Conjugate Direction Methods

GIAN Short Course on Optimization: Applications, Algorithms, and Computation

Sven Leyffer

Argonne National Laboratory

September 12-24, 2016

Outline

Exact Line-Search for Quadratics

Analysis uses exact line-search arguments.
Consider quadratic

$$q(x) = \frac{1}{2}x^T Gx + b^T x$$

and perform an exact line-search: $\hat{x} + \alpha s$:

$$\underset{\alpha \geq 0}{\text{minimize}} \;\; q(\hat{x} + \alpha s) = \tfrac{1}{2}(\hat{x} + \alpha s)^T G(\hat{x} + \alpha s) + b^T(\hat{x} + \alpha s)$$

Re-arrange quadratic as

$$q(\hat{x} + \alpha s) = \tfrac{1}{2}\alpha^2 s^T Gs + \alpha \left(s^T G\hat{x} + b^T s\right) + \tfrac{1}{2}\hat{x}^T G\hat{x} + b^T \hat{x}$$

Setting $\frac{dq}{d\alpha} = 0$ we get:

$$0 = \alpha s^T Gs + s^T (G\hat{x} + b) \quad \Leftrightarrow \quad \alpha = -\frac{s^T(G\hat{x} + b)}{s^T Gs} = \frac{-s^T \nabla q(\hat{x})}{s^T Gs}$$

Conjugate Direction Methods

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$$

Conjugate direction methods relate to a quadratic model of $f(x)$.

Definition (Conjugacy)

$m \leq n$ nonzero vectors, $s^{(1)}, \ldots, s^{(m)} \in \mathbb{R}^n$ are *conjugate wrt positive definite Hessian G*, iff $s^{(i)^T} G s^{(j)} = 0$ for all $i \neq j$.

- Conjugacy is orthogonality across positive definite Hessian, G.
- For $G = I$, get orthogonality.

Definition (Conjugacy)

A *conjugate direction method* generates conjugate directions applied to a positive definite quadratic.

Conjugate Direction Methods

Theorem (Linear Independence of Conjugate Directions)

A set of m conjugate directions is linearly independent.

Proof. $s^{(1)}, \ldots, s^{(m)} \in \mathbb{R}^n$ conjugate. Consider $\sum_{i=1}^{m} a_i s^{(i)} = 0$

... need to show $a_i = 0$ is only solution of this system

G positive definite $\Rightarrow$ G nonsingular, hence

$$\sum_{i=1}^{m} a_i s^{(i)} = 0 \quad \Leftrightarrow \quad G\left(\sum_{i=1}^{m} a_i s^{(i)}\right) = 0.$$

Pre-multiply by $s^{(j)}$:

$$s^{(j)^T} G \left(\sum_{i=1}^{m} a_i s^{(i)}\right) = 0 \quad \Leftrightarrow \quad a_j s^{(j)^T} G s^{(j)} = 0 \quad \Leftrightarrow \quad a_j = 0,$$

because G positive definite. $\square$

Conjugate Direction Methods

Theorem (Termination of Conjugate Direction Methods)

- *A conjugate direction method terminates for a positive definite quadratic in at most n exact line-searches.*
- *Each iterate, $x^{(k+1)}$ reached by $k \leq n$ descend steps along conjugate directions $s^{(1)}, \ldots, s^{(k)} \in \mathbb{R}^n$.*

Proof. Define the quadratic as

$$q(x) = \frac{1}{2} x^T G x + b^T x.$$

Conjugate direction, $s^{(k)}$, gives $k+1$ iterate as

$$x^{(k+1)} = x^{(k)} + \alpha_k s^{(k)} = \ldots = x^{(1)} + \sum_{j=1}^{k} \alpha_j s^{(j)} = x^{(i+1)} + \sum_{j=i+1}^{k} \alpha_j s^{(j)}.$$

Conjugate Direction Methods

Proof cont.

From previous page: Conjugate direction, $s^{(k)}$, give iterates

$$x^{(k+1)} = x^{(k)} + \alpha_k s^{(k)} = \ldots = x^{(1)} + \sum_{j=1}^{k} \alpha_j s^{(j)} = x^{(i+1)} + \sum_{j=i+1}^{k} \alpha_j s^{(j)}.$$

Corresponding gradient of quadratic is

$$g^{(k+1)} = Gx^{(k+1)} + b = G\left(x^{(i+1)} + \sum_{j=i+1}^{k} \alpha_j s^{(j)} \right) + b$$

$$\Rightarrow g^{(k+1)} = g^{(i+1)} + \sum_{j=i+1}^{k} \alpha_j G s^{(j)}$$

Pre-multiply by $s^{(i)}$ we get

$$s^{(i)^T} g^{(k+1)} = s^{(i)^T} g^{(i+1)} + \sum_{j=i+1}^{k} \alpha_j s^{(i)^T} G s^{(j)} = 0, \quad \forall i = 1, \ldots, k-1,$$

Conjugate Direction Methods

Proof cont.

From previous: pre-multiply by $s^{(i)}$ we get

$$s^{(i)^T} g^{(k+1)} = s^{(i)^T} g^{(i+1)} + \sum_{j=i+1}^{k} \alpha_j s^{(i)^T} G s^{(j)} = 0, \quad \forall i = 1, \ldots, k-1,$$

where

- $s^{(i)^T} g^{(i+1)} = 0$ due to exact line search.
- $s^{(i)^T} G s^{(j)} = 0$ due to conjugacy.
- $s^{(k)^T} g^{(k+1)} = 0$ due to exact line-search.

Hence,

$$s^{(i)^T} g^{(k+1)} = 0, \ \forall i = 1, \ldots, k.$$

Now, let $k = n$, then it follows that

$$s^{(i)^T} g^{(n+1)} = 0, \ \forall i = 1, \ldots, n \qquad \Rightarrow \qquad g^{(n+1)} = 0$$

because, $g^{(n+1)}$ orthogonal to n linearly independent vectors $\square$

Conjugate Direction Methods

Remark

Previous Theorem holds for all conjugate direction methods!

Methods differ how $s^{(k)}$ constructed without knowing Hessian

Conjugate Direction Line-Search Method

Given $x^{(0)}$, set $k = 0$. **repeat**

Compute the conjugate direction $s^{(k)}$.

Compute the steplength $\alpha_k := \text{Armijo}(f(x), x^{(k)}, s^{(k)})$

Set $x^{(k+1)} := x^{(k)} + \alpha_k s^{(k)}$ and $k = k + 1$.

until $x^{(k)}$ *is (local) optimum*;

... next consider different ways to create conjugate directions.

Outline

Classical Conjugate Gradient Method

Idea Behind Conjugate Gradients

Modify steepest descend so that directions are conjugate.

Start by deriving method for quadratic

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; q(x) = \frac{1}{2} x^T G x + b^T x$$

then generalize to nonlinear $f(x)$.

Start with $s^{(0)} = -g^{(0)}$, steepest descend direction
$\Rightarrow$ first step guaranteed to be downhill ... no stalling like Newton!

Classical Conjugate Gradient Method

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; q(x) = \frac{1}{2} x^T G x + b^T x$$

Start with $s^{(0)} = -g^{(0)}$, steepest descend direction
Choose $s^{(1)}$ as component of $-g^{(1)}$ conjugate to $s^{(0)}$:

$$s^{(1)} = -g^{(1)} + \beta_0 s^{(0)}$$

Look for formula for β_0 such that conjugacy holds, i.e.

$$0 = {s^{(0)}}^T G s^{(1)} = {s^{(0)}}^T G \left(-g^{(1)} + \beta_0 s^{(0)} \right).$$

Solve for β_0, and get

$$\beta_0 = \frac{{s^{(0)}}^T G g^{(1)}}{{s^{(0)}}^T G s^{(0)}},$$

where ${s^{(0)}}^T G s^{(0)} \neq 0$, because G positive definite, and $s^{(0)} \neq 0$.

Classical Conjugate Gradient Method

Simplify formula for β_0:

$$\beta_0 = \frac{{s^{(0)}}^T G g^{(1)}}{{s^{(0)}}^T G s^{(0)}},$$

Recall, that

$$x^{(1)} = x^{(0)} + \alpha_1 s^{(0)} \Leftrightarrow s^{(0)} = \left(x^{(1)} - x^{(0)}\right)/\alpha_1,$$

where $\alpha_1 \neq 0$, because of steepest descend.
Now use $G\delta = \gamma$ to write β_0 as

$$\beta_0 = \frac{\left(x^{(1)} - x^{(0)}\right)^T G g^{(1)}}{\left(x^{(1)} - x^{(0)}\right)^T G s^{(0)}} = \frac{\left(g^{(1)} - g^{(0)}\right)^T g^{(1)}}{\left(g^{(1)} - g^{(0)}\right)^T s^{(0)}}$$

Exact line-search implies $0 = {g^{(1)}}^T s^{(0)} = -{g^{(1)}}^T g^{(0)}$, and thus

$$\beta_0 = \frac{{g^{(1)}}^T g^{(1)}}{{g^{(0)}}^T g^{(0)}}.$$

Classical Conjugate Gradient Method

Consider general step, k:

$$s^{(k)} = \text{the component of } -g^{(k)} \text{ conjugate to } s^{(0)}, \ldots, s^{(k-1)}.$$

Desired conjugacy:

$$s^{(k)^T} G s^{(j)} = 0, \ \forall j < k \quad \Leftrightarrow \quad s^{(k)^T} \gamma^{(j)} = 0, \ \forall j < k,$$

Use Gram-Schmidt orthogonalization procedure to get

$$s^{(k)} = -g^{(k)} + \sum_{j=0}^{k-1} \beta_j s^{(j)} \quad \text{Can } \beta_j = 0 \text{ for } j < k???$$

For quadratic, can show that $\beta_j = 0, \ \forall j < k$. Hence get:

$$s^{(k)} = -g^{(k)} + \beta_{k-1} s^{(k-1)} \quad \text{where} \quad \beta_{k-1} = \begin{cases} 0 & \text{if } k = 0 \\ \dfrac{g^{(k)^T} g^{(k)}}{g^{(k-1)^T} g^{(k-1)}} & \text{otherwise} \end{cases}$$

Fletcher-Reeves conjugate gradient method

Classical Conjugate Gradient Method

Min. quadratic $q(x) = \frac{1}{2}x^T G x + b^T x$ with Fletcher-Reeves (FR)

$$s^{(k)} = -g^{(k)} + \beta_{k-1} s^{(k-1)} \quad \text{where} \quad \beta_{k-1} = \begin{cases} 0 & \text{if } k = 0 \\ \dfrac{g^{(k)^T} g^{(k)}}{g^{(k-1)^T} g^{(k-1)}} & \text{otherwise} \end{cases}$$

Theorem (Convergence of FR for Convex Quadratics)

FR with exact line-search terminates at stationary point, $x^{(m)}$ after $m \leq n$ iterations for a pos. definite quadratic. Moreover, for $0 \leq i \leq m-1$, we have that:

1. *Conjugate search directions:* $s^{(i)^T} G s^{(j)} = 0 \; \forall i \neq j,\; j < i$.
2. *Orthogonal gradients:* $g^{(i)^T} g^{(j)} = 0 \; \forall i \neq j,\; j = 1, \ldots, i-1$.
3. *Descend property:* $s^{(i)^T} g^{(j)} = -g^{(i)^T} g^{(j)} < 0 \; \forall i \neq j$.

Proof of Fletcher-Reeves Convergence

Theorem (Convergence of FR for Convex Quadratics)

FR with exact line-search terminates at stationary point, $x^{(m)}$ after $m \leq n$ iterations for a pos. definite quadratic Moreover, for $0 \leq i \leq m-1$, we have that:

1. *Conjugate search directions:* $s^{(i)^T} G s^{(j)} = 0 \; \forall i \neq j,\; j < i$.
2. *Orthogonal gradients:* $g^{(i)^T} g^{(j)} = 0 \; \forall i \neq j,\; j = 1, \ldots, i-1$.
3. *Descend property:* $s^{(i)^T} g^{(i)} = -g^{(i)^T} g^{(i)} < 0 \; \forall i \neq j$.

Proof. By induction over m ...
For $m = 0$, there is nothing to show.
For $m \geq 1$, show 1. to 3. of Theorem by induction over i.
For $i = 0$, observe

$$s^{(0)} = -g^{(0)} \;\Rightarrow\; s^{(0)^T} g^{(0)} = -g^{(0)^T} g^{(0)}.$$

$\Rightarrow$ 3. holds for $i = 0$, nothing to show for 1. and 2. (no $j < 0$!)

Proof of Fletcher-Reeves Convergence

Theorem (Convergence of FR for Convex Quadratics)

FR with exact line-search terminates at stationary point, $x^{(m)}$ after $m \leq n$ iterations for a pos. definite quadratic Moreover, for $0 \leq i \leq m-1$, we have that:

1. *Conjugate search directions:* $s^{(i)^T} G s^{(j)} = 0 \; \forall i \neq j, \; j < i$.
2. *Orthogonal gradients:* $g^{(i)^T} g^{(j)} = 0 \; \forall i \neq j, \; j = 1, \ldots, i-1$.
3. *Descend property:* $s^{(i)^T} g^{(i)} = -g^{(i)^T} g^{(i)} < 0 \; \forall i \neq j$.

Proof cont. Induction hypothesis: Assume that 1.-3. hold for i. Show 1.-3. also hold for $i+1$: Quadratic objective implies:

$$g^{(i+1)} = Gx^{(i+1)} + b = G\left(x^{(i)} + \alpha_i s^{(i)}\right) + b = g^{(i)} + \alpha_i G s^{(i)}$$

Exact line search α_i implies:

$$\alpha_i = \frac{-g^{(i)^T} s^{(i)}}{s^{(i)^T} G s^{(i)}} = \frac{g^{(i)^T} g^{(i)}}{s^{(i)^T} G s^{(i)}}, \quad \text{from 3. by induction}$$

Proof of Fletcher-Reeves Convergence

Now, we consider Part 2 for ${g^{(i)}}^T g^{(j)} = 0$:

$$\begin{aligned} {g^{(i+1)}}^T g^{(j)} &= {g^{(i)}}^T g^{(j)} + \alpha_i {s^{(i)}}^T G g^{(j)} \\ &= {g^{(i)}}^T g^{(j)} + \alpha_i {s^{(i)}}^T G \left(-s^{(j)} + \beta_{j-1} s^{(j-1)} \right) \end{aligned}$$

from definition of $s^{(j)} = -g^{(j)} + \beta_{j-1} s^{(j-1)}$, to get $g^{(j)}$. Thus,

$${g^{(i+1)}}^T g^{(j)} = {g^{(i)}}^T g^{(j)} - \alpha_i {s^{(i)}}^T G s^{(j)} + \alpha_i \beta_{j-1} {s^{(i)}}^T G s^{(j-1)}$$

For $i = j$ observe:

- Exact line-search $\Rightarrow \alpha = \frac{-s^T g}{s^T G s} \Rightarrow$ sum of first terms is zero
- Induction Part 1. $\Rightarrow$ last expression zero.

Proof of Fletcher-Reeves Convergence

Now, we consider Part 2 for ${g^{(i+1)}}^T g^{(j)} = 0$:

$$ {g^{(i+1)}}^T g^{(j)} = {g^{(i)}}^T g^{(j)} - \alpha_i {s^{(i)}}^T G s^{(j)} + \alpha_i \beta_{j-1} {s^{(i)}}^T G s^{(j-1)} $$

For $i < j$ observe:

- Induction Part 2. $\Rightarrow$ first expression zero
- Induction Part 1. $\Rightarrow$ last two expressions zero.

Thus, ${g^{(i+1)}}^T g^{(j)} = 0$ for $j = 1, \ldots, i$ which proves Part 2.

Proof of Fletcher-Reeves Convergence

Consider Part 1. Use $s^{(i+1)} = -g^{(i+1)} + \beta_i s^{(i)}$:

$$\begin{aligned} s^{(i+1)^T} G s^{(j)} &= -g^{(i+1)^T} G s^{(j)} + \beta_i s^{(i)^T} G s^{(j)} \\ &= \alpha_j^{-1} g^{(i+1)^T} \left(g^{(j)} - g^{(j+1)} \right) + \beta_i s^{(i)^T} G s^{(j)}, \end{aligned}$$

because $G s^{(j)} = \alpha_j^{-1} G \left(x^{(j)} - x^{(j+1)} \right) = \alpha_j^{-1} G \left(g^{(j)} - g^{(j+1)} \right)$.

For $j < i$ get:

- Part 2. $\Rightarrow$ first component is zero
- Part 1. and induction $\Rightarrow$ second component is zero

Proof of Fletcher-Reeves Convergence

Consider again

$$s^{(i+1)^T} G s^{(j)} = -g^{(i+1)^T} G s^{(j)} + \beta_i s^{(i)^T} G s^{(j)}$$
$$= \alpha_j^{-1} g^{(i+1)^T} \left(g^{(j)} - g^{(j+1)} \right) + \beta_i s^{(i)^T} G s^{(j)},$$

For $j = i$ re-write this expression as

$$s^{(j+1)^T} G s^{(j)} = \alpha_j^{-1} g^{(j+1)^T} g^{(j)} - \alpha_j^{-1} g^{(j+1)^T} g^{(j+1)} + \beta_j s^{(j+1)^T} G s^{(j)}.$$

Part 2. $\Rightarrow$ first component is zero

Use exact line-search α_j second component becomes

$$-\alpha_j^{-1} g^{(j+1)^T} g^{(j+1)} + \beta_j s^{(j+1)^T} G s^{(j)}$$
$$= -s^{(j+1)^T} G s^{(j)} \frac{g^{(j+1)^T} g^{(j+1)}}{g^{(j)^T} g^{(j)}} + \beta_j s^{(j+1)^T} G s^{(j)} = 0,$$

from β_j formula.

$\Rightarrow$ $s^{(i+1)^T} G s^{(j)} = 0$ for all $j = 1, \ldots, i$, which proves Part 1.

Quadratic termination follows from Part 1., and conjugate directions, $s^{(1)}, \ldots, s^{(m)}$. □

Conjugate Gradient Method for Nonlinear Functions

Consider $\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$, then

- Cannot perform exact line-search ... approx, e.g. Armijo
- Cannot expect termination after n steps
  $\Rightarrow$ re-start $s^{(n+1)} = -g^{(n+1)}$ or re-orthogonalize

Other Conjugate Gradient Schemes

$$\beta_k^{PR} = \frac{\left(g^{(k+1)} - g^{(k)}\right)^T g^{(k)}}{g^{(k-1)^T} g^{(k-1)}}$$

and

$$\beta_k^{DY} = \frac{s^{(k)^T} g^{(k)}}{s^{(k-1)^T} g^{(k-1)}}$$

Dai-Yuan better than Polak-Ribiere better than Fletcher-Reeves

Outline

The Barzilai-Borwein Method

Recent renewed interest in a simpler two-step gradient method

- Satisfy quasi-Newton in least-squares sense.

Barzilai-Borwein Method

Given $x^{(0)}$, set $k = 0$.

repeat

Set the step-size α_k using one of BB schemes below.

Set $x^{(k+1)} := x^{(k)} - \alpha_k g^{(k)}$ and $k = k + 1$. [Steepest Descend]

until $x^{(k)}$ *is (local) optimum*;

Surprise: No Line Search

- Barzilai-Borwein Algorithm has no line-search
- Success relies on non-monotone behavior (may increase $f(x)$)

The Barzilai-Borwein Method

Popular formulas for BB step size

$$\alpha_k^{BB} = \frac{\delta^{(k-1)}\delta^{(k-1)}}{\delta^{(k-1)}\gamma^{(k-1)}} \tag{1}$$

$$\alpha_k^{BBs} = \frac{\delta^{(k-1)}\gamma^{(k-1)}}{\gamma^{(k-1)}\gamma^{(k-1)}} \tag{2}$$

$$\alpha_k^{aBB} = \begin{cases} \alpha_k^{BB} & \text{for odd } k \\ \alpha_k^{BBs} & \text{for even } k \end{cases} \tag{3}$$

- Can reset the step length to steepest-descend
- Generalized to bound-constrained optimization using projection

Summary of Conjugate Direction Methods

Methods for unconstrained optimization:

$$\underset{x}{\text{minimize}} \;\; f(x)$$

- Conjugacy is orthogonality across Hessian G, i.e.

$$s^{(i)^T} G s^{(j)} = 0 \quad \forall i \neq j$$

- Conjugate direction methods terminate finitely for quadratic
- Good alternative to quasi-Newton
- Recently, interest in Barzilai-Borwein schemes