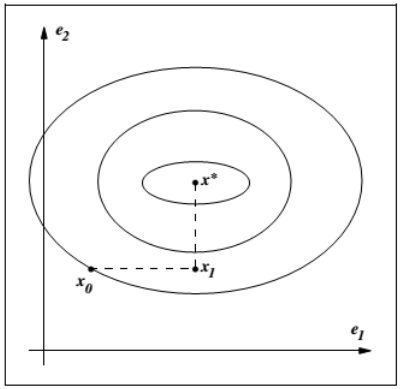
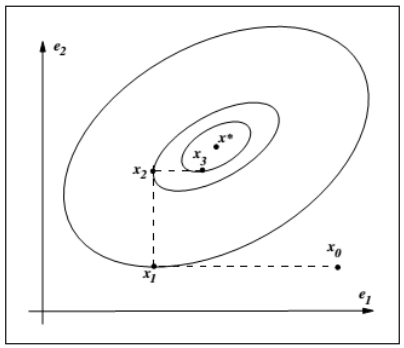
Chapter 10 Conjugate Direction Methods

1. Conjugate Direction Method
2. Conjugate Gradient Method
3. Conjugate Gradient Method for Nonquadratic Problems

Conjugate Direction



when successively minimizing along the coordinate directions, we can reach the solution of a "simple" quadratic function within n iterations.



when successively minimizing along the coordinate directions, we can not reach the solution of a "general" quadratic function within n iterations.

★ "general" quadratic function can be modified by alerting the searching directions.

Conjugate Direction

unconstrained quadratic optimization

$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$, where $\boldsymbol{Q} \succ 0$, $\boldsymbol{b} \in \mathbb{R}^n$.

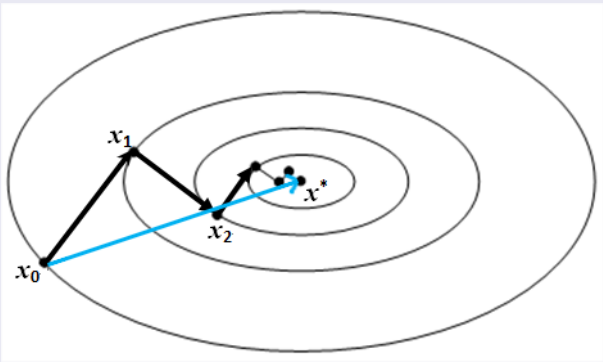
revisit of steepest descent and Newton's methods

- steepest descent method:

$$\begin{cases} \alpha_k = -\frac{\|\boldsymbol{g}^{(k)}\|^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2} \text{ is exact stepsize.} \\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}, \end{cases}$$

- Newton's method:

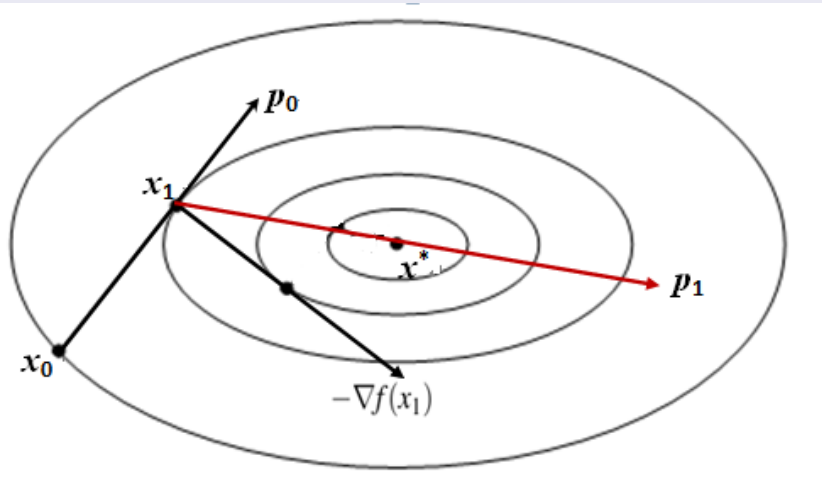
$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)}$$



- steepest descent method: less computation and CPU memory at each iteration; but a large number of iterations to reach the minimizer $\boldsymbol{x}^*$.
- Newton's method: large computation and CPU memory at each iteration; but only fewer iteration can well approximate the minimizer $\boldsymbol{x}^*$.

Conjugate Direction

motivation of conjugate direction



$$x^* = x^{(1)} + tp^{(1)}, \quad t \in \mathbb{R}$$

$$Qx^* - b = Qx^{(1)} - b + tQp^{(1)}$$

$$\nabla f(x^*) = \nabla f(x^{(1)}) + tQp_1^{(1)}$$

$$0 = \nabla f(x_1) + tQp_1^{(1)}$$

$$0 = p^{(0)\top}\nabla f(x_1) + tp^{(0)\top}Qp^{(1)}$$

$$0 = p^{(0)\top}Qp^{(1)}$$

Conjugate Direction

Notation:

$\mathbb{S}^n$: set of $n \times n$ symmetric matrices.

$\mathbb{S}^n_+$: set of $n \times n$ symmetric positive semidefinite matrices.

$\mathbb{S}^n_{++}$: set of $n \times n$ symmetric positive definite matrices.

Definition (conjugate direction)

Let $\boldsymbol{Q} \in \mathbb{S}^n$. The vectors $\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(m)}$ in $\mathbb{R}^n$ are $\boldsymbol{Q}$-conjugate if $\boldsymbol{p}^{(i)\top} \boldsymbol{Q} \boldsymbol{p}^{(j)} = 0$ for all $i \neq j$.

Lemma

Let $\boldsymbol{Q} \in \mathbb{S}^n_{++}$. If the vectors $\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(k)}$ are nonzero and $\boldsymbol{Q}$-conjugate, then they are linearly independent.

proof: analogously to the proof of orthogonal vectors.

Question: given a $\boldsymbol{Q} \in \mathbb{S}^n_{++}$, how to generate $\boldsymbol{Q}$-conjugate vectors?

Conjugate Direction: Three Methods

Method I

Given $\boldsymbol{Q} = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}$,

let the $\boldsymbol{Q}$-conjugate vectors be $\begin{cases} \boldsymbol{p}^{(0)} = [1,0,0]^\top \\ \boldsymbol{p}^{(1)} = [p_1^{(1)}, p_2^{(1)}, p_3^{(1)}]^\top \\ \boldsymbol{p}^{(2)} = [p_1^{(2)}, p_2^{(2)}, p_3^{(2)}]^\top \end{cases}$

$$\xrightarrow{\text{definition}} \begin{cases} \boldsymbol{p}^{(0)\top}\boldsymbol{Q}\boldsymbol{p}^{(1)} = 0 \\ \boldsymbol{p}^{(0)\top}\boldsymbol{Q}\boldsymbol{p}^{(2)} = 0 \\ \boldsymbol{p}^{(1)\top}\boldsymbol{Q}\boldsymbol{p}^{(2)} = 0 \end{cases} \Longrightarrow \cdots \Longrightarrow \begin{cases} \boldsymbol{p}^{(0)} = [1,0,0]^\top \\ \boldsymbol{p}^{(1)} = [1,0,-3]^\top \\ \boldsymbol{p}^{(2)} = [1,4,-3]^\top \end{cases}$$

Conjugate Direction: Three Methods

Method II: Gram-Schmidt process

Given $\boldsymbol{Q} \in \mathbb{S}^n_{++}$ and linear independent vectors $\{\boldsymbol{a}^{(i)}\}_{i=0}^{n-1}$ in $\mathbb{R}^n$, Gram-Schmidt process can transform $\{\boldsymbol{a}^{(i)}\}_{i=0}^{n-1}$ into $\boldsymbol{Q}$-conjugate vectors

Gram-Schmidt process:
$$\begin{cases} \boldsymbol{p}^{(0)} = \boldsymbol{a}^{(0)}. \\ \boldsymbol{p}^{(k)} = \boldsymbol{a}^{(k)} - \sum_{i=0}^{k-1} \frac{\boldsymbol{a}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(i)}}{\boldsymbol{p}^{(i)\top}\boldsymbol{Q}\boldsymbol{p}^{(i)}}\boldsymbol{p}^{(i)}, \quad k = 1, \ldots, n-1. \end{cases}$$

The vectors $\boldsymbol{p}^{(0)}, \cdots, \boldsymbol{p}^{(n-1)}$ are $\boldsymbol{Q}$-conjugate.

★ Particularly, if $\boldsymbol{Q} = \boldsymbol{I}$, $\boldsymbol{Q}$-conjugate reduces to the orthogonal. Thus, the above process reduces to the Gram-Schmidt orthogonalization.

Conjugate Direction: Three Methods

Method III: eigenvalue decomposition (see exercise 10.4)

Let $\boldsymbol{Q} \in \mathbb{S}^n$. Then, there exists $\boldsymbol{Q}$-conjugate vectors $\{\boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(n)}\}$ such that each $\boldsymbol{p}^{(i)}$ $(i = 1, \ldots, n)$ is an eigenvector of $\boldsymbol{Q}$.

proof. Use the fact that for any $\boldsymbol{Q} \in \mathbb{S}^n$, there exists vectors $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\}$ of its eigenvectors such that $\boldsymbol{v}_i^\top \boldsymbol{v}_j = 0$ for all $i, j = 1, \ldots, n, i \neq j$.

Theorem

Let $\boldsymbol{Q} \in \mathbb{S}^n_{++}$. If nonzeros vectors $\{\boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(n)}\}$ are $\boldsymbol{Q}$-conjugate and also orthogonal, then each $\boldsymbol{p}^{(i)}, i = 1, \ldots, n$, is an eigenvector of $\boldsymbol{Q}$.

★ orthogonal eigenvectors of a positive definite matrix are the $\boldsymbol{Q}$-conjugate. Conversely, if $\{\boldsymbol{p}^{(1)}, ..., \boldsymbol{p}^{(n)}\}$ are orthogonal and $\boldsymbol{Q}$-conjugate, then they are eigenvectors of $\boldsymbol{Q}$.

Basic Conjugate Direction Algorithm

basic conjugate direction algorithm

To minimizing a quadratic function $\min\ f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$, where $\boldsymbol{Q} \in \mathbb{S}^n_{++}$ and $\boldsymbol{b} \in \mathbb{R}^n$, given the initial point $\boldsymbol{x}^{(0)}$ and $\boldsymbol{Q}$-conjugate directions $\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(n-1)}$, the iteration scheme reads

1. $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}$,
2. $\alpha_k = -\frac{\boldsymbol{g}^{(k)\top}\boldsymbol{p}^{(k)}}{\boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}}$,
3. $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}$.

Basic Conjugate Direction Algorithm

Example (minimize f by basic conjugate direction algorithms)

$\min\ f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix} \boldsymbol{x} - \boldsymbol{x}^\top \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. Given the $\boldsymbol{Q}$-conjugate directions $\boldsymbol{p}^{(0)} = [1, 0]^\top, \boldsymbol{p}^{(1)} = \left[-\frac{3}{8}, \frac{3}{4}\right]^\top$ and the initial point $\boldsymbol{x}^{(0)} = [0, 0]^\top$.

Ans: $\because\ \nabla f(\boldsymbol{x}) = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix} \boldsymbol{x} - \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, $\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix} =: \boldsymbol{Q}$

S1. $\boldsymbol{g}^{(0)} = -\boldsymbol{b} = [1, -1]^\top$, $\alpha_0 = -\frac{\boldsymbol{g}^{(0)\top}\boldsymbol{p}^{(0)}}{\boldsymbol{p}^{(0)\top}\boldsymbol{Q}\boldsymbol{p}^{(0)}} = -\frac{1}{4}$,
$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha_0 \boldsymbol{p}^{(0)} = [-\frac{1}{4}, 0]^\top$.

S2. $\boldsymbol{g}^{(1)} = \boldsymbol{Q}\boldsymbol{x}^{(1)} - \boldsymbol{b} = [0, -\frac{3}{2}]^\top$, $\alpha_1 = -\frac{\boldsymbol{g}^{(1)\top}\boldsymbol{p}^{(1)}}{\boldsymbol{p}^{(1)\top}\boldsymbol{Q}\boldsymbol{p}^{(1)}} = 2$,
$\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} + \alpha_1 \boldsymbol{p}^{(1)} = [-1, \frac{3}{2}]^\top$.

$\because$ f is a 2D quadratic function,

$\therefore$ $\boldsymbol{x}^* = \boldsymbol{x}^{(2)}$ is a minimizer.

Basic Conjugate Direction Algorithm

Theorem

For any initial point $\boldsymbol{x}^{(0)}$, the basic conjugate direction algorithm converges to the unique $\boldsymbol{x}^*$ (indeed, $\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{b}$) in n steps, i.e., $\boldsymbol{x}^{(n)} = \boldsymbol{x}^*$.

proof. $\because$ The vectors $\{\boldsymbol{p}^{(i)}\}_{i=0}^{n-1}$ are $\boldsymbol{Q}$-conjugate,
$\therefore$ $\{\boldsymbol{p}^{(i)}\}_{i=0}^{n-1}$ are linear independent, $\therefore$ $\{\boldsymbol{p}^{(i)}\}_{i=0}^{n-1}$ is a basis of $\mathbb{R}^n$.
$\therefore$ The vector $\boldsymbol{x}^* - \boldsymbol{x}^{(0)} \in \mathbb{R}^n$ can be linearly expressed as

$$\boldsymbol{x}^* - \boldsymbol{x}^{(0)} = \beta_0 \boldsymbol{p}^{(0)} + \cdots + \beta_{n-1}\boldsymbol{p}^{(n-1)}. \tag{1}$$

Premultiply ${\boldsymbol{p}^{(k)}}^{\top}\boldsymbol{Q}$ to both sides of the above equation

$$\boldsymbol{p}^{(k)\top}\boldsymbol{Q}(\boldsymbol{x}^* - \boldsymbol{x}^{(0)}) = \beta_k \boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)} \implies \beta_k = \frac{\boldsymbol{p}^{(k)\top}\boldsymbol{Q}(\boldsymbol{x}^* - \boldsymbol{x}^{(0)})}{\boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}}.$$

On the other hand, the iterate $\boldsymbol{x}^{(k)}$ can be written as

$$\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(0)} = \alpha_0 \boldsymbol{p}^{(0)} + \cdots + \alpha_{k-1}\boldsymbol{p}^{(k-1)}. \tag{2}$$

$$\therefore \boldsymbol{x}^{(n)} - \boldsymbol{x}^{(0)} = \alpha_0 \boldsymbol{p}^{(0)} + \cdots + \alpha_{n-1}\boldsymbol{p}^{(n-1)} \text{ with } \alpha_k = -\frac{\boldsymbol{g}^{(k)\top}\boldsymbol{p}^{(k)}}{\boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}}. \tag{3}$$

By comparing (1) and (3), $\boldsymbol{x}^{(n)} = \boldsymbol{x}^{(0)} \iff \boxed{\beta_k = \alpha_k}$ for $k = 0, \ldots, n-1$.

Basic Conjugate Direction Algorithm

Indeed, the numerators of β_k and α_k are equal.

$$
\begin{aligned}
&\because \quad \boldsymbol{p}^{(k)\top}\boldsymbol{Q}(\boldsymbol{x}^* - \boldsymbol{x}^{(0)}) \\
&= \boldsymbol{p}^{(k)\top}\boldsymbol{Q}(\boldsymbol{x}^* - \boldsymbol{x}^{(k)} + \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(0)}) \\
&= \boldsymbol{p}^{(k)\top}\boldsymbol{Q}(\boldsymbol{x}^* - \boldsymbol{x}^{(k)}) + \underbrace{\boldsymbol{p}^{(k)\top}\boldsymbol{Q}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(0)})}_{=0,\ \text{(why?)}} \\
&= \boldsymbol{p}^{(k)\top}(\boldsymbol{b} - \boldsymbol{Q}\boldsymbol{x}^{(k)}) = -\boldsymbol{p}^{(k)\top}\boldsymbol{g}^{(k)}.
\end{aligned}
$$

$\therefore\ \beta_k = -\frac{\boldsymbol{p}^{(k)\top}\boldsymbol{g}^{(k)}}{\boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}} = \alpha_k$, which implies that $\boldsymbol{x}^* = \boldsymbol{x}^{(n)}$.

Lemma (accurate stepsize: $\alpha_k = \arg\min f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{p}^{(k)})$)

In exact line search, the searching direction orthogonal to gradient, i.e., $\boldsymbol{g}^{(k+1)\top}\boldsymbol{p}^{(k)} = 0$. (see Chapter 8 for proof)

Basic Conjugate Direction Algorithm

Lemma

In conjugate direction method, given a k satisfying $0 \le k \le n-1$, then $\boldsymbol{g}^{(k+1)\top}\boldsymbol{p}^{(i)} = 0$ for all $0 \le i \le k$.

proof. $\because \quad \boldsymbol{g}^{(k)} = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}$.

$\therefore \quad \boldsymbol{g}^{(k+1)} - \boldsymbol{g}^{(k)} = \boldsymbol{Q}(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) \overset{\text{recursion}}{=\!=\!=} \alpha_k \boldsymbol{Q}\boldsymbol{p}^{(k)}$.

(By induction) It follows from the above lemma that $\boldsymbol{g}^{(1)\top}\boldsymbol{p}^{(0)} = 0$.

(Induction hypothesis) If the assertion holds for $k-1$, i.e., $\boldsymbol{g}^{(k)\top}\boldsymbol{p}^{(i)} = 0$ for all $0 \le i \le k-1$.

We now prove that assertion holds for k, i.e., $\boldsymbol{g}^{(k+1)\top}\boldsymbol{p}^{(i)} = 0$ for all $0 \le i \le k$.

- It follows from the above lemma on accurate step, $\boldsymbol{g}^{(k+1)\top}\boldsymbol{p}^{(k)} = 0$.
- Furthermore, it suffice to prove $\boldsymbol{g}^{(k+1)\top}\boldsymbol{p}^{(i)} = 0$ for all $0 \le i \le k-1$.

  $\because \quad \boldsymbol{g}^{(k+1)} = \boldsymbol{g}^{(k)} + \alpha_k \boldsymbol{Q}\boldsymbol{p}^{(k)}$.

  $\therefore \quad \boldsymbol{g}^{(k+1)\top}\boldsymbol{p}^{(i)} = \boldsymbol{g}^{(k)\top}\boldsymbol{p}^{(i)} + \alpha_k \boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(i)} = 0$ for all $0 \le i \le k-1$ (why?)

Thus, given a $0 \le k \le n-1$ $\boldsymbol{g}^{(k+1)\top}\boldsymbol{p}^{(i)} = 0$ for all $0 \le i \le k$.

★ $\boldsymbol{g}^{(k+1)} \perp \text{span}[\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(k)}]$.

Expending Subspace Theorem

Theorem

The conjugate direction method satisfies $f(\boldsymbol{x}^{(k+1)}) = \min\limits_{\alpha} f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{p}^{(k)})$*, and it can also satisfies* $f(\boldsymbol{x}^{(k+1)}) = \min\limits_{a_0,\cdots,a_k} f(\boldsymbol{x}^{(0)} + \sum_{i=0}^{k} a_i \boldsymbol{p}^{(i)})$*. In another word, let* $\mathcal{V}_k = \boldsymbol{x}^{(0)} + \mathrm{span}[\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(k)}]$*. Then* $f(\boldsymbol{x}^{(k+1)}) = \min\limits_{\boldsymbol{x} \in \mathcal{V}_k} f(\boldsymbol{x})$.

proof. let matrix $\boldsymbol{D}^{(k)} = [\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(k)}]$.

$\therefore \quad \mathcal{V}_k = \boldsymbol{x}^{(0)} + \mathrm{span}[\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(k)}] = \boldsymbol{x}^{(0)} + \mathcal{R}(\boldsymbol{D}^{(k)})$.

$\because \quad \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(0)} + \sum\limits_{i=0}^{k} \alpha_i \boldsymbol{p}^{(i)} = \boldsymbol{x}^{(0)} + \boldsymbol{D}^{(k)} \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = [\alpha_0, \cdots, \alpha_k]^\top$.

$\therefore \quad \boldsymbol{x}^{(k+1)} \in \boldsymbol{x}^{(0)} + \mathcal{R}(\boldsymbol{D}^{(k)}) = \mathcal{V}_k$.

For any $\boldsymbol{x} \in \mathcal{V}_k$, there exists an $\boldsymbol{\alpha} \in \mathbb{R}^{k+1}$ such that $\boldsymbol{x} = \boldsymbol{x}^{(0)} + \boldsymbol{D}^{(k)} \boldsymbol{a}$.

Expending Subspace Theorem

Let $\phi_k(\boldsymbol{\alpha}) := f(\boldsymbol{x}^{(0)} + \boldsymbol{D}^{(k)}\boldsymbol{\alpha})$.

Then, ϕ_k is a quadratic function with unique minimum.

By the chain rules,

$$D\phi_k(\boldsymbol{\alpha}) = \nabla f(\boldsymbol{x}^{(0)} + \boldsymbol{D}^{(k)}\boldsymbol{\alpha})^\top \boldsymbol{D}^{(k)} = \nabla f(\boldsymbol{x}^{(k+1)})^\top \boldsymbol{D}^{(k)} = \boldsymbol{g}^{(k+1)\top}\boldsymbol{D}^{(k)}. \quad (4)$$

It follows from the above Lemma 10.2 that $\boldsymbol{g}^{(k+1)\top}\boldsymbol{D}^{(k)} = \boldsymbol{0}^\top$.

$\therefore$ (4) $\implies$ $D\phi_k(\boldsymbol{\alpha}) = 0 \implies$ $\boldsymbol{\alpha}$ is the minimum of ϕ_k, i.e.,

$$f(\boldsymbol{x}^{(k+1)}) = \min_{\boldsymbol{\alpha}} f(\boldsymbol{x}^{(0)} + \boldsymbol{D}^{(k)}\boldsymbol{\alpha}) = \min_{\boldsymbol{x}\in\mathcal{V}_k} f(\boldsymbol{x}).$$

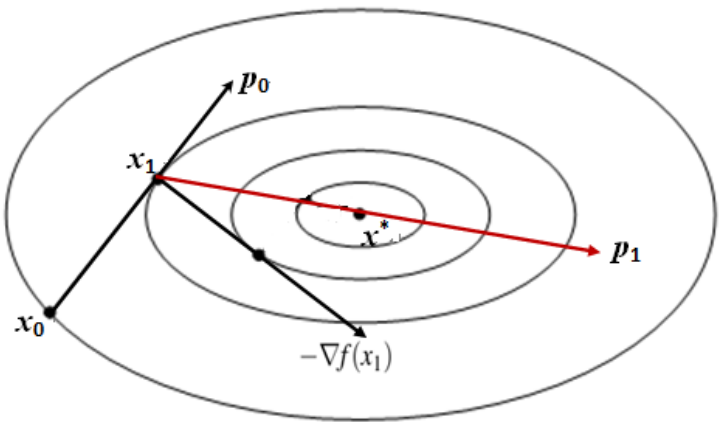
★ As k increases, the subspace span$[\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(k)}]$ continues to expand until it fills the entire $\mathbb{R}^n$ (provided that the vectors $\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots$ are linear independent).

Conjugate Gradient Method

motivation of conjugate gradient (CG) method

- CG method does not use prespecified conjugate directions, but computes the directions as method progresses. At kth iteration, $\boldsymbol{p}^{(k)}$ is calculated as a linear combination of $\boldsymbol{p}^{(k-1)}$ and the current gradient $\boldsymbol{g}^{(k)}$, in such a way that all the directions are $\boldsymbol{Q}$-conjugate.
- CG method exploits the fact that for a quadratic function of n variables, we can locate the minimizer by performing n searches along conjugate directions.

Analysis: given initial point $\boldsymbol{x}^{(0)}$ and exact stepsize criterion.



- step 0: use gradient $\boldsymbol{p}^{(0)} = -\boldsymbol{g}^{(0)}$ as the search direction to produce the new iterate

$$\alpha_0 = -\frac{\boldsymbol{g}^{(0)\top}\boldsymbol{p}^{(0)}}{\boldsymbol{p}^{(0)\top}\boldsymbol{Q}\boldsymbol{p}^{(0)}}$$

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha_0\boldsymbol{p}^{(0)}.$$

Conjugate Gradient Method

Analysis: given initial point $\boldsymbol{x}^{(0)}$ and exact stepsize criterion.

- step 1: let $\boldsymbol{p}^{(1)}$ be the linear combination of $\boldsymbol{g}^{(1)}$ and $\boldsymbol{p}^{(0)}$, i.e.,

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

  $\xrightarrow{\boldsymbol{Q}\text{-conjugacy of } \boldsymbol{p}^{(1)}, \boldsymbol{p}^{(0)}} \beta_0 = \frac{\boldsymbol{g}^{(1)\top} \boldsymbol{Q} \boldsymbol{d}^{(0)}}{\boldsymbol{p}^{(0)\top} \boldsymbol{Q} \boldsymbol{d}^{(0)}}.$

  use this $\boldsymbol{p}^{(1)}$ as the search direction to generate the new iterate

  $\alpha_1 = -\frac{\boldsymbol{g}^{(1)\top} \boldsymbol{p}^{(1)}}{\boldsymbol{p}^{(1)\top} \boldsymbol{Q} \boldsymbol{p}^{(1)}}$

  $\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} + \alpha_1 \boldsymbol{p}^{(1)}.$

- step k: compute the search direction

$$\boldsymbol{p}^{(k)} = -\boldsymbol{g}^{(k)} + \beta_{k-1} \boldsymbol{p}^{(\boldsymbol{k-1})}, \quad \text{where} \quad \beta_{k-1} = \frac{\boldsymbol{g}^{(k)\top} \boldsymbol{Q} \boldsymbol{d}^{(k-1)}}{\boldsymbol{p}^{(k-1)\top} \boldsymbol{Q} \boldsymbol{d}^{(k-1)}}. \tag{5}$$

Conjugate Gradient Method

Proposition

The vectors $\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(n-1)}$ *produced by* (5) *are* $\boldsymbol{Q}$*-conjugate.*

proof. (induction)

(i) First, prove $\boldsymbol{p}^{(0)\top}\boldsymbol{Q}\boldsymbol{p}^{(1)} = 0$.

$$\because \quad \boldsymbol{p}^{(0)\top}\boldsymbol{Q}\boldsymbol{p}^{(1)} = \boldsymbol{p}^{(0)\top}\boldsymbol{Q}(-\boldsymbol{g}^{(1)} + \beta_0 \boldsymbol{p}^{(0)}) \xrightarrow{\beta_0 = \frac{\boldsymbol{g}^{(1)\top}\boldsymbol{Q}\boldsymbol{p}^{(0)}}{\boldsymbol{p}^{(0)\top}\boldsymbol{Q}\boldsymbol{p}^{(0)}}} 0.$$

(ii) Inductive hypothesis: $\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(k)}$ are $\boldsymbol{Q}$-conjugate.

$$\xRightarrow{\text{By Lemma 10.2}} \boldsymbol{g}^{(k+1)} \perp \text{span}[\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(k)}].$$

We first prove $\boldsymbol{g}^{(k+1)\top}\boldsymbol{g}^{(j)} = 0$ for all $0 \le j \le k$.

For a given $0 \le j \le k$, we have $\boldsymbol{p}^{(j)} = -\boldsymbol{g}^{(j)} + \beta_{j-1}\boldsymbol{p}^{(j-1)}$.

$$\therefore \quad 0 = \boldsymbol{g}^{(k+1)\top}\boldsymbol{p}^{(j)} = -\boldsymbol{g}^{(k+1)\top}\boldsymbol{g}^{(j)} + \beta_{j-1}\underbrace{\boldsymbol{g}^{(k+1)\top}\boldsymbol{p}^{(j-1)}}_{0 \text{ (why?)}}$$

$$\therefore \quad \boldsymbol{g}^{(k+1)\top}\boldsymbol{g}^{(j)} = 0 \text{ for all } 0 \le j \le k.$$

Conjugate Gradient Method

We now prove $\boldsymbol{p}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(j)} = 0, \quad j = 0, \cdots, k$

$$\because \ \boldsymbol{p}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(j)} = (-\boldsymbol{g}^{(k+1)} + \beta_k \boldsymbol{p}^{(k)})^\top \boldsymbol{Q}\boldsymbol{p}^{(j)}$$
$$= -\boldsymbol{g}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(j)} + \beta_k \boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(j)}$$

Case 1: when $j < k$, inductive hypothesis $\Longrightarrow \ \boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(j)} = 0.$

$\therefore \ \boldsymbol{p}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(j)} = -\boldsymbol{g}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(j)}.$

Since we know that

$\boldsymbol{g}^{(j+1)} = \boldsymbol{g}^{(j)} + \alpha_j \boldsymbol{Q}\boldsymbol{p}^{(j)}$ and $\boldsymbol{g}^{(k+1)\top}\boldsymbol{g}^{(i)} = 0, \ i = 0, \cdots, k$

$$\therefore \ \boldsymbol{p}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(j)} = -\boldsymbol{g}^{(k+1)\top}\frac{(\boldsymbol{g}^{(j+1)} - \boldsymbol{g}^{(j)})}{\alpha_j} = 0.$$

$$\therefore \ \boldsymbol{p}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(j)} = 0, j = 0, \cdots, k-1.$$

Case 2: when $j = k$, $\boldsymbol{p}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(j)} \overset{\because j=k}{=\!=} \boldsymbol{p}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)} =$

$(-\boldsymbol{g}^{(k+1)} + \beta_k \boldsymbol{p}^{(k)})^\top \boldsymbol{Q}\boldsymbol{p}^{(k)} \overset{\because \beta_k}{=\!=} 0.$

Overall, it follows from cases 1 and 2 that $\boldsymbol{p}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(j)} = 0, \quad j = 0, \cdots, k$

Pseudo-Code of CG Method

Pseudo-code of CG method

Require: initial point $\boldsymbol{x}^{(0)}$, tolerance ϵ,
set $k = 0$ and $\boldsymbol{p}^{(0)} = -\nabla f(\boldsymbol{x}^{(0)})$

1: **repeat**
2: $\quad \alpha_k = -\frac{\boldsymbol{g}^{(k)\top}\boldsymbol{p}^{(k)}}{\boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}}$;
3: $\quad \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}$;
4: $\quad \boldsymbol{g}^{(k+1)} = \nabla f(\boldsymbol{x}^{(k+1)})$;
5: $\quad \beta_k = \frac{\boldsymbol{g}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}}{\boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}}$;
6: $\quad \boldsymbol{p}^{(k+1)} = -\boldsymbol{g}^{(k+1)} + \beta_k \boldsymbol{p}^{(k)}$;
7: **until** $\|\boldsymbol{g}^{(k)}\| < \epsilon$.

other choices of β_k

$$\beta_k = \frac{\boldsymbol{g}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}}{\boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}}$$

$$= \frac{\boldsymbol{g}^{(k+1)\top}[\boldsymbol{g}^{(k+1)} - \boldsymbol{g}^{(k)}]}{\boldsymbol{p}^{(k)\top}[\boldsymbol{g}^{(k+1)} - \boldsymbol{g}^{(k)}]} \quad \text{(Hestenes-Stiefel)}$$

$$= \frac{\boldsymbol{g}^{(k+1)\top}[\boldsymbol{g}^{(k+1)} - \boldsymbol{g}^{(k)}]}{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}} \quad \text{(Polak-Ribiere)}$$

$$= \frac{\boldsymbol{g}^{(k+1)\top}\boldsymbol{g}^{(k+1)}}{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}} \quad \text{(Fletcher-Reeves)}$$

$$= \frac{\boldsymbol{g}^{(k+1)\top}\boldsymbol{g}^{(k+1)}}{\boldsymbol{p}^{(k)\top}[\boldsymbol{g}^{(k+1)} - \boldsymbol{g}^{(k+1)}]} \quad \text{(Dai-Yuan).}$$

★ For quadratic problems, all the above formulae of β_k are equivalent. But, for nonquadratic problems, they yield different solution.

★ Powell proves that the Fletcher-Peeves formula is more stable.

Conjugate Gradient Method

Example (minimize f by CG method)

$f(x_1, x_2) = x_1^2 + 2x_2^2$ with initial point $\boldsymbol{x}^{(0)} = [5, 5]^\top$.

- Step 1: $\boldsymbol{g}^{(0)} = \nabla f(\boldsymbol{x}^{(0)}) = [10, 20]^\top$, $\boldsymbol{p}^{(0)} = -\boldsymbol{g}^{(0)} = [-10, -20]^\top$.

  $\alpha_0 = -\frac{\boldsymbol{g}^{(0)\top}\boldsymbol{p}^{(0)}}{\boldsymbol{p}^{(0)\top}\boldsymbol{Q}\boldsymbol{p}^{(0)}} = \frac{5}{18}$,

  $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha_0 \boldsymbol{p}^{(0)} = [\frac{20}{9}, \frac{-5}{9}]^\top$.

- Step 2: $\boldsymbol{g}^{(1)} = \nabla f(\boldsymbol{x}^{(1)}) = [\frac{-40}{9}, \frac{20}{9}]^\top$.

  $\beta_0 = \frac{\boldsymbol{g}^{(1)\top}\boldsymbol{Q}\boldsymbol{d}^{(0)}}{\boldsymbol{p}^{(0)\top}\boldsymbol{Q}\boldsymbol{d}^{(0)}} = \frac{4}{81}$.

  $\boldsymbol{p}^{(1)} = -\boldsymbol{g}^{(1)} + \beta_0 \boldsymbol{p}^{(0)} = [\frac{400}{81}, \frac{100}{81}]^\top$.

  $\alpha_1 = -\frac{\boldsymbol{g}^{(1)\top}\boldsymbol{p}^{(1)}}{\boldsymbol{p}^{(1)\top}\boldsymbol{Q}\boldsymbol{p}^{(1)}} = \frac{9}{20}$,

  $\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} + \alpha_1 \boldsymbol{p}^{(1)} = [0, 0]^\top$.

  $\because \ \boldsymbol{g}^{(3)} = \nabla f(\boldsymbol{x}^{(3)}) = \boldsymbol{0}, \ \ \therefore \ \boldsymbol{x}^* = \boldsymbol{x}^{(2)}$ is the minimizer.

Conjugate Gradient Method

Example (minimize f by CG method)

$f(x_1, x_2, x_3) = \frac{3}{2}x_1^2 + 2x_2^2 + \frac{3}{2}x_3^2 + x_1x_3 + 2x_2x_3 - 3x_1 - x_3$ with initial point $\boldsymbol{x}^{(0)} = [0, 0, 0]^\top$.

- Step 1: $\boldsymbol{g}^{(0)} = \nabla f(\boldsymbol{x}^{(0)}) = [-3, 0, -1]^\top$, $\boldsymbol{p}^{(0)} = -\boldsymbol{g}^{(0)}$.
  $\alpha_0 = -\frac{\boldsymbol{g}^{(0)\top}\boldsymbol{p}^{(0)}}{\boldsymbol{p}^{(0)\top}\boldsymbol{Q}\boldsymbol{p}^{(0)}} = \frac{10}{36}$, $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha_0\boldsymbol{p}^{(0)} = [0.8333, 0, 0.2778]^\top$.
- Step 2: $\boldsymbol{g}^{(1)} = \nabla f(\boldsymbol{x}^{(1)}) = [-0.2222, 0.5556, 0.6667]^\top$.
  $\beta_0 = \frac{\boldsymbol{g}^{(1)\top}\boldsymbol{Q}\boldsymbol{d}^{(0)}}{\boldsymbol{p}^{(0)\top}\boldsymbol{Q}\boldsymbol{d}^{(0)}} = 0.08025.$
  $\boldsymbol{p}^{(1)} = -\boldsymbol{g}^{(1)} + \beta_0\boldsymbol{p}^{(0)} = [0.4630, -0.5556, -0.5864]^\top$.
  $\alpha_1 = -\frac{\boldsymbol{g}^{(1)\top}\boldsymbol{p}^{(1)}}{\boldsymbol{p}^{(1)\top}\boldsymbol{Q}\boldsymbol{p}^{(1)}} = 0.2187,$
  $\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} + \alpha_1\boldsymbol{p}^{(1)} = [0.9346, -0.1215, 0.1495]^\top$.
- Step 3: $\boldsymbol{g}^{(2)} = \nabla f(\boldsymbol{x}^{(2)}) = [-0.04673, -0.1869, 0.1402]^\top$.
  $\beta_1 = \frac{\boldsymbol{g}^{(2)\top}\boldsymbol{Q}\boldsymbol{d}^{(1)}}{\boldsymbol{p}^{(1)\top}\boldsymbol{Q}\boldsymbol{d}^{(1)}} = 0.07075.$
  $\boldsymbol{p}^{(2)} = -\boldsymbol{g}^{(2)} + \beta_1\boldsymbol{p}^{(1)} = [0.07948, 0.1476, -0.1817]^\top$.
  $\alpha_2 = -\frac{\boldsymbol{g}^{(2)\top}\boldsymbol{d}^{(2)}}{\boldsymbol{d}^{(2)\top}\boldsymbol{Q}\boldsymbol{d}^{(2)}} = 0.8231$, $\boldsymbol{x}^{(3)} = \boldsymbol{x}^{(2)} + \alpha_2\boldsymbol{p}^{(2)} = [1.0, 0.0, 0.0]^\top$.
  $\because\ \boldsymbol{g}^{(3)} = \nabla f(\boldsymbol{x}^{(3)}) = \boldsymbol{0}$, $\therefore\ \boldsymbol{x}^* = \boldsymbol{x}^{(3)}$ is the minimizer.

Extension of CG to Nonquadratic Problems

Extension: Theoretically, $\boldsymbol{Q}$ in CG method is replaced by Hessian of objective function. However, for general nonlinear functions, each iteration must refresh the Hessian, which requires a large amount of computation.

Pseudo-code of CG method for nonquadratic problems

1. Set $k = 0$, select the initial value $\boldsymbol{x}^{(0)}$.
2. $\boldsymbol{g}^{(0)} = \nabla f\left(\boldsymbol{x}^{(0)}\right)$, if $\boldsymbol{g}^{(0)} = \boldsymbol{0}$, stop; else, set $\boldsymbol{p}^{(0)} = -\boldsymbol{g}^{(0)}$.
3. $\alpha_k = -\frac{\boldsymbol{g}^{(k)\top}\boldsymbol{p}^{(k)}}{\boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}} \longrightarrow \boxed{\alpha_k = \arg\min_{a\geq 0} f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{p}^{(k)})}$.
4. $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}$.
5. $\boldsymbol{g}^{(k+1)} = \nabla f(\boldsymbol{x}^{(k+1)})$. If $\boldsymbol{g}^{(k+1)} = \boldsymbol{0}$, stop.
6. $\beta_k = \frac{\boldsymbol{g}^{(k+1)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}}{\boldsymbol{p}^{(k)\top}\boldsymbol{Q}\boldsymbol{p}^{(k)}}. \longrightarrow \boxed{\text{HS/PR/FR formula}}$
7. $\boldsymbol{p}^{(k+1)} = -\boldsymbol{g}^{(k+1)} + \beta_k \boldsymbol{p}^{(k)}$.
8. Set $k = k + 1$, go to step 3.

- ★ The search directional conjugacy and orthogonality disappear in the nonquadratic conjugate gradient method. [reason?]
- ★ No longer has the n-step convergence. Generally need a restrat operation.

Extension of CG to Nonquadratic Problems

1. CG method is related to the Krylov subspace method. Krylov subspace method (proposed by Hestenes, Stiefel and Lanczos), which was rated as one of the ten most influential methods in science and engineering of the 20th century.
2. Characteristics of CG method.
   1. For the nD quadratic problem, the results can be obtained in n steps.
   2. CG method not need to calculate the Hessian matrix.
   3. There is no need to store $n \times n$ matrices and not perform inverse operations.
3. In terms of calculation efficiency, CG and gradient descent method is between the fastest descent method and the Newton method. CG method is suitable to solve with large-scale problems.

Exercise

Exercise in the textbook: 10.4, 10.10.

Note: 10.4(b) should modify:

$\boldsymbol{Q}$ is a positive and definite matrix, $\{\boldsymbol{p}^{(1)}, ..., \boldsymbol{p}^{(n)}\}$ is an orthogonal vectors, if $\{\boldsymbol{p}^{(1)}, ..., \boldsymbol{p}^{(n)}\}$ is $\boldsymbol{Q}$-conjugated, so $\{\boldsymbol{p}^{(1)}, ..., \boldsymbol{p}^{(n)}\}$ must be the eigenvector of matrix $\boldsymbol{Q}$.