The Conjugate Gradient Algorithm

Optimization over a Subspace

Consider the problem

$$\begin{array}{l} \min f(x) \\ \text{subject to } x \in x_0 + S, \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and S is the subspace $S := \mathrm{Span}\{v_1, \ldots, v_k\}$.

Optimization over a Subspace

Consider the problem

$$\begin{array}{l} \min f(x) \\ \text{subject to } x \in x_0 + S, \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and S is the subspace $S := \mathrm{Span}\{v_1, \ldots, v_k\}$.

If $V \in \mathbb{R}^{n\times k}$ has columns $v_1, \ldots, v_k$, then this problem is equivalent

$$\begin{array}{l} \min f(x_0 + Vz) \\ \text{subject to } z \in \mathbb{R}^k \ . \end{array}$$

Subspace Optimality Condition

$$\begin{array}{l} \min f(x_0 + Vz) \\ \text{subject to } z \in \mathbb{R}^k \ . \end{array}$$

Set $\hat{f}(z) = f(x_0 + Vz)$. If $\bar{z}$ solves this problem, then

$$V^T \nabla f(x_0 + V\bar{z}) = \nabla \hat{f}(\bar{z}) = 0.$$

Subspace Optimality Condition

$$\begin{array}{l} \min f(x_0 + Vz) \\ \text{subject to } z \in \mathbb{R}^k \ . \end{array}$$

Set $\hat{f}(z) = f(x_0 + Vz)$. If $\bar{z}$ solves this problem, then

$$V^T \nabla f(x_0 + V\bar{z}) = \nabla \hat{f}(\bar{z}) = 0.$$

Set $\bar{x} = x_0 + V\bar{z}$, we note that $\bar{x}$ solves the original problem if and only if $\bar{z}$ solves the problem above.

Subspace Optimality Condition

$$\begin{array}{l} \min f(x_0 + Vz) \\ \text{subject to } z \in \mathbb{R}^k \ . \end{array}$$

Set $\hat{f}(z) = f(x_0 + Vz)$. If $\bar{z}$ solves this problem, then

$$V^T \nabla f(x_0 + V\bar{z}) = \nabla \hat{f}(\bar{z}) = 0.$$

Set $\bar{x} = x_0 + V\bar{z}$, we note that $\bar{x}$ solves the original problem if and only if $\bar{z}$ solves the problem above.

Then $V^T \nabla f(\bar{x}) = 0$, or equivalently, $v_i^T \nabla f(\bar{x}) = 0$ for $i = 1, 2, \ldots, k$, go $\nabla f(\bar{x}) \in \text{Span}(v_1, \ldots, v_k)^{\perp}$.

Subspace Optimality Theorem

Consider the problem

$$\mathcal{P}_S \qquad \begin{array}{l} \min f(x) \\ \text{subject to } x \in x_0 + S, \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and S is the subspace $S := \mathrm{Span}\{v_1, \ldots, v_k\}$. If $\bar{x}$ solves $\mathcal{P}_S$, then $\nabla f(\bar{x}) \perp S$.

Subspace Optimality Theorem

Consider the problem

$$\mathcal{P}_S \qquad \begin{array}{l} \min f(x) \\ \text{subject to } x \in x_0 + S, \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and S is the subspace $S := \text{Span}\{v_1, \ldots, v_k\}$. If $\bar{x}$ solves $\mathcal{P}_S$, then $\nabla f(\bar{x}) \perp S$.

If it is further assumed that f is conves, then $\bar{x}$ solves $\mathcal{P}_S$ if and only if $\nabla f(\bar{x}) \perp S$.

Conjugate Direction Algorithm

$$\mathcal{P}: \quad \begin{array}{l} \text{minimize } f(x) \\ \text{subject to } x \in \mathbb{R}^n \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is C^2 is given by

$$f(x) := \frac{1}{2}x^T Qx - b^T x$$

with Q is a symmetric positive definite.

Conjugate Direction Algorithm

DEFINITION [Conjugacy]
Let $Q \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. We say that the vectors $x, y \in \mathbb{R}^n \setminus \{0\}$ are Q-conjugate (or Q-orthogonal) if $x^T Q y = 0$.

Conjugate Direction Algorithm

DEFINITION [Conjugacy]
Let $Q \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. We say that the vectors $x, y \in \mathbb{R}^n \backslash \{0\}$ are Q-conjugate (or Q-orthogonal) if $x^T Qy = 0$.

PROPOSITION [Conjugacy implies Linear Independence]
If $Q \in \mathbb{R}^{n\times n}$ is positive definite and the set of nonzero vectors d_0, $d_1, \ldots, d_k$ are (pairwise) Q-conjugate, then these vectors are linearly independent.

Conjugate Direction Algorithm

[Conjugate Direction Algorithm]
Let $\{d_i\}_{i=0}^{n-1}$ be a set of nonzero Q-conjugate vectors. For any $x_0 \in \mathbb{R}^n$ the sequence $\{x_k\}$ generated according to

$$x_{k+1} := x_k + \alpha_k d_k, \quad k \geq 0$$

with

$$\alpha_k := \arg\min\{f(x_k + \alpha d_k) : \alpha \in \mathbb{R}\}$$

converges to the unique solution, x^* of $\mathcal{P}$ after n steps, that is $x_n = x^*$.

Conjugate Direction Algorithm

[CONJUGATE DIRECTION ALGORITHM]
Let $\{d_i\}_{i=0}^{n-1}$ be a set of nonzero Q-conjugate vectors. For any $x_0 \in \mathbb{R}^n$ the sequence $\{x_k\}$ generated according to

$$x_{k+1} := x_k + \alpha_k d_k, \quad k \geq 0$$

with

$$\alpha_k := \arg\min\{f(x_k + \alpha d_k) : \alpha \in \mathbb{R}\}$$

converges to the unique solution, x^* of $\mathcal{P}$ after n steps, that is $x_n = x^*$.

We have already shown that $\alpha_k = -\nabla f(x_k) i^T d_k / d_k Q d_k$.

Expanding Subspace Theorem

Let $\{d_i\}_{i=0}^{n-1}$ be a sequence of nonzero Q-conjugate vectors in $\mathbb{R}^n$. Then for any $x_0 \in \mathbb{R}^n$ the sequence $\{x_k\}$ generated according to

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k d_k \\ \alpha_k &= -\frac{g_k^T d_k}{d_k^T Q d_k} \end{aligned}$$

has the property that $f(x) = \frac{1}{2}x^T Qx - b^T x$ attains its minimum value on the affine set $x_0 +$ Span $\{d_0, \ldots, d_k\}$ at the point x_k.

Expanding Subspace Theorem: Proof

Let $\bar{x}$ solve

$$\min \left\{ f(x) \;|x \in x_0 + \text{ Span } \{d_0, \ldots, d_k\} \right\}.$$

The Subspace Optimality Theorem tells us that
$\nabla f(\bar{x})^T d_i = 0, \;\; i = 0, \ldots, k.$

Expanding Subspace Theorem: Proof

Let $\bar{x}$ solve

$$\min \left\{ f(x) \mid x \in x_0 + \text{ Span } \{d_0, \dots, d_k\} \right\}.$$

The Subspace Optimality Theorem tells us that
$\nabla f(\bar{x})^T d_i = 0, \ i = 0, \dots, k.$
Since $\bar{x} \in x_0 + \text{ Span } \{d_0, \dots, d_k\}$, there exist $\beta_i \in \mathbb{R}$ such that

$$\bar{x} = x_0 + \beta_0 d_0 + \beta_2 d_2 + \cdots + \beta_k d_k.$$

Expanding Subspace Theorem: Proof

Let $\bar{x}$ solve

$$\min \{ f(x) \ | x \in x_0 + \text{ Span } \{d_0, \ldots, d_k\} \} .$$

The Subspace Optimality Theorem tells us that
$\nabla f(\bar{x})^T d_i = 0, \ i = 0, \ldots, k$.
Since $\bar{x} \in x_0 + \text{ Span } \{d_0, \ldots, d_k\}$, there exist $\beta_i \in \mathbb{R}$ such that

$$\bar{x} = x_0 + \beta_0 d_0 + \beta_2 d_2 + \cdots + \beta_k d_k .$$

Therefore,

$$\begin{aligned} 0 &= \nabla f(\bar{x})^T d_i \\ &= (Q(x_0 + \beta_0 d_0 + \beta_2 d_2 + \cdots + \beta_k d_{k-1}) + g)^T d_i \\ &= (Qx_0 + g)^T d_i + \beta_0 d_0^T Q d_i + \beta_2 d_2^T Q d_i + \cdots + \beta_k d_k^T Q d_i \\ &= \nabla f(x_0)^T d_i + \beta_i d_i^T Q d_i \ . \end{aligned}$$

Expanding Subspace Theorem: Proof

$$0 = \nabla f(x_0)^T d_i + \beta_i d_i^T Q d_i, \ \ i = 0, 1, \ldots, k.$$

Hence,

$$\beta_i = -\nabla f(x_0)^T d_i / d_i^T Q d_i, \ \ i = 0, \ldots, k.$$

Expanding Subspace Theorem: Proof

$$0 = \nabla f(x_0)^T d_i + \beta_i d_i^T Q d_i, \;\; i = 0, 1, \dots, k.$$

Hence,

$$\beta_i = -\nabla f(x_0)^T d_i / d_i^T Q d_i, \;\; i = 0, \dots, k.$$

Similarly, the iteration

$$\begin{aligned} x_{i+1} &= x_i + \alpha_i d_i \\ \alpha_i &= -\frac{\nabla f(x_i)^T d_i}{d_i^T Q d_i} \end{aligned}$$

gives

$$x_{k+1} = x_0 + \alpha_0 d_0 + \alpha_2 d_2 + \cdots + \alpha_k d_k.$$

Expanding Subspace Theorem: Proof

$$0 = \nabla f(x_0)^T d_i + \beta_i d_i^T Q d_i, \;\; i = 0, 1, \ldots, k.$$

Hence,

$$\beta_i = -\nabla f(x_0)^T d_i / d_i^T Q d_i, \;\; i = 0, \ldots, k.$$

Similarly, the iteration

$$\begin{aligned} x_{i+1} &= x_i + \alpha_i d_i \\ \alpha_i &= -\frac{\nabla f(x_i)^T d_i}{d_i^T Q d_i} \end{aligned}$$

gives

$$x_{k+1} = x_0 + \alpha_0 d_0 + \alpha_2 d_2 + \cdots + \alpha_k d_k.$$

So we need to show

$$\beta_i = -\frac{\nabla f(x_0)^T d_i}{d_i^T Q d_i} = -\frac{\nabla f(x_i)^T d_i}{d_i^T Q d_i} = \beta_i, \;\; i = 0, \ldots, k \;.$$

Expanding Subspace Theorem: Proof

$$
\begin{aligned}
\nabla f(x_i)^T d_i &= (Q(x_0 + \alpha_0 d_0 + \cdots + \alpha_{i-1} d_{i-1}) + g)^T d_i \\
&= (Qx_0 + g)^T d_i + \alpha_0 d_0^T Q d_i + \cdots + \alpha_{i-1} d_{i-1}^T Q d_i \\
&= \nabla f(x_0)^T d_i
\end{aligned}
$$

The Conjugate Gradient Algorithm

Initialization: $x_0 \in \mathbb{R}^n$, $d_0 = -g_0 = -\nabla f(x_0) = b - Qx_0$.

For $k = 0, 1, 2, \ldots$

$$
\begin{array}{lll}
\alpha_k & := -g_k^T d_k / d_k^T Q d_k & \\
x_{k+1} & := x_k + \alpha_k d_k & \\
g_{k+1} & := Q x_{k+1} - b & \text{(STOP if } g_{k+1} = 0) \\
\beta_k & := g_{k+1}^T Q d_k / d_k^T Q d_k & \\
d_{k+1} & := -g_{k+1} + \beta_k d_k & \\
k & := k + 1. &
\end{array}
$$

The Conjugate Gradient Theorem

The C-G algorithm is a conjugate direction method. If it does not terminate at x_k ($g_k \neq 0$), then

1. Span $[g_0, g_1, \ldots, g_k] =$ span $[g_0, Qg_0, \ldots, Q^k g_0]$
2. Span $[d_0, d_1, \ldots, d_k] =$ span $[g_0, Qg_0, \ldots, Q^k g_0]$
3. $d_k^T Q d_i = 0$ for $i \leq k - 1$
4. $\alpha_k = g_k^T g_k / d_k^T Q d_k$
5. $\beta_k = g_{k+1}^T g_{k+1} / g_k^T g_k$.

The Conjugate Gradient Theorem: Proof

Prove (1)-(3) by induction. (1)-(3) true for $k = 0$. Suppose true up to k and show true for $k + 1$.

(1) Span $[g_0, g_1, \ldots, g_k] =$ Span $[g_0, Qg_0, \ldots, Q^k g_0]$

The Conjugate Gradient Theorem: Proof

Prove (1)-(3) by induction. (1)-(3) true for $k = 0$. Suppose true up to k and show true for $k + 1$.

(1) Span $[g_0, g_1, \ldots, g_k] =$ Span $[g_0, Qg_0, \ldots, Q^k g_0]$

Since

$$g_{k+1} = g_k + \alpha_k Q d_k,$$

$g_{k+1} \in$ Span$[g_0, \ldots, Q^{k+1} g_0]$ (ind. hyp.).

The Conjugate Gradient Theorem: Proof

Prove (1)-(3) by induction. (1)-(3) true for $k = 0$. Suppose true up to k and show true for $k + 1$.

(1) Span $[g_0, g_1, \ldots, g_k] =$ Span $[g_0, Qg_0, \ldots, Q^k g_0]$

Since

$$g_{k+1} = g_k + \alpha_k Q d_k,$$

$g_{k+1} \in$ Span$[g_0, \ldots, Q^{k+1} g_0]$ (ind. hyp.).
Also $g_{k+1} \notin$ Span $[d_0, \ldots, d_k]$ otherwise $g_{k+1} = 0$ (by The Subspace Optimality Theorem) since the method is a conjugate direction method up to step k (ind. hyp.). So
$g_{k+1} \notin$ Span $[g_0, \ldots, Q^k g_0]$ and
Span $[g_0, g_1, \ldots, g_{k+1}] =$ Span $[g_0, \ldots, Q^{k+1} g_0]$ proving (1).

The Conjugate Gradient Theorem: Proof

(2) Span $[d_0, d_1, \ldots, d_k] =$ Span $[g_0, Qg_0, \ldots, Q^k g_0]$

The Conjugate Gradient Theorem: Proof

(2) Span $[d_0, d_1, \ldots, d_k] =$ Span $[g_0, Qg_0, \ldots, Q^k g_0]$

To prove (2) write

$$d_{k+1} = -g_{k+1} + \beta_k d_k$$

so that (2) follows from (1) and the induction hypothesis on (2).

The Conjugate Gradient Theorem: Proof

(3) $d_k^T Q d_i = 0$ for $i \leq k-1$

The Conjugate Gradient Theorem: Proof

(3) $d_k^T Q d_i = 0$ for $i \leq k-1$

To see (3) observe that

$$d_{k+1}^T Q d_i = -g_{k+1} Q d_i + \beta_k d_k^T Q d_i.$$

For $i = k$ the right hand side is zero by the definition of β_k.

The Conjugate Gradient Theorem: Proof

(3) $d_k^T Q d_i = 0$ for $i \le k-1$

To see (3) observe that

$$d_{k+1}^T Q d_i = -g_{k+1} Q d_i + \beta_k d_k^T Q d_i .$$

For $i = k$ the right hand side is zero by the definition of β_k.

For $i < k$ both terms vanish.
The term $g_{k+1}^T Q d_i = 0$ by the Expanding Subspace Theorem since $Q d_i \in$ Span$[d_0, \ldots, d_k]$ by (1) and (2).
The term $d_k^T Q d_i$ vanishes by the induction hypothesis on (3).

The Conjugate Gradient Theorem: Proof

(4) $\alpha_k = g_k^T g_k / d_k^T Q d_k$

The Conjugate Gradient Theorem: Proof

(4) $\alpha_k = g_k^T g_k / d_k^T Q d_k$

$$-g_k^T d_k = g_k^T g_k - \beta_{k-1} g_k^T d_{k-1}$$

where $g_k^T d_{k-1} = 0$ by the Expanding Subspace Theorem.
So

$$\alpha_k = -g_k^T d_k / d_k^T Q d_k = g_k^T g_k / d_k^T Q d_k \ .$$

The Conjugate Gradient Theorem: Proof

(5) $\beta_k = g_{k+1}^T g_{k+1} / g_k^T g_k$

The Conjugate Gradient Theorem: Proof

(5) $\beta_k = g_{k+1}^T g_{k+1} / g_k^T g_k$

$g_{k+1}^T g_k = 0$ by the Expanding Subspace Theorem because $g_k \in \text{Span}[d_0, \ldots, d_k]$.

The Conjugate Gradient Theorem: Proof

(5) $\beta_k = g_{k+1}^T g_{k+1} / g_k^T g_k$

$g_{k+1}^T g_k = 0$ by the Expanding Subspace Theorem because $g_k \in$ Span$[d_0, \ldots, d_k]$.

Hence

$$g_{k+1}^T Q d_k = g_{k+1}^T Q\left(\frac{x_{k+1} - x_k}{\alpha_k}\right) = \frac{1}{\alpha_k} g_{k+1}^T [g_{k+1} - g_k] = \frac{1}{\alpha_k} g_{k+1}^T g_{k+1}.$$

The Conjugate Gradient Theorem: Proof

(5) $\beta_k = g_{k+1}^T g_{k+1} / g_k^T g_k$

$g_{k+1}^T g_k = 0$ by the Expanding Subspace Theorem because $g_k \in$ Span$[d_0, \ldots, d_k]$.

Hence

$$g_{k+1}^T Q d_k = g_{k+1}^T Q\left(\frac{x_{k+1} - x_k}{\alpha_k}\right) = \frac{1}{\alpha_k} g_{k+1}^T [g_{k+1} - g_k] = \frac{1}{\alpha_k} g_{k+1}^T g_{k+1}.$$

Therefore,

$$\beta_k = \frac{g_{k+1}^T Q d_k}{d_k^T Q d_k} = \frac{1}{\alpha_k} \frac{g_{k+1}^T g_{k+1}}{d_k^T Q d_k} = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}.$$

Comments on the CG Algorithm

The C–G method decribed above is a descent method since the values

$$f(x_0),\ f(x_1),\ \ldots, f(x_n)$$

form a decreasing sequence. Moreover, note that

$$\nabla f(x_k)^T d_k = -g_k^T g_k \quad \text{and} \quad \alpha_k > 0\ .$$

Thus, the C–G method behaves very much like the descent methods discussed peviously.

Comments on the CG Algorithm

Due to the occurrence of round-off error the C-G algorithm is best implemented as an iterative method. That is, at the end of n steps, f may not attain its global minimum at x_n and the intervening directions d_k may not be Q-conjugate. But it is also possible for the CG algorithm to terminate early.

Comments on the CG Algorithm

Due to the occurrence of round-off error the C-G algorithm is best implemented as an iterative method. That is, at the end of n steps, f may not attain its global minimum at x_n and the intervening directions d_k may not be Q-conjugate. But it is also possible for the CG algorithm to terminate early.

Consequently, at each step one should check the value $\|\nabla f(x_{k+1})\|$ and the size of the step $\|x_{k+1} - x_k\|$. If either is sufficiently small, then accept x_k as the point at which f attains its global minimum value; otherwise, continue to iterate regardless of the iteration count (up to a maximum acceptable number of iterations). Since CG is a descent method, continued progress is assured.

The Non-Quadratic CG Algorithm

Initialization: $x_0 \in \mathbb{R}^n$, $g_0 = \nabla f(x_0)$, $d_0 = -g_0$, $0 < c < \beta < 1$.
Having x_k otain x_{k+1} as follows:
Check restart criteria. If a restart condition is satisfied, then reset $x_0 = x_n$, $g_0 = \nabla f(x_0)$, $d_0 = -g_0$; otherwise, set

$$
\begin{array}{lll}
\alpha_k & \in & \left\{ \lambda \;\middle|\; \begin{array}{c} \lambda > 0, \nabla f(x_k + \lambda d_k)^T d \geq \beta \nabla f(x_k)^T d_k, \text{ and} \\ f(x_k + \lambda d_k) - f(x_k) \leq c \lambda \nabla f(x_k)^T d_k \end{array} \right\} \\
x_{k+1} & := & x_k + \alpha_k d_k \\
g_{k+1} & := & \nabla f(x_{k+1}) \\
\beta_k & := & \begin{cases} \dfrac{g_{k+1}^T g_{k+1}}{g_k^T g_k} & \text{Fletcher-Reeves} \\ \max\left\{0, \dfrac{g_{k+1}^T (g_{k+1} - g_k)}{g_k^T g_k}\right\} & \text{Polak-Ribiere} \end{cases} \\
d_{k+1} & := & -g_{k+1} + \beta_k d_k \\
k & := & k + 1.
\end{array}
$$

The Non-Quadratic CG Algorithm

Restart Conditions

1. $k = n$
2. $|g_{k+1}^T g_k| \geq 0.2 g_k^T g_k$
3. $-2 g_k^T g_k \geq g_k^T d_k \geq -0.2 g_k^T g_k$

The Non-Quadratic CG Algorithm

The Polak-Ribiere update for β_k has a demonstrated experimental superiority. One way to see why this might be true is to observe that

$$g_{k+1}^T(g_{k+1} - g_k) \approx \alpha_k g_{k+1}^T \nabla^2 f(x_k) d_k$$

thereby yielding a better second–order approximation. Indeed, the formula for β_k in in the quadratic case is precisely

$$\frac{\alpha_k g_{k+1}^T \nabla^2 f(x_k) d_k}{g_k^T g_k} .$$