Conjugate Gradients

Michael Kazhdan

(600.657)

Anouncements

Information about the Seminar (600.757) have been posted online:

http://www.cs.jhu.edu/~misha

Tech Specs:

- Meet on Tuesday afternoon.
- Two papers discussed each week.
- Votes for next week's candidate papers due in by Thursday evening.

Outline

Review of Steepest Descent

Conjugate Gradients

Steepest Descent

Review:

The idea behind this approach is to interpret the solution of the equation $Ax=b$ as the minimization of the function:

$$F(x)=\frac{x^t Ax}{2}-b^t x$$

Steepest Descent

Review:

The idea behind this approach is to interpret the solution of the equation $Ax=b$ as the minimization of the function:

$$F(x)=\frac{x^t Ax}{2}-b^t x$$

Given a guess for the solution, x_i, the next guess, x_{i+1}, is generated by taking a step in the direction opposite to the direction in which F increases:

$$x_{i+1}=x_i-t\cdot\nabla F(x_i)$$

Steepest Descent

Review:

Since the gradient of F at x_i is the residual:

$$\nabla F(x_i)=Ax_i-b:=r_i$$

this gives the update rule:

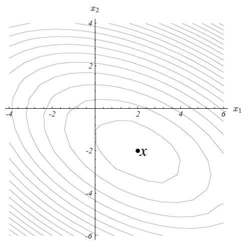
$$x_{i+1}=x_i-tr_i \qquad \text{with} \qquad t=\frac{r_i^t r_i}{r_i^t Ar_i}$$

Steepest Descent

Example:

$$A = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \qquad b = \begin{pmatrix} 2 \\ -8 \end{pmatrix}$$

For this matrix A and this vector b, the plot of the iso-contours of the function $F(x)$ is shown on the right.



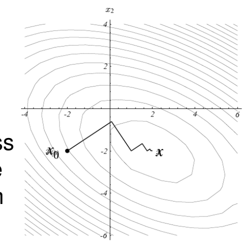
Shewchuk, 1994

Steepest Descent

Example:

$$A = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \qquad b = \begin{pmatrix} 2 \\ -8 \end{pmatrix}$$

Starting with an initial guess x_0, if we iterate through the steepest descent algorithm we make the steps:



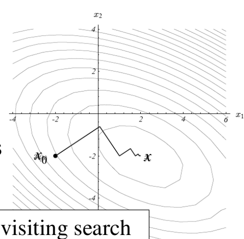
Shewchuk, 1994

Steepest Descent

Example:

$$A = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \qquad b = \begin{pmatrix} 2 \\ -8 \end{pmatrix}$$

Starting with an initial guess x_0, if we iterate through the steepest descent algorithm we make



We often end up re-visiting search directions we had already tried

Shewchuk, 1994

Outline

Review of Steepest Descent

Conjugate Gradients

Conjugate Gradients

Goal:

To define an iterative approach that:

1. Gets us closer and closer to the solution.
2. Ensures we do not visit the same direction twice.

Conjugate Gradients

To do this, we will think of working with the sequence of errors $\{e_0,\ldots,e_i,\ldots\}$ rather than the sequence of guesses $\{x_0,\ldots,x_i,\ldots\}$:

$$e_i = x - x_i$$

Conjugate Gradients

To do this, we will think of working with the sequence of errors $\{e_0,\dots,e_i,\dots\}$ rather than the sequence of guesses $\{x_0,\dots,x_i,\dots\}$:

$$e_i = x - x_i$$

That is, rather than trying to generate a sequence of guesses with:

$$\lim_{i\to\infty} x_i = x$$

Conjugate Gradients

To do this, we will think of working with the sequence of errors $\{e_0,\dots,e_i,\dots\}$ rather than the sequence of guesses $\{x_0,\dots,x_i,\dots\}$:

$$e_i = x - x_i$$

That is, rather than trying to generate a sequence of guesses with:

$$\lim_{i\to\infty} x_i = x$$

We try to generate a sequence of errors with:

$$\lim_{i\to\infty} e_i = 0$$

Conjugate Gradients

Note:

If we think of an update rule as adding some vector ε_i to x_i to give us x_{i+1}:

$$x_{i+1} = x_i + \varepsilon_i$$

Conjugate Gradients

Note:

If we think of an update rule as adding some vector ε_i to x_i to give us x_{i+1}:

$$x_{i+1} = x_i + \varepsilon_i$$

This is equivalent to subtracting the vector ε_i from e_i to give us e_{i+1}:

$$e_{i+1} = e_i - \varepsilon_i$$

Conjugate Gradients (First Pass)

Approach:

Suppose that we have an initial guess x_0 and we have a set of orthonormal directions $\{d_0,\dots,d_{n-1}\}$.

Conjugate Gradients (First Pass)

Approach:

Suppose that we have an initial guess x_0 and we have a set of orthonormal directions $\{d_0,\dots,d_{n-1}\}$.

We would like to design an algorithm that defines the (i+1)-st error by removing the component of the error lying along the d_i direction.

$$e_{i+1} = e_i - \langle e_i, d_i \rangle d_i$$

Conjugate Gradients (First Pass)

Claim:

This method is guaranteed to get the right answer after *n* iterations.

Conjugate Gradients (First Pass)

Proof:

Since the $\{d_0,\ldots,d_{n-1}\}$ are orthonormal, we can write the error in the initial guess as:

$$e_0 = \sum_{i=0}^{n-1} \langle e_0, d_i \rangle d_i$$

Conjugate Gradients (First Pass)

Proof:

Since the $\{d_0,\ldots,d_{n-1}\}$ are orthonormal, we can write the error in the initial guess as:

$$e_0 = \sum_{i=0}^{n-1} \langle e_0, d_i \rangle d_i$$

After the first iteration we have:

$$e_1 = e_0 - \langle e_0, d_0 \rangle d_0$$

Conjugate Gradients (First Pass)

Proof:

Since the $\{d_0,\ldots,d_{n-1}\}$ are orthonormal, we can write the error in the initial guess as:

$$e_0 = \sum_{i=0}^{n-1} \langle e_0, d_i \rangle d_i$$

After the first iteration we have:

$$e_1 = e_0 - \langle e_0, d_0 \rangle d_0 = \sum_{i=1}^{n-1} \langle e_0, d_i \rangle d_i$$

Conjugate Gradients (First Pass)

Proof:

Since the $\{d_0,\ldots,d_{n-1}\}$ are orthonormal, we can write the error in the initial guess as:

$$e_0 = x_0 - x = \sum_{i=0}^{n-1} \langle e_0, d_i \rangle d_i$$

After the *k*-th iteration we have:

$$e_k = \sum_{i=k}^{n-1} \langle e_0, d_i \rangle d_i$$

Conjugate Gradients (First Pass)

Proof:

Since the $\{d_0,\ldots,d_{n-1}\}$ are orthonormal, we can write the error in the initial guess as:

$$e_0 = x_0 - x = \sum_{i=0}^{n-1} \langle e_0, d_i \rangle d_i$$

And after the *n*-th iteration we have:

$$e_n = \sum_{i=n}^{n-1} \langle e_0, d_i \rangle d_i = 0$$

Conjugate Gradients (First Pass)

Problem:

We don't know the correct solution *x*…

Conjugate Gradients (First Pass)

Problem:

We don't know the correct solution *x*…

We don't know the value of $e_0=x-x_0$…

Conjugate Gradients (First Pass)

Problem:

We don't know the correct solution *x*…

We don't know the value of $e_0=x-x_0$…

We can't figure out what the component of the error in direction d_i is:

$$\langle e_0, d_i \rangle = ?$$

Conjugate Gradients

Solution:

To address this problem, we will change our notion of "distance" so that we can compute the component of the error in direction d_i without ever knowing the value of *x*.

Conjugate Gradients

Observation:

If we have a symmetric positive definite matrix *A*, we can think of the matrix as defining a new inner-product:

$$\langle u, v \rangle_A = \langle u, Av \rangle$$

Conjugate Gradients

Observation:

If we have a symmetric positive definite matrix *A*, we can think of the matrix as defining a new inner-product:

$$\langle u, v \rangle_A = \langle u, Av \rangle$$

This new inner product has the same properties that the traditional inner product has:

1. Symmetry: $\langle u, v \rangle_A = \langle v, u \rangle_A$
2. Positivity: $\langle u, u \rangle_A \geq 0$
3. Definiteness: $\langle u, u \rangle_A = 0 \Leftrightarrow u = 0$

Conjugate Gradients

Key Idea:

Although we cannot compute the dot-product:

$$\langle e_0, d_i \rangle = \langle x - x_0, d_i \rangle$$

using the traditional inner-product…

Conjugate Gradients

Key Idea:

Although we cannot compute the dot-product:

$$\langle e_0, d_i \rangle = \langle x - x_0, d_i \rangle$$

using the traditional inner-product…

We can compute it using the inner-product defined by *A*:

$$\langle e_0, d_i \rangle_A = \langle x - x_0, d_i \rangle_A$$

Conjugate Gradients

Key Idea:

Although we cannot compute the dot-product:

$$\langle e_0, d_i \rangle = \langle x - x_0, d_i \rangle$$

using the traditional inner-product…

We can compute it using the inner-product defined by *A*:

$$\begin{aligned}\langle e_0, d_i \rangle_A &= \langle x - x_0, d_i \rangle_A \\ &= \langle A(x - x_0), d_i \rangle\end{aligned}$$

Conjugate Gradients

Key Idea:

Although we cannot compute the dot-product:

$$\langle e_0, d_i \rangle = \langle x - x_0, d_i \rangle$$

using the traditional inner-product…

We can compute it using the inner-product defined by *A*:

$$\begin{aligned}\langle e_0, d_i \rangle_A &= \langle x - x_0, d_i \rangle_A \\ &= \langle A(x - x_0), d_i \rangle \\ &= \langle b - Ax_0, d_i \rangle\end{aligned}$$

Conjugate Gradients

Approach:

If the vectors $\{d_0,\ldots,d_{n-1}\}$ are *A*-orthonormal:

$$\langle d_i, d_j \rangle_A = \delta_{ij}$$

Conjugate Gradients

Approach:

If the vectors $\{d_0,\ldots,d_{n-1}\}$ are *A*-orthonormal:

$$\langle d_i, d_j \rangle_A = \delta_{ij}$$

We can define an analogous algorithm, starting with an initial error e_0 we generate the errors e_i by successively removing the error component in direction d_i:

$$e_{i+1} = e_i - \langle e_0, d_i \rangle_A d_i$$

Conjugate Gradients

Approach:

If the vectors $\{d_0,\ldots,d_{n-1}\}$ are A-orthonormal:

$$\langle d_i, d_j \rangle_A = \delta_{ij}$$

We can define an analogous algorithm, starting with an initial error e_0 we generate the errors e_i by successively removing the error component in direction d_i:

$$e_{i+1} = e_i - \langle e_0, d_i \rangle_A d_i$$

As before, this method is guaranteed to give the correct answer after n iterations.

Conjugate Gradients

Approach:

If the vectors $\{d_0,\ldots,d_{n-1}\}$ are A-orthonormal:

$$\langle d_i, d_j \rangle_A = \delta_{ij}$$

We can define an analogous algorithm, starting with an initial error e_0 we generate the errors e_i by successively removing the error component in direction d_i:

$$e_{i+1} = e_i - \langle e_0, d_i \rangle_A d_i$$

As before, this method is guaranteed to give

However, it does not require knowing the vector x in advance, only b.

Conjugate Gradients

Conceptually:

Since we don't know the solution x, we cannot really talk about updating the error e_i.

Conjugate Gradients

Conceptually:

Since we don't know the solution x, we cannot really talk about updating the error e_i.

However, we can talk about updating the residual:

$$r_i = Ae_i = b - Ax_i$$

Conjugate Gradients

Conceptually:

In this context, the update step becomes:

$$e_{i+1} = e_i - \langle e_i, d_i \rangle_A d_i$$

Conjugate Gradients

Conceptually:

In this context, the update step becomes:

$$e_{i+1} = e_i - \langle e_i, d_i \rangle_A d_i$$

⇓

$$Ae_{i+1} = A\left(e_i - \langle e_i, d_i \rangle_A d_i\right)$$

Conjugate Gradients

Conceptually:

In this context, the update step becomes:

$$e_{i+1} = e_i - \langle e_i, d_i \rangle_A d_i$$

$$\Downarrow$$

$$Ae_{i+1} = A\left(e_i - \langle e_i, d_i \rangle_A d_i\right)$$

$$\Downarrow$$

$$r_{i+1} = r_i - \langle r_i, d_i \rangle A d_i$$

Conjugate Gradients

Question:

How do we generate a good set of search directions $\{d_0,\ldots,d_{n-1}\}$?

- The directions are *A*-orthonormal.
- The directions have the property that most of the convergence happens early on (so we don't have to run a full *n* iterations).

Conjugate Gradients

$$F(x) = \frac{x^t A x}{2} - b^t x$$

$$\nabla F(x_i) = Ax_i - b$$

Choosing Directions:

Choosing the first direction d_0 is easy.

Conjugate Gradients

$$F(x) = \frac{x^t A x}{2} - b^t x$$

$$\nabla F(x_i) = Ax_i - b$$

Choosing Directions:

Choosing the first direction d_0 is easy.

Given the guess x_0, we want to choose a direction to update in order to minimize $F(x)$.

Conjugate Gradients

$$F(x) = \frac{x^t A x}{2} - b^t x$$

$$\nabla F(x_i) = Ax_i - b = r_0$$

Choosing Directions:

Choosing the first direction d_0 is easy.

Given the guess x_0, we want to choose a direction to update in order to minimize $F(x)$.

Using the fact that the gradient at x_0 is:

$$r_0 = \nabla F(x_0)$$

Conjugate Gradients

$$F(x) = \frac{x^t A x}{2} - b^t x$$

$$\nabla F(x_i) = Ax_i - b = r_0$$

Choosing Directions:

Choosing the first direction d_0 is easy.

Given the guess x_0, we want to choose a direction to update in order to minimize $F(x)$.

Using the fact that the gradient at x_0 is:

$$r_0 = \nabla F(x_0)$$

this gives:

$$d_0 = \frac{r_0}{\|r_0\|_A}$$

Conjugate Gradients

$$F(x) = \frac{x^t A x}{2} - b^t x$$

$$\nabla F(x_i) = Ax_i - b = r_0$$

Choosing Directions:

To choose the next direction d_1, we start with the gradient direction:

$$d_1 \approx \nabla F(x_1) = r_1$$

and update it so that $\{d_0, d_1\}$ are A-orthonormal:

$$d_1 = \frac{r_1 - \langle r_1, d_0 \rangle_A d_0}{\left\| r_1 - \langle r_1, d_0 \rangle_A d_0 \right\|_A}$$

Conjugate Gradients

$$F(x) = \frac{x^t A x}{2} - b^t x$$

$$\nabla F(x_i) = Ax_i - b = r_0$$

Choosing Directions:

To choose the next direction d_1, we start with the

The problem with this approach is that it smacks of Gram-Schmidt orthogonalization.

and update it so that $\{d_0, d_1\}$ are A-orthonormal:

$$d_1 = \frac{r_1 - \langle r_1, d_0 \rangle_A d_0}{\left\| r_1 - \langle r_1, d_0 \rangle_A d_0 \right\|_A}$$

Conjugate Gradients

$$F(x) = \frac{x^t A x}{2} - b^t x$$

$$\nabla F(x_i) = Ax_i - b = r_0$$

Choosing Directions:

To choose the next direction d_1, we start with the

The problem with this approach is that it smacks of Gram-Schmidt orthogonalization.

ar

Generating the vector d_i requires computing the dot-product with all d_j, where $j<i$.

$$d_1 = \frac{r_1 - \langle r_1, d_0 \rangle_A d_0}{\left\| r_1 - \langle r_1, d_0 \rangle_A d_0 \right\|_A}$$

Conjugate Gradients

$$F(x) = \frac{x^t A x}{2} - b^t x$$

$$\nabla F(x_i) = Ax_i - b = r_0$$

Choosing Directions:

To choose the next direction d_1, we start with the

The problem with this approach is that it smacks of Gram-Schmidt orthogonalization.

ar

Generating the vector d_i requires computing

The complexity of computing the first i directions is $O(i^2 n)$.

$$d_1 = \frac{\cdots}{\left\| r_1 - \langle r_1, d_0 \rangle_A d_0 \right\|_A}$$

Conjugate Gradients

$$F(x) = \frac{x^t A x}{2} - b^t x$$

$$\nabla F(x_i) = Ax_i - b = r_0$$

Choosing Directions:

Turns out that life is not so bad.

Conjugate Gradients

$$F(x) = \frac{x^t A x}{2} - b^t x$$

$$\nabla F(x_i) = Ax_i - b = r_0$$

Choosing Directions:

Turns out that life is not so bad.

For any $j<i$, the residual r_{i+1} satisfies the property:

$$\langle r_{i+1}, d_j \rangle_A = 0$$

Conjugate Gradients

$$F(x)=\frac{x^t Ax}{2}-b^t x$$

$$\nabla F(x_i)=Ax_i-b=r_0$$

Choosing Directions:

Turns out that life is not so bad.

For any $j<i$, the residual r_{i+1} satisfies the property:

$$\langle r_{i+1},d_j\rangle_A=0$$

Thus, performing the Gram-Schmidt orthogonalization only requires two dot-products.

$$d_{i+1}=\frac{r_{i+1}-\langle r_{i+1},d_j\rangle_A d_j}{\left\|r_{i+1}-\langle r_{i+1},d_j\rangle_A d_j\right\|_A}$$

Conjugate Gradients

Proof:

To show this, we will use two facts:

1. The i-th residual, r_i, is orthogonal (in the traditional sense) to all directions d_k where $k<i$.
2. The vector Ad_k can be expressed as the linear sum of the vectors $\{d_0,\dots,d_{k+1}\}$.

Conjugate Gradients

Proof:

To show this, we will use two facts:

1. The i-th residual, r_i, is orthogonal (in the traditional sense) to all directions d_k where $k<i$.
2. The vector Ad_k can be expressed as the linear sum of the vectors $\{d_0,\dots,d_{k+1}\}$.

Assume True:

Then for any $k<i$, we have:

$$\langle r_{i+1},d_k\rangle_A=\langle r_{i+1},Ad_k\rangle$$

Conjugate Gradients

Proof:

To show this, we will use two facts:

1. The i-th residual, r_i, is orthogonal (in the traditional sense) to all directions d_k where $k<i$.
2. The vector Ad_k can be expressed as the linear sum of the vectors $\{d_0,\dots,d_{k+1}\}$.

Assume True:

Then for any $k<i$, we have:

$$\langle r_{i+1},d_k\rangle_A=\langle r_{i+1},Ad_k\rangle$$

$$=\sum_{j=0}^{k+1}\alpha_j\langle r_{i+1},d_j\rangle$$

Conjugate Gradients

Proof:

To show this, we will use two facts:

1. The i-th residual, r_i, is orthogonal (in the traditional sense) to all directions d_k where $k<i$.
2. The vector Ad_k can be expressed as the linear sum of the vectors $\{d_0,\dots,d_{k+1}\}$.

Assume True:

Then for any $k<i$, we have:

$$\langle r_{i+1},d_k\rangle_A=\langle r_{i+1},Ad_k\rangle$$

$$=\sum_{j=0}^{k+1}\alpha_j\langle r_{i+1},d_j\rangle=0$$

Conjugate Gradients

Claim 1:

The i-th residual, r_i, is orthogonal (in the traditional sense) to all directions d_k where $k<i$.

Conjugate Gradients

Proof:

Since we have:

$$r_i = r_0 - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle A d_j$$

Conjugate Gradients

Proof:

Since we have:

$$r_i = r_0 - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle A d_j$$

We know that for *k<i:*

$$\langle r_i, d_k \rangle = \langle r_0, d_k \rangle - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle \langle A d_j, d_k \rangle$$

Conjugate Gradients

Proof:

Since we have:

$$r_i = r_0 - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle A d_j$$

We know that for *k<i:*

$$\langle r_i, d_k \rangle = \langle r_0, d_k \rangle - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle \langle A d_j, d_k \rangle$$

$$= \langle r_0, d_k \rangle - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle \langle d_j, d_k \rangle_A$$

Conjugate Gradients

Proof:

Since we have:

$$r_i = r_0 - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle A d_j$$

We know that for *k<i:*

$$\langle r_i, d_k \rangle = \langle r_0, d_k \rangle - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle \langle A d_j, d_k \rangle$$

$$= \langle r_0, d_k \rangle - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle \langle d_j, d_k \rangle_A$$

$$= \langle r_0, d_k \rangle - \langle r_0, d_k \rangle$$

Conjugate Gradients

Proof:

Since we have:

$$r_i = r_0 - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle A d_j$$

We know that for *k<i:*

$$\langle r_i, d_k \rangle = \langle r_0, d_k \rangle - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle \langle A d_j, d_k \rangle$$

$$= \langle r_0, d_k \rangle - \sum_{j=0}^{i-1} \langle r_0, d_j \rangle \langle d_j, d_k \rangle_A$$

$$= \langle r_0, d_k \rangle - \langle r_0, d_k \rangle = 0$$

Conjugate Gradients

Claim 2:

The vector Ad_k can be expressed as the linear sum of the vectors $\{d_0,\ldots,d_{k+1}\}$.

Conjugate Gradients

Claim 2:

The vector Ad_k can be expressed as the linear sum of the vectors $\{d_0,\ldots,d_{k+1}\}$.

Proof:

Let us denote by D^i the vector sub-space:

$$D^i = \mathrm{Span}\{d_0,\ldots,d_i\}$$

Conjugate Gradients

Claim 2:

The vector Ad_k can be expressed as the linear sum of the vectors $\{d_0,\ldots,d_{k+1}\}$.

Proof:

Let us denote by D^i the vector sub-space:

$$D^i = \mathrm{Span}\{d_0,\ldots,d_i\}$$

We would like to show that $Ad_k \subset D^{k+1}$.

Conjugate Gradients

Proof:

$$D^i = \mathrm{Span}\{d_0,\ldots,d_i\}$$

Since d_i is obtained by computing the component of r_i orthogonal to $\{d_0,\ldots,d_{i-1}\}$, we have:

$$D^i = \mathrm{Span}\{D^{i-1}, r_i\}$$

Conjugate Gradients

Proof:

$$D^i = \mathrm{Span}\{d_0,\ldots,d_i\}$$

Since d_i is obtained by computing the component of r_i orthogonal to $\{d_0,\ldots,d_{i-1}\}$, we have:

$$D^i = \mathrm{Span}\{D^{i-1}, r_i\}$$

Continuing in a recursive fashion, we know that:

$$D^i = \mathrm{Span}\{r_0,\ldots,r_i\}$$

Conjugate Gradients

Proof:

But we also know that:

$$r_{i+1} = r_i - \langle r_i, d_i \rangle A d_i$$

Conjugate Gradients

Proof:

But we also know that:

$$\underset{D^{i+1}}{\underset{\Cap}{r_{i+1}}} = \underset{D^{i}}{\underset{\Cap}{r_i}} - \langle r_i, d_i \rangle A d_i$$

So that if $\langle r_i, d_i \rangle \neq 0$, we must have $Ad_i \subset D^{i+1}$.

Conjugate Gradients

Proof:

But we also know that:

$$\underset{D^{i+1}}{\underset{\Cap}{r_{i+1}}} = \underset{D^{i}}{\underset{\Cap}{r_i}} - \langle r_i, d_i \rangle A d_i$$

So that if $\langle r_i, d_i \rangle \neq 0$, we must have $Ad_i \subset D^{i+1}$.

(If $\langle r_i, d_i \rangle = 0$, this implies that the *i*-th residual is zero and we have reached the solution at step *i*.)