An Introduction to the Conjugate Gradient Method Without the Agonizing Pain

abbreviated from note by

Jonathan Richard Shewchuk,
School of Computer Science,
Carnegie Mellon University,
Pittsburgh, PA 15213

by Chen-Shiung Hsue, National Tsing-Hua Univ. Taiwan

1 Introduction

The Conjugate Gradient Method (CG) is the most popular iteration method for solving large systems of linear equations. CG is effective for systems of the form

$$A \cdot \mathbf{x} = \mathbf{b} \tag{1}$$

where $\mathbf{x}$ is an unknown vector, $\mathbf{b}$ is a known vector, and A is a known, square, symmetric, positive-definite matrix. A matrix A is positive-definite, if, for every nonzero vector $\mathbf{x}$,

$$\mathbf{x} \cdot A \cdot \mathbf{x} = \sum_{i,j=1}^{n} x_i A_{ij} x_j > 0 \tag{2}$$

Iterative methods like CG are suited for use with sparse matrices. If A is dense, your best course of action is probably to factor A and solve the equation by back substitution. The time spent factoring a dense A is

roughly equivalent to the time spent solving the system iteratively; and once A is factored, the system can be back-solved quickly for multiple values of $\mathbf{b}$.

2 The Quadratic Form

A quadratic form is simply a scalar, quadratic function of a vector with the form

$$f(\mathbf{x}) = \frac{1}{2}\,\mathbf{x} \cdot A \cdot \mathbf{x} - \mathbf{b} \cdot x + c \tag{3}$$

For A positive-definite, the surface defined by $f(\mathbf{x})$ is shaped like a paraboloid bowl. Further, $f(\mathbf{x})$ is minimized by the solution to $A \cdot \mathbf{x} = \mathbf{b}$. It is readily shown that

$$\nabla f(\mathbf{x}) = \frac{1}{2}A^T \cdot \mathbf{x} + \frac{1}{2}A \cdot \mathbf{x} - \mathbf{b} \tag{4}$$

If A symmetric, this equation reduces to

$$\nabla f(\mathbf{x}) = A \cdot \mathbf{x} - \mathbf{b} \tag{5}$$

Setting the gradient to zero, we obtain Equation (1), the linear system we wish to solve. Therefore, the solution to $A \cdot \mathbf{x} = \mathbf{b}$ is a critical point of $f(\mathbf{x})$. If A is positive-definite as well as symmetric, then this solution is a minimum of $f(\mathbf{x})$, so $A \cdot \mathbf{x} = \mathbf{b}$ can be solved by finding an $\mathbf{x}$ that minimizes $f(x)$.

Consider the relation between f at some arbitrary point $\mathbf{p}$ and at the solution point $\mathbf{x} = A^{-1} \cdot \mathbf{b}$. From Equation (3) one can show that if A is symmetric,

$$f(\mathbf{p}) = f(\mathbf{x}) + \frac{1}{2}(\mathbf{p} - \mathbf{x}) \cdot A \cdot (\mathbf{p} - \mathbf{x}) \tag{6}$$

If A is positive-definite as well, then the latter term on the right hand side is positive. It follows that $\mathbf{x}$ is a global minimum of f.

3 The Method of Steepest Descent

In the method of Steepest Descent, we start at an arbitrary point $\mathbf{x}_{(0)}$ and slide down to the bottom of the paraboloid. We take a series of steps

$\mathbf{x}_{(0)}, \mathbf{x}_{(1)}, \ldots$ until we are satisfied that we are close enough to the solution $\mathbf{x}$. In each step, we choose the direction in which f decreases most quickly, which is the direction opposite to $\nabla f(\mathbf{x}_{(i)})$. This direction is

$$\nabla f(\mathbf{x}_{(i)}) = \mathbf{b} - A\mathbf{x}_{(i)}.$$

Let us introduce a few definitions here. The *error*

$$\mathbf{e}_{(i)} = \mathbf{x}_{(i)} - \mathbf{x}$$

is a vector that indicates how far we are from the solution. The *residual*

$$\mathbf{r}_{(i)} = \mathbf{b} - A \cdot \mathbf{x}_{(i)}$$

indicates how far we are from the correct value of $\mathbf{b}$. It is easy to see that

$$\mathbf{r}_{(i)} = -A\mathbf{e}_{(i)}$$

and you should think of the residue as being the error transformed by A into the same space as $\mathbf{b}$. More importantly,

$$\mathbf{r}_{(i)} = \nabla f(\mathbf{x}_{(i)}),$$

and you should also think of the residual as the direction of steepest descent. For nonlinear problems, as discussed later, only the latter definition applies. So remember, whenever you read "residual", think "direction of steepest descent."

As our first step, we will choose a point along the direction of steepest descent.

$$\mathbf{x}_{(1)} = \mathbf{x}_{(0)} + \alpha\ \mathbf{r}_{(0)} \tag{7}$$

The question is, how big a step should we take?

A *line search* is a procedure that chooses α to minimize f along a line. From basic calculus, α minimizes f when the *directional derivative* $\frac{d}{d\alpha} f(x_{(1)})$ is equal to zero. By the chain rule, this is $\nabla f(x_{(1)}) \cdot \mathbf{r}_{(0)}$. Setting

this expression to zero, we find that α should be chosen so that $\mathbf{r}_{(0)}$ and ∇f are orthogonal.

To determine α, note that $\nabla f(x_{(1)}) = -\mathbf{r}_{(0)}$, and we have

$$\begin{aligned}
\mathbf{r}_{(1)} \cdot \mathbf{r}_{(0)} &= 0 \\
(\mathbf{b} - A \cdot \mathbf{x}_{(1)}) \cdot \mathbf{r}_{(0)} &= 0 \\
(\mathbf{b} - A \cdot (\mathbf{x}_{(0)} + \alpha \mathbf{r}_{(0)})) \cdot \mathbf{r}_{(0)} &= 0 \\
(\mathbf{b} - A \cdot \mathbf{x}_{(0)}) \cdot \mathbf{r}_{(0)} - \alpha (\mathbf{r}_{(0)} \cdot A) \cdot \mathbf{r}_{(0)} &= 0 \\
(\mathbf{b} - A \cdot \mathbf{x}_{(0)}) \cdot \mathbf{r}_{(0)} &= \alpha (\mathbf{r}_{(0)} \cdot A) \cdot \mathbf{r}_{(0)} \\
r_{(0)} \cdot \mathbf{r}_{(0)} &= \alpha \ \mathbf{r}_{(0)} \cdot A \cdot \mathbf{r}_{(0)} \\
\alpha &= \frac{\mathbf{r}_{(0)} \cdot \mathbf{r}_{(0)}}{\mathbf{r}_{(0)} \cdot A \cdot \mathbf{r}_{(0)}} \ .
\end{aligned}$$

Putting it all together, the method of Steepest Descent is:

$$\mathbf{r}_{(i)} = \mathbf{b} - A \cdot \mathbf{x}_{(i)} \ , \tag{8}$$

$$\alpha_i = \frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\mathbf{r}_{(i)} \cdot A \cdot \mathbf{r}_{(i)}} \ . \tag{9}$$

$$\tag{10}$$

$$\mathbf{x}_{(i+1)} = \mathbf{x}_{(i)} + \alpha_i \mathbf{r}_{(i)} \tag{11}$$

The procedure continues until it converges. Note the zigzag path, which appears because each gradient is orthogonal to the previous gradient.

The algorithm, as written above, requires two matrix vector multiplications per iteration. The computational cost of Steepest Descent is dominated by matrix-vector products: fortunately, one can be eliminated. By premultiplying both sides of Equation (11) by $-A$ and adding $\mathbf{b}$, we have

$$\mathbf{r}_{(i+1)} = \mathbf{r}_{(i)} - \alpha_i A \cdot \mathbf{r}_{(i)}. \tag{12}$$

Although Equation (8) is still needed to compute $\mathbf{r}_{(0)}$. Equation (12) can be used for every iteration thereafter. The product $A \cdot \mathbf{r}$, which occurs in both Equation (9) and (12) , need only be computed once. The disadvantage

of using this recurrence is that the sequence defined by Equation (12) is generated without any feedback from the value of $\mathbf{x}_{(i)}$, so that accumulation of floating point roundoff error may cause $\mathbf{x}_{(i)}$ to converge to some point near $\mathbf{x}$. This effect can be avoided by periodically using Equation (8) to recompute the correct residual.

4 Thinking with Eigenvectors and Eigenvalues

If B is symmetric, then there exists a set of n linearly independent eigenvectors of B, denoted $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ with the corresponding eigenvalues, denoted $\lambda_1, \lambda_2, \ldots, \lambda_n$. This set of eigenvectors is not unique, because each eigenvector can be scaled by an arbitrary nonzero constant.

Any vector $\mathbf{x}$ can be decomposed as a linear superposition of the eigenvectors

$$\mathbf{x} = \sum_i v_i$$

Repeated applying B to $\mathbf{x}$ yields

$$B^l\mathbf{x} = \sum_i B^l\mathbf{v}_i = \sum_i \lambda_i^l\mathbf{v}_i \ .$$

If the magnitudes of *all* the eigenvalues are smaller than one, $B^l\mathbf{x}$ will converge to zero. If one of the eigenvalues has magnitude greater than one, $\mathbf{x}$ will diverge to infinity. This is why numerical analysts attach importance to the *spectral radius* $\rho(B)$ of a matrix:

$$\rho(B) = max|\lambda_i|, \quad \lambda_i \text{ is an eigenvalue of } \ B$$

If we want $B^l\mathbf{x}$ to converge to zero quickly, $\rho(B)$ should be less than one, and preferably as small as possible.

Here is a useful fact: the eigenvalues of a positive-definite matrix are all positive.

4.1 Jacobi iterations

Consider the Jacobi Method for solving $A \cdot \mathbf{x} = \mathbf{b}$. The matrix A is split into two parts: D, whose diagonal elements are identical to those of A, and whose off-diagonal elements are zero, and E, whose diagonal elements are zero, and whose off-diagonal elements are identical to those of A. Thus, $A = D + E$. We derive the Jacobi Method:

$$\begin{aligned} A \cdot \mathbf{x} &= \mathbf{b} \\ D \cdot \mathbf{x} &= -E \cdot \mathbf{x} + \mathbf{b} \\ \mathbf{x} &= -D^{-1}E \cdot \mathbf{x} + D^{-1} \cdot \mathbf{b} \\ \mathbf{x} &= B \cdot \mathbf{x} + z, \quad \text{where } B = -D^{-1}E, \quad \text{and } \mathbf{z} = D^{-1} \cdot \mathbf{b} \end{aligned} \tag{13}$$

Because D is diagonal, it is easy to invert. This identity can be converted into an iterative method by forming the recurrence

$$\mathbf{x}_{(i+1)} = B \cdot \mathbf{x}_{(i)} + \mathbf{z} \tag{14}$$

Given a starting vector $\mathbf{x}_{(0)}$ this formula generates a sequence of vectors. Our hope is that each successive vector will be closer to the solution $\mathbf{x}$ than the last. $\mathbf{x}$ is called a *stationary point* of Equation (14), because if $\mathbf{x}_{(i)} = \mathbf{x}$, then $\mathbf{x}_{(i+1)}$ will also equal $\mathbf{x}$.

This derivation may seem quite arbitrary. We could have formed any number of identities for $\mathbf{x}$ instead of Equation (13). In fact, simply by splitting A differently - that is, by choosing a different D and E - we could have derived the Gauss-Seidel method, or the method of Successive Over-Relaxation (SOR). Our hope is that we have chosen a splitting for which B has a small spectral radius. The choice of the Jacobi splitting is for simplicity.

Express each iterate $\mathbf{x}_{(i)}$ as the sum of the exact solution $\mathbf{x}$ and the error term $\mathbf{e}_{(i)}$. Then, Equation (14) becomes

$$\mathbf{x}_{(i+1)} = B \cdot \mathbf{x}_i + \mathbf{z}$$

$$
\begin{aligned}
&= B \cdot (\mathbf{x} + \mathbf{e}_i) + \mathbf{z} \\
&= B \cdot \mathbf{x} + \mathbf{z} + B \cdot \mathbf{e}_i \\
&= \mathbf{x} + B \cdot \mathbf{e}_i \quad \text{(by Equation (13))}, \\
\mathbf{e}_{i+1} &= B \cdot \mathbf{e}_i. \qquad (15)
\end{aligned}
$$

Each iteration does not affect the "correct part" of $\mathbf{x}_{(i)}$ (because $\mathbf{x}$ is a stationary point); but each iteration does affect the error term. It is apparent from equation (15) that if $\rho(B) < 1$, then the error term $\mathbf{e}_{(i)}$ will converge to zero as i approaches infinity. Hence, the initial vector $\mathbf{x}_{(0)}$ has no effect on the inevitable outcome!

Thus the spectral radius $\rho(B)$ determines the speed of convergence. Suppose that $\mathbf{v}_j$ is the eigenvector of B with the largest eigenvalue (so that $\rho(B) = \lambda_j$). If the initial error $\mathbf{e}_{(0)}$, expressed as a linear combination of eigenvectors, includes a component in the direction of $\mathbf{v}_j$, this component will be the slowest to converge.

B is not generally symmetric (even if A is); and may even be defective. However, the rate of convergence of the Jacobi Method depends largely on $\rho(B)$, which depends on A. Unfortunately, the Jacobi Method does not converge for every A, or even for every positive-definite A.

5 Convergence Analysis of Steepest Descent

5.1 Instant Results

Let's first consider the case where $\mathbf{e}_{(i)}$ is an eigenvector with eigenvalue λ_e. Then, the residual $\mathbf{r}_{(i)} = -A \cdot \mathbf{e}_{(i)} = -\lambda_e \mathbf{e}_{(i)}$ is also an eigenvector. Thus

$$
\begin{aligned}
\mathbf{e}_{(i+1)} &= \mathbf{e}_{(i)} + \frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\mathbf{r}_{(i)} \cdot A \cdot \mathbf{r}_{(i)}} \mathbf{r}_{(i)} \\
&= \mathbf{e}_{(i)} + \frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\lambda_e \mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}} (-\lambda_e \mathbf{e}_{(i)}) \\
&= 0.
\end{aligned}
$$

The point $\mathbf{x}_{(i)}$ lies on one of the axes of the ellipsoid, and so the residual points directly to the center of the ellipsoid. Choosing $\alpha_i = \lambda^{-1}$ gives us instant convergence.

For a more general analysis, we must express $\mathbf{e}_{(i)}$ as a linear combination of eigenvectors. If A is symmetric, we can choose the eigenvectors $\mathbf{v}_i$ to be orthonormal:

$$\mathbf{v}_i \cdot \mathbf{v}_j = \delta_{i,j} \tag{16}$$

We can now express the error term as a linear combination of eigenvectors:

$$\mathbf{e}_{(i)} = \sum_{j=1}^{n} \xi_j \mathbf{v}_j, \tag{17}$$

We then have the following identities:

$$\begin{aligned}
\mathbf{r}_{(i)} &= -A \cdot \mathbf{e}_{(i)} = -\sum_{j=1}^{n} \xi_j \lambda_j \mathbf{v}_j, && (18)\\
|\mathbf{e}_{(i)}|^2 = \mathbf{e}_{(i)} \cdot \mathbf{e}_{(i)} &= \sum_{j=1}^{n} \xi_j^2, && (19)\\
\mathbf{e}_{(i)} \cdot A \cdot \mathbf{e}_{(i)} &= \sum_{j=1}^{n} \xi_j^2 \lambda_j, && (20)\\
|\mathbf{r}_{(i)}|^2 = \mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)} &= \sum_{j=1}^{n} \xi_j^2 \lambda_j^2, && (21)\\
\mathbf{r}_{(i)} \cdot A \cdot \mathbf{r}_{(i)} &= \sum_{j=1}^{n} \xi_j^2 \lambda_j^3, && (22)
\end{aligned}$$

Equation (18) shows that $\mathbf{r}_{(i)}$ too can be expressed as a sum of eigenvector components, and the length of these components are $-\xi_j \lambda_j$.

Now we can proceed with the analysis.

$$\begin{aligned}
\mathbf{e}_{(i+1)} &= \mathbf{e}_{(i)} + \frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\mathbf{r}_{(i)} \cdot A \cdot \mathbf{r}_{(i)}} \mathbf{r}_{(i)} \\
&= \mathbf{e}_{(i)} + \frac{\sum_{j=1}^{n} \xi_j^2 \lambda_j^2}{\sum_{j=1}^{n} \xi_j^2 \lambda_j^3} \mathbf{r}_{(i)} && (23)
\end{aligned}$$

We saw in the last example that, if $\mathbf{e}_{(i)}$ has only one eigenvector component, then convergence is achieved in one step by choosing $\alpha_i = \lambda^{-1}$. Now let's

examine the case where all the eigenvalues are equal to λ. We have

$$\begin{aligned}\mathbf{e}_{(i+1)} &= \mathbf{e}_{(i)} + \frac{\lambda^2 \sum_{j=1}^n \xi_j^2}{\lambda^3 \sum_{j=1}^n \xi_j^2} \ (-\lambda \ \mathbf{e}_{(i)}) \\ &= 0\end{aligned}$$

Once again, there is instant convergence. Because all the eigenvalues are equal, the ellipsoid is spherical; hence, no matter what point we start at, the residual must point to the center of the sphere.

However, if there are several unequal, nonzero eigenvalues, then no choice of α_i will eliminate all the eigenvector components, and our choice becomes a sort of compromise.

5.2 General Convergence

We shall define the *energy norm*:

$$||e||_A = (\mathbf{e} \cdot A \cdot e)^{1/2}$$

Examination shows that minimizing $||e||_A$ is equivalent to minimizing $f(\mathbf{x}_{(i)})$. with this norm we have

$$\begin{aligned}||e_{(i+1)}||_A^2 &= \mathbf{e}_{(i+1)} \cdot A \cdot \mathbf{e}_{(i+1)} \\ &= (\mathbf{e}_{(i)} + \alpha_i \ \mathbf{r}_{(i)}) \cdot A \cdot (\mathbf{e}_{(i)} + \alpha_i \ \mathbf{r}_{(i)}) \\ &= \mathbf{e}_{(i)} \cdot A \cdot \mathbf{e}_{(i)} + 2\alpha_i \ \mathbf{r}_{(i)} \cdot A \cdot \mathbf{e}_{(i)} + \alpha_i^2 \ \mathbf{r}_{(i)} \cdot A \cdot \mathbf{r}_{(i)} \\ &= ||e_{(i)}||_A^2 + 2 \ \frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\mathbf{r}_{(i)} \cdot A \cdot \mathbf{r}_{(i)}} \ (-\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}) + \left(\frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\mathbf{r}_{(i)} \cdot A \cdot \mathbf{r}_{(i)}} \right)^2 \ \mathbf{r}_{(i)} \cdot A \cdot \mathbf{r}_{(i)} \\ &= ||e_{(i)}||_A^2 - \frac{(\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)})^2}{\mathbf{r}_{(i)} \cdot A \cdot \mathbf{r}_{(i)}} \\ &= ||e_{(i)}||_A^2 \ \left(1 - \frac{(\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)})^2}{(\mathbf{r}_{(i)} \cdot A \cdot \mathbf{r}_{(i)})(\mathbf{e}_{(i)} \cdot A \cdot \mathbf{e}_{(i)})} \right) \\ &= ||e_{(i)}||_A^2 \ \left(1 - \frac{\sum_{j=1}^n (\xi_j^2 \lambda_j^2)^2}{(\sum_{j=1}^n \xi_j^2 \lambda_j^3)(\sum_{j=1}^n \xi_j^2 \lambda_j)} \right) \\ &= ||e_{(i)}||_A^2 \ \omega^2\end{aligned}$$

where

$$\omega^2 = 1 - \frac{\sum_{j=1}^{n}(\xi_j^2\lambda_j^2)^2}{(\sum_{j=1}^{n}\xi_j^2\lambda_j^3)(\sum_{j=1}^{n}\xi_j^2\lambda_j)} \tag{24}$$

The analysis depends on finding an upper bound for ω. Take $n = 2$. Assume that $\lambda_1 \leq \lambda_2$. The *spectral condition number* of A is defined to be $\kappa = \lambda_1/\lambda_2 \geq 1$. The *slope* of $e_{(i)}$ (relative to the coordinate system defined by the eigenvectors), which depends on the starting point, is denoted $\mu = \xi_2/\xi_1$. We have

$$\begin{aligned} \omega^2 &= 1 - \frac{(\xi_1^2\lambda_1^2 + \xi_2^2\lambda_2^2)^2}{(\xi_1^2\lambda_1 + \xi_2^2\lambda_2)(\xi_1^2\lambda_1^3 + \xi_2^2\lambda_2^3)} \\ &= 1 - \frac{(\kappa^2 + \mu^2)^2}{(\kappa + \mu^2)(\kappa^3 + \mu^2)} \end{aligned} \tag{25}$$

An upper bound for ω(corresponding to the worst-case starting points) is found by setting $\mu^2 = \kappa^2$:

$$\omega^2 \leq 1 - \frac{4\kappa^4}{\kappa^5 + 2\kappa^4 + \kappa^3} = \frac{(\kappa - 1)^2}{(\kappa + 1)^2}$$

or

$$\omega = \frac{\kappa - 1}{\kappa + 1}. \tag{26}$$

Equation (26) can be proved to be valid for $n > 2$, where the *condition number* of a symmetric, positive-definite matrix is defined as

$$\kappa = \lambda_{max}/\lambda_{min},$$

the ratio of the largest to smallest eigenvalue. The more *ill-conditioned* the matrix(that is, the larger its condition number κ), the slower the convergence of Steepest Descent. The convergence results for Steepest Descent are

$$||e_{(i)}||_A \leq \omega = \left(\frac{\kappa - 1}{\kappa + 1}\right)^i \ ||e_{(0)}||_A \tag{27}$$

In other word,

$$\frac{f(x_{(i)}) - f(x)}{f(x_{(0)}) - f(x)} = \frac{\frac{1}{2}\ e_{(i)} \cdot A \cdot e_{(i)}}{\frac{1}{2}\ e_{(0)} \cdot A \cdot e_{(0)}} = \left(\frac{\kappa - 1}{\kappa + 1}\right)^i \ .$$

6 The Method of Conjugate Directions

6.1 Conjugacy

Steepest Descent often finds itself taking steps in the same direction as earlier steps. An obvious solution is to pick a set of orthogonal *search directions* $\mathbf{d}_{(0)}, \mathbf{d}_{(1)}, \ldots, \mathbf{d}_{(n-1)}$. In each search direction, we'll take exactly one step, After n steps, we'll be done. In general, for each step we choose a point

$$x_{(i+1)} = x_{(i)} + \alpha_i \mathbf{d}_{(i)} \ . \tag{28}$$

To find the value of α_i, use the fact that $e_{(i+1)}$ should be orthogonal to $d_{(i)}$, so that we need never step in the direction of $d_{(i)}$ again. Using this condition, we have

$$\begin{aligned} \mathbf{d}_{(i)} \cdot \mathbf{d}_{(i+1)} &= 0 \\ \mathbf{d}_{(i)} \cdot (\mathbf{e}_{(i)} + \alpha_i \ \mathbf{d}_{(i)}) &= 0 \\ \alpha_i &= -\frac{\mathbf{d}_{(i)} \cdot \mathbf{e}_{(i)}}{\mathbf{d}_{(i)} \cdot \mathbf{d}_{(i)}} \ . \end{aligned} \tag{29}$$

Unfortunately, we can't compute α_i without knowing $\mathbf{e}_{(i)}$.

The solution is to make the search directions A-orthogonal instead of orthogonal. Two vectors $\mathbf{d}_{(i)}$ and $\mathbf{d}_{(j)}$ are *A-orthogonal*, or *conjugate*, if

$$\mathbf{d}_{(i)} \cdot A \cdot \mathbf{d}_{(j)} = 0 \ .$$

Our new requirement is that $\mathbf{e}_{(i+1)}$ be A-orthogonal to $\mathbf{d}_{(i)}$. Not coincidentally, this orthogonality condition is equivalent to finding the minimum point along the search direction $\mathbf{d}_{(i)}$, as in the Method of Steepest Descent.

$$\begin{aligned} \frac{d}{d\alpha} f(\mathbf{x}_{(i+1)} &= \nabla f(\mathbf{x}_{(i+1)}) \cdot \frac{d}{d\alpha} \mathbf{x}_{(i+1)} = 0 \\ -\ \mathbf{r}_{(i+1)} \cdot \mathbf{d}_{(i)} &= 0 \\ \mathbf{d}_{(i)} \cdot A \cdot \mathbf{e}_{(i+1)} &= 0 \end{aligned}$$

Following the derivation of Equation (29), when the search directions are A-orthogonal, we have:

$$\alpha_i = -\frac{\mathbf{d}_{(i)} \cdot A \cdot \mathbf{e}_{(i)}}{\mathbf{d}_{(i)} \cdot A \cdot \mathbf{d}_{(i)}} \tag{30}$$

$$= \frac{\mathbf{d}_{(i)} \cdot \mathbf{r}_{(i)}}{\mathbf{d}_{(i)} \cdot A \cdot \mathbf{d}_{(i)}} \tag{31}$$

Unlike Equation (29), we can calculate this expression. Note that if the search vector were the residual, this formula would be identical to the formula used by Steepest Descent.

To prove that this procedure really does compute $\mathbf{x}$ in n steps, express the error term as a linear combination of search directions; namely,

$$\mathbf{e}_{(0)} = \sum_{j=0}^{n-1} \delta_j \ \mathbf{d}_{(j)} \tag{32}$$

Because the search directions are A-orthogonal, we have:

$$\mathbf{d}_{(k)} \cdot A \cdot \mathbf{e}_{(0)} = \sum_{j=0} \delta_j \ \mathbf{d}_{(k)} \cdot A \cdot \mathbf{d}_{(j)} = \delta_k \ \mathbf{d}_{(k)} \cdot A \cdot \mathbf{d}_{(k)}$$

$$\delta_k = \frac{\mathbf{d}_{(k)} \cdot A \cdot \mathbf{e}_{(0)}}{\mathbf{d}_{(k)} \cdot A \cdot \mathbf{d}_{(k)}}$$

$$= \frac{\mathbf{d}_{(k)} \cdot A \cdot (\mathbf{e}_{(0)} + \sum_{j=0}^{k-1} \alpha_j \ \mathbf{d}_{(j)})}{\mathbf{d}_{(k)} \cdot A \cdot \mathbf{d}_{(k)}} \quad \text{(by } A\text{-orthogonality of d vectors)} \tag{33}$$

$$= \frac{\mathbf{d}_{(k)} \cdot A \cdot \mathbf{e}_{(k)}}{\mathbf{d}_{(k)} \cdot A \cdot \mathbf{d}_{(k)}} \tag{34}$$

where we have used the relation

$$\mathbf{e}_{(k)} = \mathbf{x}_{(k)} - \mathbf{x} = (\mathbf{x}_{(0)} + \sum_{j=0}^{k-1} \alpha_j \ \mathbf{d}_{(j)}) - \mathbf{x} = (\mathbf{e}_{(0)} + \sum_{j=0}^{k-1} \alpha_j \ \mathbf{d}_{(j)})$$

which is a direct consequence of Equation (28). And from Equation (30), we find that :

$$\alpha_i = -\delta_i$$

This fact gives us a new way to look at the error term. The process of building up $\mathbf{x}$ component by component can also be viewed as a process of

cutting down the error term component by component.

$$\begin{aligned} \mathbf{e}_{(i)} &= \mathbf{e}_{(0)} + \sum_{j=0}^{i-1} \alpha_j \ \mathbf{d}_{(j)} = \sum_{j=0}^{n-1} \delta_j \ \mathbf{d}_{(j)} - \sum_{j=0}^{i-1} \delta_j \ \mathbf{d}_{(j)} \\ &= \sum_{j=i}^{n-1} \delta_j \ \mathbf{d}_{(j)} \end{aligned} \tag{35}$$

After n iterations, every component is cut away, and $\mathbf{e}_{(n)} = 0$; the proof is complete.

6.2 Gram-Schmidt Conjugation

There is a simple way, called *conjugate Gram-Schmidt process*, to generate a set of A-orthogonal search directions $\left\{\mathbf{d}_{(j)}\right\}$. Let $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$ be a set of n linearly independent vectors. Set $\mathbf{d}_{(0)} = \mathbf{u}_0$, and for $i > 0$, set

$$\mathbf{d}_{(i)} = \mathbf{u}_i + \sum_{k=0}^{i-1} \beta_{ik} \ \mathbf{d}_{(k)} \ , \tag{36}$$

where the β_{ik} are defined for $i > k$. To find their values, use the A-orthogonality of $\mathbf{d}_{(k)}$. We have for $i > k$:

$$\begin{aligned} 0 &= \mathbf{d}_{(i)} \cdot A \cdot \mathbf{d}_{(j)} = \mathbf{u}_i \cdot A \cdot \mathbf{d}_{(j)} + \sum_{k=0}^{i-1} \beta_{ik} \ \mathbf{d}_{(k)} \cdot A \cdot \mathbf{d}_{(j)} \\ &= \mathbf{u}_i \cdot A \cdot \mathbf{d}_{(j)} + \beta_{ij} \ \mathbf{d}_{(j)} \cdot A \cdot \mathbf{d}_{(j)} \end{aligned}$$

Or

$$\beta_{ij} = - \ \frac{\mathbf{u}_i \cdot A \cdot \mathbf{d}_{(j)}}{\mathbf{d}_{(j)} \cdot A \cdot \mathbf{d}_{(j)}} \tag{37}$$

The difficulty with using Gram-Schmidt conjugation in the method of Conjugate Directions is that all the old search vectors must be kept in memory to construct each new one, and furthermore $O(n^3)$ operations are required to generate the full set. As a result, the method of Conjugate Directions enjoyed little use until the discovery of CG - which *is* a method of Conjugate Directions -cured these disadvantages.

6.3 Optimality of the Error Term

Conjugate Directions has an interesting property: it finds at every step the best solution within the bounds of where it's been allowed to explore. Let $\mathcal{D}_i$ be the i-dimensional subspace span$\Big\{\mathbf{d}_{(0)}, \mathbf{d}_{(1)}, \ldots, \mathbf{d}_{(j-1)}\Big\}$; the value $\mathbf{e}_{(i)}$ is chosen from $\mathbf{e}_{(0)} + \mathcal{D}_i$ that minimizes $|e_{(i)}|_A$. In fact, some authors derive CG by trying to minimize $|e_{(i)}|_A$ within $\mathbf{e}_{(0)} + \mathcal{D}_i$. The energy norm can be expressed as a summation:

$$\begin{aligned} |\mathbf{e}_{(i)}|_A &= \sum_{j=i}^{n-1} \sum_{k=i}^{n-1} \delta_j \delta_k \ \mathbf{d}_{(j)} \cdot A \cdot \mathbf{d}_{(k)} \\ &= \sum_{j=i}^{n-1} \delta_j^2 \mathbf{d}_{(j)} \cdot A \cdot \mathbf{d}_{(j)} \end{aligned}$$

Each term in this summation is associated with a search direction that has not yet been traversed. Any other vector $\mathbf{e}$ chosen from $\mathbf{e}_{(0)} + \mathcal{D}_i$ must have these same terms in its expansion, which proves that $\mathbf{e}_{(i)}$ mush have the minimum energy norm.

Another important property is the fact that: at each step the residual $\mathbf{r}_{(i)}$ is orthogonal to $\mathcal{D}_i$:

$$\mathbf{d}_{(i)} \cdot \mathbf{r}_{(j)} = -\mathbf{d}_{(i)} \cdot A \cdot \mathbf{e}_{(j)} = -\sum_{j=i}^{n-1} \delta_j \mathbf{d}_{(i)} \cdot A \cdot \mathbf{d}_{(j)} \tag{38}$$

and hence by A-orthogonality of the d-vectors, we have for $i < j$,

$$\mathbf{d}_{(i)} \cdot \mathbf{r}_{(j)} = 0 \tag{39}$$

Because the search directions $\mathbf{d}_{(i)}$ are constructed from the $\mathbf{u}$ vectors, the subspace spanned by $\Big\{\mathbf{u}_{(0)}, \mathbf{u}_{(1)}, \ldots, \mathbf{u}_{(i-1)}\Big\}$ is $\mathcal{D}_i$, and the residual $\mathbf{r}_{(i)}$ is orthogonal to these previous $\mathbf{r}$ vectors as well:

$$\mathbf{d}_{(i)} \cdot \mathbf{r}_{(j)} = \mathbf{u}_{(i)} \cdot \mathbf{r}_{(j)} + \sum_{k=i}^{i-1} \beta_{ik} \ \mathbf{d}_{(k)} \cdot \mathbf{r}_{(j)} \tag{40}$$

We have for $i < j$:

$$0 = \mathbf{u}_{(i)} \cdot \mathbf{r}_{(j)} \tag{41}$$

From Equation (40), we have the identity which will be used later.

$$\mathbf{d}_{(i)} \cdot \mathbf{r}_{(i)} = \mathbf{u}_{(i)} \cdot \mathbf{r}_{(i)} \tag{42}$$

To conclude this section, note that as with the method of Steepest Descent, the number of matrix-vector products per iteration can be reduced to one by using a recurrence :

$$\begin{aligned} \mathbf{r}_{(i+1)} &= -A \cdot \mathbf{e}_{(i+1)} \\ &= -A \cdot (\mathbf{e}_{(i)} + \alpha_i \mathbf{d}_{(i)}) \\ &= \mathbf{r}_{(i)} - \alpha_i A \cdot \mathbf{d}_{(i)} \end{aligned} \tag{43}$$

7 The Method of Conjugate Gradients

CG is simply the method of Conjugate Directions where the search directions are constructed by conjugation of the residuals (that is, by setting $\mathbf{u}_i = \mathbf{r}_{(i)}$).

Because the search vectors are built from the residuals, the subspace $\text{span}\left\{\mathbf{r}_{(0)}, \mathbf{r}_{(1)}, \ldots, \mathbf{r}_{(i-1)}\right\}$ is equal to $\mathcal{D}_i$. As each residual is orthogonal to the previous search directions, it is also orthogonal to the previous residuals; thus:

$$\mathbf{r}_{(i)} \cdot \mathbf{r}_{(j)} = 0, \quad i \neq j. \tag{44}$$

Equation (43) shows that each new residual $\mathbf{r}_{(i)}$ is just a linear combination of the previous residual and $A \cdot \mathbf{d}_{(i-1)}$. Recalling that $\mathbf{d}_{(i-1)} \in \mathcal{D}_i$, this fact implies that each new subspace $\mathcal{D}_{i+1}$ is formed from the union of the previous subspace $\mathcal{D}_i$ and the subspace $A \cdot \mathcal{D}_i$. Hence

$$\begin{aligned} \mathcal{D}_i &= \mathit{span}\left\{\mathbf{d}_{(0)}, A \cdot \mathbf{d}_{(0)}, A^2 \cdot \mathbf{d}_{(0)}, \ldots, A^{i-1} \cdot \mathbf{d}_{(0)}\right\} \\ &= \mathit{span}\left\{\mathbf{r}_{(0)}, A \cdot \mathbf{r}_{(0)}, A^2 \cdot \mathbf{r}_{(0)}, \ldots, A^{i-1} \cdot \mathbf{r}_{(0)}\right\} \ . \end{aligned}$$

This subspace is called a *Krylov subspace*, a subspace created by repeatedly applying a matrix to a vector. It has a pleasing property: because

$A \cdot \mathcal{D}_i \subset \mathcal{D}_{i+1}$, the fact that the next residual $\mathbf{r}_{(i+1)}$ is orthogonal to $\mathcal{D}_{i+1}$ (Equation (39)) implies that $\mathbf{r}_{(i+1)}$ is A-orthogonal to $\mathcal{D}_i$. Gram-Schmidt conjugation becomes easy, because $\mathbf{r}_{(i+1)}$ is already A-orthogonal to all of the previous search directions except $\mathbf{d}_i$!

We may simplify the Gram-Schmidt constants $\beta_{ij} = -\frac{\mathbf{r}_{(i)} \cdot A \cdot \mathbf{d}_{(j)}}{\mathbf{d}_{(j)} \cdot A \cdot \mathbf{d}_{(j)}}$, which are defined for $i > j$ only.

Taking the inner product of $\mathbf{r}_{(i)}$ with Equation (43),

$$\begin{aligned} \mathbf{r}_{(i)} \cdot \mathbf{r}_{(j+1)} &= \mathbf{r}_{(i)} \cdot \mathbf{r}_{(j)} - \alpha_j \ \mathbf{r}_{(i)} \cdot A \cdot \mathbf{d}_{(j)} \\ \alpha_j \ \mathbf{r}_{(i)} \cdot A \cdot \mathbf{d}_{(j)} &= \mathbf{r}_{(i)} \cdot \mathbf{r}_{(j)} - \mathbf{r}_{(i)} \cdot \mathbf{r}_{(j+1)} \end{aligned}$$

or

$$\mathbf{r}_{(i)} \cdot A \cdot \mathbf{d}_{(j)} = \begin{cases} \frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\alpha_i} & i = j \\ -\frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\alpha_{i-1}} & i = j+1 \\ 0 & \text{otherwise} \end{cases}$$

Hence we obtain

$$\beta_{ij} = \begin{cases} \frac{1}{\alpha_{i-1}} \ \frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\mathbf{d}_{(i-1)} \cdot A \cdot \mathbf{d}_{(i-1)}} & i = j+1 \\ 0 & i > j+1 \end{cases} \tag{45}$$

As if by magic, most of the β_{ij} terms have disappeared. It is no longer necessary to store old search vectors. Both the space complexity and time complexity per iteration are reduced from $O(n^2)$ to $O(m)$, where m is the number of nonzero entries of A. Henceforth, we shall use the abbreviation $\beta_i = \beta_{i,i-1}$. We have

$$\begin{aligned} \beta_i &= \frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\mathbf{d}_{(i-1)} \cdot \mathbf{r}_{(i-1)}} \\ &= \frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\mathbf{r}_{(i-1)} \cdot \mathbf{r}_{(i-1)}} \end{aligned}$$

Putting all together, the method of Conjugate Gradients is:

$$\mathbf{d}_{(0)} = \mathbf{r}_{(0)} = \mathbf{b} - A \cdot \mathbf{x}_{(0)} \tag{46}$$

$$\alpha_i = \frac{\mathbf{r}_{(i)} \cdot \mathbf{r}_{(i)}}{\mathbf{d}_{(i)} \cdot A \cdot \mathbf{d}_{(i)}} \tag{47}$$

$$\mathbf{x}_{(i+1)} = \mathbf{x}_{(i)} + \alpha_i \ \mathbf{d}_{(i)} \tag{48}$$

$$\mathbf{r}_{(i+1)} = \mathbf{r}_{(i)} - \alpha_i \ A \cdot \mathbf{d}_{(i)} \tag{49}$$

$$\beta_{i+1} = \frac{\mathbf{r}_{(i+1)} \cdot \mathbf{r}_{(i+1)}}{\mathbf{d}_{(i)} \cdot \mathbf{r}_{(i)}} \tag{50}$$

$$\mathbf{d}_{(i+1)} = \mathbf{r}_{(i)} + \beta_{i+1} \ \mathbf{d}_{(i)} \tag{51}$$

$$\tag{52}$$