New bounds for the block Gaussian sketch and project method

Elizaveta Rebrova[*,1] and Deanna Needell[*,2]

[*]Department of Mathematics, University of California, Los Angeles, Los Angeles, California
[1]rebrova@math.ucla.edu
[2]deanna@math.ucla.edu

***Abstract*— We analyze a block Gaussian version of a well-known Kaczmarz iterative solver. The Kaczmarz method is one of the most popular methods for solving large-scale over-determined linear systems due to its speed and simplicity. The block Gaussian version enjoys the regularization properties of Gaussian sketching, combined with the accelerated convergence of the block versions of the method. We prove that the method converges to the solution exponentially fast in expectation, and benefits from the non-trivial size of the block, regardless of the structure of the original system. We also provide numerical experiments supporting our theoretical analysis of the method.**

I. Introduction

The Kaczmarz method [5] is an iterative method for solving large-scale over-determined linear systems. Being simple, efficient and well-adapted to the large amounts of data (due to its iterative nature), the Kaczmarz method is widely used in a variety of applications, from image reconstruction to signal processing [13], [7], [3]. Given a consistent system of linear equations of the form

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

the original Kaczmarz method starts with some initial guess x_0, and then iteratively projects the previous approximation x_k onto the solution spaces of the next equation in the system. Namely, if $a_1, \dots, a_m \in \mathbb{R}^n$ are the rows of A, then the k-th step of the algorithm is

$$x_k = x_{k-1} + \frac{b_i - \langle a_i, x_{k-1}\rangle}{\|a_i\|^2} a_i,$$

where $b = (b_1, \dots, b_n) \in \mathbb{R}^m$ is the right hand side of the system, $i = k \mod m$ and $x_{k-1} \in \mathbb{R}^n$ is the approximation of a solution x_* obtained in the previous step. The process continues until it triggers an appropriate convergence criterion.

To provide theoretical guarantees for the convergence of the method, Strohmer and Vershynin [14] proposed to choose the next row a_i at random with the probabilities weighted proportionally to the L_2 norms of the rows a_i. The authors have shown that this *randomized Kaczmarz algorithm* is guaranteed to converge exponentially in expectation, namely,

$$\mathbb{E}\,\|x_k - x_*\|_2^2 \le \left(1 - \frac{1}{R}\right)^k \|x_0 - x_*\|_2^2, \tag{I.2}$$

where x_* is the solution of the system (I.1) and R is a constant depending only on the matrix A, namely, $R = \|A\|_F^2 / s_{min}^2(A)$.

Here and further, we denote by $s_{min}(A)$ and $s_{max}(A)$ the smallest and largest singular values of the matrix A (that is, eigenvalues of the matrix $\sqrt{A^*A}$). Then, $\|A\|_F := trace(\sqrt{A^*A})$ (Frobenius, or Hilbert-Shmidt, norm of the matrix) and $\|A\| := \sup_{\|x\|_2=1} \|Ax\|_2$ (operator norm of the matrix). Moreover, we always assume that the matrix A has full column rank, so that $s_{min}(A) > 0$ and the convergence rate is non-trivial.

There is a variety of extensions and refinements of the first randomized Kaczmarz method. They include specializations of the method to some other classes of problems (like solving inconsistent linear systems [8], phase retrieval [15], stochastic gradient descent [11], etc); improvements in the weighting of the rows (from the one based on $\|a_i\|_2$ to some better "optimal" probabilities, see, e.g., [2]), and new hybrid methods based on Kaczmarz [6]. We omit a detailed discussion of such related work but refer the reader to those mentioned and others therein.

A. Block Kaczmarz and sketch-and-project ideas

The extension that will be of our major interest throughout the paper is a version of the Kaczmarz algorithm that uses blocks of the rows for iterative projections (rather than individual rows one by one). Namely, the $k+1$-st iteration has the form

$$x_{k+1} = x_k + (A_\tau)^\dagger (b_\tau - A_\tau x_k),$$

where A_τ and b_τ denote the restriction onto the (row) indices from the subset $\tau \subset \{1, \dots, m\}$ and $(A_\tau)^\dagger$ denotes the Moore-Penrose inverse of the matrix A_τ.

This framework was initially proposed by Elfving [4], and its randomized version was presented and analyzed in the paper by Needell and Tropp [9]. In the randomized version, the matrix A is split into several row blocks, and at each iteration one of these blocks is chosen uniformly at random with replacement. The authors prove the exponential convergence of the method with a strong convergence constant,

$$\mathbb{E}\,\|x_k - x_*\|_2^2 \le \left(1 - \frac{s_{min}^2(A)}{C\|A\|^2 \log(m+1)}\right)^k \|x_0 - x_*\|_2^2,$$

if we manage to choose a "good" row block partition, and under an assumption that all the rows are standardized, namely, $\|a_i\|_2 = 1$; see [9] for details.

Although the existence of this "good" partition is theoretically guaranteed, it is not always straightforward how to find such partition (e.g., if A has coherent rows). However, experimental evidence shows that the block Kaczmarz method still exhibits fast convergence even in these cases. This observation is especially interesting since coherent matrices are precisely the examples for which standard randomized Kaczmarz does not perform well (as projections at each step follow roughly the same direction, which might not be a direction toward the true solution x_*). Some theoretical analysis of this improvement for blocks of size two is available in [10].

A unified view on both regular and block Kaczmarz methods, along with many other randomized iterative solvers, was proposed by Gower and Richtárik in [2]. The main idea of their *sketch-and-project* framework is the following. One can observe that the random selection of a row (or a row block) can be represented as a sketch, that is, left multiplication by a random vector (or a matrix), thereby pre-processing every iteration of the method, which is represented by a projection onto the image of the sketch.

Thus, the iteration can be written as

$$x_{k+1} = (\mathrm{Id} - (S^T A)^\dagger S^T A) x_k + (S^T A)^\dagger S^T b, \quad \text{(I.3)}$$

where S is the sketch matrix, taken from some random matrix model at each step. For brevity, we will denote $A_S := S^T A$.

Clearly, in the case of block Kaczmarz, sketch matrices S are just shifted identity matrices tabbed by zeros for the correct size (m by block size). Standard Kaczmarz can be, of course, considered as a special case of a block method with the block size one.

The sketch-and-project viewpoint suggests a natural idea to generalize the methods by adopting some other sketch matrices S. Gower and Richtárik propose to take S to be a standard Gaussian matrix with independent entries. The authors show exponential convergence with the standard rate (I.2) with $R = 2\|A\|_F^2/\pi s_{min}^2(A)$ in the one dimensional case (when S is a Gaussian vector in $\mathbb{R}^m$).

B. Block Gaussian Kaczmarz and organization of the paper

We continue the study of the Gaussian sketch matrices in application to the Kaczmarz methods: we prove an exponential convergence rate for any block size b, analyze the dependence between b and the convergence rate in iteration and in time, compare the Gaussian Kaczmarz methods (block and one-dimensional versions) to the standard ones, and discuss various ways to select Gaussian sketches.

First, in Section II-A we prove

Theorem 1.1: Suppose A is a $m \times n$ matrix with full column rank ($m \ge n$), such that its condition number $\kappa^2(A) := s_{max}^2(A)/s_{min}^2(A) \le e^{m/4}/3$, and let x_* be a solution of the system $Ax = b$. For any initial estimate x_0, the block Gaussian Kaczmarz method (iteration (I.3) with S being an $m \times b$ random matrix with i.i.d. standard normal entries) produces a sequence $\{x_k, k \ge 0\}$ of iterates that satisfy

$$\mathbb{E}\,\|x_k - x_*\|_2^2 \le \left(1 - \frac{b}{15m\kappa^2(A)}\right)^k \|x_0 - x_*\|_2^2. \quad \text{(I.4)}$$

Remark 1.2: Note that the condition $\kappa^2(A) := s_{max}^2(A)/s_{min}^2(A) \le e^{m/4}/3$ naturally holds for many standard classes of matrices. For example, random Gaussian matrices have condition numbers $\kappa(A) \sim m$. Random matrices with i.i.d. elements having only two finite moments still have polynomial condition numbers $\kappa(A)$ with high probability.

An alternative (although very similar) estimate can be obtained for all matrices, without a condition number assumption, in trade of the absolute constants:

Theorem 1.3: Suppose A is a $m \times n$ matrix with full column rank ($m \ge n$) and let x_* be a solution of the system $Ax = b$. For any initial estimate x_0, the block Gaussian Kaczmarz method (iteration (I.3) with S being an $m \times b$ random matrix with i.i.d. standard normal entries) produces a sequence $\{x_k, k \ge 0\}$ of iterates that

satisfy

$$\mathbb{E}\|x_k - x_*\|_2^2 \leq \left(1 - \left[\frac{\sqrt{b}s_{min}(A)}{9\sqrt{b}\|A\| + C\|A\|_F}\right]^2\right)^k \|x_0 - x_*\|_2^2. \tag{I.5}$$

Here, $C > 0$ is an absolute constant.

This theorem is proved in Section II-B. It is usually stronger than Theorem 1.1 (note that since $\|A\|_F \leq m\|A\|$ for any matrix, Theorem 1.1 might give tighter results than Theorem 1.3 for some matrices A with $\|A\|_F \approx m\|A\|$, but its advantage will be at most by a constant multiple in the convergence rate R).

From Theorems 1.1 and 1.3 we can see that the expected gain of the block iterations is linear in the size of the block b (if we look at the per iteration gain). Clearly, for larger b, both the sketching step (computing S^TA) and the inversion step (computing $(S^TA)^\dagger$) become slower. In Section III we study this tradeoff numerically.

A natural question is whether we actually need to generate a new Gaussian matrix at each step. In some settings, if memory is not an issue, we may rather have a finite set of matrices (potentially, generated in advance) and sample from it. For comparison, in the regular Kaczmarz methods we have only a finite set of the sketch matrices (and the cardinality of this set is the number of rows of the matrix divided by the number of blocks). We prove that in the Gaussian case we can be satisfied with a finite collection of sketches as well. In Section II-C we prove the following.

Theorem 1.4: Suppose A is a $m \times n$ matrix with full column rank ($m \geq n$) and let x_* be a solution of the system $Ax = b$. Let N be such that $Cm^2 \log m \leq N \leq \exp(m/3)$ (for a large absolute constant C). Let $\mathcal{S} = \{S_1, \ldots, S_N\}$ be a random set of $m \times b$ random matrices with i.i.d. standard normal entries. Then, with probability at least $1 - 3/m$, for any initial estimate x_0, finite block Gaussian Kaczmarz method (iteration (I.3) with S being chosen at random from a set $\mathcal{S}$) produces a sequence $\{x_k, k \geq 0\}$ of iterates that satisfy

$$\mathbb{E}\,\|x_k - x_*\|_2^2 \leq \left(1 - \frac{b}{36m\kappa^2(A)}\right)^k \|x_0 - x_*\|_2^2.$$

Thus, the convergence rate is as good as in the case of taking a new sketch at each iteration (Theorem 1.3). However, the size of a pre-selected set $\mathcal{S}$ required by Theorem 1.4 ($N \gg m^2 \log m$) is likely too big to be practical. Our experiments show that in practice the size $N \sim m/b$ (number of rows divided by the block size, like in the regular block Kaczmarz case) is enough to demonstrate the same convergence.

See Section III for all the related numerical experiments. We next turn to the proofs of these main theorems.

II. Proofs of Main Results

A. Proof of Theorem 1.1, convergence estimate via condition number

The following lemma is standard and is proved in, e.g., [[12], Proposition 4.4]:

Lemma 2.1: Let $X = (X_{ij})$ be an $m \times n$ random matrix, $m \geq n$, whose entries are independent copies of a standard normal random variable. Then for all $s \geq 0$

$$\mathbb{P}\{s_{max}(X) > (2+s)\sqrt{m}\} \leq \exp(-s^2m/2).$$

Lemma 2.2: Let S be $m \times b$ matrix with i.i.d. standard normal entries and A is $m \times n$ fixed matrix. Let $\mathcal{E}$ be any event such that $\mathbb{P}(\mathcal{E}) \geq 1 - e^{-cm}$ for some $c \in (0, 1/2]$. Then for any fixed $u \in S^{n-1}$ and large enough m,

$$\mathbb{E}(\|A_Su\|_2^2|\mathcal{E}) \geq \mathbb{E}(\|A_Su\|_2^2) - e^{-cm/2}\|A\|^2.$$

If $c = 1/2$, it is enough to take $m \geq 25$ for the statement to hold.

Proof: For any $t > 0$,

$$\mathbb{P}(\|A_Su\|_2^2 > t|\mathcal{E}) = 1 - \mathbb{P}(\|A_Su\|_2^2 \leq t|\mathcal{E}) \geq 1 - \frac{\mathbb{P}(\|A_Su\|_2^2 \leq t)}{\mathbb{P}(\mathcal{E})} \geq \mathbb{P}(\|A_Su\|_2^2 > t) - e^{-cm},$$

since $\mathbb{P}(\mathcal{E}) \geq 1 - e^{-cm}$. Then,

$$\begin{aligned}\mathbb{E}(\|A_Su\|_2^2|\mathcal{E}) &\geq \int_{t=0}^{9m\|A\|^2} \mathbb{P}(\|A_Su\|_2^2 > t|\mathcal{E})dt \\ &\geq \int_{t=0}^{9m\|A\|^2} \mathbb{P}(\|A_Su\|_2^2 > t)dt - \int_{t=0}^{9m\|A\|^2} e^{-cm}dt \\ &\geq \mathbb{E}(\|A_Su\|_2^2) - \int_{t_0}^{\infty} \mathbb{P}(\|A_Su\|_2^2 > t)dt - \frac{9m\|A\|^2}{e^{cm}},\end{aligned}$$

where $t_0 = 9m\|A\|^2$. To bound the integral term, note that a trivial inequality $\|S^TA\| \leq \|S^T\|\|A\|$ implies

$$\mathbb{P}(\|A_Su\|_2^2 > t) \leq \mathbb{P}(\|A_S\|_2^2 > t) \leq \mathbb{P}(\|S^T\| > \frac{\sqrt{t}}{\|A\|}).$$

This allows as to bound

$$\int_{t=9m\|A\|^2}^{\infty} \mathbb{P}(\|S^TAu\|_2^2 > t)dt \tag{II.1}$$

$$\leq \int_{q=3}^{\infty} \mathbb{P}(\|S^T\| > q\sqrt{m})2qm\|A\|^2dq \leq 6e^{-m/2}\|A\|^2, \tag{II.2}$$

using change of variable $q = \sqrt{t}/\sqrt{m}\|A\|$, Lemma 2.1 and the fact that $q \le 3(q-2)$ for $q \ge 3$. As a result,

$$\begin{aligned}\mathbb{E}(\|A_S u\|_2^2|\mathcal{E}) &\ge \mathbb{E}(\|A_S u\|_2^2) - e^{-cm}\|A\|^2(9m+6)\\ &\ge \mathbb{E}(\|A_S u\|_2^2) - e^{-cm/2}\|A\|^2\end{aligned}$$

for m large enough. Note that $e^{m/4} \ge 9m+6$ for all $m \ge 25$. Lemma 2.2 is proved. ∎

Proposition 2.3: Suppose A is a $m \times n$ matrix with full column rank ($m \ge n$) and let x_* be a solution of the system $Ax = b$. Let x_k be a fixed vector in $\mathbb{R}^n$. Let S be $m \times b$ matrix with i.i.d. standard normal entries and x_{k+1} is obtained by iteration (I.3). Then,

$$\begin{aligned}&\mathbb{E}_S \|x_{k+1} - x_*\|_2^2\\ &\qquad \le (1 - \frac{b}{10m\kappa^2(A)} - \frac{e^{-m/4}}{10m})\|x_k - x^*\|_2^2.\end{aligned}$$

Proof: Since

$$\begin{aligned}x_{k+1} - x_* &= x_k - x_* - (A_S^\dagger)(A_S x_k - A_S x_*)\\ &= (\mathrm{Id} - A_S^\dagger A_S)(x_k - x_*),\end{aligned}$$

we have

$$\mathbb{E}\,\|x_{k+1} - x_*\|_2^2 = \mathbb{E}\,\|(\mathrm{Id} - A_S^\dagger A_S)(x_k - x_*)\|_2^2.$$

To prove the proposition, we are going to show that there exists a constant $c > 0$ such that for any fixed $u \in \mathbb{R}^n$

$$\mathbb{E}\,\|(\mathrm{Id} - A_S^\dagger A_S)u\|_2^2 \le (1-r)\|u\|_2^2, \tag{II.3}$$

where $r = b/10m\kappa^2(A) + e^{-m/4}/10m$.

Since $A_S^\dagger A_S$ is an orthogonal projector

$$\mathbb{E}\,\|(\mathrm{Id} - A_S^\dagger A_S)u\|_2^2 = \|u\|_2^2 - \mathbb{E}\,\|A_S^\dagger A_S u\|_2^2.$$

So, our goal is to prove that for a fixed $u \in S^{n-1}$

$$\mathbb{E}\,\|A_S^\dagger A_S u\|_2^2 \ge r. \tag{II.4}$$

Now, for any $\gamma > 0$, by the total expectation theorem,

$$\begin{aligned}\mathbb{E}\,\|A_S^\dagger A_S u\|_2^2 &\ge \mathbb{E}(s_{min}^2(A_S^\dagger)\cdot\|A_S u\|_2^2)\\ &\ge \mathbb{E}(s_{min}^2(A_S^\dagger)\|A_S u\|_2^2 \,\big|\, s_{min}^2(A_S^\dagger) \ge \frac{1}{\gamma^2})\\ &\qquad\qquad \cdot \mathbb{P}(s_{min}^2(A_S^\dagger) \ge \frac{1}{\gamma^2})\\ &\ge \frac{1}{\gamma^2}\,\mathbb{E}(\|A_S u\|_2^2|\mathcal{E})\mathbb{P}(\mathcal{E}),\end{aligned} \tag{II.5}$$

where $\mathcal{E} := \{\|A_S\| \le \gamma\}$. Now, with $\gamma = 3\sqrt{m}\|A\|$, we have

1) $\mathbb{P}(\mathcal{E}) = \mathbb{P}\{\|S^T A\| \le 3\sqrt{m}\|A\|\} \ge \mathbb{P}\{\|S^T\| \le 3\sqrt{m}\}$ since $\|S^T A\| \le \|S^T\|\|A\|$, and by Lemma 2.1

$$\mathbb{P}\{\|S^T\| \le 3\sqrt{m}\} \ge 1 - \exp(-m/2).$$

2) Then, by Lemma 2.2 applied to the event $\mathcal{E}$,

$$\mathbb{E}(\|A_S u\|_2^2|\mathcal{E}) \ge \mathbb{E}(\|A_S u\|_2^2) - e^{-m/4}\|A\|^2.$$

3) Finally, unconditional expectation can be computed directly:

$$\mathbb{E}(\|A_S u\|_2^2) = \mathbb{E}\sum_{i=1}^b \langle S_i^T, (Au)\rangle^2 = b s_{min}^2(A).$$

Combining three estimates above, we obtain

$$\begin{aligned}\mathbb{E}\,\|A_S^\dagger A_S u\|_2^2 &\ge \frac{(b s_{min}^2(A) - e^{-m/4}\|A\|^2)(1-e^{-m/2})}{9m\|A\|^2}\\ &\ge \frac{b s_{min}^2(A)}{10m s_{max}^2(A)} - \frac{e^{-m/4}}{10m},\end{aligned} \tag{II.6}$$

for any $m \ge 25$. This concludes the proof of Proposition 2.3. ∎

Proof of Theorem 1.1. Note the due to the condition number assumption, $\kappa^2(A) \le e^{m/4}/3$, the exponential term in the one step estimate from the Proposition 2.3 become negligible:

$$1 - \frac{b}{10m\kappa^2(A)} - \frac{e^{-m/4}}{10m} \le 1 - \frac{b}{15m\kappa^2(A)}.$$

Thus,

$$\begin{aligned}\mathbb{E}\,\|x_k - x_*\|_2^2 &= \mathbb{E}_{S_1}\mathbb{E}_{S_2}\ldots\mathbb{E}_{S_k}\|x_k - x_*\|_2^2\\ &\le \left[1 - \frac{b}{15m\kappa^2(A)}\right]^k \|x_0 - x_*\|_2^2.\end{aligned}$$

Theorem 1.1 is proved.

B. Proof of Theorem 1.3, convergence estimate via the mix of Frobenius and operator norms

The first auxiliary lemma is a direct corollary of a matrix deviation inequality (see, e.g., [[16], Theorem 9.1.1]). We will use it to make an estimate for the norm $\|S^T A\|$ (more refined than a trivial estimate $\|S^T A\| \le \|S^T\|\|A\|$ that was used in the proof of Theorem 1.1):

Lemma 2.4: Let S be $m \times b$ matrix with i.i.d. standard normal entries and A is $m \times n$ fixed matrix. Also, S^{n-1} denotes a unit sphere in $\mathbb{R}^n$. Then the following holds with some absolute constant $C > 0$:

$$\mathbb{E}(\sup_{w \in AS^{n-1}} \|S^T w\|_2) \le \sqrt{b}\|A\| + C\|A\|_F.$$

Proof: The matrix deviation inequality [[16], Theorem 9.1.1] states that for any $U \subset \mathbb{R}^n$

$$\mathbb{E}\sup_{x \in U}\Big|\|S^T x\|_2 - \sqrt{b}\|x\|_2\Big| \le C\gamma(U),$$

where $\gamma(U)$ is a Gaussian complexity defined by

$$\gamma(U) := \mathbb{E}\sup_{g \sim N(0, I_n)} |\langle g, x\rangle|.$$

Now, in the case when U is an ellipsoid $U = AS^{n-1}$, Gaussian complexity as well as the L_2-norm bound for the element in U are bounded in terms of the norms of the matrix A, namely, $\gamma(U) \lesssim \|A\|_F$ (see, e.g., [[16], Section 7.6]) and $\sup_{x\in U} \|x\|_2 = \sup_{y\in S^{n-1}} \|Ay\|_2 = \|A\|$. Thus,

$$\mathbb{E}(\sup_{x\in AS^{n-1}} \|S^T x\|_2) \le \sqrt{b}\sup_{x\in T}\|x\|_2 + C\gamma(U) \le \sqrt{b}\|A\| + C\|A\|_F.$$

This concludes Lemma 2.4. ∎

The second auxiliary lemma estimates the L_2-norm $\|S^T Ax\|_2$. It relies on the following version of Cramér's concentration inequality (see, e.g., [1])

Theorem 2.5 (Cramér's Theorem): Let X be a random variable, such that for all $\lambda \in \mathbb{R}$ $\mathbb{E}\, e^{\lambda X} < +\infty$. Let $X_1, \dots X_n$ be i.i.d. copies of X, and set $S = \sum_{i=1}^n X_i$. Then for any $a < \mathbb{E}\, X$ we have

$$\mathbb{P}\{\frac{1}{n}S < \alpha\} \le \exp(-I(\alpha)\cdot n),$$

where the function $I : \mathbb{R} \to [0, +\infty]$ is defined by

$$I(\alpha) = \sup_{t\in\mathbb{R}}(t\alpha - \log \mathbb{E}\exp(tX)). \tag{II.7}$$

Lemma 2.6: Let S be $m\times b$ matrix with i.i.d. standard normal entries and $v \in \mathbb{R}^m$ is a fixed vector. Then

$$\mathbb{P}(\|S^T v\|_2^2 > \|v\|^2 b/10) \ge 0.5.$$

Proof: Note that a random variable $Z := \|S^T v\|_2^2/\|v\|_2^2$ has a distribution of a sum of the squares of b independent standard normal Gaussian random variables

$$Z = \sum_{i=1}^b (\sum_{j=1}^m S_{ij}^T \frac{v_j}{\|v\|_2^2})^2 \sim \sum_{i=1}^b Z_i^2.$$

So, for any $i = 1, \dots, b$ a random variable Z_i^2 has chi-squared distribution with one degree of freedom, and a direct computation involving its moment generating function shows that for any $\alpha < 1$, the function $I(\alpha)$ defined by (II.7) is

$$I(\alpha) = \frac{\alpha - 1 + \ln(\alpha^{-1})}{2}.$$

Therefore, Cramer's Theorem 2.5 with $b \ge 1$ gives

$$\mathbb{P}(\|S^T v\|_2^2/\|v\|^2 \le \frac{b}{10}) \le \exp(-\frac{b(\ln(10) - 0.9)}{2}) \le \frac{1}{2}.$$

∎

Proposition 2.7: Suppose A is a $m \times n$ matrix with full column rank ($m \ge n$) and let x_* be a solution of the system $Ax = b$. Let x_k be a fixed vector in $\mathbb{R}^n$. Let S be $m \times b$ matrix with i.i.d. standard normal entries and x_{k+1} is obtained by iteration (I.3). Then,

$$\mathbb{E}_S\|x_{k+1} - x_*\|_2^2 \le (1 - \frac{bs_{min}^2(A)}{80(\sqrt{b}\|A\| + C\|A\|_F)^2})\|x_k - x_*\|_2^2.$$

Proof: Exactly like in the first part of the proof of Theorem 2.3, it is enough to show that for a fixed $u \in S^{n-1}$

$$\mathbb{E}\,\|A_S^\dagger A_S u\|_2^2 \ge \frac{bs_{min}^2(A)}{64(\sqrt{b}\|A\| + C\|A\|_F)^2}. \tag{II.8}$$

This time we apply total expectation theorem, conditioning on the norm of $\|A_S u\|_2$. For any parameter $\gamma^2 > 0$,

$$\begin{aligned}\mathbb{E}\,\|A_S^\dagger A_S u\|_2^2 &= \mathbb{E}(\big\|A_S^\dagger \frac{A_S u}{\|A_S u\|_2^2}\big\|_2^2 \cdot \|A_S u\|_2) \\ &\ge \mathbb{E}(\big\|A_S^\dagger \frac{A_S u}{\|A_S u\|_2}\big\|_2^2 \|A_S u\|_2^2 \,\big|\, \|A_S u\|_2^2 \ge \gamma^2) \\ &\quad \cdot \mathbb{P}(\|A_S u\|_2^2 \ge \gamma^2) \\ &\ge \gamma^2\,\mathbb{E}(\inf_{v\in S^{n-1}} \|A_S^\dagger v\|_2^2 \,\big|\, \mathcal{E}_\gamma)\mathbb{P}(\mathcal{E}_\gamma) \end{aligned} \tag{II.9}$$

where the event $\mathcal{E}_\gamma = \{\|A_S u\|_2^2 \ge \gamma^2\}$. Futhermore,

$$\begin{aligned}\mathbb{E}(\inf_{v\in S^{n-1}} \|A_S^\dagger v\|_2^2 \,\big|\, \mathcal{E}_\gamma) &= \mathbb{E}(\frac{1}{\sup\limits_{v\in S^{n-1}} \|A_S v\|_2^2} \,\big|\, \mathcal{E}_\gamma) \\ &\ge \frac{1}{\mathbb{E}^2(\sup\limits_{v\in S^{n-1}} \|A_S v\|_2 \,\big|\, \mathcal{E}_\gamma)}\end{aligned}$$

by the fact that $f(x) = x^2$ is a monotone function on $x \ge 0$ (and so $\sup \|.\|^2 = (\sup \|.\|)^2$) and Jensen's inequality applied to a convex function $g(x) = x^{-2}$. To estimate the denominator from above, we can use total probability theorem again, namely, for any event $\mathcal{E}_\gamma$

$$\mathbb{E}(\sup_{v\in S^{n-1}} \|A_S v\|_2 \,\big|\, \mathcal{E}_\gamma) \le \frac{\mathbb{E}(\sup \|A_S v\|_2)}{\mathbb{P}(\mathcal{E}_\gamma)}.$$

Finally, $\mathbb{E}(\sup\limits_{v\in S^{n-1}} \|A_S v\|_2)$ can be estimated by Lemma 2.4 as

$$\mathbb{E}\sup_{w\in AS^{n-1}} \|S^T w\|_2 \le \sqrt{b}\|A\| + C\|A\|_F.$$

Combining the last two estimates with (II.9), we obtain

$$\mathbb{E}\,\|A_S^\dagger A_S u\|_2^2 \ge \frac{\gamma^2\mathbb{P}^3(\|A_S u\|_2^2 \ge \gamma^2)}{(\sqrt{b}\|A\| + C\|A\|_F)^2}.$$

The numerator can be estimated by the Lemma 2.6 if we take $v = Au$ and $\gamma^2 = \|Au\|_2^2 b/8$:

$$\begin{aligned}\gamma^2\mathbb{P}^3(\|A_S u\|_2^2 \ge \gamma^2) &\ge \frac{\|Au\|_2^2 b}{10}\mathbb{P}^3(\|A_S u\|_2^2 \ge \frac{\|Au\|_2^2 b}{10}) \\ &\ge \frac{\|Au\|_2^2 b}{80} \ge \frac{s_{min}^2(A)b}{80}.\end{aligned}$$

So,

$$\mathbb{E}\,\|A_S^\dagger A_S u\|_2^2 \geq \frac{\gamma^2 \mathbb{P}^3(\|A_S u\|_2^2 \geq \gamma^2)}{L^2_{\|A\|,\|A\|_F}} \geq \frac{s^2_{min}(A) b}{80 L^2_{\|A\|,\|A\|_F}},$$

where $L_{\|A\|,\|A\|_F} = \sqrt{b}\|A\| + C\|A\|_F$. This concludes the proof of Proposition 2.7. ∎

Proof of Theorem 1.3. Theorem 1.3 follows immediately from the result of Proposition 2.7, since

$$\mathbb{E}\,\|x_k - x_*\|_2^2 = \mathbb{E}_{S_1}\mathbb{E}_{S_2}\ldots\mathbb{E}_{S_k}\|x_k - x_*\|_2^2 \leq \left[1 - c\frac{b s^2_{min}(A)}{(\sqrt{b}\|A\| + C\|A\|_F)^2}\right]^k \|x_0 - x_*\|_2^2.$$

C. Proof of Theorem 1.4, when we select a sketch from a pre-selected finite set of Gaussian matrices

The following lemma is a standard Bernstein's inequality for sub-exponential random variables (its proof can be found in, e.g., [[16], Corollary 2.8.3])

Lemma 2.8: Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have

$$\mathbb{P}\left\{\mathbb{P}\Big|\frac{1}{N}\sum_{i=1}^N X_i\Big| \geq t\right\} \leq 2\exp(-c\min(\frac{t^2}{K^2}, \frac{t}{K})N),$$

where $K = \max_i \|X_i\|_{\phi_1}$ is maximum sub-exponential norm of X_i.

Remark 2.9: We will also use the following easy facts about sub-exponential random variables (their proofs can be also found in [16]):

1) If a random variable X is standard Gaussian then X^2 is sub-exponential. Moreover, $\|X^2\|_{\phi_1} = \|X\|^2_{\phi_2} = const$
2) If a random variables X, Y are standard Gaussians then XY is a sub-exponential random variable. Moreover, $\|XY\|_{\phi_1} = \|X\|_{\phi_2}\|Y\|_{\phi_2} = const$
3) Centering of a sub-exponential random variable X produces another sub-exponential random variable, such that $\|X - \mathbb{E}\,X\|_{\phi_1} \leq C\|X\|_{\phi_1}$.

Definition 2.10 (Good collection): We will call a set $\mathcal{S} = \{S_1, \ldots, S_N\}$ of $m \times b$ matrices good, if the following conditions hold:

1) all $S_k \in \mathcal{S}$: $\|S_k\| \leq 3\sqrt{m}$;
2) all $(i,j) \neq (i,u) \in [m]\times[b]$: $|\sum_{k=1}^N s^k_{ji} s^k_{ui}| \leq \frac{N}{4m}$;
3) all $(i,j) \in [m]\times[b]$: $|\sum_{k=1}^N (s^k_{ij})^2| \geq \frac{N}{2}$

Here, S^k_{ij} denotes (i,j)-entry of the matrix S_k and $N = |\mathcal{S}|$.

Intuitively, conditions (2) and (3) from the definition of a good collection mean that, if in the process of sampling entries of the matrices in the collection $\mathcal{S}$ (uniformly with replacement), the sample covariance matrix obtained would be reasonably close to identity. Now we will show that a a random collection of standard Gaussian matrices is likely a good collection:

Proposition 2.11: Let $S_1, \ldots, S_N$ be independent, mean zero $m \times b$ random matrices with i.i.d. standard normal entries. If $(m^2 \ln m) \ll N \leq \exp(m/3)$, then with probability $1 - 3/m$ they form a good collection $\mathcal{S} = \{S_1, \ldots, S_N\}$ (in the sense of Definition 2.10).

Proof: Let us compute the probability that a random set of N standard Gaussian matrices is not good, namely, one of the conditions is not satisfied.

By Lemma 2.1 combined with the union bound,

$$\mathbb{P}\{\exists S \in \mathcal{S} : \|S\| > 3\sqrt{m}\} \leq N\exp(-m/2).$$

For $j \neq u$, for any matrix S_k from the collection, its (j,i) and (u,i) entries are independent Gaussian random variables, hence, there product is an exponential random variable. Moreover, all $\xi^k_{j,i,u} = s^k_{ji}s^k_{ui}$ are mean zero (as a product of two independent mean zero random variables) and independent for the non-coinciding j, i, u. So, by Lemma 2.8,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=k}^N \xi^k_{j,i,u}\right| \geq \frac{1}{4m}\right) \leq 2\exp(-c_1 N/16m^2).$$

Taking union bound over all pairs of indices (j,i) and (u,i), the probability that (2) does not hold for for S is bounded by $2m^2 b\exp(-c_1 N/16m^2)$.

Finally, $(s^k_{ij})^2$ are sub-exponential random variables, $\mathbb{E}(s^k_{ij})^2 = 1$. By Lemma 2.8 again,

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=k}^N (s^k_{ij})^2 \leq \frac{1}{2}\right) \leq \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=k}^N (s^k_{ij})^2 - 1\right| \geq \frac{1}{2}\right) \leq 2\exp(-c_2 N/4).$$

Taking union bound over all pairs of indices (j,i), the probability that (3) does not hold for for S is bounded by $2mb\exp(-c_2 N/4)$.

Therefore, combining three probabilities of the exceptional events, if $C_3 m^2 \ln m \leq N \leq \exp(m/3)$ with $C_3 \geq 48/c_1$, the probability that a random collection $\mathcal{S}$ of cardinality N is good is

$$1 - 2m^2 b e^{-c_1 N/16m^2} - 2e^{-c_2 N/4} - N e^{-m/2} \geq 1 - \frac{3}{m}.$$

This concludes the proof of Proposition 2.11. ∎

Now we will prove that given a good collection $\mathcal{S}$, iterative process (I.3) (choosing a sketch S from $\mathcal{S}$ randomly at each iteration) converges as fast as the analogous process, that samples a new Gaussian matrix at every step.

Proposition 2.12: Suppose A is a $m \times n$ matrix with full column rank ($m \geq n$) and let x_* be a solution of the system $Ax = b$. Let x_k be a fixed vector in $\mathbb{R}^n$. Let $\mathcal{S} = \{S_1, \dots, S_N\}$ be a good set of $m \times b$ matrices S_i (in the sense of Definition 2.10). At every iteration we choose a random matrix S uniformly at random from $\mathcal{S}$ (with replacement), and iterate according to (I.3). Then,

$$\mathbb{E}\,\|x_{k+1} - x_*\|_2^2 \leq \left[1 - \frac{b}{36m\kappa^2(A)}\right] \|x_k - x_*\|_2^2.$$

Proof: Exactly like in the first part of the proof of Proposition 2.3, it is enough to show that for a fixed $u \in S^{n-1}$

$$\mathbb{E}\,\|A_S^\dagger A_S u\|_2^2 \geq \frac{b}{36m\kappa^2(A)}, \tag{II.10}$$

where expectation is taken over the random choices of $S \in \mathcal{S}$.

Since $\mathcal{S}$ is a good collection, for any $S \in \mathcal{S}$ we have $1/s_{min}^2(A_S^\dagger) = s_{max}^2(S^T A) \leq \|S^T\|^2 \|A\|^2 \leq 9m\|A\|^2$. Thus,

$$\begin{aligned}\mathbb{E}\,\|A_S^\dagger A_S u\|_2^2 &\geq \mathbb{E}(s_{min}^2(A_S^\dagger) \cdot \|A_S u\|_2^2) \\ &\geq \frac{1}{9m\|A\|^2}\,\mathbb{E}(\|A_S u\|_2^2) \end{aligned} \tag{II.11}$$

Now, to estimate the expectation term, with the notation $v = Au$,

$$\begin{aligned}\mathbb{E}(\|A_S u\|_2^2) &= \mathbb{E}\sum_{i=1}^{b}\langle S_i^T, v\rangle^2 = \sum_{i=1}^{b}\mathbb{E}(\sum_{j=1}^{n} S_{ij}^T v_j)^2 \\ &= \sum_{i=1}^{b}\left[\sum_{j=1}^{m}\mathbb{E}(S_{ji}^2)v_j^2 + \sum_{j\neq r}\mathbb{E}(S_{ji}S_{ri})v_j v_r\right] \\ &= \sum_{i=1}^{b}\left[\sum_{j=1}^{m}\frac{1}{N}\sum_{k=1}^{N}(s_{ji}^k)^2 v_j^2 + \sum_{j\neq r}\frac{1}{N}\sum_{k=1}^{N}(s_{ji}^k s_{ri}^k)v_j v_r\right] \\ &\geq \sum_{i=1}^{b}\left[\sum_{j=1}^{m}\frac{1}{N}\sum_{k=1}^{N}(s_{ji}^k)^2 v_j^2\right] \\ &\quad - \sum_{i=1}^{b}\left[\sum_{j\neq r}\frac{1}{N}\Big|\sum_{k=1}^{N} s_{ji}^k s_{ri}^k\Big| \cdot |v_j v_r|\right] \qquad \text{(II.12)} \\ &\geq \sum_{i=1}^{b}\left[\sum_{j=1}^{m}\frac{1}{2}v_j^2\right] - \sum_{i=1}^{b}\left[\sum_{j,r=1}^{m}\frac{1}{4m}|v_j v_r|\right] \geq \frac{b}{4}\|v\|^2. \end{aligned} \tag{II.13}$$

In (II.12), we used the properties (2) and (3) of a good collection, and the last line holds since

$$\sum_{j,r}|v_j v_r| \leq 0.5\sum_{j,r}(v_j^2 + v_r^2) \leq n\|v\|_2^2.$$

Now recall that $\|v\|^2 = \|Au\|^2 \geq s_{min}^2(A)$. Combining (II.11) with (II.13) we conclude that

$$\mathbb{E}\,\|A_S^\dagger A_S u\|_2^2 \geq \frac{1}{9m\|A\|^2}\frac{b}{4}s_{min}^2(A).$$

■

Proof of Theorem 1.4. Given the constraints on the size of the collection N, a random collection $\mathcal{S}$ of standard normal matrices will be a good set (in the sense of Definition 2.10) with the probability at least $1 - 3/m$. Conditioned on this high probability event, the iteration process will converge exponentially fast as promised by statement of Theorem 1.4, since

$$\begin{aligned}&\mathbb{E}(\|x_k - x_*\|_2^2 |\, \mathcal{S} \text{ is good }) \\ &= \mathbb{E}_{S_1}\mathbb{E}_{S_2}\dots\mathbb{E}_{S_k}(\|x_k - x_*\|_2^2 |\, \mathcal{S} \text{ is good }) \\ &\leq \left[1 - \frac{b}{36m\kappa^2(A)}\right]^k \|x_0 - x_*\|_2^2.\end{aligned}$$

Here, $\mathbb{E}_{S_1}, \dots, \mathbb{E}_{S_k}$ refer to the randomness of choosing a matrix $S_i \in \mathcal{S}$ (uniformly at random with replacement). The last inequality is guaranteed by Proposition 2.12.

III. Numerical experiments

In this section, we present some numerical experiments to complement the discussion of the theoretical performance of the Gaussian block Kaczmarz method. Everything was coded and run in MATLAB R2018b. In comparing different methods, we run the iteration process until the fastest method reaches relative error 1e-4, or until 1000 iterations. Relative error is defined as $\|x_k - x_*\|_2^2/\|x_*\|_2^2$. Time is always measured in seconds. We generate the solution of a system as a random vector (and define the left hand side b accordingly), so we do not need to worry about the case when $\|x_*\|_2 = 0$. We use $x_0 = 0$ as an initial point.

We consider two main models of matrices A: the first one is an incoherent Gaussian matrix with i.i.d. $N(0,1)$ elements ("Gaussian model"), the second one models a coherent matrix with almost collinear rows, $A_{ij} \sim Unif[0.8, 1]$ ("flat model"). Unless otherwise stated, we consider matrices of size $m = 50000$ and $n = 500$.

A. Comparison with the other Kaczmarz methods

First, we compare the rate of convergence of four Kaczmarz methods. As mentioned earlier, all of them can be described in terms of iteration (I.3). Regular Kaczmarz method uses sketches $S = (0, \dots, 0, 1, 0, \dots, 0)^T$, where the position of 1 is chosen randomly st each iteration. Block Kaczmarz uses $m \times b$ sketches

$$S = (\text{zeros}(b, \text{shift}), I(b,b), \text{zeros}(b, m - \text{shift} - b))^T,$$

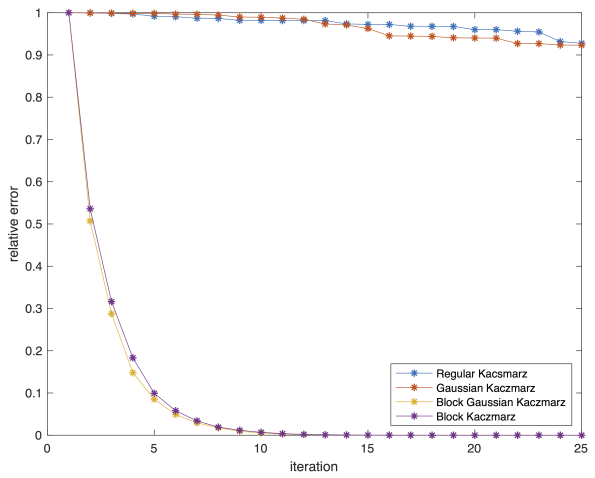

Fig. 1: Gaussian model. Decay of the relative error of four Kaczmarz methods in 25 iterations (until block methods reach the error 1e-4).

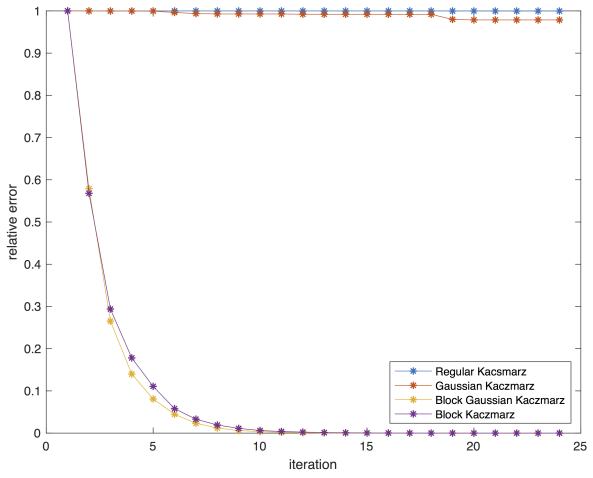

Fig. 2: Flat model. Decay of the relative error of four Kaczmarz methods in 25 iterations (until block methods reach the error 1e-4).

where b is the block size and shift $= bz$, $z \in \{1, 2, ..., \lfloor m/b \rfloor\}$ is selected randomly at each step. For the sake of efficiency, we realize these methods by selection of rows (or row blocks) for projection rather than by the sketching procedure described above. Gaussian Kaczmarz uses sketches $S = \xi$, where the vectors $\xi \in \mathbb{R}^m$ have $N(0, 1)$ independent coordinates, and Gaussian block Kaczmarz uses sketches S, where the matrices $S \in \mathbb{R}^{m \times b}$ have $N(0, 1)$ independent coordinates.

Figures 1 and 2 show that the block methods converge much faster in iteration on both coherent and incoherent matrices. Figure 3 shows that the same holds for the rate of convergence in time. Moreover, in iteration, block Gaussian Kaczmarz and regular randomized block Kaczmarz behave almost identically (the Gaussian version converges slightly faster), but, because of the heavy step of multiplication by Gaussian matrices (rather than the light step of selecting a row block), in time, regular block Kaczmarz converges faster. We used block size $b = 223 \sim \sqrt{m}$.

We can also observe that both one-dimensional methods perform much worse on the flat (coherent) matrices, whereas block methods do not seem to distinguish between coherent and incoherent case. The effect of Gaussian sketching is especially impressive in the one-dimensional case and a flat model, when the Gaussian method has decent convergence, and the regular randomized Kaczmarz is barely making progress toward the solution. See also the related discussion in Section IV.

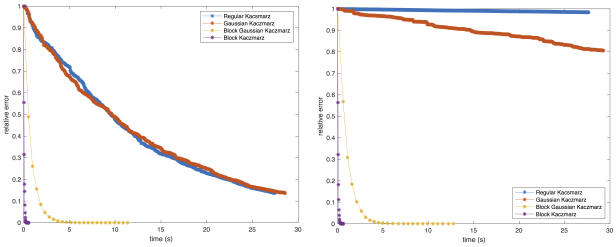

Fig. 3: Left: Gaussian model; right: flat model. Decay of the relative error of four Kaczmarz methods in time.

B. *Dependence on the size of the block*

Then, confirming the fact that convergence rates (I.4) and (I.5) are better with larger b, we observe that the error decays faster for larger values of b (See Figure 4 (left) where we plot the decay of the relative error with the iterations for the Gaussian model, varying block sizes $b = 1, 10, 50, 500$). Note that since the dimension of the solution x_k is $n = 500$ and A have full column rank, for the block sizes larger or equal than 500, the process converges in one iteration. Moreover, the same trend preserves when we look at the convergence rate in time (Figure 4, right): it is worth taking larger block sizes for faster convergence. However, in some cases (see Figure 5) it might be practical to use smaller block sizes to achieve a reasonable error before the completion of the first step of a slower iteration (corresponding to larger b).

Figure 6 confirms the advantage of taking bigger block sizes (as long as $b \leq n$), and also illustrates that

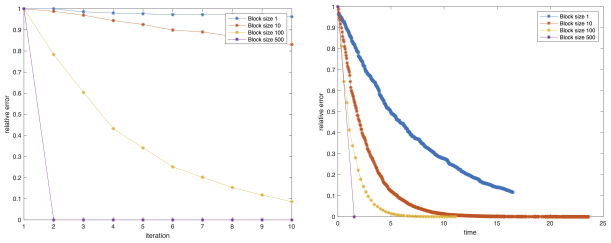

Fig. 4: Iteration (left) and time (right) vs error; block sizes 1 (blue); 10 (red); 100 (yellow) and 500 (purple)

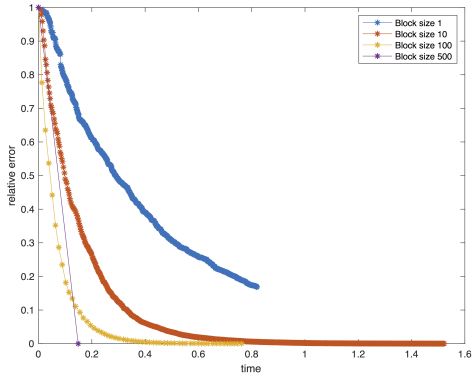

Fig. 5: Time vs error; $A \sim randn(2000, 500)$.

iterations become slower as b increases (for $b \geq 500$ convergence to the error requires exactly one iteration of the algorithm).

C. Finite number of samples

The final part of the experiments is related to the second way of sketching, when we pre-select a set of Gaussian matrices $\mathcal{S}$ and take sketch matrices from it. Although our theoretical analysis requires the cardinality of $\mathcal{S}$ to be at least $Cm^2 \log m$, the numerical experiments show that in practice the cardinality of $\lceil m/b \rceil$ is enough to reproduce the per iteration convergence rate of the

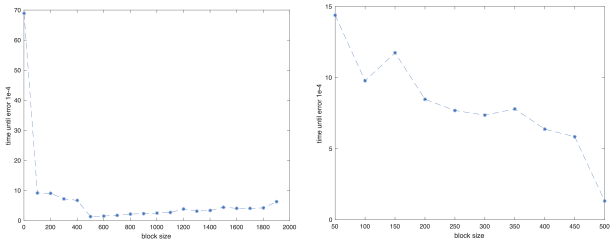

Fig. 6: Block size vs time until error 1e-4; block size ranges 1:2000 (left); 50:500 (right)

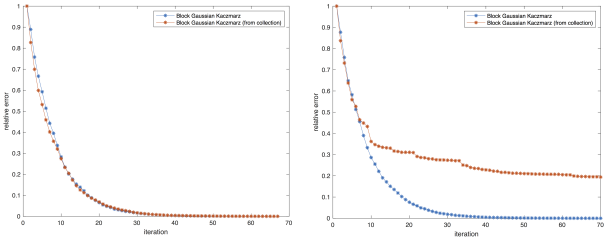

Fig. 7: Blue: Block Gaussian Kaczmarz selecting new gaussian sketch each time. Red: Block Gaussian Kaczmarz selecting sketches from the finite collection. Left: block size = 70; $|\mathcal{S}| = 70 \sim 5000/70$; right: block size = 70; $|\mathcal{S}| = 7$ (too small). Matrix $A \sim randn(5000, 500)$

original Gaussian block Kaczmarz. However, for very small collections $\mathcal{S}$ the method stops converging too far from the solution (see Figure 7).

IV. Conclusions and Future Directions

To the best of our knowledge, our paper presents the first theoretical analysis of the exponential convergence properties of Gaussian block Kaczmarz algorithm for arbitrary block size (Theorems 1.1 and 1.3). In the trivial block size case $b = 1$, Theorem 1.3 recovers the expected exponential convergence rate with the factor $1 - cs^2_{min}(A)/\|A\|^2_F$, which coincides with the results of Gower and Richtárik [2] for the one-dimensional Gaussian Kaczmarz (and the standard convergence rate of the randomized Kaczmarz, proved by Strohmer and Vershynin [14]).

We also propose a way that allows an approach to avoid potentially infinite generation of the Gaussian sketch matrices: one could rather be satisfied by sampling sketches from a pre-determined collection of Gaussian matrices. We prove (Theorem 1.4) that with high probability, a random collection of Gaussian matrices will provide the same rate of convergence as the original random sampling approach.

Unlike the analysis of the randomized block Kaczmarz method, the convergence guarantees for the Gaussian version (Theorems 1.3 and 1.4) do not require any additional structural assumptions on the matrix (such as regularized rows), or any non-trivial preprocessing (to find a special "good" partition of the rows into the paving row blocks). Namely, the Gaussian algorithm is theoretically justified in its most straightforward execution version.

Our estimates show clear positive correlation between the convergence rate per iteration with the block size. Numerical experiments support this: per iteration convergence gets faster with the increasing block size. Our

simulations also show that despite the fact that with bigger blocks every iteration of the algorithm becomes slower, the relative error still decays faster in time with bigger block size. However, for some combinations of the desired time and error, average sized blocks b might be preferable.

The block versions, as well as Gaussian versions (even one-dimensional), overcome a well-known issue of the original randomized Kaczmarz method, namely, its poor performance on coherent matrices (see also Figure 3, right). One of the next directions of the current work is to explain this observation theoretically. Some other interesting related questions are: to give a better upper bound on the required size of the collection from Theorem 1.4, to investigate the effect of sampling with and without replacement on the convergence, and to consider some "lighter" versions of the sketch matrices (such as sparsified Gaussian matrices) for faster computation.

REFERENCES

[1] Boucheron, S., Lugosi, G. and Massart, P., 2013. Concentration inequalities: A non-asymptotic theory of independence. Oxford university press.

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

[3] Feichtinger, H.G., Cenker, C., Mayer, M., Steier, H. and Strohmer, T., 1992, November. New variants of the POCS method using affine subspaces of finite codimension with applications to irregular sampling. In Visual Communications and Image Processing'92 (Vol. 1818, pp. 299-311). International Society for Optics and Photonics.

[4] Elfving, T., 1980. Block-iterative methods for consistent and inconsistent linear equations. Numerische Mathematik, 35(1), pp.1–12.

[5] Kaczmarz, S., 1937. Angenaherte auflosung von systemen linearer gleichungen. Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat., pp.355–357.

[6] De Loera, J.A., Haddock, J. and Needell, D., 2017. A Sampling Kaczmarz–Motzkin Algorithm for Linear Feasibility. SIAM Journal on Scientific Computing, 39(5), pp.S66–S87.

[7] Natterer, F., 1986. The mathematics of computerized tomography (Vol. 32). Siam.

[8] Needell, D., 2010. Randomized Kaczmarz solver for noisy linear systems. BIT Numerical Mathematics, 50(2), pp.395-403.

[9] Needell, D. and Tropp, J.A., 2014. Paved with good intentions: analysis of a randomized block Kaczmarz method. Linear Algebra and its Applications, 441, pp.199–221.

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

[11] Needell, D., Ward, R. and Srebro, N., 2014. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In Advances in Neural Information Processing Systems (pp. 1017-1025).

[12] Rudelson, M., Lecture notes on non-asymptotic random matrix theory

[13] Sezan, M.I. and Stark, H., 1987. Incorporation of a priori moment information into signal recovery and synthesis problems. Journal of mathematical analysis and applications, 122(1), pp.172–186.

[14] Strohmer, T. and Vershynin, R., 2009. A randomized Kaczmarz algorithm with exponential convergence. Journal of Fourier Analysis and Applications, 15(2), p.262.

[15] Tan, Y.S. and Vershynin, R., 2017. Phase retrieval via randomized Kaczmarz: theoretical guarantees. Information and Inference: A Journal of the IMA.

[16] Vershynin, R., High-Dimensional Probability, An Introduction with Applications in Data Science