Bretherton – Amath 585

The Conjugate Gradient Method: Supplement to RJL 4.3.3

The conjugate gradient (CG) method is an iterative method for solving $A\mathbf{u} = \mathbf{f}$ when A is a sparse, positive definite $m \times m$ matrix. This type of problem arises commonly in FDA and FEM discretizations of Poisson's equation or other elliptic BVPs.

This is a summary and suppliement to the discussion of CG in RJL, which is a bit lengthy and skips some key points. Like steepest descents, the strategy is to minimize the functional

$$\phi(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T A\mathbf{u} - \mathbf{u}^T\mathbf{f} \tag{1}$$

Both CG and steepest descents can be applied without modification if A is negative definite rather than positive definite, in which case this is a maximization problem.

Visualizing the functional ϕ

The minimum is at the exact solution $\mathbf{u}^*$ satisfying $A\mathbf{u}^* = \mathbf{f}$. Let the error of some general $\mathbf{u}$ from $\mathbf{u}^*$ be

$$\delta = \mathbf{u} - \mathbf{u}^*. \tag{2}$$

Then, noting $\mathbf{u}^{*T}A = (A^T\mathbf{u}^*)^T = (A\mathbf{u}^*)^T = f^T$,

$$\begin{aligned}
\phi(\mathbf{u}) &= \frac{1}{2}(\delta + (\mathbf{u}^*)^T A(\delta + (\mathbf{u}^*) - (\delta + \mathbf{u}^*)^T\mathbf{f} \\
&= \frac{1}{2}(\delta^T A\delta + \delta^T A\mathbf{u}^* + \mathbf{u}^{*T}A\delta + \mathbf{u}^{*T}A\mathbf{u}^*) - \delta^T\mathbf{f} - \mathbf{u}^{*T}\mathbf{f} \\
&= \frac{1}{2}(\delta^T A\delta + \delta^T\mathbf{f} + \mathbf{f}^T\delta + \mathbf{u}^{*T}\mathbf{f}) - \delta^T\mathbf{f} - \mathbf{u}^{*T}\mathbf{f} \\
&= \frac{1}{2}\delta^T A\delta + C, \qquad C = -\frac{1}{2}\mathbf{u}^{*T}\mathbf{f}
\end{aligned} \tag{3}$$

Since A is spd, it has the diagonalization $A = E\Lambda E^T$, where E is the matrix whose columns are the the eigenvectors corresponding to its eigenvalues λ_p , and $\Lambda = \text{diag}(\lambda_p)$. Setting $\nu = E^T\delta$, with components ν_p, we can write

$$\phi(\mathbf{u}) = \frac{1}{2}\sum_{p=1}^{m]}\lambda_p\nu_p^2 + C \tag{4}$$

This implies that ϕ is a paraboloidal function of $\mathbf{u}$ centered on $\mathbf{u}^*$ and that the isosurfaces of ϕ are ellipsoids with principal axes along the eigenvectors. The longest axis of the ellipsoid corresponds to the smallest eigenvalue λ_1 and the shortest axis of the ellipsoid corresponds to the largest eigenvalue λ_m. The maximum ratio between the longest and shortest axis of an ellipsoidal isosurface of ϕ is equal to the condition number $\kappa = \lambda_m/\lambda_1$ of A.

If $\mathbf{u}$ is restricted to any subspace, the isosurfaces of ϕ within this subspace will also be ellipsoidal, with a unique $\hat{\mathbf{u}}$ that minimizes ϕ over the subspace.

Use of A-conjugate search directions

Through sparse matrix multiplications $A\mathbf{u}$, we want to discover and make use of the structure of ϕ as we iterate toward a minimum, and to do so more efficiently than using steepest descents. Rather than using a downgradient search direction, CG makes use of the following key realization. Let $\mathbf{p}_{k-1}$ be the search direction at iteration $k-1$ and let $\mathbf{u}_k$ be the point along this search direction which minimizes ϕ. At this point, $\mathbf{p}_{k-1}$ must be tangent to the ϕ isosurface. Thus, the downgradient direction, which is along the residual $\mathbf{r}_k = f - A\mathbf{u}_k$, must be orthogonal to $\mathbf{p}_{k-1}$.

The ideal new search direction would be exactly in the direction $\mathbf{u}^* - \mathbf{u}_k$. We don't know $\mathbf{u}^*$. However, we do know that

$$\begin{aligned} 0 &= \mathbf{p}_{k-1}^T\mathbf{r}_k \\ &= \mathbf{p}_{k-1}^T(\mathbf{f} - A\mathbf{u}_k) \\ &= \mathbf{p}_{k-1}^T A(\mathbf{u}^* - \mathbf{u}_k) \qquad (5) \end{aligned}$$

That is, the ideal search direction is *A-conjugate* to the prior search direction $\mathbf{p}_{k-1}$. Although we don't know this ideal search direction, this motivates always choosing a search direction $\mathbf{p}_k$ that is A-conjugate to the prior search direction $\mathbf{p}_{k-1}$.

Now suppose that starting with an initial guess $\mathbf{u}_0$, we could somehow sequentiallly define a set of search directions $\mathbf{p}_k$ for line minimization of ϕ such that each new search direction is A-conjugate to all the prior search directions $\mathbf{p}_j, j = 0..., k-1$. If we let S_k be the k-dimensional subspace that includes the current and all prior iterates $\mathbf{u}_j, j = 0, ..., k$, then we show below that $\mathbf{u}_k$ will minimize ϕ over that entire subspace (not just along the search line). Thus we are guaranteed to reach the exact solution in m iterations, when we will have minimized ϕ over the entire m-dimensional space R^m.

The proof is by induction. For $k = 1$, S_1 consists of the single search direction $\mathbf{p}_1$ away from the initial guess $\mathbf{u}_0$, and $\mathbf{u}_1$ is constructed to minimize ϕ along this line, Now assume that $\mathbf{u}_{k-1}$ minimizes ϕ over the subspace S_{k-1}. Also assume the new search direction $\mathbf{p}_{k-1}$ is A-conjugate to all the prior search directions $\mathbf{p}_j, j = 0..., k-2$. Then we must prove $\mathbf{u}_k$ minimizes ϕ over the subspace S_k.

To show this, it suffices to show that $-\nabla\phi(\mathbf{u}_k)$ has no projection into S_k, i. e. that $\mathbf{r}_k = -\nabla\phi(\mathbf{u}_k)$ is orthogonal to a set of k independent basis vectors that define S_k. One such set is the search directions $\mathbf{p}_j, j = 0, ..., k-1$. Thus, we will show that

$$0 = \mathbf{p}_j^T \mathbf{r}_k, \qquad j = 0, ..., k-1$$

This claim can be verified as follows. Because of the line minimization, $\mathbf{r}_k$ is orthogonal to $\mathbf{p}_{k-1}$. Since $\mathbf{u}_{k-1}$ minimizes ϕ over the subspace S_{k-1},

$$0 = \mathbf{p}_j^T \mathbf{r}_{k-1}, \qquad j = 0, ..., k-2$$

Hence, for $j = 0, ..., k-2$,

$$\begin{aligned} \mathbf{p}_j^T \mathbf{r}_k &= \mathbf{p}_j^T \mathbf{r}_{k-1} + \mathbf{p}_j^T (\mathbf{r}_k - \mathbf{r}_{k-1}) \\ &= 0 - \mathbf{p}_j^T A (\mathbf{u}_k - \mathbf{u}_{k-1}) \\ &= -\alpha_{k-1} \mathbf{p}_j^T A \mathbf{p}_{k-1} = 0 \end{aligned} \tag{6}$$

by the assumed A-conjugacy of the search directions. This shows $\mathbf{u}_k$ minimizes ϕ over the subspace S_k and completes the induction step.

The CG algorithm is a simple way of choosing the successive search directions to have this A-conjugacy property.

The CG iteration

Starting at the initial guess $\mathbf{u}_0$, we choose an initial search direction $\mathbf{p}_0 = \mathbf{r}_0$ down the gradient of ϕ.

For each succeeding iteration $k = 1, 2, ..., m$, loop through the following steps:

1. Find the α_{k-1} for which $\mathbf{u}_k = \mathbf{u}_{k-1} + \alpha_{k-1}\mathbf{p}_{k-1}$ minimizes ϕ along the search path $\mathbf{p}_{k-1}$.

2. Calculate the residual $\mathbf{r}_k$

3. Declare convergence and exit loop if $\mathbf{r}_k$ is small enough.

4. Otherwise, use search direction $\mathbf{p}_k = \mathbf{r}_k + \beta_{k-1}\mathbf{p}_{k-1}$ with β_{k-1} chosen to make $\mathbf{p}_k$ A-conjugate to $\mathbf{p}_{k-1}$

What we need to show is that this choice of $\mathbf{p}_k$ is also A-conjugate to all the previous search directions $\mathbf{p}_j, j = 0, ..., k-2$. Consider the expressions for the residual and the new search direction,

$$\begin{aligned} \mathbf{r}_j &= \mathbf{f} - A\mathbf{u}_j = \mathbf{f} - A(\mathbf{u}_{j-1} - \alpha_{j-1}p_{j-1}) = \mathbf{r}_{j-1} - \alpha_{j-1}Ap_{j-1} && (7) \\ \mathbf{p}_j &= \mathbf{r}_j + \beta_{j-1}\mathbf{p}_{j-1} && (8) \end{aligned}$$

Starting with $j = 0$, for which $\mathbf{p}_0 = \mathbf{r}_0$, (7) implies $\mathbf{r}_1$ is a linear combination of $\mathbf{r}_0$ and $A\mathbf{r}_0$, then (8) implies this is also true for $\mathbf{p}_1$. Iterating in j, we deduce that $\mathbf{r}_j$ and $\mathbf{p}_j$ are each linear combinations (i. e. in the span) of $\mathbf{r}_0, ..., A^j\mathbf{r}_0$. This type of subspace of R^m generated by increasing powers of A acting on a vector is called a *Krylov space*.

With this background, we use induction to prove $\mathbf{p}_k$ is A-conjugate to all the previous search directions $\mathbf{p}_j, j = 0, ..., k-1.$. For $k = 1$, $\mathbf{p}_1$ is A-conjugate to the only previous search direction $\mathbf{p}_0$ by construction. Assume that $\mathbf{p}_{k-1}$ is A-conjugate to all the previous search directions $\mathbf{p}_j, j = 0, ..., k-2$. Then by (8), for each of these j's,

$$\mathbf{p}_k^T A\mathbf{p}_j = \mathbf{r}_k^T A\mathbf{p}_j + \beta_{k-1}\underbrace{\mathbf{p}_{k-1}^T A\mathbf{p}_j}_{0} \qquad (9)$$

Thus to show $\mathbf{p}_k$ is A-conjugate to each $\mathbf{p}_j$, it suffices to show that the residual $\mathbf{r}_k$ is orthogonal to $A\mathbf{p}_j$.

Now $A\mathbf{p}_j$ is in the span of $A\mathbf{r}_0, ..., A^{j+1}\mathbf{r}_0$, which is a subspace of the span of $\mathbf{r}_0, A\mathbf{r}_0, ..., A^{k-1}\mathbf{r}_0$, which is also the span of $\mathbf{p}_0, ..., \mathbf{p}_{k-1}$. By the argument in the previous section, since $\mathbf{u}_k$ minimizes ϕ over this subspace, the residual $\mathbf{r}_k = -\nabla\phi(\mathbf{u}_k)$ must be orthogonal to all the $\mathbf{p}_j$. This shows that $\mathbf{p}_k$ is A-conjugate to the search directions $\mathbf{p}_j, j = 0, ..., k-2$. By construction, it is also A-conjugate to $\mathbf{p}_{k-1}$, so the induction step is proved.

Computation of α_{k-1} and β_{k-1}

We choose α_{k-1} to minimize ϕ along the line $\mathbf{u}_k = \mathbf{u}_{k-1} + \alpha\mathbf{p}_{k-1}$. Defining $\mathbf{w}_{k-1} = A\mathbf{u}_{k-1}$, this gives RJL (4.40):

$$\alpha_{k-1} = \frac{\mathbf{p}_{k-1}^T\mathbf{r}_{k-1}}{\mathbf{p}_{k-1}^T\mathbf{w}_{k-1}} \tag{10}$$

The numerator can be simplified by noting $\mathbf{p}_{k-1} = \mathbf{r}_{k-1} + \beta_{k-2}\mathbf{p}_{k-2}$:

$$\mathbf{p}_{k-1}^T\mathbf{r}_{k-1} = \mathbf{r}_{k-1}^T\mathbf{r}_{k-1} + \beta_{k-2}\underbrace{\mathbf{p}_{k-2}^T\mathbf{r}_{k-1}}_{0} \tag{11}$$

We choose β_{k-1} to make $\mathbf{p}_k = \mathbf{r}_k - \beta_{k-1}\mathbf{p}_{k-1}$ A-conjugate to $\mathbf{p}_{k-1}$:

$$\beta_{k-1} = -\frac{\mathbf{r}_k^T A\mathbf{p}_{k-1}}{\mathbf{p}_{k-1}^T A\mathbf{p}_{k-1}} \tag{12}$$

Recalling

$$\alpha_{k-1}A\mathbf{p}_{k-1} = A(\mathbf{u}_k - \mathbf{u}_{k-1}) = -(\mathbf{r}_k - \mathbf{r}_{k-1}) \tag{13}$$

and $\mathbf{r}_k^T\mathbf{r}_{k-1} = \mathbf{r}_k^T\mathbf{p}_{k-1} = 0$, this can be simplified to the form

$$\beta_{k-1} = -\frac{\mathbf{r}_k^T(\mathbf{r}_k - \mathbf{r}_{k-1})}{\mathbf{p}_{k-1}^T(\mathbf{r}_k - \mathbf{r}_{k-1})} = \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{p}_{k-1}^T\mathbf{r}_{k-1}} = \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}} \tag{14}$$

The Matlab script `CG.m` on the class web page implements these forms (10) (with the simplification (11)) and (14).

Convergence rate of CG

Although CG is only guaranteed to converge in m iterations, for most A's it converges much faster. RJL 4.3.4 gives some theory that suggests that it typically converges to adequate tolerance in $O(\kappa^{1/2})$ iterations, where κ is the condition number of A. If $\kappa \gg 1$ this is much faster than the $O(\kappa)$ iterations required for convergence of steepest descents. For a FDA or FEM to a Poisson problem in one or more dimensions, $\kappa = O(m^2)$ so CG will converge in $O(m)$ iterations.

The convergence of CG can be improved by *preconditioning* the matrix A to reduce its condition number. RJL 4.3.5-6 discusses some popular choices, including use of an *incomplete Cholesky decomposition* of A.