\beta}_{t+1} - \boldsymbol{y}) = 0$ by the optimality condition $\partial L/\partial\boldsymbol{\beta}^j = 0$). Then again by the Pythagorean theorem,

$$\|\boldsymbol{X}\boldsymbol{\beta}_{t+1} - \boldsymbol{X}\boldsymbol{\beta}^\star\|_2^2 = \|\boldsymbol{X}\boldsymbol{\beta}_t - \boldsymbol{X}\boldsymbol{\beta}^\star\|_2^2 - \|\boldsymbol{X}\boldsymbol{\beta}_{t+1} - \boldsymbol{X}\boldsymbol{\beta}_t\|_2^2. \qquad (3.9)$$

The rest of the proof follows by simply substituting for the last term in the above two equations, and is presented in the following table for easy comparison. Note $\boldsymbol{\Sigma} = \boldsymbol{X}^*\boldsymbol{X}$ is the full-rank covariance matrix and we first take expectations with respect to the randomness at the $(t+1)$st step, conditioning on all randomness up to the tth step. We later iterate this expectation.

Randomized Kaczmarz: $\mathbb{E}_t\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^\star\|_2^2$	Randomized Gauss-Seidel: $\mathbb{E}_t\|\boldsymbol{X}\boldsymbol{\beta}_{t+1} - \boldsymbol{X}\boldsymbol{\beta}^\star\|_2^2$
$= \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\|_2^2 - \mathbb{E}\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|_2^2$	$= \|\boldsymbol{X}\boldsymbol{\beta}_t - \boldsymbol{X}\boldsymbol{\beta}^\star\|_2^2 - \mathbb{E}\|\boldsymbol{X}\boldsymbol{\beta}_{t+1} - \boldsymbol{X}\boldsymbol{\beta}_t\|_2^2$
$= \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\|_2^2$	$= \|\boldsymbol{X}\boldsymbol{\beta}_t - \boldsymbol{X}\boldsymbol{\beta}^\star\|_2^2$
$\quad - \sum_i \frac{\|\boldsymbol{X}^i\|_2^2}{\|\boldsymbol{X}\|_F^2}\frac{(\boldsymbol{X}^i(\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star))^2}{(\|\boldsymbol{X}^i\|_2^2)^2}\|\boldsymbol{X}^i\|_2^2$	$\quad - \sum_j \frac{\|\boldsymbol{X}_{(j)}\|_2^2}{\|\boldsymbol{X}\|_F^2}\frac{(\boldsymbol{X}^*_{(j)}\boldsymbol{X}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star))^2}{(\|\boldsymbol{X}_{(j)}\|_2^2)^2}\|\boldsymbol{X}_{(j)}\|_2^2$
$= \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\|_2^2\left(1 - \frac{\|\boldsymbol{X}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star)\|_2^2}{\|\boldsymbol{X}\|_F^2\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\|_2^2}\right)$	$= \|\boldsymbol{X}\boldsymbol{\beta}_t - \boldsymbol{X}\boldsymbol{\beta}^\star\|_2^2\left(1 - \frac{\|\boldsymbol{X}^*\boldsymbol{X}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star)\|_2^2}{\|\boldsymbol{X}\|_F^2\|\boldsymbol{X}\boldsymbol{\beta}_t - \boldsymbol{X}\boldsymbol{\beta}^\star\|_2^2}\right)$
$\leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\|_2^2\left(1 - \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{Tr(\boldsymbol{\Sigma})}\right)$	$\leq \|\boldsymbol{X}\boldsymbol{\beta}_t - \boldsymbol{X}\boldsymbol{\beta}^\star\|_2^2\left(1 - \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{Tr(\boldsymbol{\Sigma})}\right)$

Here, $\lambda_{\min}(\boldsymbol{\Sigma})\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\|_2^2 \leq \|\boldsymbol{X}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star)\|_2^2$ i.e. $\lambda_{\min}(\boldsymbol{\Sigma})$ is the smallest eigenvalue of $\boldsymbol{\Sigma}$ (singular value of $\boldsymbol{X}$). It follows that

$$\text{(RK)} \qquad \mathbb{E}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\|_2^2 \leq \left(1 - \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{Tr(\boldsymbol{\Sigma})}\right)^t\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^\star\|_2^2 \qquad (3.10)$$

$$\text{(RGS)} \qquad \mathbb{E}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\|_{\boldsymbol{\Sigma}}^2 \leq \left(1 - \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{Tr(\boldsymbol{\Sigma})}\right)^t\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^\star\|_{\boldsymbol{\Sigma}}^2,$$

where $\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}^2 = \boldsymbol{w}^*\boldsymbol{\Sigma}\boldsymbol{w} = \|\boldsymbol{X}\boldsymbol{w}\|_2^2$ is the norm induced by $\boldsymbol{\Sigma}$. Since $\boldsymbol{\Sigma}$ is invertible when $m > n$ and $\boldsymbol{X}$ has full column rank, the last equation also implies linear convergence

of $\mathbb{E}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^\star\|_2^2$. The final results exist in Strohmer and Vershynin [26], Leventhal and Lewis [15] but there is utility in seeing the two proofs in a form that differs from their original presentation, side by side. In this setting, both RK and RGS are essentially equivalent (without computational considerations).

3.2. Overconstrained System, Inconsistent. Here, we will assume that $m > n$, $\boldsymbol{X}$ is full column rank, and the system is inconsistent, so $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_{LS} + \boldsymbol{r}$, where $\boldsymbol{r}$ is such that $\boldsymbol{X}^*\boldsymbol{r} = 0$. It is easy to see this condition, because as mentioned earlier,

$$\boldsymbol{\beta}_{LS} = (\boldsymbol{X}^*\boldsymbol{X})^{-1}\boldsymbol{X}^*\boldsymbol{y},$$

implying that $\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{\beta}_{LS} = \boldsymbol{X}^*\boldsymbol{y}$. Substituting $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_{LS} + \boldsymbol{r}$ gives that $\boldsymbol{X}^*\boldsymbol{r} = 0$.

In this setting, RK is known to not converge to the least squares solution, as is easily verified experimentally and geometrically. The tightest convergence upper bounds known are by [18] and [30] who show that

$$\mathbb{E}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{LS}\|_2^2 \leq \left(1 - \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{Tr(\boldsymbol{\Sigma})}\right)^t \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}_{LS}\|_2^2 + \frac{\|\boldsymbol{r}\|_2^2}{\lambda_{\min}(\boldsymbol{\Sigma})}$$

$$= \left(1 - \frac{\sigma_{\min}^2(\boldsymbol{X})}{\|\boldsymbol{X}\|_F^2}\right)^t \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}_{LS}\|_2^2 + \frac{\|\boldsymbol{r}\|_2^2}{\sigma_{\min}^2(\boldsymbol{X})},$$

where we write $\sigma_{\min}(\boldsymbol{X})$ to denote the smallest (non-zero) singular value of $\boldsymbol{X}$ and again $\|\boldsymbol{X}\|_F$ its Frobenius norm. Attempting the previous proof, (3.8) no longer holds – the Pythagorean theorem fails because $\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_{LS}$ is no longer orthogonal to $\boldsymbol{X}^i$ since $\boldsymbol{X}^i(\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_{LS}) = y^i - \boldsymbol{X}^i\boldsymbol{\beta}_{LS} \neq 0$. Intuitively, the reason RK does not converge is that every update of RK (say of row i) is a projection onto the "wrong" hyperplane that has constant y^i (where the "right" hyperplane would involve projecting onto a parallel hyperplane with constant $y^i - r^i$ where $\boldsymbol{r}$ was defined above). An alternative intuition is that all RK updates are in the span of the rows, but $\boldsymbol{\beta}_{LS}$ is not in the row span. These intuitive explanations are easily confirmed by experiments seen in [30, pp. 787–788],[18, pp. 402]. Zouzias and Freris [30] alleviate this issue with the REK algorithm, whose convergence obeys

$$\mathbb{E}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{LS}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(\boldsymbol{X})}{\|\boldsymbol{X}\|_F^2}\right)^{\lfloor t/2 \rfloor} \left(1 + 2\frac{\sigma_{\min}^2(\boldsymbol{X})}{\sigma_{\max}^2(\boldsymbol{X})}\|\boldsymbol{\beta}_{LS}\|_2^2\right). \qquad (3.11)$$

However, the fate of RK doesn't hold for RGS. Almost magically, in the previous proof, the Pythagorean theorem still holds in equation (3.9) because

$$\boldsymbol{X}_{(j)}^*(\boldsymbol{X}\boldsymbol{\beta}_{t+1} - \boldsymbol{X}\boldsymbol{\beta}_{LS}) = \boldsymbol{X}_{(j)}^*(\boldsymbol{X}\boldsymbol{\beta}_{t+1} - \boldsymbol{y}) + \boldsymbol{X}_{(j)}^*(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{LS}) = 0. \qquad (3.12)$$

The first term is 0 by the optimality condition for $\boldsymbol{\beta}_{t+1}$ i.e. $\boldsymbol{X}_{(j)}^*(\boldsymbol{X}\boldsymbol{\beta}_{t+1} - \boldsymbol{y}) = \partial L/\partial\boldsymbol{\beta}^j = 0$. The second term is zero by the global optimality of $\boldsymbol{\beta}_{LS}$, i.e. $\boldsymbol{X}^*(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{LS}) = \nabla L = 0$. Also, $\boldsymbol{\Sigma}$ is full rank as before. Indeed, RGS works in the space of fitted values $\boldsymbol{X}\boldsymbol{\beta}$ and not the iterates $\boldsymbol{\beta}$.

In summary, RK does not converge to the LS solution, but RGS does at the same linear rate. This is what motivated the development of Randomized Extended Kaczmarz (REK) by Zouzias and Freris [30] which, as discussed earlier, is a modification of RK designed to converge to $\boldsymbol{\beta}_{LS}$ by randomly projecting out r. An independent paper by Dumitrescu [5] argues however that in this setting RGS is preferable to REK in terms of computational convergence.

3.3. Underconstrained System, Infinite Solutions. Here, we will assume that $m < n$, $\boldsymbol{X}$ is full row rank and the system is consistent with infinitely many solutions. As mentioned earlier, it is easy to show that

$$\boldsymbol{\beta_{LN}} = \boldsymbol{X}^*(\boldsymbol{X}\boldsymbol{X}^*)^{-1}\boldsymbol{y}$$

(which clearly satisfies $\boldsymbol{X}\boldsymbol{\beta_{LN}} = \boldsymbol{y}$). Every other consistent solution can be expressed as

$$\boldsymbol{\beta} = \boldsymbol{\beta_{LN}} + \boldsymbol{z} \quad \text{where} \quad \boldsymbol{X}\boldsymbol{z} = 0.$$

Clearly any such $\boldsymbol{\beta}$ would also satisfy $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta_{LN}} = y$. Since $\boldsymbol{X}\boldsymbol{z} = 0$, $\boldsymbol{z} \perp \boldsymbol{\beta_{LN}}$ implying $\|\boldsymbol{\beta}\|^2 = \|\boldsymbol{\beta_{LN}}\|^2 + \|\boldsymbol{z}\|^2$, showing that $\boldsymbol{\beta_{LN}}$ is indeed the minimum norm solution as claimed.

In this case, RK has good behavior, and starting from $\boldsymbol{\beta}_0 = 0$, it does converge linearly to $\boldsymbol{\beta_{LN}}$. Intuitively, $\boldsymbol{\beta_{LN}} = \boldsymbol{X}^*\boldsymbol{\alpha}$ (for $\boldsymbol{\alpha} = (\boldsymbol{X}\boldsymbol{X}^*)^{-1}\boldsymbol{y}$) and hence is in the row span of $\boldsymbol{X}$. Starting from $\boldsymbol{\beta}_0 = 0$, RK only adds multiples of rows to its iterates, and hence will never have any component orthogonal to the row span of $\boldsymbol{X}$. There is exactly one solution with no component orthogonal to the row span of $\boldsymbol{X}$, and that is $\boldsymbol{\beta_{LN}}$, and hence RK converges linearly to the required point, where the rate can be bounded in exactly the same way as (3.10). It is important not to start from an arbitrary $\boldsymbol{\beta}_0$ since the RK updates can never eliminate any component of $\boldsymbol{\beta}_0$ that is perpendicular to the row span of $\boldsymbol{X}$. Of course, the same properties are shared by REK for this case as well. It is noted in Zouzias and Freris [30, Sec. 2.1] that the REK converges at the same rate for underdetermined systems as it does overdetermined systems.

Mathematically, the previous earlier proof works because the Pythagorean theorem holds since it is a consistent system. Now, $\boldsymbol{\Sigma}$ is not full rank but note that since both $\boldsymbol{\beta_{LN}}$ and $\boldsymbol{\beta}_t$ are in the row span, $\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}}$ has no component orthogonal to $\boldsymbol{X}$ (unless it equals zero in which case the algorithm has already converged). Hence $\lambda_{\min}(\boldsymbol{\Sigma})\|\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}}\|^2 \leq \|\boldsymbol{X}(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|^2$ holds, where $\lambda_{\min}(\boldsymbol{\Sigma})$ is now understood to be the smallest positive eigenvalue of $\boldsymbol{\Sigma}$. To summarize, the exact same bound (3.10) still holds in this case, with the appropriate understanding of $\lambda_{\min}(\boldsymbol{\Sigma})$ and under the assumption that the initialized $\boldsymbol{\beta}_0$ is in the row span of $\boldsymbol{X}$.

RGS unfortunately suffers the opposite fate. The iterates do not converge to $\boldsymbol{\beta_{LN}}$, even though $\boldsymbol{X}\boldsymbol{\beta}_t$ does converge to $\boldsymbol{X}\boldsymbol{\beta_{LN}}$. Mathematically, the convergence proof still carries forward as before, but in the last step when $\boldsymbol{X}^*\boldsymbol{X}$ cannot be inverted because it is not full rank. Hence we get convergence of the residual to zero, without getting convergence of the iterates to the least norm solution. Intuitively, the iterates of RGS add components to the estimates that are orthogonal to the row span of $\boldsymbol{X}$. These components are never eliminated because in minimizing the residual, they are ignored. Therefore, RGS is able to minimize the residual without finding the least norm solution.

Unfortunately, when each update is cheaper for RK than RGS (due to matrix size), RGS is preferred for reasons of convergence and when it is cheaper for RGS than RK, RK is preferred.

4. REGS. We next introduce an extension of RGS, analogous to the extension REK of RK. The purpose of extending RK was to allow for convergence to the least squares solution. Now, the purpose of extending RGS is to allow for convergence to the least norm solution. We view this method as a completion to the unified analysis of these approaches, and it may also possess advantages in its own right.

4.1. The algorithm. Consider the linear system (1.1) with $m < n$. Let $\boldsymbol{\beta_{LN}}$ denote the least norm solution of the underdetermined system as described in (1.3). The REGS algorithm is described by the following pseudo-code. Analogous to the role $\boldsymbol{z}$ plays in REK, $\boldsymbol{z}$ iteratively approximates the component in $\boldsymbol{\beta}$ orthogonal to the row-span of $\boldsymbol{X}$. By iteratively removing this component, we converge to the least norm solution. Note that outputting $\boldsymbol{\beta}_t$ instead of $\boldsymbol{\beta}_t^{LN} = \boldsymbol{\beta}_t - \boldsymbol{z}_t$ in Algorithm 1 recovers the RGS algorithm. This may be preferable in the overdetermined setting.

Algorithm 1 Randomized Extended Gauss-Seidel (REGS)

1: **procedure** $(\boldsymbol{X}, \boldsymbol{y}, \text{maxIter})$ ▷ $m \times n$ matrix $\boldsymbol{X}$, $\boldsymbol{y} \in \mathbb{C}^m$, maximum iterations T
2:    Initialize $\boldsymbol{\beta}_0 = \boldsymbol{0}$, $\boldsymbol{z}_0 = \boldsymbol{0}$
3:    **for** $t = 1, 2, \ldots, T$ **do**
4:       Choose column $\boldsymbol{X}_{(j)}$ with probability $\frac{\|\boldsymbol{X}_{(j)}\|_2^2}{\|\boldsymbol{X}\|_F^2}$
5:       Choose row $\boldsymbol{X}^i$ with probability $\frac{\|\boldsymbol{X}^i\|_2^2}{\|\boldsymbol{X}\|_F^2}$
6:       Set $\boldsymbol{\gamma}_t = \frac{\boldsymbol{X}_{(j)}^* (\boldsymbol{X}\boldsymbol{\beta}_{t-1} - \boldsymbol{y})}{\|\boldsymbol{X}_{(j)}\|_2^2} \boldsymbol{e}_{(j)}$
7:       Set $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\gamma}_t$
8:       Set $\boldsymbol{P}_i = \boldsymbol{Id}_n - \frac{(\boldsymbol{X}^i)^* \boldsymbol{X}^i}{\|\boldsymbol{X}^i\|_2^2}$ ▷ $\boldsymbol{Id}_n$ denotes the $n \times n$ identity matrix
9:       Update $\boldsymbol{z}_t = \boldsymbol{P}_i(\boldsymbol{z}_{t-1} + \boldsymbol{\gamma}_t)$
10:      Update $\boldsymbol{\beta}_t^{LN} = \boldsymbol{\beta}_t - \boldsymbol{z}_t$
11:   **end for**
12:   Output $\boldsymbol{\beta}_t^{LN}$
13: **end procedure**

4.2. Main result. Our main result for the REGS method shows linear convergence to the least norm solution.

THEOREM 4.1. *The REGS algorithm outputs an estimate $\boldsymbol{\beta}_T^{LN}$ such that*

$$\mathbb{E}\|\boldsymbol{\beta}_T^{LN} - \boldsymbol{\beta_{LN}}\|_2^2 \leq \alpha^T \|\boldsymbol{\beta}^{LN}\|_2^2 + 2\alpha^{\lfloor T/2 \rfloor} \frac{B}{1-\alpha} \tag{4.1}$$

where $B = \frac{\|\boldsymbol{X}\boldsymbol{\beta_{LN}}\|_2^2}{\|\boldsymbol{X}\|_F^2}$ and $\alpha = \left(1 - \frac{\sigma_{min}^2(\boldsymbol{X})}{\|\boldsymbol{X}\|_F^2}\right)$.

Proof: We devote the remainder of this section to the proof of Theorem 4.1.

Let $\mathbb{E}_{t-1}$ denote the expected value conditional on the first $t-1$ iterations, and instate the notation of the theorem. That is, $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot|i_1, j_1, i_2, j_2, \ldots i_{t-1}, j_{t-1}]$ where i_{t^*} is the t^{*th} row chosen and j_{t^*} is the t^{*th} column chosen. We denote conditional expectation with respect to the choice of column as $\mathbb{E}_{t-1}^j[\cdot] = \mathbb{E}[\cdot \mid i_1, j_1, \ldots i_{t-1}, j_{t-1}, i_t]$. Similarly, we denote conditional expectation with respect to the choice of row as $\mathbb{E}_{t-1}^i[\cdot] = \mathbb{E}[\cdot \mid i_1, j_1, \ldots i_{t-1}, j_{t-1}, j_t]$. Then note by the law of total expectation we have that $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}_{t-1}^i[\mathbb{E}_{t-1}^j[\cdot]]$. We will use the following elementary facts and lemmas.

FACT 1. *([30, Fact 3]) For any $\boldsymbol{P}_i$ as in the algorithm, $\mathbb{E}\|\boldsymbol{P}_i \boldsymbol{w}\|_2^2 \leq \alpha \|\boldsymbol{w}\|_2^2$ for any $\boldsymbol{w}$.*

LEMMA 4.2. *([15, Thm. 3.6]) We have that*

$$\mathbb{E}_{t-1}\|\boldsymbol{X}\boldsymbol{\beta}_t - \boldsymbol{X}\boldsymbol{\beta_{LN}}\|_2^2 \leq \alpha \|\boldsymbol{X}\boldsymbol{\beta}_{t-1} - \boldsymbol{X}\boldsymbol{\beta_{LN}}\|_2^2$$

9

and that

$$\mathbb{E}\|\boldsymbol{X}\boldsymbol{\beta}_t - \boldsymbol{X}\boldsymbol{\beta_{LN}}\|_2^2 \leq \alpha^t \|\boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\boldsymbol{\beta_{LN}}\|_2^2.$$

Now we first consider $\|\boldsymbol{\beta}_t^{LN} - \boldsymbol{\beta_{LN}}\|_2^2$:

$$
\begin{aligned}
\|\boldsymbol{\beta}_t^{LN} - \boldsymbol{\beta_{LN}}\|_2^2 &= \|\boldsymbol{\beta}_t - \boldsymbol{z}_t - \boldsymbol{\beta_{LN}}\|_2^2 \\
&= \|\boldsymbol{\beta}_t - \boldsymbol{P}_i(\boldsymbol{z}_{t-1} + \boldsymbol{\gamma}_t) - \boldsymbol{P}_i\boldsymbol{\beta_{LN}} - (\boldsymbol{Id}_n - \boldsymbol{P}_i)\boldsymbol{\beta_{LN}}\|_2^2 \\
&= \|\boldsymbol{\beta}_t - \boldsymbol{P}_i(\boldsymbol{z}_{t-1} + \boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}) - \boldsymbol{P}_i\boldsymbol{\beta_{LN}} - (\boldsymbol{Id}_n - \boldsymbol{P}_i)\boldsymbol{\beta_{LN}}\|_2^2 \\
&= \|(\boldsymbol{Id}_n - \boldsymbol{P}_i)\boldsymbol{\beta}_t + \boldsymbol{P}_i(\boldsymbol{\beta}_{t-1} - \boldsymbol{z}_{t-1}) - \boldsymbol{P}_i\boldsymbol{\beta_{LN}} - (\boldsymbol{Id}_n - \boldsymbol{P}_i)\boldsymbol{\beta_{LN}}\|_2^2 \\
&= \|(\boldsymbol{Id}_n - \boldsymbol{P}_i)\boldsymbol{\beta}_t + \boldsymbol{P}_i\boldsymbol{\beta}_{t-1}^{LN} - \boldsymbol{P}_i\boldsymbol{\beta_{LN}} - (\boldsymbol{Id}_n - \boldsymbol{P}_i)\boldsymbol{\beta_{LN}}\|_2^2 \\
&= \|\boldsymbol{P}_i(\boldsymbol{\beta}_{t-1}^{LN} - \boldsymbol{\beta_{LN}}) + (\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2 \\
&= \|\boldsymbol{P}_i(\boldsymbol{\beta}_{t-1}^{LN} - \boldsymbol{\beta_{LN}})\|_2^2 + \|(\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2. \qquad (4.2)
\end{aligned}
$$

So far, we have only used substitution of variables as defined for the algorithm and that $\boldsymbol{\beta_{LN}} = \boldsymbol{P}_i\boldsymbol{\beta_{LN}} + (\boldsymbol{Id}_n - \boldsymbol{P}_i)\boldsymbol{\beta_{LN}}$ is an orthogonal decomposition. We first focus on the expected value of the second term.

LEMMA 4.3. *We also have that*

$$\mathbb{E}_{t-1}\|(\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2 \leq \frac{\alpha\|\boldsymbol{X}(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta_{LN}})\|_2^2}{\|\boldsymbol{X}\|_F^2}.$$

Proof:

$$
\begin{aligned}
\mathbb{E}_{t-1}&\|(\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2 \\
&= \mathbb{E}_{t-1}[(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})^*(\boldsymbol{Id}_n - \boldsymbol{P}_i)^*(\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})] \\
&= \mathbb{E}_{t-1}[(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})^*(\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})] \\
&= \mathbb{E}_{t-1}\left[(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})^*\left(\frac{(\boldsymbol{X}^i)^*\boldsymbol{X}^i}{\|\boldsymbol{X}^i\|_2^2}\right)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\right] \\
&= \mathbb{E}_{t-1}\left[\frac{\|\boldsymbol{X}^i(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2}{\|\boldsymbol{X}^i\|_2^2}\right] \\
&= \mathbb{E}_{t-1}^j\left[\mathbb{E}_{t-1}^i\frac{\|\boldsymbol{X}^i(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2}{\|\boldsymbol{X}^i\|_2^2}\right] \\
&= \mathbb{E}_{t-1}^j\left[\sum_{i=1}^m \frac{\|\boldsymbol{X}^i(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2}{\|\boldsymbol{X}^i\|_2^2} \cdot \frac{\|\boldsymbol{X}^i\|_2^2}{\|\boldsymbol{X}\|_F^2}\right] \\
&= \mathbb{E}_{t-1}^j\left[\frac{\|\boldsymbol{X}(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2}{\|\boldsymbol{X}\|_F^2}\right] \\
&\leq \frac{\alpha\|\boldsymbol{X}(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta_{LN}})\|_2^2}{\|\boldsymbol{X}\|_F^2}.
\end{aligned}
$$

The first line follows by expanding the norm, the second line since $(\boldsymbol{Id}_n - \boldsymbol{P}_i)$ is a projection matrix, the third line from the definition of $\boldsymbol{P}_i$, the fourth line is

computation, the fifth line follows from the law of total expectation, the next two lines are computation, and finally the last line follows by Lemma 4.2. Notice that in the seventh line, $\mathbb{E}_{t-1}^{j} = \mathbb{E}_{t-1}$ because the random variable $\boldsymbol{\beta}_t$ only depends on the choice of columns. □

We want to control the term $r_t = \mathbb{E}\|(\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2$ by bounding it by some α and B such that $r_t \leq \alpha^t B$. We calculate this here:

$$
\mathbb{E}\|(\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2 = \mathbb{E}[\mathbb{E}_{t-1}\|(\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2]
$$
$$
\leq \frac{\alpha \mathbb{E}\|\boldsymbol{X}(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta_{LN}})\|_2^2}{\|\boldsymbol{X}\|_F^2}
$$
$$
\leq \alpha^t \frac{\|\boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\boldsymbol{\beta_{LN}}\|_2^2}{\|\boldsymbol{X}\|_F^2}.
$$

The first line follows by definition, the second is by Lemma 4.3, and the third by Lemma 4.2.

Finally, we take the expected value of $\|\boldsymbol{\beta}_t^{LN} - \boldsymbol{\beta_{LN}}\|_2^2$. From equation (4.2) and using Fact 1 we obtain:

$$
\mathbb{E}\|\boldsymbol{\beta}_t^{LN} - \boldsymbol{\beta_{LN}}\|_2^2 = \mathbb{E}\|\boldsymbol{P}_i(\boldsymbol{\beta}_{t-1}^{LN} - \boldsymbol{\beta_{LN}})\|_2^2 + \mathbb{E}\|(\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2
$$
$$
\leq \alpha \mathbb{E}\|(\boldsymbol{\beta}_{t-1}^{LN} - \boldsymbol{\beta_{LN}})\|_2^2 + \mathbb{E}\|(\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2.
$$

We complete the proof using the following lemma from [30]:

LEMMA 4.4. *([30, Thm. 8]) Suppose that for some $\alpha, \bar{\alpha} < 1$, the following bounds hold for all $t^* \geq 0$:*

$$
\mathbb{E}\|\boldsymbol{\beta}_{t^*}^{LN} - \boldsymbol{\beta_{LN}}\|_2^2 \leq \alpha \mathbb{E}\|\boldsymbol{\beta}_{t^*-1}^{LN} - \boldsymbol{\beta_{LN}}\|_2^2 + r_{t^*} \text{ and } r_{t^*} \leq \bar{\alpha}^{t^*} B.
$$

Then for any $T > 0$,

$$
\mathbb{E}\|\boldsymbol{\beta}_T^{LN} - \boldsymbol{\beta_{LN}}\|_2^2 \leq \alpha^T \|\boldsymbol{\beta}_0^{LN} - \boldsymbol{\beta_{LN}}\|_2^2 + (\alpha^{\lfloor T/2 \rfloor} + \bar{\alpha}^{\lfloor T/2 \rfloor}) \frac{B}{1-\alpha}.
$$

Letting $\alpha = \bar{\alpha} = \alpha$, $r_t^* = \mathbb{E}\|(\boldsymbol{Id}_n - \boldsymbol{P}_i)(\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}})\|_2^2$, $B = \frac{\|\boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\boldsymbol{\beta_{LN}}\|_2^2}{\|\boldsymbol{X}\|_F^2}$, and noting that $\boldsymbol{\beta}_0^{LN} = \boldsymbol{\beta}_0 = 0$, we complete the proof of Theorem 4.1.
□

REMARK 1. *Here we note that the same proof works for overdetermined systems. In particular, this works because Lemma 4.2 holds for $\boldsymbol{\beta_{LS}}$ and $\boldsymbol{\beta}^*$ also (see Thm. 3.6 in [15]). Also, Lemma 4.3 follows for both overdetermined consistent systems (see table in Section 3.1) as well as overdetermined inconsistent systems (from (3.12) and subsequent arguments).*

4.3. Comparison. Theorem 4.1 shows that, like the RK and REK methods, REGS converges linearly to the least-norm solution in the underdetermined case. We believe it serves to complement existing analysis and completes the theory of these iterative methods in all three cases of interest. For that reason, we compare the three approaches for the underdetermined setting here. For ease of comparison, set α as in Theorem 4.1, and write $\kappa = \sigma_{\max}(\boldsymbol{X})/\sigma_{\min}(\boldsymbol{X})$ for the condition number of $\boldsymbol{X}$. From the convergence rate bounds for RK [26] and REK [30] given in Section 3, and after applying elementary bounds to (4.1) of Theorem 4.1, we have:

$$\text{(RK)} \qquad \mathbb{E}\|\boldsymbol{\beta}_t - \boldsymbol{\beta_{LN}}\|_2^2 \quad \leq \quad \alpha^t \|\boldsymbol{\beta_{LN}}\|_2^2 \qquad\qquad (4.3)$$

$$\text{(REK)} \qquad \mathbb{E}\|\boldsymbol{\beta}_{2t} - \boldsymbol{\beta_{LN}}\|_2^2 \quad \leq \quad \alpha^t (1 + 2\kappa^2) \|\boldsymbol{\beta_{LN}}\|_2^2 \qquad (4.4)$$

$$\text{(REGS)} \qquad \mathbb{E}\|\boldsymbol{\beta}_{2t} - \boldsymbol{\beta_{LN}}\|_2^2 \quad \leq \quad \alpha^t (1 + 2\kappa^2) \|\boldsymbol{\beta_{LN}}\|_2^2. \qquad (4.5)$$

We find similar results in the overdetermined, inconsistent setting. Using the convergence rate bounds for RGS [15], REK [30], and REGS (Theorem 4.1), also given in section 3, we have:

$$\text{(RGS)} \qquad \mathbb{E}\|\boldsymbol{\beta}_t - \boldsymbol{\beta_{LS}}\|_2^2 \quad \leq \quad \alpha^t \|\boldsymbol{\beta_{LS}}\|_2^2 \qquad\qquad (4.6)$$

$$\text{(REK)} \qquad \mathbb{E}\|\boldsymbol{\beta}_{2t} - \boldsymbol{\beta_{LS}}\|_2^2 \quad \leq \quad \alpha^t (1 + 2\kappa^2) \|\boldsymbol{\beta_{LS}}\|_2^2 \qquad (4.7)$$

$$\text{(REGS)} \qquad \mathbb{E}\|\boldsymbol{\beta}_{2t} - \boldsymbol{\beta_{LS}}\|_2^2 \quad \leq \quad \alpha^t (1 + 2\kappa^2) \|\boldsymbol{\beta_{LS}}\|_2^2. \qquad (4.8)$$

Thus, up to constant terms (which are likely artifacts of the proofs), the bounds provide the same convergence rate α, which is not surprising in light of the connections between the methods. In the next section, we compare these approaches experimentally.

5. Empirical Results. In this section we present our experimental results. The code used to run these experiments can be found at [4]. For each experiment, we initialize a matrix $\boldsymbol{X}$ and vector $\boldsymbol{\beta}$ with independent standard normal entries and run 50 trials. The right hand side $\boldsymbol{y}$ is taken to be $\boldsymbol{X}\boldsymbol{\beta}$. At each iteration t, we keep track of the ℓ_2-error $\|\boldsymbol{\beta}_t^{LN} - \boldsymbol{\beta_{LN}}\|_2^2$ and fix the stopping criterion to be $\|\boldsymbol{\beta}_t^{LN} - \boldsymbol{\beta_{LN}}\|_2^2 < 10^{-6}$ (of course in practice one chooses a more practical criterion). In each plot, the solid blue line represents the median ℓ_2-error at iteration t, the light blue shaded region captures the range of error across trials, and the red line represents the theoretical upper bound at each iteration. In Figure 5.1, we show the convergence of $\boldsymbol{\beta}_t^{LN}$ for varying sized underdetermined linear systems. In Figure 5.2, we show the convergence of a matrix $\boldsymbol{X}$ of size 700x1000 and its theoretical upper bound. As it turns out, the REGS algorithm often converges much faster than the theoretical worst-case bound.

We also tested REGS on tomography problems using the Regularization toolbox by Hansen [11] (http://www.imm.dtu.dk/~pcha/Regutools/). For the 2D tomography problem $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$ with $\boldsymbol{X}$ an $m \times n$ matrix where $n = dN^2$ and $m = N^2$, we use $N = 20$ and $d = 3$ for our experiments. Here, $\boldsymbol{X}$ consists of samples of absorption along a random line on an $N \times N$ grid and d is the oversampling factor. The results from this experiment are shown in Figure 5.2.

We also compare the performance of all four algorithms (RK, REK, RGS, REGS) under the different settings discussed in this paper. Each line in each plot represents the median ℓ_2-error at that iteration or CPU time over 50 trials using a stopping criterion of 10^{-6}. For the underdetermined case, $\boldsymbol{X}$ is a 50×500 Gaussian matrix and a 500×50 Gaussian matrix for the overdetermined cases. In the overdetermined, inconsistent case, we set $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{r}$ where $\mathbf{r} \in \text{null}(\mathbf{X}^*)$ (computed in Matlab using the null() function). Figure 5.3, Figure 5.4, and Figure 5.5 show the empirical results for the underdetermined, overdetermined inconsistent, and overdetermined consistent cases respectively. Note we only plot the methods which actually converge to the desired solution in each case. Looking at iterations to convergence, it seems that RK and RGS converge faster than their extended counterparts while REGS and REK converge to the desired solution at about the same rate.
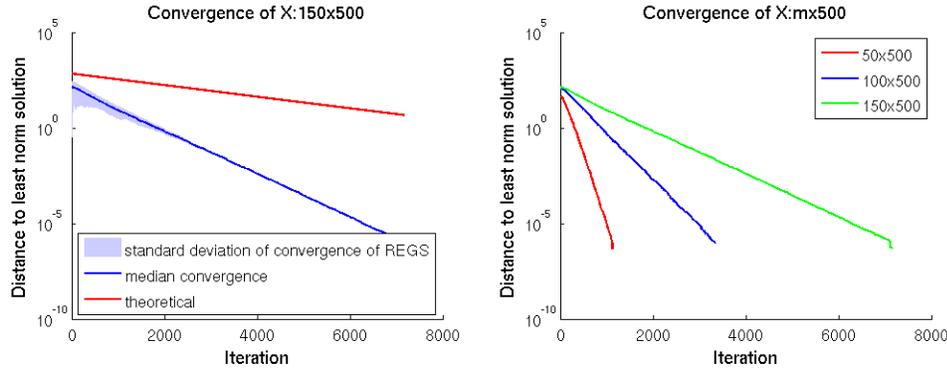
FIGURE 5.1. *Left: ℓ_2-error (log-scale) of REGS on a 150×500 matrix and its the theoretical bound. Right: Comparison of ℓ_2-error (log-scale) of REGS for $m \times 500$ sized matrices with $m = 50, 100, 150$.*
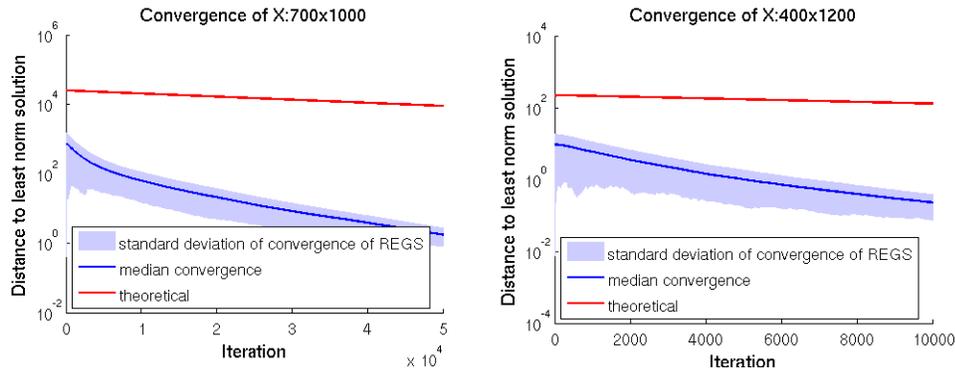

FIGURE 5.2. *Left: ℓ_2-error (log-scale) of REGS on a 700×1000 matrix and its the theoretical bound. Right: ℓ_2-error (log-scale) of REGS on the tomography problem with a 400×1200 matrix.*

6. Conclusion. The Kaczmarz and Gauss-Seidel methods operate in two different spaces (i.e. row versus column space), but share many parallels. In this paper we drew connections between these two methods, highlighting the similarities and differences in convergence analysis. The approaches possess conflicting convergence properties; RK converges to the desired solution in the underdetermined case but not the inconsistent overdetermined setting, while RGS does the exact opposite. The extended method REK in the Kaczmarz framework fixes this issue, converging to the solution in both scenarios. Here, we present the REGS method, a natural extension of RGS, which completes the overall picture. We hope that our unified analysis of all four methods will assist researchers working with these approaches.
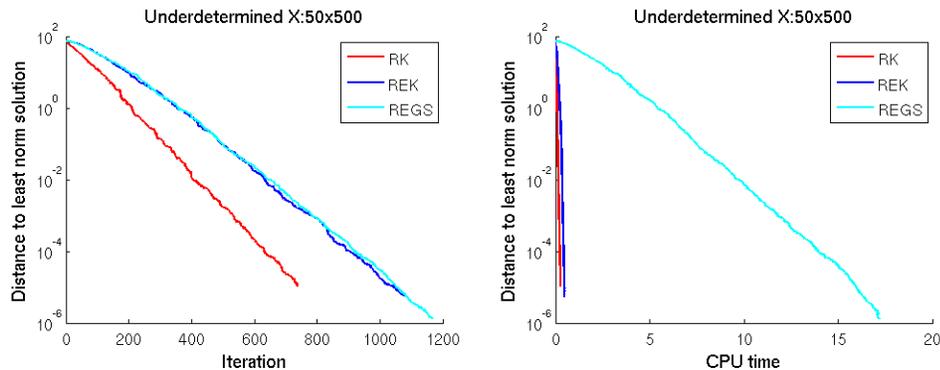
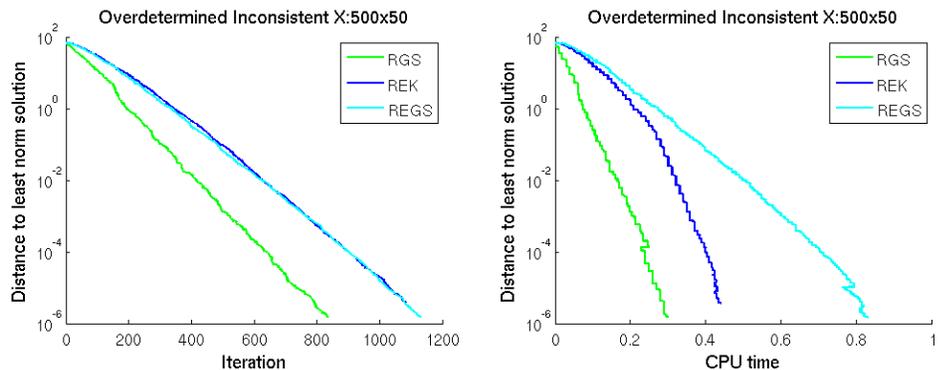FIGURE 5.3. *Comparison of median ℓ_2-error (log-scale) of RK, REK, and REGS for an underdetermined system.*

FIGURE 5.4. *Comparison of median ℓ_2-error (log-scale) of RGS, REK, and REGS for an overdetermined, inconsistent system.*

References.

[1] Byrne, C. L. [2008], *Applied iterative methods*, A K Peters Ltd., Wellesley, MA.

[2] Censor, Y., Eggermont, P. P. B. and Gordon, D. [1983], 'Strong underrelaxation in Kaczmarz's method for inconsistent systems', *Numerische Mathematik*
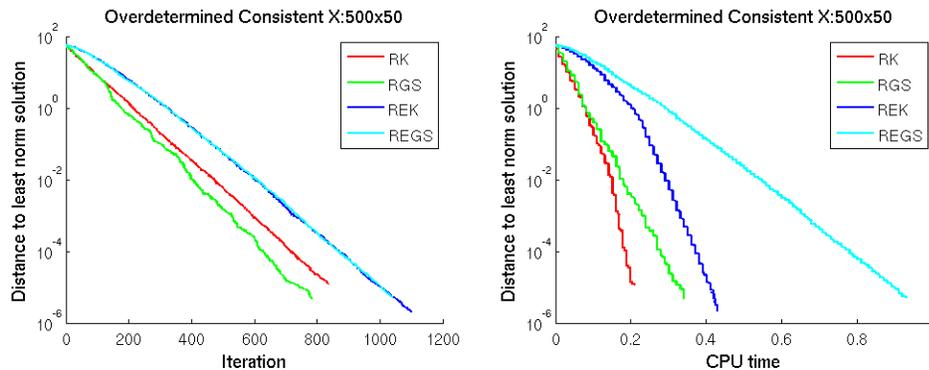
FIGURE 5.5. *Comparison of median ℓ_2-error (log-scale) of RK, RGS, REK, and REGS for an overdetermined, consistent system*

41(1), 83–92.

[3] Chen, X. and Powell, A. [2012], 'Almost sure convergence of the Kaczmarz algorithm with random measurements', *J. Fourier Anal. Appl.* pp. 1–20. 10.1007/s00041-012-9237-2.
URL: *http://dx.doi.org/10.1007/s00041-012-9237-2*

[4] *Code link* [n.d.]. http://www.cmc.edu/pages/faculty/DNeedell/regs.zip.

[5] Dumitrescu, B. [2014], 'On the relation between the randomized extended kaczmarz algorithm and coordinate descent', *BIT Numerical Mathematics* pp. 1–11.

[6] Eldar, Y. C. and Needell, D. [2011], 'Acceleration of randomized Kaczmarz method via the Johnson-Lindenstrauss lemma', *Numer. Algorithms* **58**(2), 163–177.
URL: *http://dx.doi.org/10.1007/s11075-011-9451-z*

[7] Elfving, T. [1980], 'Block-iterative methods for consistent and inconsistent linear equations', *Numer. Math.* **35**(1), 1–12.
URL: *http://dx.doi.org/10.1007/BF01396365*

[8] Gordon, R., Bender, R. and Herman, G. T. [1970], 'Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography', *J. Theoret. Biol.* **29**, 471–481.

[9] Hamaker, C. and Solmon, D. C. [1978], 'The angles between the null spaces of X-rays', *J. Math. Anal. Appl.* **62**(1), 1–23.

[10] Hanke, M. and Niethammer, W. [1990], 'On the acceleration of Kaczmarz's method for inconsistent linear systems', *Linear Algebra and its Applications* **130**, 83–98.

[11] Hansen, P. C. [1994], 'Regularization tools: A matlab package for analysis and solution of discrete ill-posed problems', *Numerical algorithms* **6**(1), 1–35.

[12] Herman, G. and Meyer, L. [1993], 'Algebraic reconstruction techniques can be made computationally efficient', *IEEE Trans. Medical Imaging* **12**(3), 600–609.

[13] Herman, G. T. [2009], *Fundamentals of computerized tomography: image reconstruction from projections*, Springer.

[14] Kaczmarz, S. [1937], 'Angenäherte auflösung von systemen linearer gleichungen',

Bull. Int. Acad. Polon. Sci. Lett. Ser. A pp. 335–357.

[15] Leventhal, D. and Lewis, A. S. [2010], 'Randomized methods for linear constraints: convergence rates and conditioning', *Math. Oper. Res.* **35**(3), 641–654.
   URL: *http://dx.doi.org/10.1287/moor.1100.0456*

[16] Liu, J., Wright, S. J. and Sridhar, S. [2014], 'An asynchronous parallel randomized kaczmarz algorithm', *arXiv preprint arXiv:1401.4780* .

[17] Natterer, F. [2001], *The mathematics of computerized tomography*, Vol. 32 of *Classics in Applied Mathematics*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Reprint of the 1986 original.
   URL: *http://dx.doi.org/10.1137/1.9780898719284*

[18] Needell, D. [2010], 'Randomized Kaczmarz solver for noisy linear systems', *BIT* **50**(2), 395–403.
   URL: *http://dx.doi.org/10.1007/s10543-010-0265-5*

[19] Needell, D., Sbrero, N. and Ward, R. [2014], 'Stochastic gradient descent and the randomized kaczmarz algorithm', *Math. Program. Series A* . to appear.

[20] Needell, D. and Tropp, J. A. [2013], 'Paved with good intentions: Analysis of a randomized block kaczmarz method', *Linear Algebra and its Applications* .

[21] Needell, D. and Ward, R. [2013], 'Two-subspace projection method for coherent overdetermined linear systems', *Journal of Fourier Analysis and Applications* **19**(2), 256–269.

[22] Nesterov, Y. [2012], 'Efficiency of coordinate descent methods on huge-scale optimization problems', *SIAM J. Optimiz.* **22**(2), 341–362.

[23] Popa, C. [1998], 'Extensions of block-projections methods with relaxation parameters to inconsistent and rank-deficient least-squares problems', *BIT* **38**(1), 151–176.
   URL: *http://dx.doi.org/10.1007/BF02510922*

[24] Popa, C., Preclik, T., Köstler, H. and Rüde, U. [2012], 'On Kaczmarz's projection iteration as a direct solver for linear least squares problems', *Linear Algebra and Its Applications* **436**(2), 389–404.

[25] Richtárik, P. and Takáč, M. [2012], 'Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function', *Math. Program.* pp. 1–38.

[26] Strohmer, T. and Vershynin, R. [2009], 'A randomized Kaczmarz algorithm with exponential convergence', *J. Fourier Anal. Appl.* **15**(2), 262–278.
   URL: *http://dx.doi.org/10.1007/s00041-008-9030-4*

[27] Tanabe, K. [1971], 'Projection method for solving a singular system of linear equations and its applications', *Numerische Mathematik* **17**(3), 203–214.

[28] Whitney, T. M. and Meany, R. K. [1967], 'Two algorithms related to the method of steepest descent', *SIAM Journal on Numerical Analysis* **4**(1), 109–118.

[29] Xu, J. and Zikatanov, L. [2002], 'The method of alternating projections and the method of subspace corrections in Hilbert space', *J. Amer. Math. Soc.* **15**(3), 573–597.
   URL: *http://dx.doi.org/10.1090/S0894-0347-02-00398-3*

[30] Zouzias, A. and Freris, N. M. [2013], 'Randomized extended kaczmarz for solving least squares', *SIAM Journal on Matrix Analysis and Applications* **34**(2), 773–793.