Conjugate Gradients

MJ Rutter

Easter 2016, revised June 2020

1 Introduction

There are three main methods for minimising high-dimensional functions with quadratic minima: steepest descents, conjugate gradients, and BFGS (Broyden Fletcher Goldfarb Shanno). All three assume that it is possible to calculate the function's gradient. For a function of n dimensions, BFGS has a storage requirement of n^2, and so is used in most electronic structure codes for relaxing the atoms only, where n is maybe a few hundred. Conjugate gradients has a storage requirement of merely n, so it is appropriate for wavefunctions expressed in a plane wave basis where n can be many thousands. Steepest descents requires no storage at all between iterations.

1.1 Steepest Descents and Information

Steepest descents is an obvious, but poor, iterative method of minimising a multidimensional function. It simply requires one to find the direction of steepest descent (the gradient), move along it until one reaches the minimum, and then repeat. The reason it is disliked is that, although fairly robust, even in simple cases it can require a very large number of steps.

Consider minimising a function of two dimensions. If it is quadratic around that minimum, then it can be expanded about the minimum, at $\mathbf{x_0}$, as

$$F(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x_0})\mathbf{A}(\mathbf{x}-\mathbf{x_0}) + c' = \frac{1}{2}\mathbf{x}\mathbf{A}\mathbf{x} + \mathbf{b}\mathbf{x} + c$$

The unknowns in this equation are the coefficients of $\mathbf{b}$, $\mathbf{A}$, and the constant c. As the coefficents of $\mathbf{A}$ always appear in pairs, $a_{ij} + a_{ji}$, there are only $0.5n^2 + 0.5n$ of them, and the total number of unknowns is $0.5n^2 + 1.5n + 1$. Evalulating $\mathbf{x}\mathbf{A}\mathbf{x}$ and $\mathbf{A}\mathbf{x}$ yields $n+1$ pieces of information, so one might hope for convergence in around $0.5n + 1$ trials.

In practice, if, in steepest descents one is lucky and picks a starting point on one of the axes of $\mathbf{A}$ it will converge in a single step. Otherwise its convergence is never absolute, although it will reach any desired precision within a finite number of steps (proof by assertion).

The problem is made clear in figure 1. The cause is that the gradient direction never points to the minimum unless one starts on one of the axes of the ellipse. A step will never finish on one of the axes of the ellipse unless it is in a direction perpendicular to that axis, which again will not occur (unless it started on the other axis).

Without proof, I state that, for a positive definite matrix, for which a condition number, κ, can be defined as the ratio of the largest to the smallest eigenvalue, the error in the eigenvector as found by steepest descent changes by at least a factor of $\frac{\kappa-1}{\kappa+1}$ on each iteration, and the error in the eigenvalue by the square of this. (These factors are less than one, showing that the error decreases on each iteration. For all eigenvalues equal, the factor is zero, and the algorithm converges perfectly in a single step.)

1.2 Conjugate Gradients

The aim of the conjugate gradients method is to produce a better search direction for steps after the first step. Steepest descents could also be called orthogonal gradients, in that each step is guaranteed to be orthogonal to the previous, as the new gradient vector can have no component along the old after a successful line minimisation.

The idea of conjugate gradients is simple: a step should not 'undo' the minimisation done by the previous step, in that it should not require a further step in the direction of the previous step.

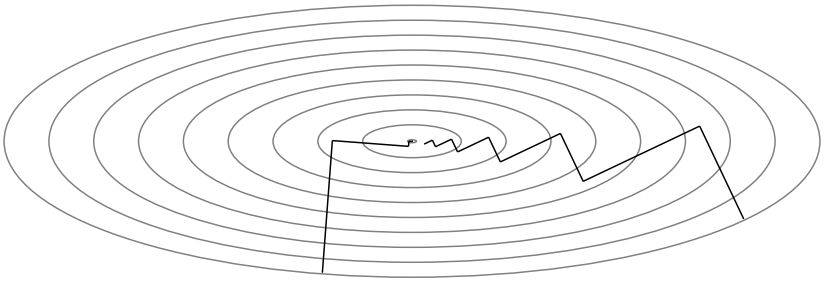

Figure 1: Convergence of steepest descents. Contour lines of a quadratic minimum shown, and the first ten iterations from two distinct starting points. Although the starting points are at similar function values, one gets closer to the minimum in two steps than the other manages in ten.

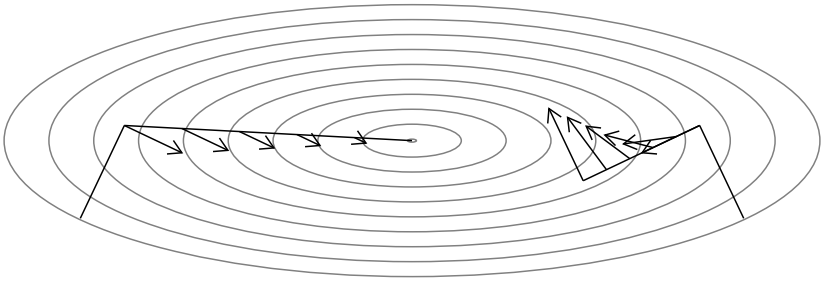

Figure 2: On the right are shown the first two steps of the steepest descents algorithm, together with the gradient vector at points along the second step. On the left are the first two steps of the conjugate gradients algorithm, starting from an equivalent point. Again the gradient vector is shown at intervals along the second search direction.

How is this concept expressed? If the previous step was in the direction $\mathbf{p}$, then it will have ended when the gradient of $F(\mathbf{x})$ was perpendicular to $\mathbf{p}$. The direction of the next step, $\mathbf{q}$, should be such that, as one moves along the step, the gradient remains perpendicular to $\mathbf{p}$.

The gradient of $F(\mathbf{x})$ is $\mathbf{Ax} + \mathbf{b}$, and the previous step ended at the point at which $\mathbf{p}.(\mathbf{Ax} + \mathbf{b}) = 0$. For a small displacement q, the change in the gradient is

$$\mathbf{A}(\mathbf{x} + \mathbf{q}) - \mathbf{Ax} = \mathbf{Aq}$$

We wish to retain the condition that $\mathbf{p}.\mathbf{Ax} = -\mathbf{b}.\mathbf{q}$ so the displacement should be in a direction $\mathbf{q}$ such that $\mathbf{p}.\mathbf{Aq} = 0$, not the direction $\mathbf{Ax}$ as steepest descents would yield. Note that the condition $\mathbf{p}.\mathbf{Aq} = 0$ is equivalent to $\mathbf{q}.\mathbf{Ap} = 0$, which can easily be seen by writing the expression in summation form:

$$\begin{aligned}
\mathbf{p}.\mathbf{Aq} &= \sum_i p_i \sum_j A_{ij} q_j \\
&= \sum_{i,j} p_i A_{ij} q_j \\
&= \sum_{i,j} q_j A_{ji} p_i \\
&= \sum_j q_j \sum_i A_{ji} p_i \\
&= \mathbf{q}.\mathbf{Ap}
\end{aligned}$$

using the fact that $\mathbf{A}$ is symmetric.

Note too that this condition alone will not yield a unique direction in problems of more than two dimensions.

Figure 2 shows the gradient vectors along the second step of the steepest descents and conjugate gradients algorithm for minimising a quadratic function in two dimensions. A line minimisation will stop when the gradient is either orthogonal to the line direction, or zero. In two dimensions the condition that the gradient remains orthogonal to the previous step fixes its direction to a direction not orthogonal to the search direction of the second step. Thus the second step will end when the gradient is zero, i.e. the global minimum found. This is only true if the function is precisely described by a quadratic form. If it is not, then the gradient will not remain perpendicular to the first step at all points along the second search direction. In higher dimensions the gradient can rotate to remain perpendicular to the previous search direction and to become perpendicular to the current search direction.

Just as two vectors for which $\mathbf{q}.\mathbf{p} = 0$ are said to be orthogonal, two vectors for which $\mathbf{q}.\mathbf{Ap} = 0$ are said to be conjugate. In practice one wishes to construct a whole set of conjugate directions for the search directions, so that $\mathbf{q_i}.\mathbf{Aq_j} = \delta_{ij}$.

That the set of n vectors $\mathbf{q}_i$ spans the n dimensional space is readily proven. If it did not, one of the vectors must be linearly dependent on the others, so one can find a non-trivial set of λ_i such that

$$\sum_i \lambda_i \mathbf{q}_i = \mathbf{0}$$

One can multiply both sides by $\mathbf{A}$, and then form the dot product with $\mathbf{q}_j$:

$$\sum_i \lambda_i \mathbf{q}_j \mathbf{A} \mathbf{q}_i = 0$$

Given that $\mathbf{q_i}.\mathbf{Aq_j} = \delta_{ij}$ this reduces to

$$\lambda_j = 0 \ \forall j$$

which is a contradiction.

If an n dimensional minimum is precisely described by a quadratic form, then this method will converge exactly in n steps.

1.2.1 Construction of Conjugate Directions

The difficulty is constructing the required conjugate directions, especially as $\mathbf{A}$ might not be known. This is usually done by considering the following.

In the method of steepest descents, one moves through a set of approximations to the minimum vector, and these approximations can be denoted $\mathbf{x}_i$. At each of these trials, there is an associated gradient vector $\mathbf{g}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}$. The method of steepest descents sets

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \alpha_i \mathbf{g}_i$$

with α_i chosen such that $\mathbf{g}_{i+1}$ is perpendicular to $\mathbf{g}_i$.

$$\begin{aligned} \mathbf{g}_{i+1}.\mathbf{g}_i &= 0 \\ (\mathbf{A}(\mathbf{x}_i - \alpha_i \mathbf{g}_i) + \mathbf{b}).\mathbf{g}_i &= 0 \\ (\mathbf{g}_i - \alpha_i \mathbf{A}\mathbf{g}_i).\mathbf{g}_i &= 0 \\ \alpha_i &= \frac{\mathbf{g}_i.\mathbf{g}_i}{\mathbf{g}_i \mathbf{A} \mathbf{g}_i} \end{aligned} \tag{1}$$

The method of conjugate gradients can be considered as follows. The set $\mathbf{p}_i$ of conjugate directions have been shown to span an n dimensional space, so any other vector can be expressed as a linear sum of them.

If $\mathbf{x}^*$ is the position of the minimum, and $\mathbf{x}_1$ an initial guess, then

$$\mathbf{x}^* - \mathbf{x}_1 = \sum_{i=1}^{n} \alpha_i \mathbf{p}_i$$

This can be premultiplied by $\mathbf{A}$ and then the scalar product taken with $\mathbf{p}_r$

$$\begin{aligned} \mathbf{p}_r \mathbf{A}(\mathbf{x}^* - \mathbf{x}_1) &= \mathbf{p}_r \mathbf{A} \sum_{i=1}^{n} \alpha_i \mathbf{p}_i \\ -\mathbf{p}_r \mathbf{b} - \mathbf{p}_r \mathbf{A} \mathbf{x}_1 &= \mathbf{p}_r \mathbf{A} \alpha_r \mathbf{p}_r \\ \alpha_r &= \frac{-\mathbf{p}_r \mathbf{g}_1}{\mathbf{p}_r \mathbf{A} \mathbf{p}_r} \end{aligned} \tag{2}$$

given that at $\mathbf{x}^*$ the gradient is zero, so $\mathbf{A}\mathbf{x}^* = -\mathbf{b}$.

This shows that $\mathbf{x}^*$ can be reached in n steps from $\mathbf{x}_1$ with

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{p}_i$$

or

$$\mathbf{x}_r = \mathbf{x}_1 + \sum_{i=1}^{r-1} \alpha_i \mathbf{p}_i$$

If this is premultiplied by $\mathbf{A}$ one obtains

$$\mathbf{g}_r = \mathbf{g}_1 + \sum_{i=1}^{r-1} \alpha_i \mathbf{A} \mathbf{p}_i$$

and taking the scalar product with $\mathbf{p}_r$ yields

$$\mathbf{p}_r.\mathbf{g}_1 = \mathbf{p}_r.\mathbf{g}_r$$

With this result, it can be seen that the calculation of α_i for a conjugate gradient step in equation 2 is identical to the calculation of α_i for a line minimisation step in a given direction given by equation 1. So for each step from $\mathbf{x}_1$ to $\mathbf{x}_n$ one proceeds to the minimum along the direction given by the next conjugate direction.

One can also take the scalar product with $\mathbf{p}_s$ $(s < r)$ yielding

$$\begin{aligned}\mathbf{p}_s\mathbf{g}_r &= \mathbf{p}_s\mathbf{g}_1 + \alpha_s\mathbf{p}_s\mathbf{A}\mathbf{p}_s \\ &= \mathbf{p}_s\mathbf{g}_1 - \frac{\mathbf{p}_s\mathbf{g}_1}{\mathbf{p}_s\mathbf{A}\mathbf{p}_s}\mathbf{p}_s\mathbf{A}\mathbf{p}_s \\ &= 0\end{aligned} \tag{3}$$

So the gradient direction is orthogonal to all previous search directions at each point $\mathbf{x}_i$.

All that remains is to construct a set of conjugate directions $\mathbf{p}_i$.

The initial direction, $\mathbf{p}_1$, is simply the gradient direction at $\mathbf{x}_i$, that is to say $\mathbf{g}_1$. For subsequent steps, one starts with the gradient direction at that point, and then, using a method similar to Gram-Schmidt orthogonalisation, makes this vector conjugate to all previous vectors. This would seem to require the storage of all previous vectors, but in fact this is unnecessary.

$$\begin{aligned}\mathbf{p}_1 &= -\mathbf{g}_1 \\ \mathbf{p}_{r+1} &= -\mathbf{g}_{r+1} + \sum_{i=1}^{r}\gamma_i^r\mathbf{p}_i\end{aligned} \tag{4}$$

where

$$\gamma_i^r = \frac{\mathbf{g}_{r+1}\mathbf{A}\mathbf{p}_i}{\mathbf{p}_i\mathbf{A}\mathbf{p}_i}$$

That this Gram-Schmidt process has the desired properties can be proved by induction. Assuming that the vectors up to $\mathbf{p}_r$ are conjugate, one can construct $\mathbf{p}_{r+1}$ and check its conjugacy against some $\mathbf{p}_s$ for $s < r + 1$:

$$\begin{aligned}\mathbf{p}_s\mathbf{A}\mathbf{p}_{r+1} &= \mathbf{p}_s\mathbf{A}\mathbf{g}_{r+1} - \sum_{i=1}^{r}\frac{\mathbf{g}_{r+1}\mathbf{A}\mathbf{p}_i}{\mathbf{p}_i\mathbf{A}\mathbf{p}_i}\mathbf{p}_s\mathbf{A}\mathbf{p}_i \\ &= \mathbf{p}_s\mathbf{A}\mathbf{g}_{r+1} - \frac{\mathbf{g}_{r+1}\mathbf{A}\mathbf{p}_s}{\mathbf{p}_s\mathbf{A}\mathbf{p}_s}\mathbf{p}_s\mathbf{A}\mathbf{p}_s \\ &= \mathbf{p}_s\mathbf{A}\mathbf{g}_{r+1} - \mathbf{g}_{r+1}\mathbf{A}\mathbf{p}_s\end{aligned}$$

As $\mathbf{A}$ is symmetric, this is zero as it should be.

The gradient directions $\mathbf{g}_i$ are all orthogonal. The point $\mathbf{x}_2$ is chosen such that $\mathbf{g}_2$ is orthogonal to $\mathbf{g}_1$, being the definition of line minimisation along $\mathbf{p}_1 = -\mathbf{g}_1$.

One can then procede by induction assuming that all gradients up to r are orthogonal, and considering the scalar product of $\mathbf{g}_{r+1}$ with $\mathbf{g}_s$ for some $s \leq r$.

$$\mathbf{g}_{r+1}.\mathbf{p}_s = 0$$

from equation 3 and equation 4 gives

$$\mathbf{p}_s = -\mathbf{g}_s + \sum_{i=1}^{s-1} \gamma_i^{s-1} \mathbf{p}_i$$

Taking the scalar product with $\mathbf{g}_{r+1}$ gives

$$\begin{aligned} \mathbf{g}_{r+1}.\mathbf{p}_s &= -\mathbf{g}_{r+1}.\mathbf{g}_s + \sum_{i=1}^{s-1} \gamma_i^{s-1} \mathbf{g}_{r+1}.\mathbf{p}_i \\ 0 &= -\mathbf{g}_{r+1}.\mathbf{g}_s \end{aligned} \tag{5}$$

So the gradient vectors are orthogonal.

The final result needed returns to

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{p}_i$$

This can be premultiplied by $\mathbf{A}$ to give

$$\begin{aligned} \mathbf{A}\mathbf{x}_{i+1} - \mathbf{A}\mathbf{x}_i &= \alpha_i \mathbf{A}\mathbf{p}_i \\ \mathbf{g}_{i+1} - \mathbf{g}_i &= \alpha_i \mathbf{A}\mathbf{p}_i \end{aligned}$$

With this result one can manipulate the definition of γ_i^r

$$\begin{aligned} \gamma_i^r &= \frac{\mathbf{g}_{r+1}\mathbf{A}\mathbf{p}_i}{\mathbf{p}_i\mathbf{A}\mathbf{p}_i} \\ (\mathbf{p}_i\mathbf{A}\mathbf{p}_i)\gamma_i^r &= \frac{1}{\alpha_i}\mathbf{g}_{r+1}.(\mathbf{g}_{i+1} - \mathbf{g}_i) \end{aligned}$$

As $i \leq r$, and assuming α_i and $\mathbf{p}_i\mathbf{A}\mathbf{p}_i$ are never zero, then the only γ_i^r which is non-zero is γ_r^r, and the double index is unnecessary

Equation 5, and the definition of α_r in equation 2, gives

$$\begin{aligned} (\mathbf{p}_r\mathbf{A}\mathbf{p}_r)\gamma_r &= \frac{1}{\alpha_r}\mathbf{g}_{r+1}.\mathbf{g}_{r+1} \\ (\mathbf{p}_r\mathbf{A}\mathbf{p}_r)\gamma_r &= -\frac{\mathbf{p}_r\mathbf{A}\mathbf{p}_r}{\mathbf{p}_r\mathbf{g}_1}\mathbf{g}_{r+1}.\mathbf{g}_{r+1} \\ \gamma_r &= -\frac{\mathbf{g}_{r+1}.\mathbf{g}_{r+1}}{\mathbf{p}_r\mathbf{g}_r} \\ \gamma_r &= \frac{\mathbf{g}_{r+1}.\mathbf{g}_{r+1}}{\mathbf{g}_r\mathbf{g}_r} \end{aligned}$$

That

$$\mathbf{g}_r\mathbf{g}_r = -\mathbf{p}_r\mathbf{g}_r$$

follows from equations 4 and 3

1.2.2 Convergence

The theory of conjugate gradients depends on the minimum being quadratic much more strongly than the theory of steepest descents does. So conjugate gradients can perform badly if it is exposed to a non-quadratic minimum (e.g. started too far away from a minimum which is locally quadratic). One solution to this issue is to restart the algorithm every few steps, or in particular if convergence is slowing.

Like steepest descents, its rate of convergence depends also on the condition number κ of the matrix, defined as the ratio of the largest to the smallest eigenvalue. Conjugate gradients is often used after preconditioning, which tries to find a matrix with a lower κ, and a way of mapping this problem back to the original one.

Minimising the quadratic form

$$\frac{1}{2}\mathbf{x}\mathbf{A}\mathbf{x} + \mathbf{b}\mathbf{x} + c$$

is equivalent to finding the point at which its derivative is zero, i.e.

$$\mathbf{A}\mathbf{x} = -\mathbf{b}$$

Preconditioning attempts to rewrite this as

$$\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = -\mathbf{M}^{-1}\mathbf{b}$$

and then solves

$$(\mathbf{M}^{-1}\mathbf{A})\mathbf{x} = -\mathbf{b}'$$

However, care must be taken to ensure that $\mathbf{M}^{-1}\mathbf{A}$ remains symmetric and positive-definite. Having $\mathbf{A}$ and $\mathbf{M}$ symmetric and positive-definite is not sufficient to guarantee that $\mathbf{M}^{-1}\mathbf{A}$ is. A very simple preconditioner is given by

$$M_{ij} = \begin{cases} A_{ij}, & i = j \\ 0, & i \neq j \end{cases}$$

The convergence improvement per step usually quoted for conjugate gradients is

$$2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)$$

The prefactor of 2 means that this is worse than the corresponding expression for steepest descents of

$$\frac{\kappa - 1}{\kappa + 1}$$

However, in practice conjugate gradients is usually considerably better. I believe that the above expression is an upper bound. For $\kappa > 9$ it seems unreasonable.

1.3 Line Minimisation

For steepest descents, the quality of the line minimisation matters little. Provided that it finds a point lower than the starting point, it will converge. If it undershoots the true minimum slightly, it may even cause the algorithm to converge faster than would have been the case with an accurate line minimiser.

This is not true for conjugate gradients. The theory relies heavily on the line minimisation being done exactly at every step.