Understanding Conjugate Gradients

R.C. Luo

Oct 2020

1 Introduction

Our task now is to solve the linear equation [1]

$$Ax = b$$

where A is a **symmetric, positive-definite** matrix. And to ensure efficiency, we also hope that A has a **sparse** property. Then, solving this equation is equivalent to finding the minimum of the **quadratic form**

$$f(x) = \frac{1}{2}x^{\mathrm{T}}Ax - b^{\mathrm{T}}x + c.$$

Before proving this equivalence, we begin with a few definitions and notes on notation.

Throughout this note, unless otherwise mentioned, we focus on solving the *specific problem*:

$$A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ -8 \end{bmatrix}, \quad c = 0.$$

For the specific problem, the exact solution is

$$x = [2, -2]^{\mathrm{T}}.$$

In general, assume A is an $n \times n$ matrix, then the solution lies at the intersection point of n hyperplanes, each having dimensional $n - 1$.

The **gradient** of a quadratic form is defined to be

$$f'(x) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{bmatrix}.$$

By definition, for the quadratic form

$$\begin{aligned} f'(x) &= \frac{1}{2}A^{\mathrm{T}}x + \frac{1}{2}Ax - b \\ &= Ax - b \qquad (\text{because } A^{\mathrm{T}} = A) \end{aligned} \tag{1}$$

Furthermore, since the **Hessian**

$$f''(x) = (Ax - b)' = A^{\mathrm{T}} = A$$

is positive definite, according to basic calculas, the exact solution of $f'(x) = Ax - b = 0$ yields the exact minimum point of $f(x)$, and vice virsa.

2 Steepest Gradient Descent

In the method of Steepest Descent, we start at an arbitrary point $x_{(0)}$ and and slide down to the bottom of the paraboloid. We take a series of steps $x_{(1)}, x_{(2)}, \cdots$ until we are satisfied that we are close enough to the solution x. Note that when we take a step, we choose the direction in which f decreases most quickly, which is the direction opposite $f'(x_{(i)})$, and according to equation (1), this direction is $-f'(x_{(i)}) = b - Ax_{(i)}$.

Definition 2.1. *The* ***error*** $e_{(i)} = x_{(i)} - x$ *is a vector that indicates how far we are from the solution. The* ***residual*** $r_{(i)} = b - Ax_{(i)}$ *indicates how far we are from the correct value of* b*. It's clear that* $r_{(i)} = -Ae_{(i)}$.

Suppose we start at $x_{(0)} = [-2, 2]^{\mathrm{T}}$. Our first step is

$$x_{(1)} = x_{(0)} + \alpha r_{(0)}.$$

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

A **line search** os a procedure that chooses α to minimize f along a line. But for our task, we can analytically find the optimal α.

From basic calculas, α minimizes f when the directional derivative $\frac{d}{d\alpha} f(x_{(1)})$ is equal to zero. By the chain rule,

$$\frac{d}{d\alpha} f(x_{(1)}) = f'(x_{(1)})^{\mathrm{T}} \frac{d}{d\alpha} x_{(1)} = f'(x_{(1)})^{\mathrm{T}} r_{(0)}.$$

Setting this expression to zero, we find that α should be chosen so that $r_{(0)}$ and $f'(x_{(1)}) = -r_{(1)}$ are orthogonal:

$$\begin{aligned}
r_{(1)}^{\mathrm{T}} r_{(0)} &= 0, \\
(b - Ax_{(1)})^{\mathrm{T}} r_{(0)} &= 0, \\
(b - A(x_{(0)} + \alpha r_{(0)}))^{\mathrm{T}} r_{(0)} &= 0, \\
(b - Ax_{(0)})^{\mathrm{T}} r_{(0)} - \alpha (Ar_{(0)})^{\mathrm{T}} r_{(0)} &= 0, \\
(b - Ax_{(0)})^{\mathrm{T}} r_{(0)} &= \alpha (Ar_{(0)})^{\mathrm{T}} r_{(0)}, \\
r_{(0)}^{\mathrm{T}} r_{(0)} &= \alpha r_{(0)}^{\mathrm{T}} (Ar_{(0)}), \\
\alpha &= \frac{r_{(0)}^{\mathrm{T}} r_{(0)}}{r_{(0)}^{\mathrm{T}} A r_{(0)}}.
\end{aligned}$$

Putting these all together, the method of steepest descent is:

$$\begin{aligned} r_{(i)} &= b - Ax_{(i)}, \\ \alpha_{(i)} &= \frac{r_{(i)}^T r_{(i)}}{r_{(i)}^T A r_{(i)}}, \\ x_{(i+1)} &= x_{(i)} + \alpha_{(i)} r_{(i)}. \end{aligned}$$

3 Conjugate Directions

Steepest Descent often finds itself taking steps in the same direction as earlier steps. Wouldn't it be better if, every time we took a step, we got it right the first time? Here's an idea: let's pick a set of (A-) orthogonal search directions $d_{(0)}, d_{(1)}, \ldots, d_{(n-1)}$. In each search direction, we'll take exactly one step, and that step will be just the right length to line up evenly with x. After n steps, we'll be done.

3.1 An attempt

We try to ensure that all search directions are orthogonal to one another. Suppose for each step, we choose a point

$$x_{(i+1)} = x_{(i)} + \alpha_{(i)} d_{(i)}.$$

To find the value of $\alpha_{(i)}$, use the fact that e_{i+1} should be orthogonal to $d_{(i)}$, so we have

$$\begin{aligned} d_{(i)}^{\mathrm{T}} e_{(i+1)} &= 0, \\ d_{(i)}^{\mathrm{T}} (e_{(i)} + \alpha_{(i)} d_{(i)}) &= 0, \\ \alpha_{(i)} &= -\frac{d_{(i)}^{\mathrm{T}} e_{(i)}}{d_{(i)}^{\mathrm{T}} d_{(i)}}. \end{aligned}$$

Unfortunately, we haven't accomplished anything, because we can't compute $\alpha_{(i)}$ without knowing $e_{(i)}$; and if we knew $e_{(i)}$, the problem would already be solved.

3.2 Modification

In fact, we will make the search directions A-orthogonal instead of orthogonal. Two vectors $d_{(i)}$ and $d_{(j)}$ are **A-orthogonal**, or **conjugate**, if

$$d_{(i)}^{\mathrm{T}} A d_{(j)} = 0.$$

Now, our new requirement is that $e_{(i+1)} =$ be A-orthogonal to $d_{(i)}$. Hence, similar to former derivatives, we obtain that

$$\begin{aligned}\alpha_{(i)} &= -\frac{d_{(i)}^{\mathrm{T}} A e_{(i)}}{d_{(i)}^{\mathrm{T}} A d_{(i)}} \\ &= \frac{d_{(i)}^{\mathrm{T}} r_{(i)}}{d_{(i)}^{\mathrm{T}} A d_{(i)}}.\end{aligned} \tag{2}$$

Then, we prove that this procedure really does compute x in n steps.

Express the error term as a linear combination of search directions; namely,

$$e_{(0)} = \sum_{j=0}^{n-1} \delta_j d_{(j)}. \tag{3}$$

The values of δ_j can be found by a mathematical trick. Because the search directions are A-orthogonal, it's possible to eliminate all the δ_j values but one from expression (3) by premultiplying the expression by $d_{(k)}^{\mathrm{T}} A$:

$$d_{(k)}^{\mathrm{T}} A e_{(0)} = \sum_{j=0}^{n-1} \delta_j d_{(j)} A d_{(k)}^{\mathrm{T}} = \delta_k d_{(k)} A d_{(k)}^{\mathrm{T}},$$

$$\delta_{(k)} = \frac{d_{(k)}^{\mathrm{T}} A e_{(0)}}{d_{(k)}^{\mathrm{T}} A d_{(k)}} = \frac{d_{(k)}^{\mathrm{T}} A (e_{(0)} + \sum_{i=0}^{k-1} \alpha_{(i)} d_{(i)})}{d_{(k)}^{T} A d_{(k)}} = \cdots = \frac{d_{(k)}^{\mathrm{T}} A e_{(k)}}{d_{(k)}^{\mathrm{T}} A d_{(k)}}.$$

Hence, associated with equation (2), we find that

$$\alpha_{(i)} = -\delta_{(i)}.$$

Therefore,

$$\begin{aligned}e_{(i)} &= e_{(0)} + \sum_{j=0}^{i-1} \alpha_{(j)} d_{(j)} \\ &= \sum_{j=0}^{n-1} \delta_{(j)} d_{(j)} - \sum_{j=0}^{i-1} \delta_{(j)} d_{(j)} \\ &= \sum_{j=i}^{n-1} \delta_{(j)} d_{(j)}.\end{aligned} \tag{4}$$

So, after n iterations, we have $e_{(n)} = 0$.

4 Conjugate Gradients Method

The famous **conjugate gradients method** follows directly from conjugate directions. The only question left is how to get the search directions

$$d_{(0)}, \cdots, d_{(n-1)}.$$

Naturally, we hope the search direction d_k to be r_k, since it is proved to be the steepest direction at current point. However, the resulting directions

$$u_{(0)} = r_{(0)}, u_{(1)} = r_{(1)}, \cdots, u_{(n)} = r_{(n)}$$

don't preserve A-orthogonality.

To solve the problem, we apply **Gram-Schmidt conjugation** to

$$u_{(0)}, \cdots, u_{(n-1)},$$

thus converting them to the conjugate directions

$$d_{(0)}, \cdots, d_{(n-1)}.$$

But before doing that, let's derive some useful properties about conjugate directions.

Definition 4.1. *For $1 \le i \le n$, define the i-dimensional subspace*

$$D_i = \operatorname{span}\{d_{(0)}, d_{(1)}, \ldots, d_{(i-1)}\}.$$

Lemma 4.1. *For $1 \le i \le n$,*

$$D_i = \operatorname{span}\{r_{(0)}, r_{(1)}, \ldots, r_{(i-1)}\}.$$

Lemma 4.2. *For $1 \le i \le n$, $r_{(i)}$ is orthogonal with subspace D_i.*

Proof. It's equivalent prove that, for $j < i$

$$\langle d_{(j)}, r_{(i)} \rangle = 0.$$

It follows directly from the definition that

$$r_{(i)} = -Ae_{(i)},$$

hence it's equivalent to prove that, for $j < i$

$$\langle d_{(j)}, e_{(i)} \rangle_A = 0.$$

According to equation (4),

$$e_{(i)} = e_{(j+1)} + \sum_{k=j+1}^{i-1} \alpha_{(k)} d_{(k)}.$$

Since $d_{(j)}$ is A-orthogonal to $e_{(j+1)}$ and $d_{(j+1)}, \ldots, d_{i-1}$, this completes the proof. $\square$

Lemma 4.3. *For $1 \le i \le n$,*

$$D_i = \operatorname{span}\{r_{(0)}, Ar_{(0)}, \ldots, A^{i-1} r_{(0)}\}. \tag{5}$$

Proof. Note that

$$\begin{aligned} r_{(k+1)} &= -Ae_{(k+1)} \\ &= -A(e_{(k)} + \alpha_{(k)} d_{(k)}) \\ &= r_{(k)} - \alpha_{(k)} A d_{(k)}. \end{aligned}$$

So each new subspace D_{i+1} is formed the union of the previous subspace D_i and the subspace AD_i.

Hence,

$$\begin{aligned} D_i &= \operatorname{span}\{d_{(0)}, Ad_{(0)}, A^2 d_{(0)}, \ldots, A^{i-1} d_{(0)}\} \\ &= \operatorname{span}\{r_{(0)}, Ar_{(0)}, A^2 r_{(0)}, \ldots, A^{i-1} r_{(0)}\}. \end{aligned}$$

□

This subspace is called a **Krylov subspace**, a subspace created by repeatedly applying a matrix to a vector. It has a pleasing property: because AD_i is included in D_{i+1}, the fact that the next residual $r_{(i+1)}$ is orthogonal to D_{i+1} (from lemma 4.3) implies that $r_{(i+1)}$ is A-orthogonal to D_i. Gram-Schmidt conjugation becomes easy, because $r_{(i+1)}$ is already A-orthogonal to all of the previous directions except $d_{(i)}$!

Let's put it all together into one piece now. The method of Conjugate Gradients is:

Algorithm 1 Conjugate gradients method

$\boldsymbol{r}_0 := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$
if $\boldsymbol{r}_0$ is sufficiently small, then return $\boldsymbol{x}_0$ as the result
$\boldsymbol{d}_0 := \boldsymbol{r}_0$
$k := 0$
while true **do**
    $\alpha_k := \frac{\boldsymbol{r}_k^{\mathrm{T}}}{\boldsymbol{d}_k^{\mathrm{T}} \boldsymbol{A} \boldsymbol{d}_k}$
    $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$
    $\boldsymbol{r}_{k+1} := \boldsymbol{r}_k - \alpha_k \boldsymbol{A} \boldsymbol{d}_k$
    if $\boldsymbol{r}_{k+1}$ is sufficiently small, then exit loop
    $\beta_k := \frac{\boldsymbol{r}_{k+1}^{\mathrm{T}} \boldsymbol{r}_{k+1}}{\boldsymbol{r}_k^{\mathrm{T}} \boldsymbol{r}_k}$
    $\boldsymbol{d}_{k+1} := \boldsymbol{r}_{k+1} + \beta_k \boldsymbol{d}_k$ (Gram-Schmidt conjugation)
    $k := k + 1$
end while
return $\boldsymbol{x}_{k+1}$ as the result

References

[1] Jonathan Richard Shewchuk et al. An introduction to the conjugate gradient method without the agonizing pain, 1994.