# Primer for the Conjugate Gradient Method

Notes to help introduce and derive this method

Nick Patterson

April 22, 2011

## Contents

# 1 Introduction and Definitions

The Conjugate Gradient (CG) method is an efficient algorithm used to solve the matrix equation, $\mathbf{A}\vec{x} = \vec{b}$, where $\mathbf{A}$ is a Symmetric Positive-Definite (SPD) matrix. The size of the matrix and vectors are $\vec{x} \in \mathcal{R}^{n\times 1}$, $\vec{b} \in \mathcal{R}^{n\times 1}$, $\mathbf{A} \in \mathcal{R}^{n\times n}$, and $n$ is some integer. This method is considered efficient because, if one where to assume no round-off error, it will converge to the exact solution in at most $n$ steps[1]. Further properties of this method are rapid convergence, numerically stable, each step gives a better estimate of the solution than the previous one, each step should depend on the original data (i.e. the matrix $\mathbf{A}$), and at any step one can start over using the last estimate obtained as the initial guess[1].

Some mathematical definitions needed before moving on further are given here. A matrix (assumed to be real) is symmetric if it is equal to its transpose, $\mathbf{A} = \mathbf{A}^T$, where the superscript $T$ denotes the transpose. A positive-definite matrix, which is related conceptually to a positive scalar, is defined as a matrix where the following property holds, $\vec{x}^T\mathbf{A}\vec{x} > 0, \forall \vec{x} \neq \vec{0}$. The inner product of two vectors, $\vec{x}$ and $\vec{y}$, is defined as the scalar $\langle \vec{x}, \vec{y} \rangle = \vec{x}^T\vec{y} = x_1y_1 + x_2y_2 + \cdots + x_ny_n$. Two vectors are defined as normal or orthogonal if $\langle \vec{x}, \vec{y} \rangle = 0$. The magnitude of a vector is defined as $|\vec{x}| = \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{\langle \vec{x}, \vec{x} \rangle}$. One can include a symmetric matrix in an inner product, called an A-norm, with the following rules,

$$\begin{aligned} \langle \vec{x}, \vec{y} \rangle_{\mathbf{A}} &= \langle \vec{x}, \mathbf{A}\vec{y} \rangle \\ &= \langle \mathbf{A}^T\vec{x}, \vec{y} \rangle \\ &= \langle \mathbf{A}\vec{x}, \vec{y} \rangle . \end{aligned} \tag{1.1}$$

Two vectors are defined as conjugate, also known as A-orthogonal, if the A-norm is zero, $\langle \vec{x}, \vec{y} \rangle_{\mathbf{A}} = 0$. A basis is a set of vectors which span a vector space. While orthogonality is not a necessary property of a basis, no basis vector can be zero or parallel to another basis vector. Let the set $\{\vec{p}_i\}_{i=1}^n$ form a basis for a vector space of $n$ vectors, each with $n$ components. Any arbitrary vector $\vec{x}$ can be written in terms of a basis and a corresponding set of scalars, $\{\alpha_i\}_{i=1}^n$ as shown,

$$\vec{x} = \alpha_1\vec{p}_1 + \alpha_2\vec{p}_2 + \cdots + \alpha_n\vec{p}_n. \tag{1.2}$$

## 2 Conjugate Gradient Algorithm

With these definitions in place, we shall now present the CG algorithm. The rest of this document will discuss derivations and alternative expressions. The following algorithm will solve the matrix equation, $\mathbf{A}\vec{x} = \vec{b}$, where $\vec{p}_i$ are the basis vectors or search directions, $\vec{r}_i$ are the residual vectors, and $\alpha_i$ and $\beta_i$ are a set of scalars. The residual vector is defined as $\vec{r}_i = \vec{b} - \mathbf{A}\vec{x}_i$.

**CG Algorithm**

**Initialize**

$$\begin{aligned}\vec{x}_0 &= 0 \text{ ($x_0$ can be defined on input, default is zero)}\\ \vec{p}_0 &= \vec{r}_0 = \vec{b} - \mathbf{A}\vec{x}_0\end{aligned}$$

**Loop**

$$\begin{aligned}\alpha_i &= \frac{\langle \vec{p}_i, \vec{r}_i \rangle}{\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle}\\ \vec{x}_{i+1} &= \vec{x}_i + \alpha_i \vec{p}_i\\ \vec{r}_{i+1} &= \vec{r}_i - \alpha_i \mathbf{A}\vec{p}_i\\ \beta_i &= -\frac{\langle \vec{r}_{i+1}, \mathbf{A}\vec{p}_i \rangle}{\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle}\\ \vec{p}_{i+1} &= \vec{r}_{i+1} + \beta_i \vec{p}_i\end{aligned}$$

The initial guess can be either an input to the function, or it can be chosen randomly. An initial guess of a zero vector can be assumed for simplicity. As it is written here, one needs to store every vector $\vec{x}_i$, $\vec{p}_i$, and $\vec{r}_i$, as well as each scalar $\alpha_i$ and $\beta_i$. However, in practice, one needs to store one vector each for the solution, residual, and search-direction, as well as a single value for the scalar $\alpha$ and a single value for $\beta$. The only additional values needed to be stored is the magnitudes of the residual for both the previous and current time step. For high performance, the matrix $\mathbf{A}$ does not even need to be formed, only a procedure to store the result of multiplying the sparse matrix $\mathbf{A}$ with the current step's search direction, $\vec{p}_i$.

## 3 Pillars of the Conjugate Gradient Method

A large number of properties and relations can be written down from this algorithm, none more important than the following two. CG produces a set of orthogonal residual vectors and a set of mutually conjugate basis vectors[2], shown as follows:

$$\langle \vec{r}_i, \vec{r}_j \rangle = 0 \qquad (j < i) \tag{3.1}$$

$$\langle \vec{p}_i, \mathbf{A}\vec{p}_j \rangle = 0 \qquad (j < i). \tag{3.2}$$

## 3.1 Induction Proof of Orthogonal Residuals

The proof of Equation (3.1) comes from induction. From the algorithm, we know $\vec{r}_0 = \vec{p}_0 = \vec{b}$, and we have

$$\begin{aligned} \vec{r}_{i+1} &= \vec{r}_i - \alpha_i \mathbf{A}\vec{p}_i \\ \vec{r}_{i+1}^T &= \vec{r}_i^T - \alpha_i \vec{p}_i^T \mathbf{A}. \end{aligned} \tag{3.3}$$

### 3.1.1 Trivial Step ($j < i - 1$)

Starting for the case of ($j < i-1$), from the trivial step, when $j = 0$, we have

$$\begin{aligned} \langle \vec{r}_i, \vec{r}_0 \rangle &= \langle \vec{r}_{i-1} - \alpha_i \mathbf{A}\vec{p}_{i-1}, \vec{b} \rangle \\ &= \vec{r}_{i-1}^T \vec{b} - \alpha_i \vec{p}_{i-1}^T \mathbf{A}\vec{p}_0 \\ &= \langle \vec{r}_{i-1}, \vec{b} \rangle, \end{aligned}$$

where we have used the conjugate nature of $\vec{p}_i$ to get to the last step. Moving terms to the same side, this last step shows $\langle \vec{r}_{i-1} - \vec{r}_i, \vec{r}_0 \rangle = 0$. Manipulation of Equation (3.3) gives $\vec{r}_{i-1} - \vec{r}_i = \alpha_i \mathbf{A}\vec{p}_{i-1}$. Therefore, we have

$$\begin{aligned} \langle \vec{r}_{i-1} - \vec{r}_i, \vec{r}_0 \rangle &= \langle \alpha_i \mathbf{A}\vec{p}_{i-1}, \vec{p}_0 \rangle \\ &= \alpha_i \langle \vec{p}_{i-1}, \mathbf{A}\vec{p}_0 \rangle \\ &= 0, \end{aligned}$$

where the last statement is due to the A-orthogonality of $\vec{p}_i$, and the fact that $j = 0$ and $j < i - 1$. So, to prove that the residuals are orthogonal for $j < i - 1$, the trivial step is complete.

### 3.1.2 Induction Hypothesis ($j < i - 1$)

To take the next step in the proof, we must first make the induction hypothesis,

$$\langle \vec{r}_i, \vec{r}_j \rangle = 0$$

and then show

$$\langle \vec{r}_i, \vec{r}_{j+1} \rangle = 0.$$

By simply using Equation (3.3), we have

$$\begin{aligned}\langle \vec{r}_i, \vec{r}_{j+1} \rangle &= \langle \vec{r}_i, \vec{r}_j \rangle - \alpha_{j+1} \langle \vec{r}_i, \mathbf{A}\vec{p}_j \rangle \\ \langle \vec{r}_i, \vec{r}_{j+1} \rangle &= -\alpha_{j+1} \langle \vec{r}_i, \mathbf{A}\vec{p}_j \rangle,\end{aligned}$$

where the last step uses the induction hypothesis. To continue, we must go back to the algorithm to obtain

$$\vec{p}_{i+1} = \vec{r}_{i+1} + \beta_i \vec{p}_i, \tag{3.4}$$

which means that $\vec{r}_i = \vec{p}_{i+1} - \beta_i \vec{p}_i$. Examination of $\langle \vec{r}_i, \mathbf{A}\vec{p}_j \rangle$ shows that

$$\langle \vec{r}_i, \mathbf{A}\vec{p}_j \rangle = \langle \vec{p}_{i+1}, \mathbf{A}\vec{p}_j \rangle - \beta_i \langle \vec{p}_i, \mathbf{A}\vec{p}_j \rangle,$$

and this is identically zero since both terms on the right-hand-side are zero because $\vec{p}_i$ is conjugate and our restriction that $j < i - 1$. Hence we have proved that $\langle \vec{r}_i, \vec{r}_j \rangle = 0$ for $j < i - 1$.

### 3.1.3 Consecutive Residuals ($j = i - 1$)

For the case that $j = i - 1$, we have

$$\begin{aligned}\langle \vec{r}_i, \vec{r}_{i-1} \rangle &= \langle \vec{r}_{i-1}, \vec{r}_{i-1} \rangle - \alpha_{i-1} \langle \mathbf{A}\vec{p}_{i-1}, \vec{r}_{i-1} \rangle \\ &= \langle \vec{r}_{i-1}, \vec{r}_{i-1} \rangle - \frac{\langle \vec{r}_{i-1}, \vec{r}_{i-1} \rangle}{\langle \vec{p}_{i-1}, \mathbf{A}\vec{r}_{i-1} \rangle} \langle \mathbf{A}\vec{p}_{i-1}, \vec{r}_{i-1} \rangle \\ &= \langle \vec{r}_{i-1}, \vec{r}_{i-1} \rangle - \langle \vec{r}_{i-1}, \vec{r}_{i-1} \rangle \\ &= 0,\end{aligned}$$

where we have used $\alpha_{i-1} = \dfrac{\langle \vec{r}_{i-1}, \vec{r}_{i-1} \rangle}{\langle \vec{p}_{i-1}, \mathbf{A}\vec{r}_{i-1} \rangle}$. However, according to the algorithm, $\alpha_{i-1} = \dfrac{\langle \vec{p}_{i-1}, \vec{r}_{i-1} \rangle}{\langle \vec{p}_{i-1}, \mathbf{A}\vec{p}_{i-1} \rangle}$. If we can show that these two expression are equivalent, then Equation (3.1) is proved.

First let us consider the denominators for these expressions for $\alpha_{i-1}$, where we need to show that $\langle \vec{p}_{i-1}, \mathbf{A}\vec{r}_{i-1} \rangle = \langle \vec{p}_{i-1}, \mathbf{A}\vec{p}_{i-1} \rangle$. From Equation (3.4), we can take the A-norm of $\vec{p}_i$ to get

$$\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle = \langle \vec{r}_i, \mathbf{A}\vec{p}_i \rangle + \beta_{i-1} \langle \vec{p}_{i-1}, \mathbf{A}\vec{p}_i \rangle$$

Again using the conjugate nature of $\vec{p}_i$, this reduces simply to $\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle = \langle \vec{r}_i, \mathbf{A}\vec{p}_i \rangle$.

Next let us examine the numerator to show that $\langle \vec{r}_{i-1}, \vec{r}_{i-1} \rangle = \langle \vec{p}_{i-1}, \vec{r}_{i-1} \rangle$. By substituting Equation (3.4) into the magnitude of $\vec{r}_i$, we have

$$|\vec{r}_i| = \langle \vec{r}_i, \vec{r}_i \rangle = \langle \vec{p}_i, \vec{r}_i \rangle - \beta_{i-1} \langle \vec{p}_{i-1}, \vec{r}_i \rangle.$$

Taking a slight detour, from $\vec{r}_{i+1} = \vec{r}_i - \alpha_i \mathbf{A}\vec{p}_i$ and $\alpha_i = \dfrac{\langle \vec{p}_i, \vec{r}_i \rangle}{\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle}$ we can determine

$$\begin{aligned} \langle \vec{p}_j, \vec{r}_{i+1} \rangle &= \langle \vec{p}_j, \vec{r}_i \rangle - \alpha_i \langle \vec{p}_j, \mathbf{A}\vec{p}_i \rangle \\ &= \begin{cases} \langle \vec{p}_j, \vec{r}_i \rangle & (i \neq j) \\ 0 & (i = j) \end{cases}. \end{aligned} \tag{3.5}$$

This leads to the following interesting property,

$$\langle \vec{p}_i, \vec{r}_0 \rangle = \langle \vec{p}_i, \vec{r}_1 \rangle = \cdots = \langle \vec{p}_i, \vec{r}_i \rangle.$$

Another a consequence of Equation (3.5) is $\langle \vec{p}_{i-1}, \vec{r}_i \rangle = 0$. Hence,

$$\langle \vec{r}_i, \vec{r}_i \rangle = \langle \vec{p}_i, \vec{r}_i \rangle - \beta_{i-1} \langle \vec{p}_{i-1}, \vec{r}_i \rangle = \langle \vec{p}_i, \vec{r}_i \rangle.$$

Therefore we have found equivalent expressions for $\alpha_i$:

$$\alpha_i = \frac{\langle \vec{p}_i, \vec{r}_i \rangle}{\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle} = \frac{\langle \vec{r}_i, \vec{r}_i \rangle}{\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle} = \frac{\langle \vec{p}_i, \vec{r}_i \rangle}{\langle \vec{r}_i, \mathbf{A}\vec{p}_i \rangle} = \frac{\langle \vec{r}_i, \vec{r}_i \rangle}{\langle \vec{r}_i, \mathbf{A}\vec{p}_i \rangle}. \tag{3.6}$$

Hence Equation (3.1) is proved.

## 3.2 Induction Proof of Conjugate Search-Directions

The proof of Equation (3.2) also comes from induction. Using Equation (3.4), we can take the A-norm of $\vec{p}_i$ and $\vec{p}_j$ to obtain

$$\langle \vec{p}_i, \mathbf{A}\vec{p}_j \rangle = \langle \vec{r}_i, \mathbf{A}\vec{p}_j \rangle + \beta_{i-1} \langle \vec{p}_{i-1}, \mathbf{A}\vec{p}_j \rangle.$$

### 3.2.1 Trivial Step ($j < i - 1$)

Starting for the case of $(j < i-1)$, from the trivial step, when $j = 0$, we have

$$\begin{aligned} \langle \vec{p}_i, \mathbf{A}\vec{p}_0 \rangle &= \langle \vec{r}_i, \mathbf{A}\vec{p}_0 \rangle + \beta_{i-1} \langle \vec{p}_{i-1}, \mathbf{A}\vec{p}_0 \rangle \\ &= \langle \vec{r}_i, \mathbf{A}\vec{p}_0 \rangle + \beta_{i-1} \langle \mathbf{A}\vec{p}_{i-1}, \vec{r}_0 \rangle, \end{aligned}$$

where we have simply moved the matrix within the inner product on the second term as well as equate $\vec{p}_0 = \vec{r}_0$. We must pause here shortly to bring in two necessary relations.

If we take the inner product of two residual vectors $i$ and $j$ and use Equation (3.3), we have

$$\langle \vec{r}_i, \vec{r}_j \rangle = \langle \vec{r}_{i-1}, \vec{r}_j \rangle - \alpha_i \langle \mathbf{A}\vec{p}_{i-1}, \vec{r}_j \rangle.$$

Under the restriction that $j < i - 1$, two of these terms immediately go to zero. Because $\alpha_i$ is always non-zero, what is left is simply

$$\langle \mathbf{A}\vec{p}_{i-1}, \vec{r}_j \rangle = 0 \qquad (j < i - 1). \tag{3.7}$$

The other relation we need is again from Equation (3.3) where we choose $i = 0$. This allows us to write

$$\frac{1}{\alpha_0}(\vec{r}_0 - \vec{r}_1) = \mathbf{A}\vec{p}_0. \tag{3.8}$$

Plugging Equation (3.7) and Equation (3.8) terms back into our expression for $\langle \vec{p}_i, \mathbf{A}\vec{p}_0 \rangle$ gives us

$$\begin{aligned}
\langle \vec{p}_i, \mathbf{A}\vec{p}_0 \rangle &= \langle \vec{r}_i, \mathbf{A}\vec{p}_0 \rangle + \beta_{i-1} \langle \vec{p}_{i-1}, \mathbf{A}\vec{p}_0 \rangle \\
&= \frac{1}{\alpha_0} \langle \vec{r}_i, \vec{r}_0 - \vec{r}_1 \rangle + \beta_{i-1} \cdot 0 \\
&= \frac{1}{\alpha_0} \langle \vec{r}_i, \vec{r}_0 \rangle - \frac{1}{\alpha_0} \langle \vec{r}_i, \vec{r}_1 \rangle \\
&= 0.
\end{aligned}$$

This proves the trivial step for the proof by induction of orthogonal search-directions for $j < i - 1$.

### 3.2.2 Induction Hypothesis ($j < i - 1$)

The next step in the proof is to assume the induction hypothesis is true,

$$\langle \vec{p}_i, \mathbf{A}\vec{p}_j \rangle = 0,$$

and then to show the next iteration is true,

$$\langle \vec{p}_i, \mathbf{A}\vec{p}_{j+1} \rangle = 0.$$

If we substitute Equation (3.4) for $j+1$, we have

$$\begin{aligned}\langle \vec{p}_i, \mathbf{A}\vec{p}_{j+1}\rangle &= \langle \vec{p}_i, \mathbf{A}\vec{r}_{j+1} + \beta_j \vec{p}_j\rangle \\ &= \langle \vec{p}_i, \mathbf{A}\vec{r}_{j+1}\rangle + \beta_j \langle \vec{p}_i, \mathbf{A}\vec{p}_j\rangle.\end{aligned}$$

The first term is zero from Equation (3.7), and the second term is zero by the induction hypothesis. Accordingly, we have shown $\langle \vec{p}_i, \mathbf{A}\vec{p}_{j+1}\rangle = 0$ for $j < i-1$.

### 3.2.3 Consecutive Search-Directions ($j = i-1$)

For the case that $j = i-1$, we have

$$\begin{aligned}\langle \vec{p}_i, \mathbf{A}\vec{p}_j\rangle &= \langle \vec{r}_i, \mathbf{A}\vec{p}_j\rangle + \beta_{i-1}\langle \vec{p}_{i-1}, \mathbf{A}\vec{p}_j\rangle \\ \langle \vec{p}_i, \mathbf{A}\vec{p}_{i-1}\rangle &= \langle \vec{r}_i, \mathbf{A}\vec{p}_{i-1}\rangle + \beta_{i-1}\langle \vec{p}_{i-1}, \mathbf{A}\vec{p}_{i-1}\rangle.\end{aligned}$$

From the CG algorithm, we have $\beta_{i-1} = -\dfrac{\langle \vec{r}_i, \mathbf{A}\vec{p}_{i-1}\rangle}{\langle \vec{p}_{i-1}, \mathbf{A}\vec{p}_{i-1}\rangle}$. When this substitution is made, we are left with simply $\langle \vec{p}_i, \mathbf{A}\vec{p}_{i-1}\rangle = \langle \vec{r}_i, \mathbf{A}\vec{p}_{i-1}\rangle - \langle \vec{r}_i, \mathbf{A}\vec{p}_{i-1}\rangle \equiv 0$.

## 3.3 Completing The Proof

While the proofs in this section are not rigorously formal, they do show one more fact. To prove logical equivalence of A and B, also stated as '$A \Leftrightarrow B$', 'A iff B', and 'A if and only if B', one must do two separate proofs: 'If A, Then B' followed by 'If B, Then A', where you take one logical statement as true and use that to show the other logical statement is true. Doing that both ways completes the proof. This in fact is what we have done. First, we showed proved Equation (3.1) assuming Equation (3.2), and then we proved Equation (3.2) assuming Equation (3.1). Therefore, both hold in the CG method.

# 4 Other CG Properties and Expressions

This section will examine a few of the other properties of the CG method. The focus here is terms and formulas which I consider important for comprehension of the method or for enhancement computationally. Nothing regarding convergence or error is included since that does not meet my criteria for being necessary for the purpose of this paper.

### 4.1 Alternate Expressions for $\alpha_i$ and $\beta_i$

The scalars $\alpha_i$ and $\beta_i$ are scale factors intended to increase stability. If these values are simply set to unity, the CG method reduces simply to the Gradient Decent method. The Gradient Decent method works best when contour lines of the natural energy function (discussed later) are circular. However, the contours are only circular if the matrix $\mathbf{A}$ is the identity. In the CG method, the contours are circular in the A-norm. In other words, instead of taking the orthogonal path from the contours as one would do with Gradient Decent, one takes the conjugate path from the contours with CG.

There are two important forms for both $\alpha_i$ and $\beta_i$:

$$\alpha_i = \frac{|r_i|^2}{\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle} = \frac{\langle \vec{p}_i, \vec{r}_i \rangle}{\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle} \tag{4.1}$$

$$\beta_i = \frac{|r_{i+1}|^2}{|r_i|^2} = -\frac{\langle \vec{r}_{i+1}, \mathbf{A}\vec{p}_i \rangle}{\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle}. \tag{4.2}$$

Initial studies by Hestenes and Stiefel show that the expressions furthest to the right obtain the best results[1]. We have already shown the equivalence of $\alpha_i$ in Equation (3.6). To show the expressions for $\beta_i$, begin with taking the A-norm of $\vec{p}_i$ and $\vec{p}_{i+1}$, which of course is zero, but substitute in Equation (3.4). The complete series of steps is shown below:

$$\begin{aligned}
\langle \vec{p}_{i+1}, \mathbf{A}\vec{p}_i \rangle &= \langle \vec{r}_{i+1} + \beta_i \vec{p}_i, \mathbf{A}\vec{p}_i \rangle \\
0 &= \langle \vec{r}_{i+1}, \mathbf{A}\vec{p}_i \rangle + \beta_i \langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle \\
0 &= \frac{1}{\alpha_i} \langle \vec{r}_{i+1}, \vec{r}_i - \vec{r}_{i+1} \rangle + \beta_i \langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle \\
0 &= \frac{-1}{\alpha_i} \langle \vec{r}_{i+1}, \vec{r}_{i+1} \rangle + \beta_i \langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle \\
\beta_i &= \frac{\langle \vec{r}_{i+1}, \vec{r}_{i+1} \rangle}{\alpha_i \langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle} \\
\beta_i &= \frac{\langle \vec{r}_{i+1}, \vec{r}_{i+1} \rangle}{\langle \vec{r}_i, \vec{r}_i \rangle} \\
\beta_i &= \frac{|\vec{r}_{i+1}|^2}{|\vec{r}_i|^2}.
\end{aligned}$$

The third line comes from taking Equation (3.3) in the form $\mathbf{A}\vec{p}_i = \dfrac{\vec{r}_i - \vec{r}_{i+1}}{\alpha_i}$. Eliminating $\langle \vec{r}_i, \vec{r}_{i+1} \rangle$ and solving for $\beta_i$ gives the fifth line. The sixth line

comes from one of the alternate expressions for $\alpha_i$ in Equation (3.6).

## 4.2 Residual

The residual is computed in the CG algorithm as Equation (3.3), which is a recurrence relation. However, the residual is defined as

$$\vec{r}_i = \vec{b} - \mathbf{A}\vec{x}_i, \tag{4.3}$$

where $\vec{x}_i$ is just $\vec{x}_i = \alpha_0\vec{p}_0 + \alpha_1\vec{p}_1 + \cdots + \alpha_{i-1}\vec{p}_{i-1}$, as in Equation (1.2). Hence, we can write the residual in full as

$$\vec{r}_{i+1} = \vec{b} - (\alpha_0\mathbf{A}\vec{p}_0 + \alpha_1\mathbf{A}\vec{p}_1 + \cdots + \alpha_i\mathbf{A}\vec{p}_i)\,. \tag{4.4}$$

The difference between two consecutive residuals is also obvious, $\vec{r}_{i+1} - \vec{r}_i = -\alpha_i\mathbf{A}\vec{p}_i$, which matches Equation (3.3). This means that we need only to modify a single residual vector at each step, subtracting the $\alpha_i\mathbf{A}\vec{p}_i$, instead of having to calculate a sum that increases in length with each step.

In the CG method, the magnitude of the residual decreases with each step. However, for the more general Conjugate Directions (CD) method, the residual magnitude may increase or decrease at each step (always with a decrease at the $n^{th}$ step), meaning the residual magnitude is not the best indicator of convergence [1]. The reason for this is that the residual is not what is being minimized. There is a function that is more effective to measure the 'goodness' of a solution, and this function is what is being minimized.

## 4.3 Error Function

The 'goodness' of an estimate for the solution of $\mathbf{A}\vec{x} = \vec{b}$ is a measure of how close $\vec{x}_i$ is to the true solution, which we will call $\vec{h}$. Since $\vec{b} = \mathbf{A}\vec{h}$, we can write the residual as

$$\vec{r}_i = \vec{b} - \mathbf{A}\vec{x}_i = \mathbf{A}(\vec{h} - \vec{x}_i).$$

It is this difference, $\vec{h} - \vec{x}$, that we wish to minimize. However, the space in which we want to minimize the difference is not the standard inner product space, but the A-norm. This leads to the definition of the error function,

$$f(\vec{x}) = \langle \vec{h} - \vec{x}, \mathbf{A}(\vec{h} - \vec{x}) \rangle. \tag{4.5}$$

We can expand and simplify Equation (4.5) as shown below,

$$\begin{aligned} f(\vec{x}) &= \langle \vec{h} - \vec{x}, \mathbf{A}(\vec{h} - \vec{x}) \rangle \\ &= \langle \vec{h}, \mathbf{A}\vec{h} \rangle - \langle \vec{h}, \mathbf{A}\vec{x} \rangle - \langle \vec{x}, \mathbf{A}\vec{h} \rangle + \langle \vec{x}, \mathbf{A}\vec{x} \rangle \\ &= \langle \vec{h}, \vec{b} \rangle - \langle \vec{x}, \mathbf{A}\vec{h} \rangle - \langle \vec{x}, \vec{b} \rangle + \langle \vec{x}, \mathbf{A}\vec{x} \rangle \\ &= \langle \vec{h}, \vec{b} \rangle - 2\langle \vec{x}, \vec{b} \rangle + \langle \vec{x}, \mathbf{A}\vec{x} \rangle. \end{aligned}$$

The error function is strictly non-negative, only zero when $\vec{x} = \vec{h}$[1], stated mathematically below:

$$\begin{aligned} f(\vec{x}) &> 0 \qquad (\vec{x} \neq \vec{h}) \\ f(\vec{x}) &= 0 \qquad (\vec{x} = \vec{h}). \end{aligned}$$

It can be shown that the CG algorithm minimizes this error function at each step[2].

The calculation of this function requires knowing the true solution, $\vec{h}$. However, we cannot know the solution while we are trying to calculate it. Hestenes and Stiefel's paper discuss means to approximate this function in terms of the residual and a Rayleigh quotient[1]. Because we are employing the CG method instead of the more general CD method, it is not important to go into detail about it, and we will stick to using the magnitude of the residual as a means of quantifying the 'goodness' of the solution. Any CD-method which has mutually orthogonal residual vectors is essentially a CG-method[1], since the only restriction on a CD-method is that the search directions are mutually orthogonal.

## 4.4 Gradient of What?

While the 'Conjugate' part of 'Conjugate Gradient' is obvious, $\langle \vec{x}, \mathbf{A}\vec{y} \rangle = 0$, the origin of the word 'Gradient' comes from the error function as well as viewing the problem as one of optimization. The function being optimized, sometimes called the natural energy function[3], is defined as

$$\begin{aligned} \phi(\vec{x}) &= \frac{1}{2}\langle \vec{x}, \mathbf{A}\vec{x} \rangle - \langle \vec{x}, \vec{b} \rangle \\ &= \frac{1}{2}\vec{x}^T \mathbf{A}\vec{x} - \vec{x}^T \vec{b}. \end{aligned} \tag{4.6}$$

Taking the gradient of this function is as straightforward as if we assume we had a scalar function, $\frac{d}{dx}(\frac{1}{2}Ax^2 - bx) = Ax - b$[3]. Hence, the gradient

of the natural energy function is defined as $\mathbf{A}\vec{x} - \vec{b}$. One knows from fundamental calculus that the maximum or minimum of a function is where the derivative is zero. We know that the energy function has a minimum, not a maximum, because $\mathbf{A}$ is SPD. Therefore, the minimum of the energy function is when $\vec{x} = \vec{h}$. The function $\phi(\vec{x})$ is called the natural energy because minimizing it solves $\mathbf{A}\vec{x} = \vec{b}$. When one finds $\vec{x} = \vec{h}$, he has solved the matrix equation as well as optimized the natural energy function. For this reason, the CG method can be viewed as a linear solver and/or as an optimizer. It can be showed that the choice for $\alpha_i$ ensures that the optimal step length is chosen along each search direction, and that when the function $\phi(\vec{x})$ is minimized over $\vec{x}$, it is minimized over the entire vector space[2].

If one compares the error function, $f(\vec{x})$, with the natural energy function, $\phi(\vec{x})$, one can see that

$$\begin{aligned} f(\vec{x}) &= \langle \vec{h}, \vec{b} \rangle - 2\langle \vec{x}, \vec{b} \rangle + \langle \vec{x}, \mathbf{A}\vec{x} \rangle \\ &= 2\left( \frac{1}{2}\langle \vec{x}, \mathbf{A}\vec{x} \rangle - \langle \vec{x}, \vec{b} \rangle \right) + \langle \vec{h}, \vec{b} \rangle \\ &= 2\phi(\vec{x}) + \langle \vec{h}, \vec{b} \rangle \end{aligned}$$

Therefore, the error function is equal to a constant plus two times the natural energy function. The constant factor, $\langle \vec{h}, \vec{b} \rangle$, does not depend on $\vec{x}$, but it does depend on knowing the solution. However, since this term is constant, minimizing the error function is achieved when the natural energy function is minimized, and for both cases this occurs when $\vec{x} = \vec{h}$[2]. Hence, using the the natural energy function instead of the error function is advantageous because it can be calculated at any step $i$ with $\vec{x}_i$.

### 4.5 Krylov Subspace

The last point I would consider crucial to being aware of before using the CG method is that this method is a Krylov subspace iteration. The $n^{th}$ Krylov subspace is defined as

$$\mathcal{K}_n = \text{span}\left\{ \vec{b}, \mathbf{A}\vec{b}, \mathbf{A}^2\vec{b}, ..., \mathbf{A}^{n-1}\vec{b} \right\}. \tag{4.7}$$

A Krylov based numerical solver will add one more dimension to its vector space on each iteration. The first iteration has a very small vector space, $\mathcal{K}_1 = \{\vec{b}\}$, and the second has just two vectors to span the subspace, and so on. A typical issue facing a Krylov method is that one needs to construct a large amount of vectors, which means a large amount of matrix-vector mul-

tiplication, in order to get a useful solution. However, for the CG method, one will often need just $n$ or less iterations, and one does not need to store each of these vectors.

The CG method is a Krylov method because at each iteration another factor of $\mathbf{A}$ is multiplied. This occurs in the calculation of the residual, Equation (3.3), and is reflected in the new search-direction, Equation (3.4), because $\vec{p}_i$ is defined in terms of $\vec{r}_i$. Since $\vec{x}_{i+1}$ is defined in terms of $\vec{p}_i$, then $\vec{x}_{i+1}$, $\vec{p}_i$, and $\vec{r}_i$ all span the subspace $\mathcal{K}_i$[2]. This can be stated mathematically as

$$\begin{aligned}\mathcal{K}_n &= \text{span}\left\{\vec{b}, \mathbf{A}\vec{b}, \mathbf{A}^2\vec{b}, ..., \mathbf{A}^{n-1}\vec{b}\right\} = \text{span}\left\{\vec{x}_1, \vec{x}_2, \vec{x}_3, ..., \vec{x}_n\right\} \\ &= \text{span}\left\{\vec{p}_0, \vec{p}_1, \vec{p}_2, ..., \vec{p}_{n-1}\right\} = \text{span}\left\{\vec{r}_0, \vec{r}_1, \vec{r}_2, ..., \vec{r}_{n-1}\right\}. \end{aligned} \tag{4.8}$$

While each step increases the dimension of the Krylov space by one, it decreases the dimension of the vector space in which the solution is sought by one[1].

# 5 Summary

We have now introduced all necessary formulas and expressions and definitions needed for the CG method. In this section we will simply restate the loop part of the algorithm, and then give a written explanation for what each step is doing.

$$\alpha_i = \frac{\langle \vec{p}_i, \vec{r}_i \rangle}{\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle} \tag{5.1}$$

$$\vec{x}_{i+1} = \vec{x}_i + \alpha_i \vec{p}_i \tag{5.2}$$

$$\vec{r}_{i+1} = \vec{r}_i - \alpha_i \mathbf{A}\vec{p}_i \tag{5.3}$$

$$\beta_i = -\frac{\langle \vec{r}_{i+1}, \mathbf{A}\vec{p}_i \rangle}{\langle \vec{p}_i, \mathbf{A}\vec{p}_i \rangle} \tag{5.4}$$

$$\vec{p}_{i+1} = \vec{r}_{i+1} + \beta_i \vec{p}_i \tag{5.5}$$

Equation (5.1) is a scalar that simply weights how far along the direction $\vec{p}_i$ we need to adjust $\vec{x}_{i+1}$ for optimal stability and convergence. It is simply the ratio of the projection of $\vec{p}_i$ onto $\vec{r}_i$ with the A-norm of $\vec{p}_i$. To calculate the next approximation of the solution in the $i^{th}$ iteration, Equation (5.2), we simply add the next weighted, basis vector, $\vec{p}_i$, which was calculated in the previous iteration. This step is straightforward since $\vec{x}$ is expressed as a

weighted sum of the basis for the current Krylov subspace, as in Equation (1.2). Another way to view Equation (5.2) is that the search-direction $\vec{p}_i$ is a scalar multiple of the gradient of $\vec{x}_i$, $\alpha_i \vec{p}_i = \vec{x}_{i+1} - \vec{x}_i$. This means that the direction $\vec{p}_i$ will go in the direction that the gradient, $\Delta\vec{x}_i = \vec{x}_{i+1} - \vec{x}_i$, goes.

To calculate the next residual, Equation (5.3), we must subtract the newest, weighted, basis vector after multiplying by $\mathbf{A}$. What is happening in this step is taking the residual, the amount of how far off we are, and then removing the amount that we have improved the solution. Since $\vec{r}_i = \vec{b} - \mathbf{A}\vec{x}_i$, and we have improved the solution by $\alpha_i \vec{p}_i$, the update for the residual is simply $-\alpha_i \mathbf{A}\vec{p}_i$. This can also be looked at as the step that enforces the A-orthogonality because this is where the contribution from the $\mathbf{A}\vec{p}_i$ term is removed, and the subsequent search-directions are equated to the residual. Equation (5.4) is a scalar weight chosen to help select the next search-direction. The specific choice of $\beta_i$ allows for the optimal solution of the error function and energy function for each step, although different choices could be made for slightly different CG-like methods.

The last step, Equation (5.5), is the most subtle, however certain substitutions help illuminate what is happening. If one uses $\beta_i = \frac{|\vec{r}_{i+1}|^2}{|\vec{r}_i|^2}$ and $\vec{p}_0 = \vec{r}_0$, one can readily obtain the following by substitution of Equation (5.5),

$$\begin{aligned}
\vec{p}_1 &= \vec{r}_1 + \frac{|\vec{r}_1|^2}{|\vec{r}_0|^2}\vec{r}_0 \\
\vec{p}_2 &= \vec{r}_2 + \frac{|\vec{r}_2|^2}{|\vec{r}_1|^2}\vec{r}_1 + \frac{|\vec{r}_2|^2}{|\vec{r}_0|^2}\vec{r}_0 \\
&\vdots \\
\vec{p}_i &= \sum_{j=0}^{i} \frac{|\vec{r}_i|^2}{|\vec{r}_j|^2}\vec{r}_j \\
\vec{p}_i &= |\vec{r}_i|^2 \sum_{j=0}^{i} \hat{r}_j,
\end{aligned}$$

where $\hat{r}_j$ is the normalized residual, $\hat{r}_j = \frac{\vec{r}_j}{|\vec{r}_j|^2}$. This means that the $i^{th}$ search-direction is just a sum of the residual-basis, $\{\vec{r}_0, \vec{r}_1, ..., \vec{r}_i\}$, all multiplied by the magnitude of the $i^{th}$ residual. The residuals themselves account for the A-orthogonality. Since the residuals are mutually orthogonal,

each new residual $\vec{r}_{i+1}$ will be perpendicular to the previous Krylov space, $\mathcal{K}_i = \{\vec{r}_0, \vec{r}_1, ..., \vec{r}_i\}$. Equation (5.5) therefore creates a vector that is mutually conjugate with all previous search-directions as well as orthogonal to the previous Krylov space, making it a basis vector as well.

## References


[1] M. R. Hestenes, E. Stiefel. *Methods of Conjugate Gradients for Solving Linear Systems*. Journal of Research of the National Bureau of Standards, 49(6):409–436: 1952.

[2] L. N. Trefethen, D. B. III. *Numerical Linear Algebra* (Society for Industrial Mathematics, 1997).

[3] G. Strang. *Conjugate Gradient Method (Lecture 19)*. http://academicearth.org/lectures/conjugate-gradient-method.