RC25187 (W1107-042) July 13, 2011
Computer Science

IBM Research Report

Effective Stiffness: Generalizing Effective Resistance Sampling to Finite Element Matrices

Haim Avron
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Sivan Toledo
Tel Aviv University

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

EFFECTIVE STIFFNESS: GENERALIZING EFFECTIVE RESISTANCE SAMPLING TO FINITE ELEMENT MATRICES

Haim Avron
IBM Research
haimav@us.ibm.com

Sivan Toledo
Tel Aviv University
stoledo@tau.ac.il

July 12, 2011

Abstract. We define the notion of *effective stiffness* and show that it can used to design algorithms for solving linear systems arising from finite-element discretizations of PDEs. In particular, we show that sampling $O(n \log n)$ elements according to probabilities derived from effective stiffnesses yields an high quality preconditioner that can be used to solve the linear system efficiently. Effective stiffness generalizes the notion of effective resistance, a key ingredient of recent progress in developing nearly linear symmetric diagonally dominant (SDD) linear solvers. Solving finite elements problems is of considerably more interest than the solution of SDD linear systems, since the finite element method is frequently used to numerically solve PDEs arising in scientific and engineering applications. Unlike SDD systems, which are relatively easy to precondition, there has been limited success in designing fast solvers for finite element systems. Contrary to previous attempts, which usually target discretization of limited class of PDEs like scalar elliptic or 2D trusses, our method targets a very wide range of finite element discretizations, utilizing only some basic algebraic-combinatorial properties of the matrices arising from such discretizations.

1. Introduction

We explore the sparsification of finite element matrices using *effective stiffness sampling*. The goal of the sparsification is to reduce the number of elements in the matrix so that it can be easily factored and used as a preconditioner for an iterative linear solver. We show that sampling non-uniformly $O(n \log n)$ elements produces a matrix that is with high probability spectrally close to the original matrix, and therefore an excellent preconditioner. The sampling probability of an element is given by the largest generalized eigenvalue of the element matrix and the effective stiffness matrix of the element.

Effective stiffness generalizes the notion of effective resistance, a key ingredient in much of the recent progress in nearly optimal symmetric diagonally dominant (SDD) linear solvers [8, 2, 9]. Solving finite elements problems is of considerably more interest than the solution of SDD linear systems, since the finite element method is frequently used to numerically solve PDEs arising in scientific and engineering applications.

Unlike SDD systems, which are relatively easy to precondition, there has been limited success in designing fast solvers for finite element systems. Efforts to generalize combinatorial preconditioners to matrices that are not weighted Laplacians followed several paths, and started long before recent progresses. Gremban showed how to transform a linear system whose coefficient matrix is a signed Laplacian to a linear system of twice the size whose matrix is a weighted Laplacian. The coefficient matrix is a 2-by-2 block matrix with diagonal blocks with the same sparsity pattern as the original matrix A and with identity off-diagonal blocks. A different approach is to extend Vaidya's construction to signed graphs [3]. The class of symmetric matrices with a symmetric factorization $A = UU^T$ where columns of U have at most 2 nonzeros contains not only signed graphs, but also gain graphs, which are not diagonally dominant [4]; it turns out that these matrices can be scaled to diagonal dominance, which allows graph preconditioners to be applied to them [7].

The matrices that arise in finite-element discretization of elliptic partial differential equations (PDEs) are positive semi-definite, but in general they are not diagonally dominant. However, when the PDE is scalar (e.g., describes a problem in electrostatics), the matrices can sometimes be approximated by diagonally dominant matrices. In this scheme, the coefficient matrix A is first approximated by a diagonally-dominant matrix D, and then G_D is used to construct the graph G_B of the preconditioner B. For large matrices of this class, the first step is expensive, but because finite-element matrices have a natural representation as a sum of very sparse matrices, the diagonally-dominant approximation can be constructed for each term in the sum separately. There are at least three ways to construct these approximations: during the finite-element discretization process [5], algebraically [1], and geometrically [17]. A slightly modified construction that can accommodate terms that do not have a close diagonally-dominant approximation works well in practice [1].

Another approach for constructing combinatorial preconditioners to finite element problems is to rely on a graph that describes the relations between neighboring elements. This graph is the dual of the finite-element mesh; elements in the mesh are the vertices of the graph. Once the graph is constructed, it can be sparsified much like subset preconditioners. This approach, which is applicable to vector problems like linear elasticity, was proposed in [13]; this paper also showed how to construct the dual graph algebraically and how to construct the finite-element problem that corresponds to the sparsified dual graph. The first effective preconditioner of this class was proposed in [6]. It is not yet known how to weigh the edges of the dual graph effectively, which limits the applicability of this method. However, in applications where there is no need to weigh the edges, the method is effective [14].

Unlike previous efforts, which usually target discretization of limited class of PDEs like scalar elliptic or 2D trusses, our method targets a very wide range of finite element discretizations, utilizing only some basic algebraic-combinatorial properties of the matrices arising from such discretizations.

2. Preliminaries

2.1. **Sums of Random Matrices.** Approximating a matrix using random sampling can be viewed as a particular case of sums of random matrices. In the last few years there has been significant literature on showing concentration bounds on such sums [12, 11, 16]. We use the following bound from [10].

Theorem 2.1. [10, Theorem 1.1] *Let $A_1, A_2, \ldots$ be i.i.d matrix-valued random variables. Assume that the A_is are real and symmetric with $\|\operatorname{E}(A_i)\|_2 \leq 1$ and $\|A_i\|_2 \leq \gamma$ almost surely. Let $0 < \epsilon < 1$ and let $M = \Omega(\gamma \log(\gamma/\epsilon^2)/\epsilon^2)$. If every value in the support of A_i has rank at most M then*

$$\Pr\left(\left\|\frac{1}{M}\sum_{i=1}^{M} A_i - \operatorname{E}(A_i)\right\|_2 > \epsilon\right) \leq \frac{1}{\operatorname{poly}(M)} .$$

2.2. **Generalized eigenvalues, analysis of iterative methods and sparsification bounds.** A well known property of many iterative linear solver, including the popular conjugate gradient and the theoretically convenient Chebyshev iteration, is that their convergence rate depends on the distribution of the eigenvalues of the coefficient matrix (its spectrum). The rate depends on how much the spectrum is clustered, but it is hard to form a concise bound. A simple and useful theoretical bound for symmetric positive semidefinite matrices depends only on the ratio between the largest and smallest eigenvalue. When using preconditioned methods convergence is governed by the generalized eigenvalues.

Definition 2.2. Given two matrices A and B in $\mathbb{R}$ with the same null space $\mathbb{N}$, a *finite generalized eigenvalue* λ of (A, B) is a scalar satisfying $Ax = \lambda Bx$ for some $x \notin \mathbb{N}$. The *generalized finite*

spectrum $\Lambda(A,B)$ is the set of finite generalized eigenvalues of (A,B), and the *generalized condition number* $\kappa(A,B)$ is

$$\kappa(A,B) = \frac{\max \Lambda(A,B)}{\min \Lambda(A,B)} .$$

We define the *trace of* (A,B) (denoted by $\text{trace}(A,B)$) as the sum of finite generalized eigenvalues of (A,B).

(This definition can be generalized to the case of different null spaces, but this is irrelevant for this paper.) We will denote by $\Lambda(A)$ the set of finite non-zero eigenvalues of A (which is equal to $\Lambda(A,P_A)$, where P_A is a projection onto the range of A).

We are mainly interested in bounds on the smallest and largest generalized eigenvalues (which we denote $\lambda_{\min}(\cdot,\cdot)$ and $\lambda_{\max}(\cdot,\cdot)$ respectively), since they tell us two important properties on the pair (A,B). First, for every unit norm vector x we have

$$\lambda_{\min}(A,B) \cdot x^T B x \leq x^T A x \leq \lambda_{\max}(A,B) \cdot x^T B x .$$

Second, when B is used as a preconditioner for A, a vector x satisfying $\|x - A^+ b\|_A \leq \epsilon \|A^+ b\|_A$ is found in at most $O(\sqrt{\kappa(A,B)} \cdot \log(1/\epsilon))$ iterations where $\|x\|_A^2 = x^T A x$.

In many cases it is easier to reason about non-generalized eigenvalues. The following result from [1] relates generalized eigenvalues with regular eigenvalues of a different matrix.

Lemma 2.3. *Let $A = UU^T$ and $B = VV^T$, where U and V are real valued with the same number of rows. Assume that A and B are symmetric, positive semidefinite and $(A) = (B)$. We have*

$$\Lambda(A,B) = \Sigma^2\left(V^+ U\right)$$

and

$$\Lambda(A,B) = \Sigma^{-2}\left(U^+ V\right) .$$

In these expressions, $\Sigma(\cdot)$ is the set of nonzero singular values of the matrix within the parentheses, Σ^ℓ denotes the same singular values to the ℓth power, and V^+ denotes the Moore-Penrose pseudoinverse of V.

2.3. **Effective resistance sampling.** Recent progress of in fast SDD solvers [8, 2, 9] is based on effective resistance sampling, first suggested in [15]. Solving SDD systems can be reduced to solving a *Laplacian* system. Given a weighted undirected graph $G = ([n], E, w)$ the *Laplacian* L is given by $L = D - A$ where A is the weighted adjacency matrix $A_{ij} = w_{ij}$ and D is the diagonal matrix of weighted degrees given by $D_{ii} = \sum_{j \neq i} w_{ij}$. The *effective resistance* R_e of an edge $e = (u,v)$ is given by

$$R_e = (e_u - e_v)^T L^+ (e_u - e_v)$$

where e_u and e_v are identity vectors and L^+ is the Moore-Penrose pseudoinverse of L. The quantity is named effective resistance because R_e is equal to the potential difference induced between u and v when a unit of current is injected at u and extracted at v, when G is viewed as an electrical network with conductances given by w.

Spielman and Srivastava [15] showed that sampling sufficiently enough edges, where the probability of sampling an edge is proportional to $w_e R_e$ yields an high-quality sparsifier for G, which can be translated to an high-quality preconditioner. Koutis et al. [8, 2, 9] show that even crude approximations to the accurate effective resistances suffice, and they show how such an approximation can be computed efficiently. The asymptotically fastest solver [9] solves an n-by-n SDD linear system in time $O(m \log n \log(1/\epsilon))$ where m is the number of non-zeros in the matrix and ϵ is the accuracy of the solution.

3. Finite Element Matrices and Their Factored Form

A finite element discretization of a PDE usually leads to an algebraic system of equations $Kx = b$. The matrix K has certain properties that stem from the PDE and the specifics of how it was discretized. To make our results more general and easier to understand by a wide audience, we use the algebraic-combinatorial formulation developed in [13] rather than a PDE-derived formulation.

The matrix $K \in \mathbb{R}^{n\times n}$ is called a *stiffness matrix*, and it is a sum of *element matrices*, $K = \sum_{e=1}^{m} K_e$. Each element matrix K_e corresponds to a subset of the domain called a *finite element*. The elements are disjoint except perhaps for their boundaries and their union is the domain. We assume that each element matrix K_e is symmetric, positive semidefinite, and zero outside a small set of n_e rows and columns. In most cases n_e is uniformly bounded by a small integer. We denote the set of nonzero rows and columns of K_e by $\mathcal{N}_e$. We denote the restriction of a matrix A to indices I by $A(I)$, and denote the $\tilde{K}_e = K_e(\mathcal{N}_e)$. $\tilde{K}_e$ is the *essential element matrix* of e. Typically, in finite element discretizations both the stiffness matrix (K) and the essential element matrices ($\tilde{K}_e$s) are singular. We denote the dimension of the null space of $\tilde{K}_e$ by $d_e = \dim((\tilde{K}_e))$ and the rank of $\tilde{K}_e$ by $r_e = n_e - d_e$. For simplicity, we will assume the rank (and dimension of null space) of all the elements is the same and equal to r (d for null space dimension). The results can be easily extended for non uniform element ranks. The null space of K is denoted by $\mathbb{N}$ and we assume that its dimension is d as well.

Most, if not all, theoretical results on sampling matrices are on rectangular matrices, and their usefulness rely on the aspect ratio being high. Finite element matrices, according to our definition, are square. Luckily, K can be written as $K = F^T F$ where F has more rows than columns. The key is obtaining a factored form of $K_e = F_e^T F_e$ where F_e is $r \times n$. We can then write

$$F = \begin{pmatrix} F_1 \\ \vdots \\ F_m \end{pmatrix} \in \mathbb{R}^{mr\times n}$$

and $K = F^T F$. Many finite-element discretization techniques actually generate the element matrices in a factored form. Even if the elements are not generated in a factored form, a factored form can be easily computed. One way to do so is using the eigendecomposition $\tilde{K}_e = V_e \Sigma_e V_e^T$. Define $\tilde{F}_e = \Sigma_e^{1/2} \bar{V}_e^T$ where $\bar{V}_e$ is obtained by taking the r columns of V_e associated with non-zero eigenvalues, and let F_e be obtained by expanding the number of columns of $\tilde{F}_e$ to n by adding zero columns for columns not in $\mathcal{N}_e$. It is easy to verify that $K_e = F_e^T F_e$ and that F_e is $r \times n$.

Typically, the element matrices are *compatible* with the null space of K [13], meaning that the null space of $\tilde{K}_e$ is the restriction of the null space of K to $\mathcal{N}_e$. When all the element matrices are compatible with the null space of K, the matrices involved have useful properties [13] that we use in theorems as needed. In the Appendix we elaborate on the issue of null-space compatibility and explain why finite element matrices typically have the properties assumed in the theoretical analysis.

The main property we assume is that the rank deficiency of the factor F is minimal.

Definition 3.1. A matrix $F \in \mathbb{R}^{m\times n}$ has *minimal rank deficiency* if every set of $n - \dim((F))$ columns of F is independent.

Note that if the rank deficiency of F is minimal then every leading $l \times l$ minor of K is non-singular, as long as $l \leq n - d$.

4. Effective Stiffness of an Element

We now define the effective stiffness of an element. The stiffness matrix of an element describes the physical properties (elasticity, electrical conductivity, thermal conductivity, etc) of a piece of material called an element by showing how that piece of material responds to a load (current, mechanical force, etc) placed on the element. The *effective stiffness matrix* shows how the entire structure responds to a load that is placed on one element. Intuitively, if the stiffness matrix and the effective stiffness matrix of an element are similar, the element is important; removing it from the structure may significantly change the behavior of the overall structure. On the other hand, if the effective stiffness element has a much larger norm than the element matrix, then the element does not contribute much to the strength (or conductivity) of the overall structure, so it can be removed without changing much the overall behavior.

Algebraically, the effective stiffness matrix of e is obtained by eliminating from K all columns not associated with e.

Definition 4.1. Let $\bar{K}$ be obtained from K by an arbitrary symmetric reordering of the row and columns of K such that the last n_e rows and columns of $\bar{K}$ are $\mathcal{N}_e$ and they are ordered in ascending order (i.e., the ordering in $\bar{K}$ of the columns in $\mathcal{N}_e$ is consistent with their order in K). Suppose that $\bar{K}$ is partitioned

$$\bar{K} = \begin{pmatrix} \bar{K}_{11} & \bar{K}_{12} \\ \bar{K}_{12}^T & \bar{K}_{22} \end{pmatrix}$$

where $\bar{K}_1 \in \mathbb{R}^{(n-n_e)\times(n-n_e)}$, $\bar{K}_{12} \in \mathbb{R}^{(n-n_e)\times n_e}$ and $\bar{K}_{22} \in \mathbb{R}^{n_e \times n_e}$. If $\bar{K}_{11}$ is non singular we say that element e is *supported* . The *effective stiffness* S_e of element e is

$$S_e = \begin{cases} \bar{K}_{22} - \bar{K}_{12}^T \bar{K}_{11}^{-1} \bar{K}_{12} & e \text{ is supported} \\ \tilde{K}_e & \text{otherwise} \end{cases} .$$

It is easy to verify that the the effective stiffness is well defined in the sense that any ordering that respects the conditions of the definition gives the same S_e. Note that if the factor F has minimal rank deficiency then all elements are supported.

Before proceeding to discuss effective stiffness sampling, and stating our main result, we first show that indeed effective stiffness generalizes effective resistance by showing that effective resistance is a particular case of effective stiffness.

The Laplacian of a weighted graph $G = ([n], E, w)$ is, in fact, a finite element matrix per our definition in section 3. Given an edge $e = (u, v)$ define $K_e = w_e(e_u - e_v)(e_u - e_v)^T$. It is easy to verify that $L = \sum_{e\in E} K_e$. It is well-known that if the if the graph is connected the rank deficiency is $d = 1$ and the null space $\mathbb{N}$ is all-ones vector. L can also be written in factor form $L = F^T F$ where $F \in \mathbb{R}^{|E|\times|V|}$. Each edge $e = (u, v)$ correspond to row in F given by $F_e = \sqrt{w_e}(e_u - e_v)^T$. If the graph is connected the factor F has minimal rank deficiency, so all elements (edges) are supported.

Lemma 4.2. *Let F be the factor of a graph G. If G is connected then F has minimal rank deficiency.*

Proof. Suppose there is a non-independent size $n-1$ subset of F's columns. Let $\bar{F}$ be a reorder of F's columns such that those $n-1$ columns are the first $n-1$ columns. Since the first $n-1$ columns of $\bar{F}$ are linearly dependent there is a vector $x \neq 0$ such that

$$F \begin{pmatrix} x \\ 0 \end{pmatrix} = 0 .$$

The vector $\begin{pmatrix} x^T & 0 \end{pmatrix}^T$ is not spanned by 1_n. This contradicts our assumption that G is connected since the null space of the Laplacian of a connected graph is always span$\{1_n\}$. □

Simple calculation shows that $S_e 1_2 = 0$ and $(e_1 - e_2)^T S_e (e_1 - e_2) = R_e^{-1}$. This implies that $S_e = R_e^{-1}(e_u - e_v)(e_u - e_v)^T$.

Graph sparsification by effective resistance [15] and near-linear time linear solvers [8, 2, 9] relay on sampling edges with probability relative to $w_e R_e$. It is easy to verify that $w_e R_e = \lambda_{\max}(\tilde{K}_e, S_e)$. Our main result shows that sampling probabilities should be relative to $\lambda_{\max}(\tilde{K}_e, S_e)$ for general finite element matrices, and not only for Laplacians.

5. Effective Stiffness Sampling

The main theorem of this writeup shows how to use the effective stiffness to sample finite element matrices.

Theorem 5.1. *Let $K = F^T F = \sum_{e=1}^m K_e$ be an n-by-n finite element matrix. Assume that the factor F has minimal rank deficiency and $\mathrm{null}(S_e) = \mathrm{null}(\tilde{K}_e)$ for every element e. Let*

$$p_e = \frac{\lambda_{\max}(\tilde{K}_e, S_e)}{\sum_{i=1}^m \lambda_{\max}(\tilde{K}_i, S_i)}$$

and let $T_1, \ldots, T_M$ be a i.i.d random matrices defined by

$$T_i = p_{J_i}^{-1} K_{J_i}$$

where $J_1, \ldots, J_M$ are random integers between 1 and m which takes value e with probability p_e. In other words, T_i is a scaled version of one of the K_es, selected at random, with a scaling that is proportional to the inverse of p_e. For $M = \Omega(n \log(n))$ we have

$$\Pr\left(\kappa(K, \frac{1}{M}\sum_{i=1}^M T_i) > 2\right) \leq \frac{1}{\mathrm{poly}(M)}\,.$$

To prove Theorem 5.1 we need the following Lemma.

Lemma 5.2. *Let $U \in \mathbb{R}^{mr \times n}$ be any matrix whose columns form an orthonormal basis of $\mathrm{range}(F)$. Let $U_e \in \mathbb{R}^{r \times n}$ be the rows of U corresponding to element e. Assume that the factor F has minimal rank deficiency and $\mathrm{null}(S_e) = \mathrm{null}(\tilde{K}_e)$. The set of non-zero eigenvalues (including multiplicity) of $U_e U_e^T$ and the set of finite generalized eigenvalues of $(\tilde{K}_e, S_e)$ are the same. In particular,*

$$\lambda_{\max}(U_e U_e^T) = \lambda_{\max}(\tilde{K}_e, S_e)$$

and

$$\mathrm{trace}(U_e U_e^T) = \mathrm{trace}(\tilde{K}_e, S_e)\,.$$

Proof. We first show that we can prove the lemma by showing that it holds for a particular U. An arbitrary orthonormal basis V is related to U by $V = UZ$, where Z is an n-by-n unitary matrix. In particular, $V_e = U_e Z$ (V_e are the rows of V corresponding to element e) so $V_e V_e^T = U_e Z Z^T U_e^T = U_e U_e^T$. We obtain U from the QR factorization of $\bar{F} = \bar{U} R$ and set U to be the first $n - d$ columns of $\bar{U}$, where $\bar{F}$ is obtained from F by reordering the columns in $\mathcal{N}_e$ to the end (consistently with their ordering in F).

The last n_e columns of $\bar{F}$ are $\mathcal{N}_e$, and F_e is non-zero outside the indices of $\mathcal{N}_e$.This implies that

$$\bar{F}_e = \begin{bmatrix} 0_{r \times (n - n_e)} & \tilde{F}_e \end{bmatrix}$$

$$U_e = \begin{bmatrix} 0_{r \times (n - n_e)} & \tilde{U}_e \end{bmatrix}$$

where $\tilde{U}_e$, $\tilde{F}_e \in \mathbb{R}^{r \times n_e}$. Let us write

$$R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}$$

where $R_{11} \in \mathbb{R}^{(n-n_e)\times(n-n_e)}$, $R_{12} \in \mathbb{R}^{(n-n_e)\times n_e}$ and $R_{22} \in \mathbb{R}^{n_e \times n_e}$. Let us write $\bar{K} = \bar{F}^T\bar{F}$ and

$$\bar{K} = \begin{pmatrix} \bar{K}_{11} & \bar{K}_{12} \\ \bar{K}_{12}^T & \bar{K}_{22} \end{pmatrix}$$

where $\bar{K}_{11} \in \mathbb{R}^{(n-n_e)\times(n-n_e)}$, $\bar{K}_{12} \in \mathbb{R}^{(n-n_e)\times n_e}$ and $\bar{K}_{22} \in \mathbb{R}^{n_e\times n_e}$. Since R is the R-factor of $\bar{F}$ and $\bar{K} = \bar{F}^T\bar{F}$ it is also the Cholesky factor of $\bar{K}$. It also implies that $R_{22}^T R_{22}$ is equal to the Schur complement

$$R_{22}^T R_{22} = \bar{K}_{22} - \bar{K}_{12}^T \bar{K}_{11}^{-1} \bar{K}_{12} = S_e\,.$$

The minimal rank deficiency of F implies that that the bottom d rows of R and R_{22} are zero. Let $\bar{R}_{22} \in \mathbb{R}^{(n_e-d)\times n_e}$ be the first $n_e - d$ rows of R_{22}. It is still the case that $\bar{R}_{22}^T\bar{R}_{22} = S_e$. We have $\bar{F} = \bar{U}R$, so $\bar{F}_e = \bar{U}_e R_{22} = U_e\bar{R}_{22}$ which implies that $\tilde{F}_e = \tilde{U}_e\bar{R}_{22}$. Applying Lemma 2.3 we find that

$$\begin{aligned} \Lambda(\tilde{K}_e, S_e) &= \Lambda(\tilde{F}_e^T\tilde{F}_e, R_{22}^T R_{22}) \\ &= \Sigma^2\left((\bar{R}_{22}^T)^+ \tilde{F}_e^T\right) \\ &= \Sigma^2\left((\bar{R}_{22}^T)^+ \bar{R}_{22}^T \tilde{U}_e^T\right) \end{aligned}$$

The minimal rank deficiency of $\bar{F}$ implies that R_{22} is full rank, so R_{22}^T is a full rank matrix with more rows than columns (or equal), so $(\bar{R}_{22}^T)^+\bar{R}_{22}^T = I_{n_e}$. This implies that $(\bar{R}_{22}^T)^+\bar{R}_{22}^T\tilde{U}_e^T = \tilde{U}_e^T$ so

$$\Lambda(\tilde{K}_e, S_e) = \Sigma^2(\tilde{U}_e^T)\,.$$

$\Sigma^2(\tilde{U}_e^T)$ is exactly the set of non-zero eigenvalues of $\tilde{U}_e\tilde{U}_e^T$. Therefore, the non-zero eigenvalues of $U_eU_e^T$ are exactly the finite generalized eigenvalues of $(\tilde{K}_e, S_e)$, so

$$\lambda_{\max}(U_eU_e^T) = \lambda_{\max}(\tilde{K}_e, S_e)$$

and

$$\operatorname{trace}(U_eU_e^T) = \operatorname{trace}(\tilde{K}_e, S_e)\,.$$

□

We can now prove Theorem 5.1.

Proof. (of Theorem 5.1) We express the matrix $\frac{m}{M}\sum_{i=1}^M T_i$ *as a normal form*

$$\frac{1}{M}\sum_{i=1}^M T_i = (\mathcal{S}F)^T(\mathcal{S}F)$$

where $\mathcal{S} \in \mathbb{R}^{n_eM\times n_em}$ is a random sampling matrix and F is the factor of the stiffness matrix $K = F^TF$. If we take S to be a block matrix with $n_e \times n_e$ blocks, its blocks defined by

$$\mathcal{S}_{ie} = \begin{cases} \sqrt{\frac{1}{M}}p_e^{-1/2} I_{n_e\times n_e} & \text{if } T_i = p_e^{-1}K_e \\ 0_{n_e\times n_e} & \text{otherwise,} \end{cases}$$

then it is easy to verify that $\mathcal{S}$ indeed satisfies the identity above. Let $F = \bar{U}\bar{R}$ be a reduced QR factorization of F. The minimal rank deficiency of F implies that that the bottom d rows of R are zero. Let $R \in \mathbb{R}^{(n-d)\times n}$ be the first $n-d$ rows of $\bar{R}$, and $U \in \mathbb{R}^{mr\times(n-d)}$ be the first $n-d$ columns

of $\bar{U}$. It is easy to verify that $F = UR$ and $F^TF = R^TR$. R^T is full rank, so $(R^T)^+R^T = I_n$. Applying lemma 2.3 we have

$$\begin{aligned}
\kappa(K, \frac{1}{M}\sum_{i=1}^{M} T_i) &= \kappa(F^TF, (\mathcal{S}F)^T(\mathcal{S}F)) \\
&= \kappa(R^TR, (\mathcal{S}F)^T(\mathcal{S}F)) \\
&= \kappa^2\left((R^T)^+F^TS^T\right) \\
&= \kappa^2\left((R^T)^+R^TU^TS^T\right) \\
&= \kappa^2(U^TS^T) \\
&= \kappa^2((SU)^T) \\
&= \kappa^2(SU) \\
&= \kappa((\mathcal{S}U)^T(\mathcal{S}U))\,.
\end{aligned}$$

To bound $\kappa((\mathcal{S}U)^T(\mathcal{S}U))$ with high probability we first bound $\left\|(\mathcal{S}U)^T(\mathcal{S}U) - I_{n\times n}\right\|_2$ with high probability. Define the i.i.d random matrices $Y_1, \ldots, Y_M$ by

$$Y_i = p_{J_i}^{-1}U_{J_i}^TU_{J_i}$$

where U_e is the rows corresponding to element e in U. It is easy to verify that

$$(\mathcal{S}U)^T(\mathcal{S}U) = \frac{1}{M}\sum_{i=1}^{M} Y_i\,.$$

The expectation of the Y_i's is the identity matrix,

$$\begin{aligned}
\mathrm{E}(Y_i) &= \sum_{j=1}^{M} \Pr(T_i = p_j^{-1}K_j)p_j^{-1}U_j^TU_j \\
&= \sum_{j=1}^{M} p_jp_j^{-1}U_j^TU_j \\
&= \sum_{j=1}^{M} U_j^TU_j \\
&= U^TU = I_{n-d\times n-d}
\end{aligned}$$

and their 2-norm is bounded by

$$\begin{aligned}
\|Y_i\|_2 &\leq \max_j p_j^{-1}\left\|U_j^TU_j\right\|_2 \\
&= \max_j p_j^{-1}\lambda_{\max}(U_jU_j^T) \\
&= \max_j p_j^{-1}\lambda_{\max}(\tilde{K}_j, S_j) \\
&= \max_j \left(\left(\frac{\lambda_{\max}(\tilde{K}_j, S_j)}{\sum_{i=1}^{m}\lambda_{\max}(\tilde{K}_i, S_i)}\right)^{-1}\lambda_{\max}(\tilde{K}_j, S_j)\right) \\
&= \sum_{i=1}^{m}\lambda_{\max}(\tilde{K}_i, S_i)\,.
\end{aligned}$$

We now bound this sum of maximal generalized eigenvalues

$$\begin{aligned}\sum_{i=1}^{m}\lambda_{\max}(\tilde{K}_i,S_i) &\leq \sum_{i=1}^{m}\operatorname{trace}(\tilde{K}_i,S_i)\\ &= \sum_{i=1}^{m}\operatorname{trace}(U_iU_i^T)\\ &= \sum_{i=1}^{m}\operatorname{trace}(U_i^TU_i)\\ &= \operatorname{trace}(\sum_{i=1}^{m}U_i^TU_i)\\ &= \operatorname{trace}(U^TU)=n-d\,.\end{aligned}$$

We showed that $\|\operatorname{E}(Y_i)\|_2=\|I_{n\times n}\|=1$ and that $\|Y_i\|_2\leq n$, so we can apply Theorem 2.1 to the Y_is with $\epsilon=1/3$ and $\gamma=n$. The application of the theorem shows that for $M=\Omega(n\log(n))$,

$$\Pr\left(\left\|\frac{1}{M}\sum_{i=1}^{M}Y_i-I_{n\times n}\right\|_2>\frac{1}{3}\right)\leq\frac{1}{\operatorname{poly}(M)}\,.$$

This bounds $\left\|(\mathcal{S}U)^T(\mathcal{S}U)-I_{n\times n}\right\|_2$ with high probability. Finally, we note that for every symmetric metric A, if $\|A-I\|_2\leq t<1$ then $\kappa(A)\leq\frac{1+t}{1-t}$. Applying this to $(\mathcal{S}U)^T(\mathcal{S}U)$ with $t=1/3$ we find that whenever $\left\|(\mathcal{S}U)^T(\mathcal{S}U)-I_{n\times n}\right\|_2\leq 1/3$ (which happens with high probability), we have $\kappa((\mathcal{S}U)^T(\mathcal{S}U))\leq 2$. □

6. Sampling using approximate eigenvalues

Theorem 5.1 shows that the sampling probabilities that are proportional to $\lambda_{\max}(\tilde{K}_e,S_e)$ are effective for randomly selecting a good subset of elements to serve as a preconditioner. In practice it may be possible to obtain only estimates for the true maximum eigenvalues. The following generalization of Theorem 5.1 shows that even crude approximations to $\lambda_{\max}(\tilde{K}_e,S_e)$ suffice to get a low condition number with high probability.

Theorem 6.1. *For every element e let $\tilde{\lambda}_e$ be $(1+\delta)$-approximations to $\lambda_{\max}(\tilde{K}_e,S_e)$, that is*

$$\left|\tilde{\lambda}_e-\lambda_{\max}(\tilde{K}_e,S_e)\right|\leq\delta\cdot\lambda_{\max}(\tilde{K}_e,S_e)$$

We make the same assumptions and use the same notation as in Theorem 5.1 except that the probabilities p_e are now given by

$$p_e=\frac{\tilde{\lambda}_e}{\sum_{i=1}^{m}\tilde{\lambda}_i}\,.$$

For $M=\Omega(n\beta\log(n\beta))$, where $\beta=\frac{1+\delta}{1-\delta}$, we have

$$\Pr\left(\kappa(K,\frac{1}{M}\sum_{i=1}^{M}T_i)>2\right)\leq\frac{1}{\operatorname{poly}(M)}\,.$$

Proof. The proof is identical to the proof of Theorem 5.1 except that the bound on $\|Y_i\|_2$ needs to be modified as follows:

$$\begin{aligned}
\|Y_i\|_2 &\leq \max_j p_j^{-1} \left\|U_j^T U_j\right\|_2 \\
&= \max_j p_j^{-1} \lambda_{\max}(U_j U_j^T) \\
&= \max_j p_j^{-1} \lambda_{\max}(\tilde{K}_j, S_j) \\
&= \max_j \left(\left(\frac{\tilde{\lambda}_e}{\sum_{i=1}^m \tilde{\lambda}_i} \right)^{-1} \lambda_{\max}(\tilde{K}_j, S_j) \right) \\
&\leq \max_j \left(\left(\frac{(1-\delta)\lambda_{\max}(\tilde{K}_j, S_j)}{(1+\delta)\sum_{i=1}^m \lambda_{\max}(\tilde{K}_i, S_i)} \right)^{-1} \lambda_{\max}(\tilde{K}_j, S_j) \right) \\
&= \beta \sum_{e=1}^m \lambda_{\max}(\tilde{K}_e, S_e) \\
&\leq n\beta\,.
\end{aligned}$$

□

References

[1] Haim Avron, Doron Chen, Gil Shklarski, and Sivan Toledo. Combinatorial preconditioners for scalar elliptic finite-element problems. *SIAM J. Matrix Anal. Appl.*, 31:694–720, June 2009.

[2] Guy E. Blelloch, Anupam Gupta, Ioannis Koutis, Gary L. Miller, Richard Peng, and Kanat Tangwongsan. Near linear-work parallel sdd solvers, low-diameter decomposition, and low-stretch subgraphs. In *Proceedings of the 23rd ACM symposium on Parallelism in algorithms and architectures*, SPAA '11, pages 13–22, New York, NY, USA, 2011. ACM.

[3] Erik G. Boman, Doron Chen, Bruce Hendrickson, and Sivan Toledo. Maximum-weight-basis preconditioners. *Numerical Linear Algebra with Applications*, 11:695–721, 2004.

[4] Erik G. Boman, Doron Chen, Ojas Parekh, and Sivan Toledo. On the factor-width and symmetric H-matrices. *Numerical Linear Algebra with Applications*, 405:239–248, 2005.

[5] Erik G. Boman, Bruce Hendrickson, and Stephen Vavasis. Solving elliptic finite element systems in near-linear time with support preconditioners. *SIAM Journal on Numerical Analysis*, 46(6):3264–3284, 2008.

[6] Samuel I. Daitch and Daniel A. Spielman. Support-graph preconditioners for 2-dimensional trusses. Mar 2007.

[7] Samuel I. Daitch and Daniel A. Spielman. Faster approximate lossy generalized flow via interior point algorithms. In *STOC '08: Proceedings of the 40th annual ACM Symposium on Theory of Computing*, pages 451–460, New York, NY, USA, 2008. ACM.

[8] Ioannis Koutis, Gary L. Miller, and Richard Peng. Approaching optimality for solving sdd linear systems. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 235–244, Washington, DC, USA, 2010. IEEE Computer Society.

[9] Ioannis Koutis, Gary L. Miller, and Richard Peng. Solving sdd linear systems in time $\tilde{O}(m \log n \log(1/\epsilon))$. In *Proceedings of the 2011 IEEE 52st Annual Symposium on Foundations of Computer Science*, FOCS '11, Washington, DC, USA, 2011. IEEE Computer Society.

[10] Avner Magen and Anastasios Zouzias. Low rank matrix-valued chernoff bounds and approximate matrix multiplications. In *SODA '10: Proceedings of the twenty-second annual ACM-SIAM Symposium on Discrete Algorithm*, 2010.

[11] Roberto Imbuzerio Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electronic Communications in Probability*, 15, 2010.

[12] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54, July 2007.

[13] Gil Shklarski and Sivan Toledo. Rigidity in finite-element matrices: Sufficient conditions for the rigidity of structures and substructures. *SIAM Journal on Matrix Analysis and Applications*, 30(1):7–40, 2008.

[14] Gil Shklarski and Sivan Toledo. Computing the null space of finite element problems. *Computer Methods in Applied Mechanics and Engineering*, 198(37-40):3084 – 3095, 2009.

[15] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, STOC '08, pages 563–568, New York, NY, USA, 2008. ACM.
[16] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. June 2011.
[17] Meiqiu Wang and Vivek Sarin. Parallel support graph preconditioners. In Yves Robert, Manish Parashar, Ramamurthy Badrinath, and Viktor K. Prasanna, editors, *High Performance Computing - HiPC 2006*, volume 4297, chapter 39, pages 387–398. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

Appendix: Null space compatibility and rigidity of finite element matrices

In this section we review facts from the algebraic-combinatorial theory of rigidity of finite-element matrices developed in [13], and show that it implies typical finite-element matrices have the properties assumed in the results. This section is not necessary for the validity of the theoretical results, but it does relate them to the actual application.

Typically, the element matrices are *compatible* with the null space of K. We now define what this means exactly.

Definition 6.2. Let A be an m-by-n matrix, let $\mathcal{Z}_A$ be the set of its zero columns. We define the *essential null space* of A (enull(A)) by

$$\text{enull}(A) = \{x \,:\, Ax = 0 \quad \text{and} \quad x_i = 0 \;\text{ for }\; i \in \mathcal{Z}_A\} \,.$$

Definition 6.3. Let $\mathbb{N} \subseteq \mathbb{R}^n$ be a linear space. A matrix A is called $\mathbb{N}$-*compatible* (or compatible with $\mathbb{N}$) if every vector in enull(A) has a unique extension into a vector in $\mathbb{N}$, and if the restriction of every vector in $\mathbb{N}$ to $\mathcal{N}_A$ (setting indices outside $\mathcal{N}_A$ to zero) is always in enull(A).

A particular discretization of a PDE yields element matrices (K_es) that are compatible with some well-known null space $\mathbb{N}$, which depends on the PDE; a translation in electrostatics, translations and rotations in elasticity, and so on. Furthermore, it is usually desirable that the stiffness matrix K be *rigid* with respect to $\mathbb{N}$, which is equivalent to saying that the null space of K is exactly $\mathbb{N}$. For example, for matrix of a resistive network (the Laplacian) elements are compatible with the span of the all-ones vector. The null space of the Laplacian is exactly the span of the all-ones (i.e., the matrix is rigid) if and only if the graph is connected.

Lack of rigidity often implies that the PDE has not been discretizied correctly, and it does not make sense to solve the linear equations. This is an important scenario to detect (see [14]), but it is not the subject of this paper. We will assume the matrix K is rigid with respect to the prescribed and well know null space. The prescribed null space typically (that is, for real-life finite element matrices) implies minimal rank deficiency, which has to be proved for each case. A simple technique (which the proof of Lemma 4.2 implicitly uses) is based on the following lemma.

Lemma 6.4. *Suppose that $K = F^T F \in K^{n\times n}$ has null space* range(N) *where $N \in \mathbb{R}^{n\times d}$. If no $d \times d$ submatrix of N is singular then F has minimal rank deficiency.*

Proof. First notice that null(F) = null(K) since null(F^T) = range(F)$^\perp$. Suppose there is a set of $n - d$ columns of F which are not independent. Let $\bar{F}$ be a reordering of the columns of F such that those $n - d$ columns are first. There is a vector $x \in \mathbb{R}^{n-d}$ such that

$$\bar{F} \begin{pmatrix} x \\ 0_{d\times 1} \end{pmatrix} = 0 \,.$$

Let $\bar{N}$ be a reordering of the rows of N consistently with the reordering of the columns of F in $\bar{F}$. The vector $\begin{pmatrix} x^T & 0 \end{pmatrix}^T$ is in the null space of $\bar{F}$ so there must exist a vector $y \neq 0$ such that $\bar{N}y = \begin{pmatrix} x^T & 0 \end{pmatrix}^T$. This implies that the bottom d rows of $\bar{N}$ form a singular matrix. These rows are also rows of N, which implies that N has a $d \times d$ singular submatrix, which contradicts our assumption. □

We already saw that this technique was used to show that the factor of a connected Laplacian has minimal rank deficiency. We now show another example: elastic struts in two dimensions. In [13] it is shown that given a collection $P = \{p_i\}_{i=1}^n$of points in the plane, the null space of the rigid finite element matrix representing a collection of elastic struts between the points is spanned by the range of

$$N = \begin{pmatrix} 1 & 0 & -y_1 \\ 0 & 1 & x_1 \\ 1 & 0 & -y_2 \\ 0 & 1 & x_2 \\ \vdots & \vdots & \vdots \\ 1 & 0 & -y_n \\ 0 & 1 & x_n \end{pmatrix}.$$

The matrix N does not have singular 3-by-3 submatrix unless the points have some special properties (like three points with the same x coordinate), which they typically do not have. Even if such a property is present, a slight rotation of the point set, an operation that does not fundamentally change the physical problem, will remove it.

The minimal rank deficiency of F implies that all elements are supported. We now show that $\text{null}(S_e) = \text{null}(\tilde{K}_e)$ if K is rigid. To do so we need an additional definition and two lemmas from [13].

Definition 6.5. A matrix A is *rigid with respect to* another matrix B if for every vector in $x \in \text{enull}(A)$ there is a unique vector $y \in \text{enull}(B)$ such that $x_i = y_i$ for all $i \in \mathcal{N}_A \cap \mathcal{N}_B$. The two matrices are called *mutually rigid* if they are rigid with respect to each other.

Lemma 6.6. *(Lemma 5.5 from* [13]) Let $\mathbb{N}$ be a linear space, and let B be some matrix with no zero columns whose null space is $\mathbb{N}$. Another matrix A is $\mathbb{N}$-compatible if and only if A and B are mutually rigid.

Lemma 6.7. *(part of Lemma 3.7 from* [13]) Let A and B be n-by-n matrices of the form

$$A = \begin{bmatrix} 0 & 0 \\ 0 & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Assuming that B has no zero columns, B_{11} is non-singular and A are mutually rigid we have $\text{null}(A_{22}) = \text{null}(B_{22} - B_{21}B_{11}^{-1}B_{12})$.

It is easy to see that combining the last two Lemmas with minimal rank deficiency of F ensures that $\text{null}(S_e) = \text{null}(\tilde{K}_e)$ for every element e.

Lemma 6.8. Let $K = F^T F = \sum_{e=1}^m K_e$ be an n-by-n finite element matrix with null space $\mathbb{N}$. Assuming that F has minimal rank deficiency, and that all element are $\mathbb{N}-$compatible, then $\text{null}(S_e) = \text{null}(\tilde{K}_e)$ for every element e.