Linear Solvers

Andrew Sullivan

Department of Physics, Montana State University, Bozeman, MT 59717, USA.

October 30, 2019

1 LU Decomposition

In the regime of direct solvers, some common methods are Singular Value Decomposition (SVD) and Cholesky Decomposition for symmetric positive definite systems, QR decomposition, LU decomposition. All methods rewrite a matrix A into a product of multiple matrices with special properties that make their computation relatively easy. The special properties of these product matrices are then used to compute properties of the original matrix A. For example, a QR decomposition decomposes a real matrix A into a unitary matrix Q and an upper triangular matrix R, $\mathbf{A} = \mathbf{Q} \cdot \mathbf{R}$. Similarly a LU decomposition decomposes A into a lower triangular matrix L and an upper triangular matrix U.

$$\mathbf{A} = \mathbf{L} \cdot \mathbf{U} \tag{1}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} \alpha_{11} & 0 & 0 & 0 \\ \alpha_{21} & \alpha_{22} & 0 & 0 \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & 0 \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{bmatrix} \cdot \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ 0 & \beta_{22} & \beta_{23} & \beta_{24} \\ 0 & 0 & \beta_{33} & \beta_{34} \\ 0 & 0 & 0 & \beta_{44} \end{bmatrix} \tag{2}$$

With the special structure of $\mathbf{L}$ and $\mathbf{U}$, we can write our matrix decomposition as,

$$\mathbf{A} \cdot \mathbf{x} = (\mathbf{L} \cdot \mathbf{U}) \cdot \mathbf{x} = \mathbf{L} \cdot (\mathbf{U} \cdot \mathbf{x}) = \mathbf{b} \tag{3}$$

where,

$$\mathbf{L} \cdot \mathbf{y} = \mathbf{b}, \tag{4}$$

$$\mathbf{U} \cdot \mathbf{x} = \mathbf{y}. \tag{5}$$

These systems are trivial to solve through an iterative form of Gaussian elimination because of the triangular nature of $\mathbf{L}$ and $\mathbf{U}$. The algorithm goes as such:

- Start:
  1. Set $\alpha_{ii} = 1; i = 1, 2, \ldots, N$.
- Iterate: For $j = 1, 2, \ldots, N$,
  1. For $i = 1, 2, \ldots, j$,
     (a) $\beta_{ij} = a_{ij} - \alpha_{ik}\beta_{kj}$.
  2. For $i = j + 1, j + 2, \ldots, N$,
     (a) $\alpha_{ij} = \frac{1}{\beta_{jj}} \left(a_{ij} - \alpha_{ik}\beta_{kj} \right)$.

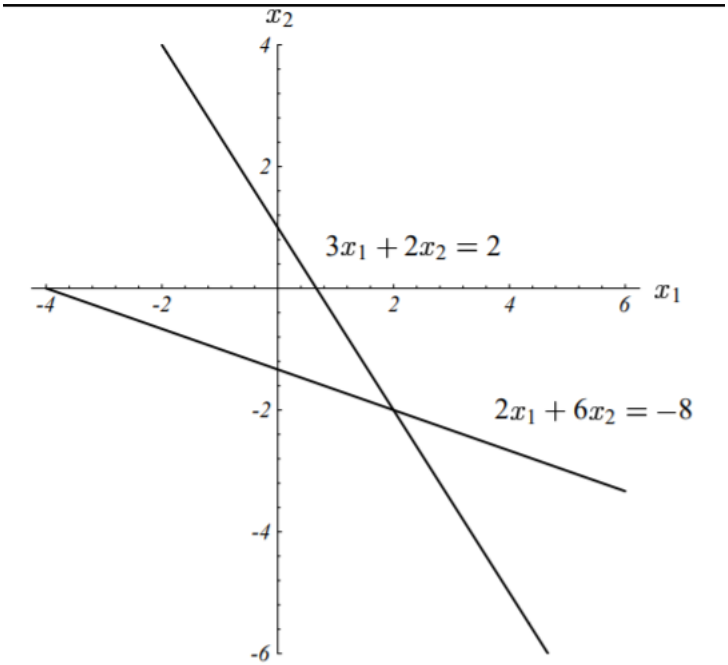

Figure 1: Example linear system.

2 Biconjugate Gradient Stabilized Method

To approach how BiCGSTAB works, it is better to start with the original conjugate gradient (CG) method which only works on symmetric positive definite systems. For example,

$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2 \\ -8 \end{bmatrix}, \qquad \mathbf{c} = \mathbf{0}, \tag{6}$$

illustrated in Fig. 1 and has the solution $\mathbf{x} = [2, -2]$.

Because $\mathbf{A}$ is symmetric, it can be written in quadratic form

$$Q(\mathbf{A}) = \mathbf{x}^T \cdot \mathbf{A} \cdot \mathbf{x}, \tag{7}$$

which is just another way of writing

$$Q(\mathbf{A}) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \cdot \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2. \tag{8}$$

So the quadratic form of our linear system is

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} + \mathbf{c} \tag{9}$$

shown in Fig. 2.

The general idea of the CG method is to find the solution of $\mathbf{x}$ through the gradient of $f(\mathbf{x})$.

$$\boldsymbol{\nabla} f(\mathbf{x}) = \frac{1}{2}\mathbf{A}^T\mathbf{x} + \frac{1}{2}\mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{A}\mathbf{x} - \mathbf{b}. \tag{10}$$

We take steps opposite the direction of the gradient at each point until we reach the minimum of $f(\mathbf{x})$. We ensure there is a minimum to reach because $\mathbf{A}$ is positive definite. This is also the same idea used in Newton's method for solving differential equations. The biconjugate gradient method is the same as CG but generalized to not require a symmetric positive definite $\mathbf{A}$ matrix and finally BiCGSTAB is BICO but with a subroutine of GMRES to stabilize performance.

In practice, the algorithm of BiCGSTAB is:

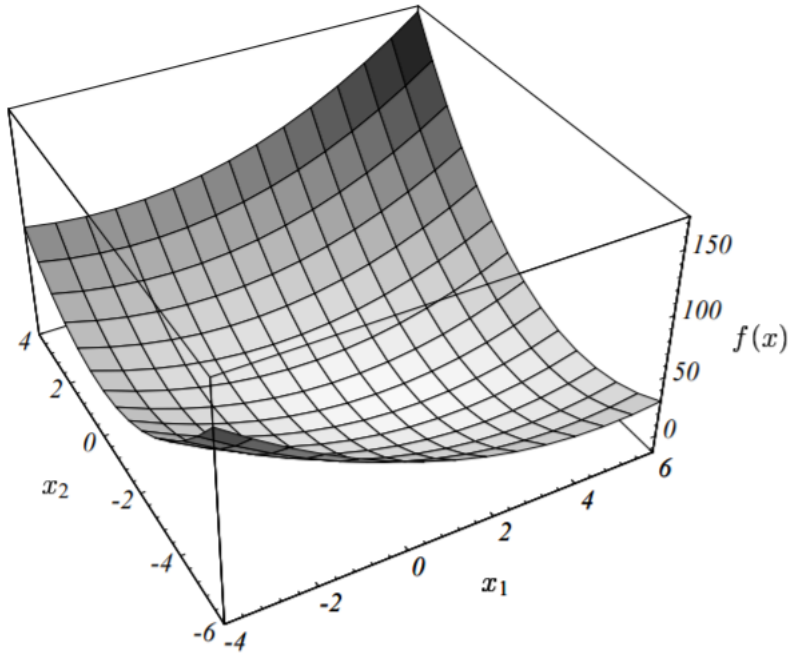

Figure 2: Quadratic form of linear system.

- Start:
  1. $\mathbf{r}_0 = \mathbf{b} - \mathbf{A} \cdot \mathbf{x}_0$.
  2. Choose $\hat{\mathbf{r}}_0 = \mathbf{r}_0$, as long as $(\hat{\mathbf{r}}_0 \cdot \mathbf{r}_0) \neq 0$.
  3. Compute $\rho_0 = \alpha = \omega_0 = 1$.
  4. Compute $\mathbf{v}_0 = \mathbf{p}_0 = \mathbf{0}$.
- Iterate: For $i = 1, 2, \ldots, i_{\max}$,
  1. $\rho_i = (\hat{\mathbf{r}}_0 \cdot \mathbf{r}_{i-1})$.
  2. $\beta = (\rho_i / \rho_{i-1})\,(\alpha / \omega_{i-1})$.
  3. $\mathbf{p}_i = \mathbf{r}_{i-1} + \beta\left(\mathbf{p}_{i-1} - \omega_{i-1}\mathbf{v}_{i-1}\right)$.
  4. $\mathbf{v}_i = \mathbf{A} \cdot \mathbf{p}_i$.
  5. $\alpha = \rho_i / (\hat{\mathbf{r}}_0 \cdot \mathbf{v}_i)$.
  6. $\mathbf{h} = \mathbf{x}_{i-1} + \alpha \mathbf{p}_i$.
  7. If $||\mathbf{h}|| <$ tol, then $\mathbf{x}_i = \mathbf{h}$ and exit.
  8. $\mathbf{s} = \mathbf{r}_{i-1} - \alpha \mathbf{v}_i$.
  9. $\mathbf{t} = \mathbf{A} \cdot \mathbf{s}$.
  10. $\omega_i = (\mathbf{t} \cdot \mathbf{s}) / (\mathbf{t} \cdot \mathbf{t})$.
  11. $\mathbf{x}_i = \mathbf{h} + \omega_i \mathbf{s}$.
  12. If $||\mathbf{A} \cdot \mathbf{x}_i - \mathbf{b}|| <$ tol, then exit.
  13. $\mathbf{r}_i = \mathbf{s} - \omega_i \mathbf{t}$.

3 Generalized Residual Method

After MINRES was developed in 1976, it was extended to include nonsymmetric systems which led to GMRES, the Generalized Residual Method. The basic idea is to use the Arnoldi method to construct a Krylov subspace defined as,

$$K_r = \text{span}\left\{b, Ab, A^2b, \ldots, A^rb\right\} \tag{11}$$

by creating the Krylov matrix

$$\mathbf{V}_k = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\} \tag{12}$$

where $\mathbf{v}_i$ are orthogonal basis vectors constructed from the Gram-Schmidt process. The Gram-Schmidt process is a method for orthonormalizing a set of vectors by successively subtracting projections of each vector from the remaining vector space. At the end of the subtraction process, the remaining vectors are guaranteed to be orthogonal to each other. The Arnoldi method is simply the successive construction of the Krylov matrix using the Gram-Schmidt process. Thus using this approach, the Krylov matrix will span the Krylov subspace. As we construct the Krylov matrix, we simultaneously compute the resulting upper triangular matrix representation of $\mathbf{A}$, called $\mathbf{H}$, in the Krylov matrix basis.

$$\mathbf{V}_k\mathbf{H}_k = \mathbf{A}\mathbf{V}_k \tag{13}$$

In practice, the matrix $\mathbf{H}_k$ will have an additional row of nonzero elements below the diagonal and will be called $\overline{\mathbf{H}}_k$. For example, with $k = 4$,

$$\overline{\mathbf{H}}_4 = \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14} \\ h_{21} & h_{22} & h_{23} & h_{24} \\ 0 & h_{12} & h_{33} & h_{34} \\ 0 & 0 & h_{43} & h_{44} \\ 0 & 0 & 0 & h_{54} \end{bmatrix} \tag{14}$$

and now

$$\mathbf{V}_{k+1}\overline{\mathbf{H}}_k = \mathbf{A}\mathbf{V}_k \tag{15}$$

Thus to solve the linear system $\mathbf{A}\mathbf{x}_0 - \mathbf{b} = \mathbf{r}_0$, with some guess $\mathbf{x}_0$, we want to minimize

$$\min||\mathbf{b} - \mathbf{A}\left(\mathbf{x}_0 + \mathbf{z}\right)|| = \min||\mathbf{r}_0 - \mathbf{A}\mathbf{z}|| \tag{16}$$

where the solution we want to find is $\mathbf{x} = \mathbf{x}_0 + \mathbf{z}$. Setting $\mathbf{z} = \mathbf{V}_k\mathbf{y}$ and $\beta = ||\mathbf{r}_0||$.

$$J(y) = ||\beta\mathbf{v}_1 - \mathbf{A}\mathbf{V}_k\mathbf{y}|| = ||\mathbf{V}_{k+1}\left[\beta\mathbf{e}_1 - \overline{\mathbf{H}}_k\mathbf{y}\right]|| \tag{17}$$

where $\mathbf{e}_1 = (1, 0, 0, \ldots, 0)$. Thus the solution to our linear system is,

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{V}_k\mathbf{y} \tag{18}$$

where $\mathbf{y}$ is the solution that minimizes,

$$J(y) = ||\beta\mathbf{e}_1 - \overline{\mathbf{H}}_k\mathbf{y}|| \tag{19}$$

The GMRES full algorithm is:

- Start:
  1. Choose initial guess, for example $\mathbf{x}_0 = \mathbf{0}$.
  2. Compute $\mathbf{r}_0 = \mathbf{b} - \mathbf{A} \cdot \mathbf{x}_0$.
  3. Compute $\mathbf{v}_1 = \mathbf{r}_0/||\mathbf{r}_0||$.
- Iterate: For $j = 1, 2, \ldots, k$,
  1. $h_{ij} = (\mathbf{A} \cdot \mathbf{v}_j) \cdot \mathbf{v}_i$; $i = 1, 2, \ldots, j$.

2. $\hat{\mathbf{v}}_{j+1} = \mathbf{A} \cdot \mathbf{v}_j - \sum_{i=1}^{j} h_{ij}\mathbf{v}_i$.
3. $h_{j+1j} = ||\hat{\mathbf{v}}_{j+1}||$.
4. $\mathbf{v}_{j+1} = \hat{\mathbf{v}}_{j+1}/h_{j+1j}$.

- Solve:

1. $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{V}_k \cdot \mathbf{y}_k$.
   - where $\mathbf{V}_k = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$.
   - and $\mathbf{y}_k$ is the solution to: $||\beta\mathbf{e}_1 - \overline{\mathbf{H}}_k \cdot \mathbf{y}||$.

One important extra step is the solution of $||\beta\mathbf{e}_1 - \overline{\mathbf{H}}_k\mathbf{y}||$ which is computed very cleverly by a QR decomposition because of the upper Hessenberg structure of $\overline{\mathbf{H}}_k$. The idea is to successively factor the each computed column of $\overline{\mathbf{H}}_k$ for each j iteration. In practice, for $j = 1$

$$\overline{\mathbf{H}}_1 = \begin{bmatrix} h_{11} \\ h_{21} \end{bmatrix} \tag{20}$$

We can turn this into an "upper triangular" matrix by applying the unitary rotation matrix,

$$\mathbf{F}_1 = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{21}$$

where $\cos\theta = h_{21}/\sqrt{h_{11}^2 + h_{21}^2}, \sin\theta = -h_{11}/\sqrt{h_{11}^2 + h_{21}^2}$.

$$\mathbf{F}_1\overline{\mathbf{H}}_1 = \begin{bmatrix} \eta_1 \\ 0 \end{bmatrix} \tag{22}$$

For $j = 2$,

$$\overline{\mathbf{H}}_2 = \begin{bmatrix} \eta_1 & h_{12} \\ 0 & h_{22} \\ 0 & h_{32} \end{bmatrix} \tag{23}$$

$$\mathbf{F}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \tag{24}$$

where $\cos\theta = h_{32}/\sqrt{h_{22}^2 + h_{32}^2}, \sin\theta = -h_{22}/\sqrt{h_{22}^2 + h_{32}^2}$.

$$\mathbf{F}_2\mathbf{F}_1\overline{\mathbf{H}}_2 = \begin{bmatrix} \eta_1 & h_{12} \\ 0 & \eta_2 \\ 0 & 0 \end{bmatrix} \tag{25}$$

The process is continued on $\overline{\mathbf{H}}_k$ until it is fully upper triangular and $\mathbf{y}$ can be directly solved for.