A fast direct solver for structured matrices

Kenneth L. Ho and Leslie Greengard
Courant Institute of Mathematical Sciences
New York University, 251 Mercer Street, New York, NY 10012
greengard@courant.nyu.edu

Introduction

We have developed a fast direct solver for structured linear systems based on multilevel matrix compression. Starting with a hierarchically block-separable matrix [2], we embed an approximation of the original matrix into a larger, but highly structured sparse one. The resulting representation allows for efficient storage, fast matrix-vector multiplication, fast matrix factorization, and fast application of the inverse.

The algorithm proceeds in two phases: a precomputation phase, consisting of matrix compression and factorization, followed by a solution phase to apply the matrix inverse. For boundary integral equations which are not too oscillatory, e.g., based on the Green’s functions for the Laplace or low-frequency Helmholtz equations, both phases typically have complexity $O(N)$ in two dimensions, where $N$ is the number of discretization points. In our current three-dimensional implementation, the corresponding costs are $O(N^{4/3})$ and $O(N \log N)$ for precomposition and solution, respectively. Extensive numerical experiments show a speedup of $\sim 100$ for the solution phase over fast multipole methods; however, the cost of precomposition remains high. Thus, the solver is particularly suited to problems where large numbers of iterations would be required.

Several closely related efforts:

- $H$-matrices (Hackbusch et al.)
- HSS matrices (Gu, Chandrasekaran, et al.)
- skeletonization-based schemes (Martinsson and Rokhlin, Greengard, Gyfüffy, Martinsson, Rokhlin).

Structured matrices

Let $Ax = f$ be written in the form

$$A_1 x_1 + A_2 x_2 + \cdots + A_p x_p = f,$$  \hspace{1cm} (1)

where $x_i$ and $f$ are complex vectors of dimension $N_i$, and $A_i \in \mathbb{C}^{N_i \times N_i}$. Assuming $N = \sum_i N_i$, solution of the full linear system by classical Gaussian elimination is well-known to require $O(N^3)$ work.

Definition

The matrix $A$ is said to be block-separable [2] if each off-diagonal submatrix $A_i$ can be decomposed as the product of three low-rank matrices:

$$A_i = L_i S_i R_i,$$ \hspace{1cm} (2)

where $L_i \in \mathbb{C}^{N_i \times k_i}$, $S_i \in \mathbb{C}^{k_i \times k_i}$, and $R_i \in \mathbb{C}^{k_i \times N_i}$, with $k_i \ll N_i$ and $k_i \ll N_j$. Note that in (2), the left matrix $L_i$ depends only on the index $i$ and the right matrix $R_i$ depends only on the index $j$.

Structural results

The following equations hold:

$$A = D + L S R,$$ \hspace{1cm} (3)

is a block-diagonal $N \times N$ matrix consisting of the diagonal blocks of $A$, and

$$D = \begin{bmatrix} A_{11} & \cdots & A_{1p} \\ \vdots & \ddots & \vdots \\ A_{p1} & \cdots & A_{pp} \end{bmatrix},$$ \hspace{1cm} (4)

is a dense $K_1 \times K_1$ matrix, where $K_1 = \sum_i k_i$ and $K_1 = \sum_k k_i'$, with zero diagonal blocks. $L$ and $R$ are given by

$$L = \begin{bmatrix} L_1 & \cdots & L_p \end{bmatrix}, \quad R = \begin{bmatrix} R_1 & \cdots & R_p \end{bmatrix},$$ \hspace{1cm} (5)

They are block-diagonal $N \times K_1$ and $K_1 \times N$ matrices, respectively. This is, of course, an effective compression of $A$ only when $K_1 \ll N$. A useful feature of the representation (3) is that it permits rapid inversion. To see this, let $z \equiv Rx$ and let $x = Sz$. We can then write the original system in the form

$$\begin{bmatrix} D & L \\ R & -I \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \beta \\ 0 \end{bmatrix},$$ \hspace{1cm} (6)

The last equation can be computed recursively if the matrix $S$ supports the block-separable property at each level of the hierarchy. The result is a telescoping representation

$$A \approx D^{(1)} L^{(1)} \bigl( D^{(2)} + L^{(2)} \cdots (D^{(3)} + L^{(3)} S^{(3)} \cdots ) \bigr) R^{(1)}.$$ \hspace{1cm} (7)

Example: Suppose $\phi(x) = \int K(x, y) \rho(y) dy$ with $K(x, y) = \log|x - y|$, discretized at $N = 8192$ points in the unit square, and compress to relative precision $\epsilon = 10^{-3}$ using a five-level quadtree-based scheme.

Real part $\Re(x)$ of the pressure field in response to a vertical plane wave for various multiple scattering geometries characterized by the separation distance $u/A$ in wavelengths.

Molecular electrostatics

Letting $\Sigma$ be a molecular surface, a classical model for the electrostatic potential $\psi$ is Poisson’s equation:

$$- \nabla \cdot [\varepsilon(\Sigma) \nabla \psi(x)] = - \sum_{i=1}^{N} \delta(x - x_i),$$ \hspace{1cm} (8)

We generated a molecular surface for a short segment of DNA consisting of $N = 19752$ triangles and solved an integral equation for the induced polarization charge to precision $\epsilon = 10^{-6}$.

The total solution time was $\sim 902 \text{ s}$, with an inverse application time of $\sim 0.8 \text{ s}$.

Conclusions

We have developed a multilevel matrix compression and solution algorithm. The matrix structure required is fairly general and relies only on the assumption that the matrix have low-rank off-diagonal blocks. It is competitive with fast iterative methods based on FMM/GMRES in both 2D and 3D, provided that the kernel of the integral equation is not too oscillatory.

A principal limitation of all current hierarchical direct solvers is the growth in skeleton size in 3D, which prohibits the scheme from achieving optimal $O(N)$ complexity. New methods to curtail this growth are under development, at least for non-oscillatory cases. Finally, although all numerical results are reported here for a single CPU core, the algorithm is naturally suited for HPC systems.

References


Acknowledgments

Supported by DOE under contract DEFG0288ER25053. Thanks to Zygmun Gburdus and NSF (K.L.H.).