A PRECONDITIONER BASED ON NON-UNIFORM ROW SAMPLING FOR LINEAR LEAST SQUARES PROBLEMS

LONG CHEN AND HUIWEN WU

ABSTRACT. Least squares method is one of the simplest and most popular techniques applied in data fitting, imaging processing and high dimension data analysis. The classic methods like QR and SVD decomposition for solving least squares problems has a large computational cost. Iterative methods such as CG and Kaczmarz can reduce the complexity if the matrix is well conditioned but failed for the ill conditioned cases. Preconditioner based on randomized row sampling algorithms have been developed but destroy the sparsity. In this paper, a new preconditioner is constructed by non-uniform row sampling with a probability proportional to the squared norm of rows. Then Gauss Seidel iterations are applied to the normal equation of the sampled matrix which aims to grab the high frequency component of solution. After all, PCG is used to solve the normal equation with this preconditioner. Our preconditioner can keep the sparsity and improve the poor conditioning for highly overdetermined matrix. Experimental studies are presented on several different simulations including dense Gaussian matrix, ‘semi Gaussian’ matrix, sparse random matrix, ‘UDV’ matrix, and random graph Laplacian matrix to show the effectiveness of the proposed least square solver.

1. INTRODUCTION

Least squares method is one of the simplest and most commonly applied techniques of data fitting. It can be applied in statistics to construct linear regression model and unbiased linear estimator [17], in imaging processing for image deblurring [3], and in high-dimensional data analysis like canonical polyadic tensor decomposition [14] etc.

Consider the overdetermined system

\[ Ax = b, \]

where \( A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^{m}, m \gg n, \) and \( \text{rank}(A) = n. \) As \( m > n, \) solutions to (1) are in general not unique. The least squares solution to the overdetermined system (1) is

\[ x_{\text{opt}} = \arg \min_{x} \| Ax - b \|_2^2, \]

where \( \| \cdot \|_2 \) is the \( l^2 \) norm of a vector. It is not difficult to show that the least squares solution \( x_{\text{opt}} \) satisfies the so-called normal equation

\[ A^T A x = A^T b. \]

We shall develop a fast iterative least squares solver for (1). There are various methods for solving the least squares problems. Classical methods such as QR and SVD decomposition that require \( O(mn^2) \) operations which is prohibitive for large size problems [12]. Kaczmarz methods [13] and randomized Kaczmarz methods [22] are effective iterative methods for consistent least squares problems while the computational costs highly depends on a scaled condition number of matrix \( A. \) Several fast least squares solvers have been developed recently utilizing randomization see, for example [8, 21, 1]. These fast
least squares solvers try to construct a spectrally equivalent but of smaller size matrix via random transformation and random sampling or mixing. Due to the random transformation, however, these methods destroy the sparsity of $A$, and thus may not be suitable for sparse matrices.

Conjugate Gradient (CG) method is an efficient iterative algorithm for solving linear systems of equations when the matrix is symmetric and positive-definite [12]. Thus CG method can be applied to solve the normal equation (3). CG is the archetype of Krylov subspace method that only needs the matrix-vector multiplication $Au$ and $A^Tv$. In order to achieve accuracy $\epsilon$, it needs $k \approx O(|\log(\epsilon)|\kappa(A))$ iterations, where the condition number $\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$ with $\sigma_{\max}(A), \sigma_{\min}(A)$ being the maximal and minimal (non-zero) singular values of $A$ respectively. When $\kappa(A)$ is large, the method converges very slowly. Preconditioned conjugate gradient (PCG) with proper preconditioners can be applied to accelerate the convergence. The key of success of PCG is a good preconditioner.

We shall use row sampling to construct an efficient preconditioner. Firstly we apply the random sampling to the rows of matrix $A$ and select $Cn \log(n)$ rows, with probability proportional to the squared norm of each row, to get a sampled matrix $A_s$. Then the preconditioner $P$ is constructed by solving the sampled normal equation (4)

$$A_s^TA_s e = r.$$

via a few symmetric Gauss-Seidel iterations. At last, we equip PCG method with this preconditioner to solve the original normal equation (3).

The main motivation of our method is to utilize the benefits of random sampling while keep the sparsity. By the random row sampling, we got a much smaller matrix $A_s$ compared to $A$ which reduces the complexity of matrix multiplication. The sampling algorithm and the chosen sampling density ensure the sampled matrix will capture the high frequency part of the original matrix $A$. The Gauss-Seidel iterations to the sampled normal system (4) will smooth out the high frequency of the error and thus improves the condition number.

The complexity of our method is $O(n^3 \log(n)) + O(kmn)$ for dense matrices. For sparse matrices, suppose the number of nonzero elements of matrix $A$ is $nnz$ and $nnz \ll mn$. The complexity is at most $O(n^2) + O(knnz)$ and could be $O(n) + O(knnz)$ depending on the sparse pattern, where $k = O(|\log(\epsilon)|\kappa(PA))$ is the iteration steps of PCG depending on the condition number of the preconditioned system $PA$. Although we cannot obtain an uniform bound of $\kappa(PA)$, we show by the numerical experiments that PCG with the new preconditioner improves the convergence significantly comparing with CG with a simple diagonal preconditioner.

The paper is organized as follows. Section 2 contains a summary of existing methods for least squares problem including classical methods like QR and SVD decompositions, Kaczmarz methods of cyclic and randomized version, and several fast least squares solvers. The details of our algorithms including how to construct the row sampling preconditioner and the complexity of algorithms are discussed in Section 3. Then numerical experiments are presented in Section 4, showing our proposed solver is efficient and effective.

2. Existing Methods

In this section, we review several existing methods for least squares problem. These methods are: the most basic ones like QR and SVD decomposition, the iterative methods such as Conjugate Gradient Descent(CG), Kaczmarz methods and also the newly developed randomized methods like randomized Kaczmarz method, and fast least squares solvers with random transformation. Convergence rate and computational complexity are
compared while complexity is mainly calculated for dense matrices. Each method has its advantages and disadvantages – none is perfect. Due to the eruption of informations, the tendency is to find more efficient methods with less computation complexity and storage.

2.1. Classical Methods. Classical methods to solve the least squares problem include direct methods such as QR decomposition, SVD decomposition, and iterative methods such as CG methods and Kaczmarz methods. The complexity for QR decomposition and SVD decomposition is $O(mn^2)$ which is huge when $m$ or $n$ is large. This is the limitation of direct methods like QR and SVD decomposition.

CG method is an iterative method to solve symmetric and positive definite (SPD) systems. It is applicable to sparse systems that are too large to be solved by a direct solver. In general for Krylov subspace methods, only matrix-vector product instead of matrix-matrix product is required and thus save the storage and reduce the complexity provided the method convergences fast. In the overdetermined case, if CG is applied to the normal equation (3), only the matrix-vector multiplication $Au$ is needed which cost $O(nnz(A))$ operations. Here $nnz(\cdot)$ is the number of nonzero entries of a matrix. In order to achieve the accuracy $\epsilon$, CG needs $O(\|\log(\epsilon)\|\sqrt{\kappa(A^T A)}) = O(\|\log(\epsilon)\|\kappa(A))$ steps, where the condition number of matrix $A$ is defined by

$$ \kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} $$

with $\sigma_{\max}(A), \sigma_{\min}(A)$ being the maximal and minimal singular values of $A$ respectively. Therefore the complexity for CG applied to (3) is $O(mn\kappa(A)\|\log(\epsilon)\|)$ for dense matrices and $O(nnz(A)\kappa(A)\|\log(\epsilon)\|)$ for sparse matrices. Tailored implementations of CG to normal equation (3) include CGLS [4] and LSQR [20].

For consistent systems, Kaczmarz method can be applied. A linear system is called consistent if there is at least one solution, i.e. $b \in \text{range}(A)$ in (1). Kaczmarz method is to project approximation onto the hyperplane $a_i^T x = b_i$ where $a_i$ is the $i$-th row of matrix $A$. Oswald and Zhou [19] obtained the convergence rate of cyclic Kaczmarz method

$$ \|x_{\text{opt}} - x_s\|_2^2 \leq \left[ 1 - \frac{1}{(\log(n) + 1)\kappa(A^T D^{-1} A)} \right]^s \|x_{\text{opt}} - x_0\|_2^2, $$

where $x_{\text{opt}}$ is the least squares solution, $x_s$ is the $s$-th iterate consisting of sweeping of all rows, $x_0$ is the initial guess and $D$ is an $m \times m$ diagonal matrix which induces a row scaling. One sweeping takes $O(mn)$ operations. In order to achieve accuracy $\epsilon$, the total complexity of Kaczmarz method is $O(mn\log(n)\kappa(A^T D^{-1} A)\|\log(\epsilon)\|)$.

The advantage of iterative methods is that they utilize the sparsity since in each iteration only matrix-vector multiplications are calculated. For example, the complexity of CG is $O(nnz(A)\kappa(A)\|\log(\epsilon)\|)$ when $A$ is sparse. For Kaczmarz method, the complexity also reduces since the cost of each projection is less than $O(n)$ depending on the sparsity of $A$. As the iteration steps of both CG and Kaczmarz for consistent systems depend crucially on the condition number $\kappa(A)$, they are slow if $A$ is ill-conditioned, i.e. $\kappa(A) \gg 1$. Preconditioner can be used to improve the condition number and in turn accelerate the convergence. One way to construct effective preconditioners is to use random sampling and random transformation, which will be discussed below.

2.2. Randomized Methods. There are several approaches to accelerate traditional least squares solvers via randomization. The main idea is either use random sampling or random projection to reduce the size of the original matrix, e.g. the randomized Kaczmarz method [22] and randomized fast solvers in [1, 9], or construct preconditioners by random
sampling to reduce the condition number which enable to apply PCG [21] or LSQR [1] of the preconditioned system.

2.2.1. Randomized Kaczmarz Methods. The convergence rate of the cyclic Kaczmaz method highly depends on the ordering of rows of $A$. In order to achieve a faster convergence which is independent of ordering of rows, choosing rows at random is a good strategy. For consistent systems, i.e. $b \in \text{Range}(A)$, Strohmer and Vershynin [22] proposed a randomized Kaczmarz (RK) method which selects the hyperplane to do projection via the probability proportional to $\|a_i\|^2_2$, and proved its exponential convergence in expectation, i.e.

$$\mathbb{E}(\|x_k - x_{\text{opt}}\|_2^2) \leq (1 - \kappa_F(A)^{-2})^k \|x_0 - x_{\text{opt}}\|_2^2,$$

where $x_k$ is the $k$-th iteration consisting of one projection only, $x_0$ is the initial guess, $x_{\text{opt}}$ is the least squares solution and $\kappa_F(A) = \|A\|_F^2 \|(A^\top A)^{-1}\|_F^2$ is a scaled condition number. Notice that here one iteration requires only $O(n)$ operations. To achieve the accuracy $\epsilon$, the expected iteration steps $O(\kappa_F^2(A) \log(\epsilon))$ and the total expected complexity is $O(n \kappa_F^2(A) \log(\epsilon))$.

For consistent systems, the randomized Kaczmarz method converges with expected exponential rate independent on the number of equations in the system. Indeed, the solver does not need to know the whole system but only a $O(n \log n)$ rows as the system is assumed to be consistent [22]. Thus it outperforms some traditional methods like CG on general extremely overdetermined system. The main limitation of RK is its inability of handling inconsistent systems. For instance, to solve $Ax = b$, where $b = y + w$, with $y = b_{\text{Range}(A)}$ is the projection of $b$ onto range of $A$ and $w = b_{(A\top A)^{-1}}$, the randomized Kaczmarz method is effective when least squares estimation is effective, i.e. the least squares error $\|w\|_2$ is negligible [25]. Extension of randomized Kaczmarz methods to inconsistent systems can be found in [16, 25, 24].

2.2.2. Fast Least Squares Solvers by Random Transformations. Drineas, Mahoney, Muthukrishnan and Sarlos [8] developed two randomized algorithms for least squares problems. Instead of solving the original least squares problem $\|Ax - b\|_2^2$, they solve an approximate least squares problem $\|X Ax - X b\|_2^2$, where $X = \text{SHD}$ for randomized sampling or $X = \text{THD}$ for randomized projection. The operator $HD$ is the randomized Hadamard transformation which aims to spread out the elements of $A$ and $S$ is the uniform sampling matrix and $T$ is the randomized projection matrix aiming to reduce the size of the original problem. The complexity of Hadamard transformation is $O(m \log(m))$ and the complexity of traditional methods to the approximated least squares problem is $O(rn^2)$ with the sample size $r = O(n/\epsilon)$ is chosen so that $XA$ is full rank but $r \ll m$. The complexity reduce to $O(m n \log(n/\epsilon) + n^3/\epsilon) \ll O(mn^2)$ provided $\epsilon$ is not too small [9].

Rokhlin and Tygert [21] proposed a fast randomized algorithm for overdetermined linear least squares regression. They constructed subsampled randomized Fourier transform (SRFT) matrix $T$ of size $r \times m$ and then apply pivoted QR decomposition to $TA = QRH$, with a $r \times n$ orthonormal matrix $Q$, an upper-triangular $n \times n$ matrix $R$ and an $n \times n$ permutation matrix $\Pi$. Then apply PCG with the right preconditioning matrix $P = R\Pi$, i.e. to minimize the preconditioned system $\|AP^{-1} y - b\|_2$.

According to theory developed in [21], $\kappa(AP^{-1}) = \kappa(TU)$ where the columns of $U$ are left singular vectors of $A$. The condition number of $TU$ can be controlled by the number of rows of $T$, i.e. $r$. In practice, $\kappa(TU) \leq 3$ when $r = 4n$. In this method, it converges fast since $\kappa(AP^{-1})$ is much smaller than $\kappa(A)$ but needs one QR decomposition which is $O(n^2 r)$. The total theoretical complexity is $O((\log(r) + \kappa(AP^{-1}) \log(\epsilon))mn) + O(n^2 r)$. 


Avron, Maymounkov and Toledo [1] developed an algorithm called BLENDENPIK, which supercharges LAPACK’s dense least-squares solver. They introduce the concept of coherence number $\mu(A)$, which is the maximum of squared norm of rows of $Q$, where the columns of $Q$ are a set of orthonormal bases of range of $A$. They find that the uniform sampling will work if the coherence number of the matrix is small and apply the row mixing to reduce $\mu(A)$ if it is large [1]. The crucial observation is that a unitary transformation preserves the condition number but changes the coherence number $\mu(A)$. After prepossessing with row mixing, the coherence number $\mu(A)$ is reduced and uniform sampling can be applied to get a sampled matrix $A_s$ with only $O(n \log(n))$ rows. Then they decomposed the sampled matrix $A_s = QR$ and used LSQR method to solve the original system $Ax = b$ with preconditioner $R^{-1}$ [1]. The complexity of row mixing is $O(mn \log(m))$ and QR decomposition of sampled matrix is $O(n^3)$. LSQR applied to the linear system $Ax = b$ with preconditioner $R^{-1}$ costs $O(mn\kappa(AR^{-1})\log(\epsilon))$. To conclude, the total complexity of BLENDENPIK method is $O(mn \log(m) + n^3 + mn\kappa(AR^{-1})\log(\epsilon))$.

The common feature of these fast least squares solvers is that they all try to use random sampling or random transformation to get a spectrally equivalent matrices but with considerably small size. Then an efficient preconditioner can be constructed via these small size matrices. However, the preprocesses of these methods transform sparse matrices into dense matrices, which cannot take the advantage of sparsity if the original matrix $A$ is.

To conclude, the traditional methods like QR and SVD decomposition need $O(mn^2)$ which is prohibitive when $m, n$ is large. Iterative methods such as CG and Kaczmarz can reduce the complexity if the matrices are well conditioned but failed for the ill conditioned cases. Preconditioner based on randomized row sampling algorithms have been developed but destroy the sparsity. Our goal is to construct a preconditioner which can keep the sparsity and improve the poor conditioning for highly overdetermined matrix.

3. Row Sampling

3.1. Preliminaries and Notation. Before walking into the details of our algorithms, we introduce some notation and concepts we may confront. Suppose $A$ is a matrix of size $m \times n$ with $m \geq n$. Denote $a_1, a_2, \cdots, a_m$ to be the row vectors of $A$ and $a^1, a^2, \cdots, a^n$ the column vectors of $A$. For any vector $v$, $\|v\| = (\sum_i v_i^2)^{\frac{1}{2}}$ is the $l_2$ norm of $v$ and is called norm of $v$ for short. For any matrix $A$, the spectral norm $\|A\| = \max_{x \neq 0} \|Ax\|/\|x\|$ is the induced matrix norm by vector $l_2$ norm and the Frobenius norm $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{\frac{1}{2}}$.

Definition 3.1 (Condition Number). The condition number $\kappa(A)$ of matrix $A$ is defined as

$$\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

with $\sigma_{\max}(A), \sigma_{\min}(A)$ are the maximal and minimal singular values of $A$ respectively.

A matrix is said to be singular if the condition number is infinite. In our setting, the matrix $A$ is of full rank. Thus the smallest singular value of $A$ is nonzero and the condition number $\kappa(A) < \infty$.

Another concept need to mention is the coherence number introduced in [1].

Definition 3.2 (Coherence Number [1]). Let $A$ be an $m \times n$ full rank matrix, and let $U$ be an $m \times n$ matrix whose columns form an orthonormal basis of the column space of $A$. The coherence of $A$ is defined as

$$\mu(A) = \max_{1 \leq i \leq m} \|u_i\|^2_2,$$

where $u_i$ is the $i$th row of matrix $U$. 

A PRECONDITIONER BASED ON NON-UNIFORM ROW SAMPLING FOR LINEAR LEAST SQUARES PROBLEMS 5
Obviously, the coherence number of a matrix is always between $n/m$ and 1. Matrices with small coherence numbers are called incoherent [1], for example, the Gaussian matrix $X$ which every element is an independent number generated by standard normal distribution. One example of semicoherent matrices is the one with large coherent number but only half rows have a large norm in the orthogonal factor, for example,

$$Y_{m \times n} = \begin{bmatrix} X_{(m-n/2) \times n/2} & 0 \\ 0 & I_{n/2} \end{bmatrix}$$

where $I_k$ is a square identity matrix of size $k \times k$. Coherent matrices are of large coherent number, for instance,

$$Z_{m \times n} = \begin{bmatrix} I_n \\ 0 \end{bmatrix}.$$

3.2. **Row Sampling.** We present a row sampling algorithm. Given a matrix $A_{m \times n}$ and a probability mass function $\{p_k, k = 1, 2, \cdots, m\}$, which will be called sampling density, randomly choose $s$ rows of $A$ via the given sampling density; see Algorithm 1.

**Input:** $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$, a probability mass function $\{p_k, k = 1, 2, \cdots, m\}$ and a sample size $s$.

**Output:** Sampled matrix $A_s \in \mathbb{R}^{s \times n}$

```markdown
for $t = 1 : s$
  Pick $i_t \in \{1, 2, \cdots, m\}$ with probability $\Pr\{i_t = k\} = p_k$ in identical and independent distributed (i.i.d.) trials.
end for
```

Let $S \in \mathbb{R}^{s \times m}$ with $S_{t,i_t} = 1/(sp_{i_t})^{1/2}$, then $A_s = SA$ is a sampling of $A$.

**Algorithm 1:** The row sampling algorithm introduced in [9].

Among various sampling densities, we chose the one proportional to the squared norm of each row:

$$p_k = \frac{\|a_k\|^2_2}{\|A\|^2_F}, k = 1, 2, \cdots, m.$$  

The naive uniform sampling $p_k = 1/m, k = 1, 2, \cdots, m$ fails when the coherence number of the matrix is large. For example, for the coherent matrix $Z$ defined in (7), we have to sample all $s$ rows from the first $n$ rows, whose probability $s!/m^s$ is very tiny, otherwise we will get a rank deficient matrix.

3.3. **Approximation property of the non-uniform sampling.** If we write the normal matrix as the summation of rank 1 matrices

$$A^T A = \sum_{i=1}^m a_i^T a_i,$$

then the approximation obtained by the random sampling is given by

$$A_s^T A_s = \frac{1}{s} \sum_{i=1}^s \frac{1}{p_k} a_{i_t}^T a_{i_t}.$$  

It is straightforward to verify that $A_s^T A_s$ is an unbiased estimator for $A^T A$, i.e.

$$\mathbb{E}[A_s^T A_s] = A^T A,$$  

Let independent random, self-adjoint matrices with dimension (Matrix Bernstein (Theorem 6.1 in [23])) Theorem 3.3 with high probability. To show this, we need the following concentration result.

\[ \exists \quad \| \sum_k \text{Var}(X_k) \| = \| \sum_k \mathbb{E}(X_k^2) \|. \]

Then for all \( t \geq 0 \)

\[ \Pr \left( \lambda_{\max}(\sum_k X_k) \geq t \right) \leq d \exp \left( -\frac{t^2/2}{\sigma^2 + Rt/3} \right). \]  

From the matrix Bernstein inequality, we can derive the following result which is slightly different with results in [18].

**Corollary 3.4 (Sum of Rank-1 Matrices).** Let \( y_1, y_2, \ldots, y_s \) be i.i.d. random column vectors in \( \mathbb{R}^n \) with

\[ \| y_k \| \leq M \quad \text{and} \quad \| \mathbb{E}[y_k y_k^\top] \| \leq \alpha^2, \]

for \( k = 1, 2, \ldots, s \). Then for any \( \epsilon \in [0, 1] \)

\[ \Pr \left( \left\| \frac{1}{s} \sum_{k=1}^s y_k y_k^\top - \mathbb{E}[y_k y_k^\top] \right\| \geq \epsilon \right) \leq 2n \exp \left( -\frac{3s\epsilon^2}{(6\alpha^2 + 2\epsilon)(M^2 + \alpha^2)} \right). \]

**Proof.** Let \( Y_k = y_k y_k^\top, A = \mathbb{E}[Y_k], \) and \( X_k = (Y_k - A)/s \) for \( k = 1, 2, \ldots, s \). Then \( \mathbb{E}[X_k] = 0 \). We bound the spectral norm of \( X_k \) as

\[ \lambda_{\max}(X_k) = \| X_k \| \leq \frac{1}{s} \left( \| Y_k \| + \| A \| \right) \leq \frac{M^2 + \alpha^2}{s}, \]

where we use the fact \( \| Y_k \| = \| y_k y_k^\top \| = \| y_k \|^2 \leq M^2 \). We then compute the variance

\[ \| \mathbb{E}[X_k^2] \| = \| \text{Var}(X_k) \| = \frac{1}{s^2} \| \text{Var}(Y_k) \| = \frac{1}{s^2} \| \mathbb{E}[Y_k^2] - A^2 \| \leq \frac{\alpha^2(M^2 + \alpha^2)}{s^2}. \]

Here we compute \( Y_k^2 = y_k y_k^\top y_k y_k^\top = \| y_k \|^2 Y_k \). Sum over \( k \) to get \( \sigma^2 \leq \alpha^2(M^2 + \alpha^2)/s \).

Plug the bound \( R \leq (M^2 + \alpha^2)/s, \sigma^2 \leq \alpha^2(M^2 + \alpha^2)/s \) into inequality (10) and rearrange the terms, we get the desired result.

\[ \square \]

We shall apply this concentration result to our row sampling scheme.

**Theorem 3.5.** Suppose \( A \) is a matrix of size \( m \times n \) with \( m > n \) and \( \| A \|_F = n \). Given \( \epsilon \in (0, 1), \delta \in (0, 1), \) let \( C = \frac{2}{\delta}(6\| A \|^2 + 2\epsilon)(1 - \log_{10}(\delta/2)) \) and \( s = C\epsilon^{-2}n \log n \). Let \( A_s \) be a sampled matrix obtained by Algorithm 1 with sampling density (8) and sample size \( s \). Then

\[ \| A_s^\top A - A^\top A \| \leq \epsilon \quad \text{with probability at least} \ 1 - \delta. \]
Remark 3.6. Constant $C$ used in Theorem 3.5 is not practical since the lower bound of $C$ is quit big. For example, when $\delta = \frac{1}{20}$, $\epsilon = \frac{1}{2} \| A \| \leq 1$ and $n = 300$, $C \epsilon^{-2}$ should be greater or equal than $\frac{56}{3}(1 + \log_{300}(40)) \approx 30.74$. In practice, however, $C \epsilon^{-2} = 4$ is good enough to get a reasonable sampling matrix. \qed

Corollary 3.7. With the same setting in Theorem 3.5, the following bound hold with high probability $1 - \delta$

$$\lambda_{\min}(A^T A) - \epsilon \leq \lambda_{\min}(A_s^T A_s) \leq \lambda_{\max}(A_s^T A_s) \leq \lambda_{\max}(A^T A) + \epsilon.$$  

Proof. By the triangle inequality, we immediately get

$$\|A_s^T A_s x\| \leq \|A^T A x\| + \epsilon \| x\| \leq (\lambda_{\max}(A^T A) + \epsilon) \| x\|,$$

which implies the desired inequality as $A_s^T A_s$ is symmetric. The lower bound of $\lambda_{\min}(A_s^T A_s)$ can be proved similarly. \qed

Notice that the spectrum bound obtained in Corollary 3.7 will not imply the bound of the preconditioned system $(A_s^T A_s)^{-1} A^T A$, which requires comparison of $(A_s^T A_s x, x)$ and $(A^T A x, x)$ for all $x \in \mathbb{R}^n$.

Next we shall establish some inequalities restricted to high frequency. We call $x \in \mathbb{R}^n$ is a high frequency of matrix $A^T A$ if the inequality

$$\lambda_{\max}(A^T A) \| x\| \leq C_h(A^T A x, x),$$

holds with a universal constant. Consider the decomposition of $x$ using the eigen-vector bases of $A^T A$. Inequality (13) implies $x$ is mainly expanded by eigen-vectors whose eigen-values are large. The constant $C_h$ in (13) is introduced to include not only the highest frequency (corresponding to the largest eigenvalue) but a range of frequencies comparable to the highest one.

**Corollary 3.8.** With the same setting in Theorem 3.5, the following bound hold with high probability $1 - \delta$: for all high frequency vectors $x$

$$
(1 - C_h \epsilon) (A^T A x, x) \leq (A^T A_s x, x) \leq (1 + C_h \epsilon) (A^T A x, x).
$$

By the triangle inequality and the definition of high frequency vectors, we will have

$$
(A^T A_s x, x) \leq (A^T A x, x) + \epsilon (x, x) \leq 1 + \frac{C_h \epsilon}{\lambda_{\max}(A^T A)} (A^T A x, x).
$$

Now we can bound $\lambda_{\max}(A^T A) \geq 1$ by

$$
n \lambda_{\max}(A^T A) \geq \sum_{i=1}^n \lambda_i(A^T A) = \sum_{i=1}^n \sigma_i^2(A) = \|A\|_F^2 = n.
$$

The lower bound can be proved similarly \hfill \Box

Corollary 3.8 implies $(A_s^T A_s)^{-1}$ is an effective smoother for $A^T A$. Since Gauss-Seidel iteration can smooth out the high frequency very quickly, we apply several symmetric Gauss-Seidel iterations instead of computing $(A_s^T A_s)^{-1}$ in practice.

To illustrate the approximation property of the sampled matrix, we plot the graph of $A^T A$ and $A^T A_s$ below. The matrix $A$ is of size $m \times n$, where $m = 9314$ and $n = 100$. The sampled matrix $A_s$ is of size $s \times n$ with $s = 1843$. The matrix $A$ is rescaled so that the diagonal of $A^T A$ is unit. The entries which have small absolute values less than a threshold $\theta = 0.125$ in the matrix $A^T A$ and $A^T A_s$ are filtered out and not shown in the graph. Each edge in the graph represents one entry in the matrix and the thickness of edge represent the magnitude respectively. From the figures, we find out that the two graphs are almost identical, which means the sampling strategy is able to capture the entries in the normal matrix with a large absolute value which is known as strong connectedness in algebraic multigrid methods [5] and related to the high frequency vectors.

4. **PCG with a Preconditioner based on Row Sampling**

In this section, we present our algorithm by constructing a fast, efficient and easy to implement randomized row sampling preconditioner and apply PCG to solve the normal equation.

4.1. **Algorithms.** We first normalize the matrix $A$ to make the column vectors have unit length, which enables all diagonal entries of $A^T A$ are one and $\|A\|_F^2 = n$. We then apply the row sampling to get a smaller matrix $A_s$ of size $s \times n$ by randomly choosing $s = O(n \log(n))$ rows of the normalized matrix $A_{m \times n}$ with sampling density $p_i = \|a_i\|_2^2/n$. We build our preconditioner by using a few steps of symmetric Gauss-Seidel (SGS) iteration to solve the approximate problem $A_s^T A_s e = r$. After all, we apply PCG to the normal equation $A^T A x = A^T b$ with this preconditioner.

For easy of understanding and completeness, the symmetric Gauss-Seidel method is presented below in Algorithm 3.
Figure 1. Graphs of matrices $A^\top A$ (left) and $A_s^\top A_s$ (right). The matrix $A$ is of size $m \times n$, where $m = 9314$ and $n = 100$. The sampled matrix $A_s$ is of size $s \times n$ with $s = 1843$ and $n = 100$. The matrix $A$ is rescaled so that the diagonal of $A^\top A$ is one. The entries which have small absolute values less than a threshold $\theta = 0.125$ in the matrix $A^\top A$ and $A_s^\top A_s$ is filtered out and not plot in the graph.

Input: $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$, convergence threshold $\epsilon \in (0, 1)$.

Output: approximated $\tilde{x}_{\text{opt}} \in \mathbb{R}^m$.

1. **Normalization**: $A \leftarrow AD^{-1}$, where $D_{jj} = \|a^j\|_2$, with $a^j$ being the $j$th column vector of $A$ for $1 \leq j \leq n$.

2. **Sampling**: Sample the row of $A$ to get $A_s$ with $s = 4n \log(n)$ by row sampling Algorithm 1 with probability (8).

3. **Preconditioner**: Construct preconditioner $e = Pr$ by solving $A_s^\top A_s e = r$ via several symmetric Gauss Seidel iterations; see Algorithm 3.

4. **PCG**: Use PCG to solve $A^\top Ax = A^\top b$ with the preconditioner constructed in Step 3. Stop when the relative residual is below $\epsilon$.

**Algorithm 2**: Randomized Sampling Preconditioned PCG

4.2. **Complexity**. We compute the complexity of PCG with the randomized sampling preconditioner for both dense matrices and sparse matrices. We use $s = 4n \log(n)$ as the default sample size. Here the factor 4 is chosen to balance the set up time and solver time. Similarly the number of SGS is set as 5 to balance the inner iteration of preconditioner and outer iteration of PCG.

For dense matrices, in the normalization step, we need $O(mn)$ to calculate the norm of each column $\|a^j\|_2$ and $O(mn)$ for the matrix multiplication $AD^{-1}$. Sampling costs $O(sn) = O(n^2 \log(n))$. The matrix multiplication $A_s^\top A_s$ costs $O(sn^2) = O(n^3 \log(n))$. The preconditioner in PCG, i.e. several symmetric Gauss Seidel iterations for $A_s^\top A_s e = r$.
Input: Sampled matrix $A_s \in \mathbb{R}^{s \times n}$, residual $r \in \mathbb{R}^n$, number of symmetric Gauss-Seidel iterations $t \in \mathbb{Z}^+$.  
Output: Correction $e \in \mathbb{R}^n$.  
for $i = 1 : t$ do  
\quad $e \leftarrow e + B^{-1}(r - A_s^T A_s e)$  
with $B$ the lower triangular part of $A_s^T A_s$.  
end for  
for $i = 1 : t$ do  
\quad $e \leftarrow e + (B^T)^{-1}(r - A_s^T A_s e)$  
with $B^T$ the upper triangular part of $A_s^T A_s$.  
end for  

Algorithm 3: The Preconditioner using Symmetric Gauss-Seidel Iterations.

|                  | Dense Matrix | Sparse Matrix |
|------------------|--------------|---------------|
| Normalization    | $O(mn)$      | $O(nnz(A))$   |
| Sampling         | $O(n^3 \log(n))$ | $O(nnz(A))$   |
| $A_s^T A_s$      | $O(n^2 \log(n))$ | $O(nnz(A_s))$ to $O(n \cdot nnz(A_s))$ |
| Preconditioner   | $O(n^2)$      | $O(nnz(A_s))$ |
| CG iteration     | $O(|\log(\epsilon)|\kappa(PA)mn)$ | $O(|\log(\epsilon)|\kappa(PA)nnz(A))$ |

needs $O(n^2)$. Finally, PCG iteration steps $k = O(|\log(\epsilon)|\kappa(PA))$ until reaching tolerance $\epsilon$ costs $O(kmn) = O(|\log(\epsilon)|\kappa(PA)mn)$. Note that since we only do matrix vector multiplication with $Ax$ and $A^T(Ax)$ instead of matrix product $A^T A$, the computation cost for each PCG step is only $O(mn)$ not $O(mn^2)$. Thus the total complexity is $O(|\log(\epsilon)|\kappa(PA)(mn + n^2)) + O(n^3 \log(n))$ when $A$ is dense. Complexity would be reduced significantly when the matrix is sparse. Let $nnz(M)$ be the number of nonzero elements of matrix $M$. In the normalization step, the cost is reduced to $O(nnz(A))$ for both the column calculation and matrix multiplication $AD^{-1}$. In the sampling step, depending on the sparse pattern, the complexity is reduced to at most $O(nnz(A))$. The matrix product of $A_s^T A_s$ costs between $O(nnz(A_s))$ and $O(n \cdot nnz(A_s))$. The preconditioner costs $O(nnz(A_s^T A_s))$. And $k = O(|\log(\epsilon)|\kappa(PA))$ PCG iterations needed. The total complexity is $O(|\log(\epsilon)|\kappa(PA)(nnz(A) + nnz(A_s^T A_s)) + O(nnz(A_s)))$ for sparse matrices, where $\alpha \in [1, n]$ depends on the sparse pattern of $A_s$. For sparse matrix $A$ with $nnz(A) \ll mn$, the proposed solver is thus more efficient.

Theoretically we cannot find a uniform control of the condition number of the preconditioned matrix $PA$ but we shall show in the next section that numerically our preconditioner is effective.

5. Numerical Results

We shall compare PCG with our randomized sampling preconditioner, denoted by suffix RS, with CG for the normalized matrix, denoted by suffix CG, which is equivalent to use PCG for the original matrix with a diagonal preconditioner. The column with prefix ‘Setup’ in tables is the CPU time for preprocess including sampling and normalization for RS and
only normalization for CG. The column ‘Time’ is the CPU time for iterative methods. Thus
the sum of these two are the CPU time for the whole procedure. The column ‘$\kappa(A^T A)$’
lists the condition number of $A^T A$ and $\mu(A)$ is the coherence number with normalized $A$.
We list the coherence number here to emphasize the weighted row sampling works well
and robust to the coherence number. It is shown in [1] that the uniform sampling fails
when the coherence number of the matrix is large. In our sampling algorithm, we use the
sampling density proportional to the squared norm of each row. Tolerance for PCG or CG
is set $10^{-7}$. Notice that iterative methods may end without reaching the tolerance.

When varying the size of the matrix, we report one realization of the sampling algo-
rithm. As the sampling contains randomness, for each example, we shall also pick up one
typical matrix and run our solver 10 times and compute the mean and standard derivation.

We tested several classes of matrices – including well conditioned matrices, ill condi-
tioned matrices, incoherent matrices and coherent matrices; see Table 2.

5.1. Gaussian Matrix. The Gaussian matrix is constructed by MATLAB command $A = \ \text{randn}(m, n)$
with each entry of $A$ being generated independent and identically by a
standard normal random variable. The matrix $A^T A$ has a small condition number followed
by Bai and Yin [2]. By Theorem 2 in [2], the limit of condition number of $A^T A$ can be

calculated as

$$
\kappa(A^T A) = \frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)} \to \frac{m(1 + \sqrt{n/m})^2}{m(1 - \sqrt{n/m})^2} = \frac{\sqrt{m + \sqrt{n}}}{\sqrt{m - \sqrt{n}}},
$$

When $m = n^2$, $\kappa(A^T A) \approx (\sqrt{n} + 1)^2/(\sqrt{n} - 1)^2 \approx 1 + O(1/\sqrt{n})$ almost surely.

Since each element is generated independent and identically, the $Q$ factor of $A$’s QR
decomposition has evenly distributed magnitude in each row. Thus the coherence number
$\mu(A)$ of Gaussian matrix is also small. In summary the Gaussian matrix belongs to the
category– ‘well conditioned and incoherent matrices’. The sampling density (8) is almost
uniform.

| $\kappa(A^T A)$ | $\mu(A)$ | Residual.CG | Iter.CG | Residual.RS | Iter.RS |
|-----------------|----------|-------------|--------|-------------|--------|
| $m(1 + \sqrt{n/m})^2$ | $m(1 - \sqrt{n/m})^2$ | $\sqrt{m + \sqrt{n}}$ | $\sqrt{m - \sqrt{n}}$ | | | |
| 109 | 3000 | 11881 | 8.39 | 0.05 | 5.03e-08 | 10 | 7.00e-08 | 11 |
| 141 | 5000 | 19881 | 8.30 | 0.01 | 8.60e-08 | 9 | 5.19e-08 | 11 |
| 200 | 10000 | 40000 | 7.70 | 0.02 | 1.70e-08 | 9 | 4.42e-08 | 11 |
| 282 | 20000 | 79524 | 7.22 | 0.05 | 4.03e-08 | 8 | 2.70e-08 | 11 |
| 400 | 40000 | 160000 | 7.40 | 0.09 | 9.62e-08 | 7 | 2.64e-08 | 11 |

5.2. ‘Semi Gaussian’ Matrix. The ‘semi Gaussian’ matrix used in [1] has the following
block structure. The left upper block $B$ is a Gaussian matrix of size $(m - n/2) \times n/2$ and

| $\kappa(A^T A)$ | $\mu(A)$ | Residual.CG | Iter.CG | Residual.RS | Iter.RS |
|-----------------|----------|-------------|--------|-------------|--------|
| $(\sqrt{m} + 1)^2/((\sqrt{m} - 1)^2) \approx 1 + O(1/\sqrt{n})$ | | | | | | |
the right lower block $I_{n/2}$ is an identity matrix of size $n/2 \times n/2$.

$$A_{m \times n} = \begin{bmatrix} B & 0 \\ 0 & I_{n/2} \end{bmatrix}.$$  

It belongs to the category—'well conditioned and coherent matrices'. For such 'semi Gaussian' matrices, the coherence number $\mu(A) = 1$ due to the presentness of an identity sub-matrix. It is shown in [1] that the uniform sampling fails for this example while our non-uniform sampling works well. To use random transformations, a small perturbation $10^{-8}$ is added to every entry of $A$ to change it to a dense matrix.

They are also well conditioned since

$$A^\top A = \begin{bmatrix} B^\top & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} B & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} B^\top B & 0 \\ 0 & I \end{bmatrix},$$

and when $\lambda_{\text{max}}(B^\top B) \geq 1$

$$\kappa(A^\top A) \leq \kappa(B^\top B).$$

The Gaussian matrix $B$ is well conditioned by the analysis in §5.1. So is $A$.

### Table 6. ‘Semi Gaussian’ Matrix: Residual and Iteration Steps

| $n$ | $m$ | nnz($A$) | $\kappa(A^\top A)$ | $\mu(A)$ | Residual.CG | Iter.CG | Residual.RS | Iter.RS |
|-----|-----|----------|---------------------|---------|-------------|--------|-------------|--------|
| 62  | 1000| 992      | 1.99e+03            | 1       | 2.25e-08    | 9      | 6.85e-08    | 11     |
| 108 | 3000| 2970     | 5.82e+03            | 1       | 6.87e-08    | 8      | 3.70e-08    | 12     |
| 140 | 5000| 4970     | 9.80e+03            | 1       | 2.43e-08    | 8      | 6.09e-08    | 11     |
| 200 | 10000| 10100   | 1.95e+04            | 1       | 6.48e-08    | 7      | 6.79e-08    | 11     |
| 282 | 20000| 20022   | 3.84e+04            | 1       | 2.28e-08    | 7      | 7.37e-08    | 11     |
Table 7. ‘Semi Gaussian’ Matrix: Elapsed CPU Time

| n  | m  | Time.CG | Setup.CG | Sum.CG | Time.RS | Setup.RS | Sum.RS |
|----|----|---------|----------|--------|---------|----------|--------|
| 62 | 1000 | 7.03e-04 | 3.96e-04 | 1.10e-03 | 1.78e-03 | 2.16e-03 | 3.94e-03 |
| 108 | 3000 | 2.31e-03 | 2.95e-03 | 5.26e-03 | 4.03e-03 | 7.25e-03 | 1.13e-02 |
| 140 | 5000 | 5.14e-03 | 6.31e-03 | 1.14e-02 | 9.95e-03 | 1.49e-02 | 2.49e-02 |
| 200 | 10000 | 1.54e-02 | 1.77e-02 | 3.31e-02 | 2.47e-02 | 3.84e-02 | 6.31e-02 |
| 282 | 20000 | 3.19e-02 | 4.28e-02 | 7.47e-02 | 4.36e-02 | 7.83e-02 | 1.22e-02 |

Table 8. ‘Semi Gaussian’ Matrix: Mean and Sample Standard Deviation

| n  | m  | Iter.Mean | Iter.Std | Time.Mean | Time.Std | Setup.Mean | Setup.Std |
|----|----|-----------|----------|-----------|----------|------------|-----------|
| 62 | 1000 | 11.7      | 0.68     | 1.18e-03  | 7.75e-05 | 1.35e-03   | 4.27e-04  |
| 108 | 3000 | 11.7      | 0.48     | 2.65e-03  | 1.30e-04 | 5.31e-03   | 3.12e-03  |
| 140 | 5000 | 11.9      | 0.57     | 6.36e-03  | 8.77e-04 | 1.14e-02   | 4.87e-03  |
| 200 | 10000 | 11.7     | 0.67     | 1.51e-02  | 2.04e-03 | 2.69e-02   | 5.00e-03  |
| 282 | 20000 | 11.3     | 0.48     | 4.51e-02  | 6.93e-03 | 7.69e-02   | 8.21e-03  |

5.3. ‘Sprand’ Matrix. The ‘sprand’ (sparse random) matrix is generated by Matlab function \( A = \text{sprand}(m, n, s, 1/c) \), where \( m \) is the number of rows, \( n \) is the number of columns, \( s \) is the sparsity and \( c \) is the estimated condition number. We can control the condition number by the input \( c \). When \( c \) is large, the generated matrix is ill conditioned. The coherence number is still small due to the randomness. Thus it belongs to the category ‘ill conditioned and incoherent matrices’.

To test the robustness to the condition number, we fix \( m = 90000, n = 300 \) and the sparsity \( s = 0.25 \) and change \( c \) to get several matrices with large condition number; see Table 9-10.

Table 9. ‘Sprand’ Matrix \( m = 90000, n = 300 \): Residual and Iteration Steps.

| \( mnz(A) \) | \( \kappa(A^T A) \) | \( \mu(A) \) | Residual.CG | Iter.CG | Residual.RS | Iter.RS |
|-------------|----------------|-------------|-------------|--------|-------------|--------|
| 87248       | 3.87e+03       | 7.17e-03    | 9.05e-08    | 98     | 9.97e-08    | 21     |
| 87208       | 1.91e+04       | 5.86e-03    | 8.70e-08    | 181    | 6.80e-08    | 38     |
| 86278       | 7.55e+04       | 3.91e-03    | 8.36e-08    | 264    | 7.94e-08    | 58     |
| 86654       | 2.89e+05       | 7.81e-03    | 9.17e-08    | 233    | 9.37e-08    | 50     |
| 86816       | 7.40e+05       | 7.71e-03    | 8.18e-07    | 296    | 4.77e-08    | 70     |

Table 10. ‘Sprand’ Matrix \( m = 90000, n = 300 \): Elapsed CPU Time

| \( mnz(A) \) | Time.CG | Setup.CG | Sum.CG | Time.RS | Setup.RS | Sum.RS |
|-------------|---------|----------|--------|---------|----------|--------|
| 87248       | 1.66    | 0.35     | 2.01   | 0.52    | 0.68     | 1.20   |
| 87208       | 3.23    | 0.43     | 3.65   | 0.88    | 0.65     | 1.53   |
| 86278       | 4.39    | 0.33     | 4.72   | 1.29    | 0.57     | 1.86   |
| 86654       | 3.92    | 0.35     | 4.27   | 1.13    | 0.56     | 1.69   |
| 86816       | 5.00    | 0.33     | 5.33   | 1.56    | 0.54     | 2.10   |
Notice that for very ill-conditioned matrices, CG without preconditioners will not reach the tolerance $10^{-7}$; see row 5 in Table 9. Although theoretically CG will result in the exact solution within at most $n$-steps, the large condition number causes the instability. Our preconditioner is effective and PCG converges within 100 steps.

We fix $m = 40,000, c = 100, s = 0.25$ and vary $n$ in Table 12. Again our preconditioned PCG works well.

We then fix $n = 200$ and $c = 100$ and vary $m$ in Table 13. The iteration steps are almost uniform to $m$. Notice that for fixed $n = 200$, the sample size $s = 4n \log n \approx 4239$ which is a small portion for large $m$.

5.4. UDV Matrix. The UDV matrices are random matrices generated by $A = UDV$, where $U$ is an $m \times n$ random orthonormal matrix, $V$ is an $n \times n$ random orthonormal matrix and $D = \text{diag}[1, 1 + (c - 1)/n, \ldots, c]$ and $c$ is the estimated condition number. For this kind of matrices, we can control the condition number by parameter $c$. When $c$ is large enough, it belongs to the category ‘ill conditioned and incoherent matrices’. Again CG fails to converge for the last three matrices in ‘UDV’ group when the condition number is large while PCG with our proposed preconditioner works well.
Table 14. ‘UDV’ Matrix \( m = 90000, n = 300, \text{nnz}(A) = 90000\):
Residual and Iteration Steps

| \( \kappa(A^\top A) \) | \( \mu(A) \) | Residual.CG | Iter.CG | Residual.RS | Iter.RS |
|-----------------|-------|-------------|-------|-------------|-------|
| 5936            | 4.81e-03 | 9.68e-08   | 116   | 7.21e-08   | 24    |
| 18853           | 4.61e-03 | 9.68e-08   | 202   | 8.81e-08   | 38    |
| 1.44e+05        | 4.66e-03 | 4.42e-07   | 294   | 8.44e-08   | 68    |
| 4.75e+05        | 4.65e-03 | 1.32e-05   | 369   | 7.92e-08   | 86    |
| 1.07e+06        | 4.72e-03 | 9.89e-06   | 253   | 4.44e-08   | 91    |

Table 15. ‘UDV’ Matrix \( m = 90000, n = 300, \text{nnz}(A) = 90000\):
Elapsed CPU Time

| \( \kappa(A^\top A) \) | \( \mu(A) \) | Time.CG | Setup.CG | Sum.CG | Time.RS | Setup.RS | Sum.RS |
|-----------------|-------|--------|----------|--------|--------|----------|--------|
| 5936            | 4.81e-03 | 2.44   | 0.42     | 2.86   | 0.49   | 0.47     | 0.96   |
| 18853           | 4.61e-03 | 3.77   | 0.43     | 4.20   | 0.71   | 0.45     | 1.16   |
| 1.44e+05        | 4.66e-03 | 5.47   | 0.42     | 5.89   | 1.30   | 0.45     | 1.75   |
| 4.75e+05        | 4.65e-03 | 5.88   | 0.44     | 6.32   | 1.57   | 0.45     | 2.02   |
| 1.07e+06        | 4.72e-03 | 5.53   | 0.43     | 5.96   | 1.65   | 0.44     | 2.09   |

Table 16. ‘UDV’ Matrix \( m = 90000, n = 300, \text{nnz}(A) = 90000\):
Mean and Sample Standard Deviation

| \( \kappa(A^\top A) \) | Iter.Mean | Iter.Std | Time.Mean | Time.Std | Setup.Mean | Setup.Std |
|-----------------|---------|--------|-----------|----------|------------|-----------|
| 5936            | 23.1    | 0.57   | 0.51      | 7.8e-02  | 0.45       | 3.79e-02  |
| 18853           | 38.8    | 0.63   | 0.76      | 7.51e-02 | 0.43       | 6.91e-03  |
| 1.44e+05        | 72      | 0.82   | 1.42      | 1.67e-01 | 0.48       | 9.76e-02  |
| 4.75e+05        | 86.4    | 0.52   | 1.86      | 2.37e-01 | 0.47       | 5.84e-02  |
| 1.07e+06        | 90.2    | 0.42   | 1.81      | 1.98e-01 | 0.48       | 5.91e-02  |

5.5. Random Graph Laplacian Matrix. In this section, we shall show our least squares solver can be applied to solve the random graph Laplacian problem which has important application in spectral clustering, text mining and web applications etc [7].

A simple graph is an undirected graph without multiple edges or loops. It can be described by a set \( V = \{v_1, \cdots, v_n\} \) of vertices and a set of edges \( E = \{(v_i, v_j)\} \). For a vertex \( v_i \), \( \deg(v_i) \) counts the number of edges attached to it. The degree matrix \( D \) of a graph is a diagonal matrix consisting of degree of each vertex, i.e. \( D_{ii} = \deg(v_i) \) for \( i = 1, 2, \ldots, n \). The adjacency matrix \( A \) of a graph is defined as

\[
A(i, j) = \begin{cases} 
1 & \text{if } (v_i, v_j) \in E; \\
0 & \text{otherwise.}
\end{cases}
\]

Given a simple graph \( G \) with \( n \) vertices, its graph Laplacian matrix \( L_{n \times n} \) is defined as

\[
L := D - A.
\]

Vertices with zero degree are isolated, i.e. not connected to other vertices. For a simple graph without isolated vertex, the normalized graph Laplacian matrix is defined by normalizing the diagonal of \( L \)

\[
L_N := I - D^{-1/2} AD^{-1/2}.
\]
Therefore the elements of $L^N$ are given by

$$L_N(i, j) = \begin{cases} 
1 & \text{if } i = j; \\
-\frac{1}{\sqrt{\deg(v_i) \deg(v_j)}} & \text{if } i \neq j \text{ and } (v_i, v_j) \in E; \\
0 & \text{otherwise}.
\end{cases}$$

We shall focus on solving the normalized graph Laplacian matrix.

We now reformulate the graph Laplacian matrix as the normal matrix of a least square problem. Given a simple graph $G$, we shall choose a sequence $\beta$ such that $B^T B$ is well conditioned and relatively easy to solve by CG. Let $m$ be the number of edges and recall that $n$ is the number of vertices. Matrix $B$ is of size $m \times n$ and each row represents a directed edge. Given any edge $e = (u, v)$, we let $s(e) = u$ be the source of $e$ and $t(e) = v$ be the target of $e$, which means the edge $e$ is leaving vertex $u$ and entering vertex $v$, respectively. Let

$$B(i, j) = \begin{cases} 
1 & \text{if } s(e_i) = v_j; \\
-1 & \text{if } t(e_i) = v_j; \\
0 & \text{otherwise.}
\end{cases}$$

It is easy to verify the relation between degree matrix, adjacency matrix and incidence matrix is

$$B^T B = D - A.$$ 

And the matrix product $B^T B$ is independent of orientation of each edge. Notice that $B$ is sparse with $n \times n$. The $eta$th power of the incidence matrix is $B^\beta$. As the sum of each row of $B$ is 0, the constant vector is in the null space of $B$, i.e. $Bc = 0$, which implies $L_N$ is singular. Furthermore the multiplicity of zero eigenvalue is the number of connected component $G$.

### 5.5.1. Random Graph Model

We shall use the random graph model presented in [6]. The spectra of adjacency matrix $A$ follows the power law which enables us to construct an ill-conditioned random graph Laplacian matrix. The construction of a random graph follows several steps. Let $(w_1, w_2, \ldots, w_n)$ be a sequence of positive numbers which denotes the expected degree of $i$-th node. Introduce two parameters $d, m$ for the average degree and maximum degree respectively.

Following [6], we shall chose a sequence $w = (w_1, \cdots, w_n)$ following a power law:

$$w_i = ci^{-(\beta - 1)}$$

where $c$ is a number determined by the average degree $d$ and $i_0$ depends on the maximum degree $m$, i.e.

$$c = \frac{\beta - 2}{\beta - 1}dn^{-1/(\beta - 1)}, \quad i_0 = n \left[ \frac{d(\beta - 2)}{m(\beta - 1)} \right]^{\beta - 1}.$$

Many massive graphs are power-law graphs with $2 \leq \beta \leq 3$. For some specific graphs like Internet graph [11], $2.1 < \beta < 2.4$. Hollywood graph [10] has exponent $\beta \approx 2.3$. The corresponding normalized graph Laplacian matrix, is well conditioned and relatively easy to solve by CG.

By our numerical experiment, the larger $\beta$ is, the larger probability we can get an ill-conditioned graph Laplacian matrix. In our test, we fix $i_0 = 11$, choose $\beta = 5, d = 30$ for the first incidence matrix $B_1$, and $\beta = 8, d = 5n$ for the second incidence matrix $B_2$, where $n$ is the number for vertices. $B_1$ is the incidence matrix related to a random graph with less edges while it is ill-conditioned. $B_2$ corresponds to a dense random graph which
is usually well-conditioned. We glue two graphs together to get an ill-conditioned graph Laplacian with $O(n^2)$ edges.

1. Construct a random graph $G$ with edge connected vertices $v_i$ and $v_j$ by probability $P(i,j)$, where

$$P(i,j) = w_i w_j \rho,$$

with $\rho = 1/\sum_{i=1}^{n} w_i$.

Then the upper triangular part ($i \leq j$) of adjacency matrix is a random Bernoulli matrix

$$A_u(i,j) = \begin{cases} 1 & \text{with probability } P(i,j); \\ 0 & \text{otherwise}. \end{cases}$$

Since the adjacency matrix for an undirected graph is symmetric, we set $A = A_u^T + A_u$.

2. Extract positions of all nonzero entries from matrix $A^T A$ by $\{i,j,s\} = \text{find}(A' * A)$ and the 3 vectors $i, j, s$ are the coordinate formate to represent a sparse matrix, i.e. $i(k) = i, j(k) = j, \text{and } s(k) = (A^T A)_{ij}$. The incidence matrix $B_1$ can be constructed by edge index $(i,j)$ and distribute $1$ for first vertex $v_i$ and $-1$ for second vertex $v_j$.

3. Construct another incidence matrix $B_2$ using the same procedure but with different parameters $\beta = 8, d = 5n$. The underlying graph of $B_2$ will have more edges while the corresponding Laplacian matrix is well conditioned.

4. Combine the two incidence matrices $B_1$ and $B_2$ with 5-overlap in vertices, i.e., construct a combination matrix $B_{\text{com}}$ by $B_{\text{com}}(1:m1, 1:n) = B1; B_{\text{com}}(m1+1:m2, \ n-4:2n-5) = B2$;

where recall that $B_1$ of size $m_1 \times n, B_2$ of size $m_2 \times n$.

5. Filter out the isolated vertices in graph corresponding to matrix $B_{\text{com}}$. The final incidence matrix $B = B_{\text{new}}$.

Our algorithm will sample $B$ to get a graph with much fewer edges. A random graph and its sampling are shown in Fig. 2.

### Table 17. Random Graph Laplacian Matrix: Residual and Iteration Steps

| $n$  | $m$    | $s$    | $\text{nnz}(B^T B)$ | $\kappa(B^T B)$ | Residual.CG | Iter.CG | Residual.RS | Iter.RS |
|------|--------|--------|----------------------|-----------------|-------------|---------|-------------|---------|
| 187  | 5078   | 3913   | 10343 | 129.22 | 7.44e-08 | 40 | 5.69e-08 | 17 |
| 355  | 19475  | 8339   | 39305 | 278.54 | 7.97e-08 | 67 | 9.61e-08 | 25 |
| 536  | 41032  | 13474  | 82600 | 467.04 | 8.09e-08 | 60 | 8.55e-08 | 18 |
| 709  | 68969  | 18616  | 138647 | 951.21 | 9.37e-08 | 93 | 9.76e-08 | 25 |
| 856  | 101829 | 23120  | 204514 | 3145.4 | 9.85e-08 | 112 | 6.03e-08 | 28 |

#### 5.5.2. Numerical Results of Graph Laplacian Matrices. In Table 17, we list the iteration steps of CG and PCG with RS precondionter. As $B^T B$ is singular, the condition number in the third column is the so-called effective condition number which is the quotient of the largest eigenvalue over the smallest nonzero eigenvalue of $B^T B$. Our least square solver is more robust compared to CG. The iterative steps of our method is less than $1/3$ compared to CG and the elapsed CPU time is comparable to CG, see Table 18.
Figure 2. Graphs of matrices $B^T B$ (left) and $B_s^T B_s$ (right). The matrix $B$ is of size $m \times n$, where $m = 68969$ and $n = 709$. The sampled matrix $B_s$ is of size $s \times n$ with $s = 18616$ and $n = 709$. The matrix $B$ is rescaled so that the diagonal of $B^T B$ is one.

Table 18. Random Graph Laplacian Matrix: Elapsed CPU Time

| $n$ | $m$  | $s$  | Time.CG | Setup.CG | Sum.CG | Time.RS | Setup.RS | Sum.RS |
|-----|------|------|---------|----------|--------|---------|----------|--------|
| 187 | 5078 | 3913 | 1.25e-02| 3.97e-03 | 1.64e-02| 1.82e-02| 1.62e-02| 3.44e-02|
| 355 | 19475| 8339 | 1.10e-02| 5.73e-03 | 1.67e-02| 1.17e-02| 1.74e-02| 2.91e-02|
| 536 | 41032| 13474| 1.55e-02| 2.78e-03 | 1.82e-02| 1.09e-02| 1.31e-02| 2.40e-02|
| 709 | 68969| 18616| 4.03e-02| 3.17e-03 | 4.35e-02| 2.31e-02| 1.98e-02| 4.29e-02|
| 856 | 101829| 23120| 8.68e-02| 7.98e-03 | 9.48e-02| 3.90e-02| 3.69e-02| 7.57e-02|

Table 19. Random Graph Laplacian Matrix: Mean and Sample Standard Deviation

| $m$  | $n$  | $\kappa(B^T B)$ | Iter.Mean | Iter.Std | Time.Mean | Time.Std | Setup.Mean | Setup.Std |
|------|------|-----------------|-----------|----------|-----------|----------|------------|-----------|
| 5078 | 187  | 1.70e+03        | 16.7      | 1.06     | 3.50e-03  | 7.88e-04 | 2.85e-03  | 3.27e-04  |
| 19475| 355  | 1.21e+04        | 21.9      | 1.79     | 9.00e-03  | 3.15e-03 | 7.01e-03  | 8.10e-04  |
| 41032| 536  | 1.85e+04        | 21        | 1.33     | 1.15e-02  | 6.91e-04 | 1.01e-02  | 1.01e-03  |
| 68969| 709  | 5.70e+04        | 26.8      | 2.86     | 2.47e-02  | 4.11e-03 | 2.00e-02  | 1.87e-03  |
| 101829| 856 | 2.75e+05        | 30.2      | 1.69     | 3.72e-02  | 2.94e-03 | 2.69e-02  | 1.72e-03  |

5.6. Summary of Numerical Results. It is well known that CG converges fast for well conditioned matrices but fails for ill-conditioned matrices. Although theoretically CG should converge in at most $n$ steps, for ill-conditioned matrices, round off errors are amplified in the orthogonalization procedure which make CG fail to converge within $n$ steps; see Tables 9 and 14. With our random sampling preconditioner, PCG converges for all of
them. For ill-conditioned matrices, the iterative steps are reduced significantly and elapsed CPU time are reduced to $1/3$ in both the `sprand` group and UDV group. We also apply our least squares solver to solve the random graph Laplacian equation and show it reduces the iteration steps. Although we cannot prove the uniform convergence, the performances listed ahead indicates that random sampling preconditioner is efficient and effective.

Although the algorithm contains randomness, the standard deviation of iterations and CPU time are acceptable. In most of examples, the ratio of standard deviation to the mean of iterative steps is within $2\%$. Only for the ‘sprand’ cases, the ratio ranges from $2\%$ to $5\%$.

6. Conclusion

In this paper, we construct a randomized row sampling method which aims to solve the least squares problems when matrix $A$ is ill conditioned, sparse, highly overdetermined matrix $m \gg n$. By row sampling with probability proportional to the squared norm of rows, we can get a sampled matrix $A_s$ with size $O(n \log n) \times n$ which can capture the high frequency of the norm matrix. Then the preconditioner is constructed by applying the symmetric Gauss Seidel iteration to the sampled normal matrix $A_s^\top A_s$. The last but not least is that our preconditioner is very easy to implement compare with the preconditioners based on random transformations.

References

[1] H. Avron, P. Maymounkov, and S. Toledo. Blendenpik: Supercharging LAPACK’s least-squares solver. SIAM Journal on Scientific Computing, 32(3):1217–1236, 2010. 1, 3, 4, 5, 6, 12, 13
[2] Z. Bai and Y. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. Advances In Statistics, pages 108–127, 2008. 12
[3] M. R. M. Biemond J, Lagendijk R L. Iterative methods for image deblurring. Proceedings of the IEEE, 78(5):856–883, 1990. 1
[4] A. Bjorck. Numerical methods for least squares problems. Siam, 1996. 3
[5] A. Brandt, S. McCormick, and J. Ruge. Algebraic multigrid (amg) for automatic multigrid solutions with application to geodetic computations. Report, Inst. for computational Studies, Fort collins, colo., 1982. 9
[6] F. Chung, L. Lu, and V. Vu. Spectra of random graphs with given expected degrees. Proceedings of the National Academy of Sciences, 100(11):6313–6318, May 2003. 17
[7] F. R. Chung. Spectral graph theory (cbms regional conference series in mathematics, no. 92). 1996. 16
[8] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. arXiv preprint arXiv:0710.1435, 2007. 1, 4
[9] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. Numerische mathematik, 117(2):219–249, 2011. 3, 4, 6
[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication - SIGCOMM ’99, 1999. 17
[11] Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. Combinatorica, Sep 1981. 17
[12] G. H. Golub and C. F. V. Loan. Matrix computations. JHU Press, 2012. 1, 2
[13] S. Kaczmarz. Angenaherte auflösung von systemen linearer gleichungen. Classe des Sciences Mathématiques et Naturelles. Série A. Sciences Mathématiques, 35(355-357), 1937. 1
[14] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. SIAM review, 51(3):455–500, 2009. 1
[15] M. W. Mahoney. Lecture notes on randomized linear algebra. arXiv preprint arXiv:1608.04481, 2016. 7
[16] D. Needell. Randomized Kaczmarz solver for noisy linear systems. BIT Numerical Mathematics, 50(2):395–403, 2010. 4
[17] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. Applied linear statistical models, volume 4. Irwin Chicago, 1996. 1
[18] R. Oliveira. Sums of random hermitian matrices and an inequality by rudelson. Electronic Communications in Probability, 15:203–212, 2010. 7
[19] P. Oswald and W. Zhou. Convergence analysis for Kaczmarz-type methods in a Hilbert space framework. *Linear Algebra and its Applications*, 478:131–161, 2015.

[20] C. C. Paige and M. A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM transactions on mathematical software*, 8(1):43–71, 1982.

[21] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008.

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

[23] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[24] C. Wang, A. Agaskar, and Y. M. Lu. Randomized Kaczmarz algorithm for inconsistent linear systems: An exact MSE analysis. In *Sampling Theory and Applications (SampTA), 2015 International Conference on*, pages 498–502. IEEE, 2015.

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