A Short Tutorial on Conjugate Gradient Method

FEM3220 Matrix Algebra Presentation

Braghadeesh Lakshminarayanan

June 22, 2022

Division of DCS, KTH Royal Institute of Technology

Table of contents

1. Introduction
2. Iterative Procedure
3. Existence of Conjugate Directions
4. Convergence of Conjugate Descent: Expanding Subspace Theorem
5. Procedure to Obtain Conjugate Directions: Gram-Schmidt Procedure
6. Conclusion

Introduction

Introduction

Introduction

- Consider the following optimization problem

Introduction

- Consider the following optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

Introduction

- Consider the following optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) \triangleq \frac{1}{2}x^\top Ax - b^\top x$$

- $f(x)$ is a convex quadratic function

Introduction

- Consider the following optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- $f(x)$ is a convex quadratic function
    - Stationary point x^*, obtained by $\nabla f(x^*) = 0$, is the global minimum

Introduction

- Consider the following optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- $f(x)$ is a convex quadratic function
  - Stationary point x^*, obtained by $\nabla f(x^*) = 0$, is the global minimum
  - $Ax^* = b$

Introduction

- Consider the following optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- $f(x)$ is a convex quadratic function
  - Stationary point x^*, obtained by $\nabla f(x^*) = 0$, is the global minimum
  - $Ax^* = b$

- Solution to the optimization problem $\iff$ solution to system of linear equation $Ax = b$

Introduction

- Consider the following optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- $f(x)$ is a convex quadratic function
  - Stationary point x^*, obtained by $\nabla f(x^*) = 0$, is the global minimum
  - $Ax^* = b$
- Solution to the optimization problem $\iff$ solution to system of linear equation $Ax = b$
- Suppose A is symmetric positive definite $\implies x^* = A^{-1} b$

Introduction

- Consider the following optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- $f(x)$ is a convex quadratic function
  - Stationary point x^*, obtained by $\nabla f(x^*) = 0$, is the global minimum
  - $Ax^* = b$

- Solution to the optimization problem $\iff$ solution to system of linear equation $Ax = b$

- Suppose A is symmetric positive definite $\implies x^* = A^{-1} b$
  - Computationally expensive $\mathcal{O}(n^3)$

Introduction

- Consider the following optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- $f(x)$ is a convex quadratic function
  - Stationary point x^*, obtained by $\nabla f(x^*) = 0$, is the global minimum
  - $Ax^* = b$
- Solution to the optimization problem $\iff$ solution to system of linear equation $Ax = b$
- Suppose A is symmetric positive definite $\implies x^* = A^{-1} b$
  - Computationally expensive $\mathcal{O}(n^3)$
- Remedy?

Introduction

- Consider the following optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- $f(x)$ is a convex quadratic function
  - Stationary point x^*, obtained by $\nabla f(x^*) = 0$, is the global minimum
  - $Ax^* = b$
- Solution to the optimization problem $\iff$ solution to system of linear equation $Ax = b$
- Suppose A is symmetric positive definite $\implies x^* = A^{-1}b$
  - Computationally expensive $\mathcal{O}(n^3)$
- Remedy?

  Adopt an iterative scheme to solve the optimization problem

Iterative Procedure

Iterative Procedure

Iterative Procedure

- Idea: Construct sequence $\{x_k\}$ such that

Iterative Procedure

- Idea: Construct sequence $\{x_k\}$ such that
  - $f(x_{k+1}) < f(x_k)$, $k = 0, \ldots$

Iterative Procedure

- Idea: Construct sequence $\{x_k\}$ such that
  - $f(x_{k+1}) < f(x_k)$, $k = 0, \ldots$
  - Stop when $\nabla f(x_k) = 0$, practical stopping criterion : $||\nabla f(x_k)|| < \epsilon$

Iterative Procedure

- Idea: Construct sequence $\{x_k\}$ such that
  - $f(x_{k+1}) < f(x_k)$, $k = 0, \ldots$
  - Stop when $\nabla f(x_k) = 0$, practical stopping criterion : $||\nabla f(x_k)|| < \epsilon$
- How do we construct such a sequence?

Iterative Procedure

- Idea: Construct sequence $\{x_k\}$ such that
  - $f(x_{k+1}) < f(x_k)$, $k = 0, \ldots$
  - Stop when $\nabla f(x_k) = 0$, practical stopping criterion : $||\nabla f(x_k)|| < \epsilon$
- How do we construct such a sequence?
  - Proposed construction : $x_{k+1} = x_k + \alpha_k d_k$, $\alpha_k \in \mathbb{R}$ and $d_k \in \mathbb{R}^n$

Iterative Procedure

- Idea: Construct sequence $\{x_k\}$ such that
  - $f(x_{k+1}) < f(x_k)$, $k = 0, \ldots$
  - Stop when $\nabla f(x_k) = 0$, practical stopping criterion : $||\nabla f(x_k)|| < \epsilon$
- How do we construct such a sequence?
  - Proposed construction : $x_{k+1} = x_k + \alpha_k d_k$, $\alpha_k \in \mathbb{R}$ and $d_k \in \mathbb{R}^n$
  - Choices of α_k and d_k such that $f(x_{k+1}) < f(x_k)$ for every k ?

Iterative Procedure

- Idea: Construct sequence $\{x_k\}$ such that
  - $f(x_{k+1}) < f(x_k)$, $k = 0, \ldots$
  - Stop when $\nabla f(x_k) = 0$, practical stopping criterion : $||\nabla f(x_k)|| < \epsilon$
- How do we construct such a sequence?
  - Proposed construction : $x_{k+1} = x_k + \alpha_k d_k$, $\alpha_k \in \mathbb{R}$ and $d_k \in \mathbb{R}^n$
  - Choices of α_k and d_k such that $f(x_{k+1}) < f(x_k)$ for every k ?
  - First order Taylor series approximation of f at x_k revelas that

Iterative Procedure

- Idea: Construct sequence $\{x_k\}$ such that
  - $f(x_{k+1}) < f(x_k)$, $k = 0, \ldots$
  - Stop when $\nabla f(x_k) = 0$, practical stopping criterion : $||\nabla f(x_k)|| < \epsilon$
- How do we construct such a sequence?
  - Proposed construction : $x_{k+1} = x_k + \alpha_k d_k$, $\alpha_k \in \mathbb{R}$ and $d_k \in \mathbb{R}^n$
  - Choices of α_k and d_k such that $f(x_{k+1}) < f(x_k)$ for every k ?
  - First order Taylor series approximation of f at x_k revelas that

$$f(x) \approx f(x_k) + (x - x_k)^\top \nabla f(x_k),$$

    where x is *sufficiently* close to x_k

Iterative Procedure

- Idea: Construct sequence $\{x_k\}$ such that
  - $f(x_{k+1}) < f(x_k)$, $k = 0, \ldots$
  - Stop when $\nabla f(x_k) = 0$, practical stopping criterion : $||\nabla f(x_k)|| < \epsilon$
- How do we construct such a sequence?
  - Proposed construction : $x_{k+1} = x_k + \alpha_k d_k$, $\alpha_k \in \mathbb{R}$ and $d_k \in \mathbb{R}^n$
  - Choices of α_k and d_k such that $f(x_{k+1}) < f(x_k)$ for every k ?
  - First order Taylor series approximation of f at x_k revelas that

    $$f(x) \approx f(x_k) + (x - x_k)^\top \nabla f(x_k),$$

    where x is *sufficiently* close to x_k
  - In particular, $x = x_{k+1}$,

Iterative Procedure

- Idea: Construct sequence $\{x_k\}$ such that
  - $f(x_{k+1}) < f(x_k)$, $k = 0, \ldots$
  - Stop when $\nabla f(x_k) = 0$, practical stopping criterion : $||\nabla f(x_k)|| < \epsilon$
- How do we construct such a sequence?
  - Proposed construction : $x_{k+1} = x_k + \alpha_k d_k$, $\alpha_k \in \mathbb{R}$ and $d_k \in \mathbb{R}^n$
  - Choices of α_k and d_k such that $f(x_{k+1}) < f(x_k)$ for every k ?
  - First order Taylor series approximation of f at x_k revelas that

    $$f(x) \approx f(x_k) + (x - x_k)^\top \nabla f(x_k),$$

    where x is *sufficiently* close to x_k
  - In particular, $x = x_{k+1}$,

    $$f(x_{k+1}) \approx f(x_k) + \alpha_k d_k{}^\top \nabla f(x_k)$$

Iterative Procedure

- Idea: Construct sequence $\{x_k\}$ such that
  - $f(x_{k+1}) < f(x_k)$, $k = 0, \ldots$
  - Stop when $\nabla f(x_k) = 0$, practical stopping criterion : $||\nabla f(x_k)|| < \epsilon$
- How do we construct such a sequence?
  - Proposed construction : $x_{k+1} = x_k + \alpha_k d_k$, $\alpha_k \in \mathbb{R}$ and $d_k \in \mathbb{R}^n$
  - Choices of α_k and d_k such that $f(x_{k+1}) < f(x_k)$ for every k ?
  - First order Taylor series approximation of f at x_k revelas that

$$f(x) \approx f(x_k) + (x - x_k)^\top \nabla f(x_k),$$

  where x is *sufficiently* close to x_k

  - In particular, $x = x_{k+1}$,

$$f(x_{k+1}) \approx f(x_k) + \alpha_k {d_k}^\top \nabla f(x_k)$$

  - Easy to see $f(x_{k+1}) < f(x_k)$ if ${d_k}^\top \nabla f(x_k) < 0$, and $\alpha_k > 0$

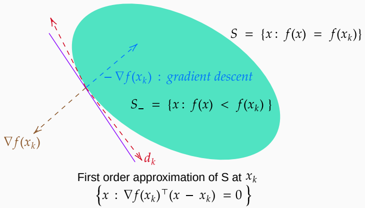
Iterative Procedure

Iterative Procedure

- Descent direction : Choose d_k such that $\nabla f(x_k)^\top d_k < 0$. More generally, $\mathcal{D} := \{d \in \mathbb{R}^n : \nabla f(x_k)^\top d < 0\}$ is the set of descent directions.

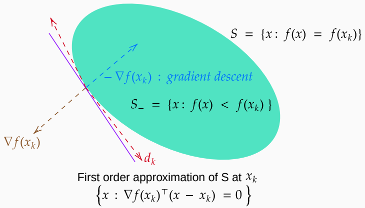
Iterative Procedure

- Descent direction : Choose d_k such that $\nabla f(x_k)^\top d_k < 0$. More generally, $\mathcal{D} := \{d \in \mathbb{R}^n : \nabla f(x_k)^\top d < 0\}$ is the set of descent directions.



Iterative Procedure

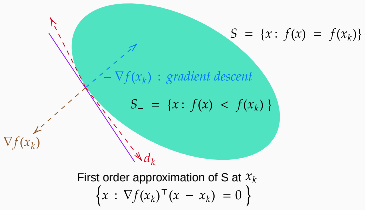
- Descent direction : Choose d_k such that $\nabla f(x_k)^\top d_k < 0$. More generally, $\mathcal{D} := \{d \in \mathbb{R}^n : \nabla f(x_k)^\top d < 0\}$ is the set of descent directions.



- For first order Taylor approximation to hold, we need α_k to be not too large.

Iterative Procedure

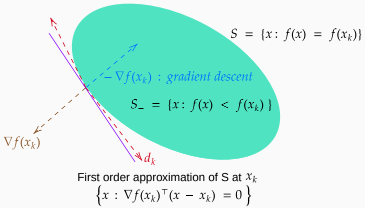
- Descent direction : Choose d_k such that $\nabla f(x_k)^\top d_k < 0$. More generally, $\mathcal{D} := \{d \in \mathbb{R}^n : \nabla f(x_k)^\top d < 0\}$ is the set of descent directions.



- For first order Taylor approximation to hold, we need α_k to be not too large.
- How do we find α_k?

Iterative Procedure

- Descent direction : Choose d_k such that $\nabla f(x_k)^\top d_k < 0$. More generally, $\mathcal{D} := \{d \in \mathbb{R}^n : \nabla f(x_k)^\top d < 0\}$ is the set of descent directions.



- For first order Taylor approximation to hold, we need α_k to be not too large.

- How do we find α_k?
  - Exact line search : step size $\alpha_k = \arg\min_{\alpha > 0} f(x_k + \alpha d_k)$

Conjugate Descent Method: Quick Overview

Conjugate Descent Method: Quick Overview

- Recall:

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x,$$

where A is symmetric positive definite matrix.

Conjugate Descent Method: Quick Overview

- Recall:

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x,$$

  where A is symmetric positive definite matrix.

- Suppose we start at some initial point $x_0 \in \mathbb{R}^n$ in the iterative procedure.

Conjugate Descent Method: Quick Overview

- Recall:

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x,$$

  where A is symmetric positive definite matrix.

- Suppose we start at some initial point $x_0 \in \mathbb{R}^n$ in the iterative procedure.
- Let $\{d_0, \ldots, d_{n-1}\}$ be a set of linearly independent directions. Note that this is a maximal linearly independent set in $\mathbb{R}^n$, and hence it forms a basis.

Conjugate Descent Method: Quick Overview

Conjugate Descent Method: Quick Overview

- $x - x_0 \in \mathbb{R}^n \implies x - x_0 = \sum_{i=0}^{n-1} \alpha_i d_i \implies x = x_0 + \sum_{i=0}^{n-1} \alpha_i d_i$

Conjugate Descent Method: Quick Overview

- $x - x_0 \in \mathbb{R}^n \implies x - x_0 = \sum_{i=0}^{n-1} \alpha_i d_i \implies x = x_0 + \sum_{i=0}^{n-1} \alpha_i d_i$
- Easy to see

$$\min_{x \in \mathbb{R}^n} f(x) \equiv \min_{\alpha \in \mathbb{R}^n} \Psi(\alpha)$$

, where
$\Psi(\alpha) =$
$\frac{1}{2}\left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right)^\top A \left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right) - b^\top \left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right)$,
$\alpha = (\alpha_0 \dots \alpha_{n-1})^T$

Conjugate Descent Method: Quick Overview

- $x - x_0 \in \mathbb{R}^n \implies x - x_0 = \sum_{i=0}^{n-1} \alpha_i d_i \implies x = x_0 + \sum_{i=0}^{n-1} \alpha_i d_i$
- Easy to see

$$\min_{x \in \mathbb{R}^n} f(x) \equiv \min_{\alpha \in \mathbb{R}^n} \Psi(\alpha)$$

, where
$\Psi(\alpha) =$
$\frac{1}{2}\left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right)^\top A \left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right) - b^\top \left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right)$,
$\alpha = (\alpha_0 \ldots \alpha_{n-1})^T$

- $\Psi(\alpha)$ is not separable in terms of α_i.

Conjugate Descent Method: Quick Overview

- $x - x_0 \in \mathbb{R}^n \implies x - x_0 = \sum_{i=0}^{n-1} \alpha_i d_i \implies x = x_0 + \sum_{i=0}^{n-1} \alpha_i d_i$
- Easy to see

$$\min_{x \in \mathbb{R}^n} f(x) \equiv \min_{\alpha \in \mathbb{R}^n} \Psi(\alpha)$$

, where
$\Psi(\alpha) =$
$\frac{1}{2}\left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right)^\top A \left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right) - b^\top \left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right),$
$\alpha = (\alpha_0 \ldots \alpha_{n-1})^T$

- $\Psi(\alpha)$ is not separable in terms of α_i.What do we do now?

Conjugate Descent Method: Quick Overview

- $x - x_0 \in \mathbb{R}^n \implies x - x_0 = \sum_{i=0}^{n-1} \alpha_i d_i \implies x = x_0 + \sum_{i=0}^{n-1} \alpha_i d_i$
- Easy to see

$$\min_{x \in \mathbb{R}^n} f(x) \equiv \min_{\alpha \in \mathbb{R}^n} \Psi(\alpha)$$

  , where
  $\Psi(\alpha) =$
  $\frac{1}{2}\left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right)^\top A \left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right) - b^\top \left(x_0 + \sum_{i=0}^{n-1} \alpha_i d_i\right),$
  $\alpha = (\alpha_0 \ldots \alpha_{n-1})^T$

- $\Psi(\alpha)$ is not separable in terms of α_i.What do we do now?
- Let $D := (d_0 | d_1 | \ldots | d_{n-1})$

Conjugate Descent Method: Quick Overview

Conjugate Descent Method: Quick Overview

- Now, $\Psi(\alpha) = \frac{1}{2}\alpha^\top \underbrace{D^\top A D}_{:=Q} \alpha + (Ax_0 - b)^\top D\alpha + \underbrace{\frac{1}{2}x_0{}^\top A x_0 - b^\top x_0}_{\text{constant}}$

Conjugate Descent Method: Quick Overview

- Now, $\Psi(\alpha) = \frac{1}{2}\alpha^\top \underbrace{D^\top A D}_{:=Q} \alpha + (Ax_0 - b)^\top D\alpha + \underbrace{\frac{1}{2}{x_0}^\top A x_0 - b^\top x_0}_{\text{constant}}$
- Let us look at the structure of Q

Conjugate Descent Method: Quick Overview

- Now, $\Psi(\alpha) = \frac{1}{2}\alpha^\top \underbrace{D^\top A D}_{:=Q} \alpha + (Ax_0 - b)^\top D\alpha + \underbrace{\frac{1}{2}{x_0}^\top A x_0 - b^\top x_0}_{\text{constant}}$
- Let us look at the structure of Q

$$Q = D^\top A D = \begin{pmatrix} {d_0}^\top A d_0 & \dots & {d_0}^\top A d_{n-1} \\ \vdots & \dots & \vdots \\ {d_{n-1}}^\top A d_0 & \dots & {d_{n-1}}^\top A d_{n-1} \end{pmatrix}$$

Conjugate Descent Method: Quick Overview

- Now, $\Psi(\alpha) = \frac{1}{2}\alpha^\top \underbrace{D^\top A D}_{:=Q} \alpha + (Ax_0 - b)^\top D\alpha + \underbrace{\frac{1}{2}x_0{}^\top A x_0 - b^\top x_0}_{\text{constant}}$
- Let us look at the structure of Q

$$Q = D^\top A D = \begin{pmatrix} d_0{}^\top A d_0 & \dots & d_0{}^\top A d_{n-1} \\ \vdots & \dots & \vdots \\ d_{n-1}{}^\top A d_0 & \dots & d_{n-1}{}^\top A d_{n-1} \end{pmatrix}$$

- Q will be diagonal if $d_i{}^\top A d_j = 0, \forall\, i \neq j$ and $\Psi(\alpha)$ will then be separable in terms of $\alpha_0, \dots, \alpha_{n-1}$.

Conjugate Descent Method: Quick Overview

- Now, $\Psi(\alpha) = \frac{1}{2}\alpha^\top \underbrace{D^\top A D}_{:=Q} \alpha + (Ax_0 - b)^\top D\alpha + \underbrace{\frac{1}{2}x_0{}^\top A x_0 - b^\top x_0}_{\text{constant}}$
- Let us look at the structure of Q

$$Q = D^\top A D = \begin{pmatrix} d_0{}^\top A d_0 & \dots & d_0{}^\top A d_{n-1} \\ \vdots & \dots & \vdots \\ d_{n-1}{}^\top A d_0 & \dots & d_{n-1}{}^\top A d_{n-1} \end{pmatrix}$$

- Q will be diagonal if $d_i{}^\top A d_j = 0, \forall\, i \neq j$ and $\Psi(\alpha)$ will then be separable in terms of $\alpha_0, \dots, \alpha_{n-1}$.

$$\Psi(\alpha) = \frac{1}{2}\sum_{i=0}^{n-1}\left[(x_0 + \alpha_i d_i)^\top A\,(x_0 + \alpha_i d_i) - 2b^\top\,(x_0 + \alpha_i d_i)\right] + \text{constant}$$

Conjugate Descent Method: Quick Overview

- Now, $\Psi(\alpha) = \frac{1}{2}\alpha^\top \underbrace{D^\top A D}_{:=Q} \alpha + (Ax_0 - b)^\top D\alpha + \underbrace{\frac{1}{2}x_0{}^\top A x_0 - b^\top x_0}_{\text{constant}}$
- Let us look at the structure of Q

$$Q = D^\top A D = \begin{pmatrix} d_0{}^\top A d_0 & \dots & d_0{}^\top A d_{n-1} \\ \vdots & \dots & \vdots \\ d_{n-1}{}^\top A d_0 & \dots & d_{n-1}{}^\top A d_{n-1} \end{pmatrix}$$

- Q will be diagonal if $d_i{}^\top A d_j = 0, \forall\, i \neq j$ and $\Psi(\alpha)$ will then be separable in terms of $\alpha_0, \dots, \alpha_{n-1}$.

$$\Psi(\alpha) = \frac{1}{2}\sum_{i=0}^{n-1}\left[(x_0 + \alpha_i d_i)^\top A\,(x_0 + \alpha_i d_i) - 2b^\top\,(x_0 + \alpha_i d_i)\right] + \text{constant}$$

Definition

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric posititve definite matrix. The vectors $\{d_0, \dots, d_{n-1}\}$ are $A - \textit{conjugate}$ if $d_i{}^T A d_j = 0, \forall\, i \neq j$.

Conjugate Descent Method: Quick Overview

Conjugate Descent Method: Quick Overview

Claim

If $\{d_0, \ldots, d_{n-1}\}$ are $A - conjugate$, then they are linearly independent.

Conjugate Descent Method: Quick Overview

Claim

If $\{d_0, \ldots, d_{n-1}\}$ are $A - conjugate$, then they are linearly independent.

Proof.

$$\sum_{i=0}^{n-1} \mu_i d_i = 0 \implies d_i^\top A \sum_{j=0}^{n-1} \mu_j d_j = 0$$
$$\implies \sum_{j=0}^{n-1} \mu_j d_i^\top A d_j = 0$$
$$\implies \mu_i d_i^\top A d_i = 0 \, (\because d_i^\top A d_j = 0 \, \forall \, i \neq j (A - conjugacy))$$
$$\implies \mu_i = 0 \, (\because A \text{ is p.d. and } \therefore d_i^\top A d_i \neq 0)$$

Therefore, $\sum_{i=0}^{n-1} \mu_i d_i = 0 \implies \mu_i = 0$, and hence, $\{d_0, \ldots, d_{n-1}\}$ is a linearly independent set. □

Conjugate Descent Method: Quick Overview

Conjugate Descent Method: Quick Overview

- $\frac{\partial \Psi}{\partial \alpha_i} = 0 \implies \alpha_i{}^* = -\frac{d_i{}^\top (Ax_0 - b)}{d_i{}^\top A d_i}$

Conjugate Descent Method: Quick Overview

- $\frac{\partial \Psi}{\partial \alpha_i} = 0 \implies \alpha_i{}^* = -\frac{d_i{}^\top (Ax_0 - b)}{d_i{}^\top A d_i}$
- Finally, $x^* = x_0 + \sum_{i=0}^{n-1} \alpha_i{}^* d_i$

Therefore, solution to the minimization of convex quadratic function is

Conjugate Descent Method: Quick Overview

- $\frac{\partial \Psi}{\partial \alpha_i} = 0 \implies \alpha_i{}^* = -\frac{d_i{}^\top (Ax_0 - b)}{d_i{}^\top A d_i}$
- Finally, $x^* = x_0 + \sum_{i=0}^{n-1} \alpha_i{}^* d_i$

Therefore, solution to the minimization of convex quadratic function is the linear combination of conjugate directions $d_0, \ldots, d_{n-1}$ and any arbitrary initial point x_0.

Conjugate Descent Method: Quick Overview

- $\frac{\partial \Psi}{\partial \alpha_i} = 0 \implies \alpha_i{}^* = -\frac{d_i{}^\top (Ax_0 - b)}{d_i{}^\top A d_i}$
- Finally, $x^* = x_0 + \sum_{i=0}^{n-1} \alpha_i{}^* d_i$

Therefore, solution to the minimization of convex quadratic function is the linear combination of conjugate directions $d_0, \ldots, d_{n-1}$ and any arbitrary initial point x_0.

Questions

- Given A, is there a set of $A - \textit{conjugate}$ vectors?

Conjugate Descent Method: Quick Overview

- $\frac{\partial \Psi}{\partial \alpha_i} = 0 \implies \alpha_i{}^* = -\frac{d_i{}^\top (Ax_0 - b)}{d_i{}^\top A d_i}$
- Finally, $x^* = x_0 + \sum_{i=0}^{n-1} \alpha_i{}^* d_i$

Therefore, solution to the minimization of convex quadratic function is the linear combination of conjugate directions $d_0, \ldots, d_{n-1}$ and any arbitrary initial point x_0.

Questions

- Given A, is there a set of $A - \textit{conjugate}$ vectors?
- If yes, how do we obtain them iteratively?

Conjugate Descent Method: Quick Overview

- $\frac{\partial \Psi}{\partial \alpha_i} = 0 \implies {\alpha_i}^* = -\frac{{d_i}^\top (Ax_0 - b)}{{d_i}^\top A d_i}$
- Finally, $x^* = x_0 + \sum_{i=0}^{n-1} {\alpha_i}^* d_i$

Therefore, solution to the minimization of convex quadratic function is the linear combination of conjugate directions $d_0, \ldots, d_{n-1}$ and any arbitrary initial point x_0.

Questions

- Given A, is there a set of $A - \mathit{conjugate}$ vectors?
- If yes, how do we obtain them iteratively?
- How does sequence x_k with $x_{k+1} = x_k + \alpha_k d_k$ converge to x^* in atmost n iterations?

Existence of Conjugate Directions

Conjugate Directions: Existence

Conjugate Directions: Existence

- A is symmetric p.d. $\implies$ A has n mutually orthogonal eigenvectors

Conjugate Directions: Existence

- A is symmetric p.d. $\implies$ A has n mutually orthogonal eigenvectors
- Suppose v_0 and v_1 are two orthogonal eigenvectors of A. Then, ${v_0}^\top v_1 = 0$.

Conjugate Directions: Existence

- A is symmetric p.d. $\implies$ A has n mutually orthogonal eigenvectors
- Suppose v_0 and v_1 are two orthogonal eigenvectors of A. Then, ${v_0}^\top v_1 = 0$.
- $Av_0 = \lambda_0 v_0 \implies {v_1}^\top A v_0 = \lambda_0 {v_1}^\top v_0 \implies {v_1}^\top A v_0 = 0 \implies v_0, v_1$ are $A - \mathit{conjugate}$.

Conjugate Directions: Existence

- A is symmetric p.d. $\implies A$ has n mutually orthogonal eigenvectors
- Suppose v_0 and v_1 are two orthogonal eigenvectors of A. Then, ${v_0}^\top v_1 = 0$.
- $Av_0 = \lambda_0 v_0 \implies {v_1}^\top A v_0 = \lambda_0 {v_1}^\top v_0 \implies {v_1}^\top A v_0 = 0 \implies v_0, v_1$ are $A - conjugate$.
  - Easy to see ${v_i}^\top A v_j = 0,\ \forall\, i \neq j$, if $v_0, \ldots, v_{n-1}$ are n orthogonal eigenvectors of A.

Conjugate Directions: Existence

- A is symmetric p.d. $\implies$ A has n mutually orthogonal eigenvectors
- Suppose v_0 and v_1 are two orthogonal eigenvectors of A. Then, ${v_0}^\top v_1 = 0$.
- $Av_0 = \lambda_0 v_0 \implies {v_1}^\top A v_0 = \lambda_0 {v_1}^\top v_0 \implies {v_1}^\top A v_0 = 0 \implies v_0, v_1$ are $A - conjugate$.
  - Easy to see ${v_i}^\top A v_j = 0,\ \forall\, i \neq j$, if $v_0, \ldots, v_{n-1}$ are n orthogonal eigenvectors of A.

$\therefore$ Conjugate directions exist!

Convergence of Conjugate Descent: Expanding Subspace Theorem

Convergence of Conjugate Descent

Convergence of Conjugate Descent

$$\min_{x\in\mathbb{R}^n} f(x) \triangleq \frac{1}{2}x^\top Ax - b^\top x$$

Convergence of Conjugate Descent

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- Let $\mathcal{B}_k$ denote the subspace spanned by $d_0, \ldots, d_{k-1}$.

Convergence of Conjugate Descent

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- Let $\mathcal{B}_k$ denote the subspace spanned by $d_0, \ldots, d_{k-1}$.
    - Easy to verify that $\mathcal{B}_k \subset \mathcal{B}_{k+1}$

Convergence of Conjugate Descent

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- Let $\mathcal{B}_k$ denote the subspace spanned by $d_0, \ldots, d_{k-1}$.
  - Easy to verify that $\mathcal{B}_k \subset \mathcal{B}_{k+1}$
- Let $x_0 \in \mathbb{R}^n$ be any arbitrary initial point and iterative scheme be $x_{k+1} = x_k + \alpha_k d_k$.

Convergence of Conjugate Descent

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- Let $\mathcal{B}_k$ denote the subspace spanned by $d_0, \ldots, d_{k-1}$.
    - Easy to verify that $\mathcal{B}_k \subset \mathcal{B}_{k+1}$
- Let $x_0 \in \mathbb{R}^n$ be any arbitrary initial point and iterative scheme be $x_{k+1} = x_k + \alpha_k d_k$.
    - α_k is obtained by exact line search: $\alpha_k = \arg\min_{\alpha > 0} f(x_k + \alpha d_k)$

Convergence of Conjugate Descent

$$\min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{2} x^\top A x - b^\top x$$

- Let $\mathcal{B}_k$ denote the subspace spanned by $d_0, \ldots, d_{k-1}$.
  - Easy to verify that $\mathcal{B}_k \subset \mathcal{B}_{k+1}$
- Let $x_0 \in \mathbb{R}^n$ be any arbitrary initial point and iterative scheme be $x_{k+1} = x_k + \alpha_k d_k$.
  - α_k is obtained by exact line search: $\alpha_k = \arg\min_{\alpha > 0} f(x_k + \alpha d_k)$

Claim

$x_k = \arg\min_{x \in x_0 + \mathcal{B}_k} f(x)$. That is, $f(x_k) \leq f(x), \forall x \in x_0 + \mathcal{B}_k$

Convergence of Conjugate Descent

$$\min_{x\in\mathbb{R}^n} f(x) \triangleq \frac{1}{2}x^\top Ax - b^\top x$$

- Let $\mathcal{B}_k$ denote the subspace spanned by $d_0, \ldots, d_{k-1}$.
  - Easy to verify that $\mathcal{B}_k \subset \mathcal{B}_{k+1}$
- Let $x_0 \in \mathbb{R}^n$ be any arbitrary initial point and iterative scheme be $x_{k+1} = x_k + \alpha_k d_k$.
  - α_k is obtained by exact line search: $\alpha_k = \arg\min_{\alpha>0} f(x_k + \alpha d_k)$

Claim

$x_k = \arg\min_{x \in x_0+\mathcal{B}_k} f(x)$. That is, $f(x_k) \leq f(x)$, $\forall x \in x_0 + \mathcal{B}_k$

Let us try to prove this claim. We shall denote the gradient $\nabla f(x_k)$ by g_k.

Convergence of Conjugate Descent

First note that $g_k = Ax_k - b$.

Convergence of Conjugate Descent

First note that $g_k = Ax_k - b$. Due to exact line search, α_k should satisfy the first order necessary condition.

Convergence of Conjugate Descent

First note that $g_k = Ax_k - b$. Due to exact line search, α_k should satisfy the first order necessary condition. That is,
$\nabla f(x_k + \alpha_k d_K)^\top d_k = 0$

Convergence of Conjugate Descent

First note that $g_k = Ax_k - b$. Due to exact line search, α_k should satisfy the first order necessary condition. That is,
$\nabla f(x_k + \alpha_k d_K)^\top d_k = 0 \implies g_{k+1}{}^\top d_k = 0, \forall\, k = 0, \ldots, n-1$

Convergence of Conjugate Descent

First note that $g_k = Ax_k - b$. Due to exact line search, α_k should satisfy the first order necessary condition. That is,
$\nabla f(x_k + \alpha_k d_K)^\top d_k = 0 \implies g_{k+1}{}^\top d_k = 0, \forall\, k = 0, \ldots, n-1$
$x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$.

Convergence of Conjugate Descent

First note that $g_k = Ax_k - b$. Due to exact line search, α_k should satisfy the first order necessary condition. That is,
$\nabla f(x_k + \alpha_k d_K)^\top d_k = 0 \implies g_{k+1}{}^\top d_k = 0, \forall\, k = 0, \dots, n-1$
$x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$. Using telescopic sum until j, we obtain
$x_k = x_j + \sum_{i=j}^{k-1} \alpha_i d_i$, where $j \in \{0, \dots, k-1\}$

Convergence of Conjugate Descent

First note that $g_k = Ax_k - b$. Due to exact line search, α_k should satisfy the first order necessary condition. That is,
$\nabla f(x_k + \alpha_k d_K)^\top d_k = 0 \implies {g_{k+1}}^\top d_k = 0, \forall\, k = 0, \ldots, n-1$
$x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$. Using telescopic sum until j, we obtain
$x_k = x_j + \sum_{i=j}^{k-1} \alpha_i d_i$, where $j \in \{0, \ldots, k-1\}$

$\therefore\ Ax_k - b = Ax_j - b + \sum_{i=j}^{k-1} \alpha_i A d_i$

Convergence of Conjugate Descent

First note that $g_k = Ax_k - b$. Due to exact line search, α_k should satisfy the first order necessary condition. That is,
$\nabla f(x_k + \alpha_k d_K)^\top d_k = 0 \implies g_{k+1}{}^\top d_k = 0, \forall\, k = 0, \ldots, n-1$
$x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$. Using telescopic sum until j, we obtain
$x_k = x_j + \sum_{i=j}^{k-1} \alpha_i d_i$, where $j \in \{0, \ldots, k-1\}$

$\therefore\ Ax_k - b = Ax_j - b + \sum_{i=j}^{k-1} \alpha_i A d_i$
$\implies\ g_k = g_j + \sum_{i=j}^{k-1} \alpha_i A d_i$

Convergence of Conjugate Descent

First note that $g_k = Ax_k - b$. Due to exact line search, α_k should satisfy the first order necessary condition. That is,
$\nabla f(x_k + \alpha_k d_K)^\top d_k = 0 \implies {g_{k+1}}^\top d_k = 0, \forall\, k = 0, \ldots, n-1$
$x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$. Using telescopic sum until j, we obtain
$x_k = x_j + \sum_{i=j}^{k-1} \alpha_i d_i$, where $j \in \{0, \ldots, k-1\}$

$$\therefore\ Ax_k - b = Ax_j - b + \sum_{i=j}^{k-1} \alpha_i A d_i$$
$$\implies g_k = g_j + \sum_{i=j}^{k-1} \alpha_i A d_i$$
$$\implies {g_k}^\top d_{j-1} = \underbrace{{g_j}^\top d_{j-1}}_{=0\text{(first order necessary condition)}} + \sum_{i=j}^{k-1} \alpha_i \underbrace{{d_i}^\top A d_{j-1}}_{=0\text{(A-conjugacy)}} = 0.$$

Convergence of Conjugate Descent

First note that $g_k = Ax_k - b$. Due to exact line search, α_k should satisfy the first order necessary condition. That is,
$\nabla f(x_k + \alpha_k d_K)^\top d_k = 0 \implies g_{k+1}{}^\top d_k = 0, \forall\, k = 0, \ldots, n-1$
$x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$. Using telescopic sum until j, we obtain
$x_k = x_j + \sum_{i=j}^{k-1} \alpha_i d_i$, where $j \in \{0, \ldots, k-1\}$

$$\therefore\ Ax_k - b = Ax_j - b + \sum_{i=j}^{k-1} \alpha_i A d_i$$
$$\implies g_k = g_j + \sum_{i=j}^{k-1} \alpha_i A d_i$$
$$\implies g_k{}^\top d_{j-1} = \underbrace{g_j{}^\top d_{j-1}}_{=0\text{(first order necessary condition)}} + \sum_{i=j}^{k-1} \alpha_i \underbrace{d_i{}^\top A d_{j-1}}_{=0\text{(A-conjugacy)}} = 0.$$
$$\therefore\ g_k{}^\top d_j = 0\ \forall j = 0, \ldots, k-1$$

Convergence of Conjugate Descent

First note that $g_k = Ax_k - b$. Due to exact line search, α_k should satisfy the first order necessary condition. That is,
$\nabla f(x_k + \alpha_k d_K)^\top d_k = 0 \implies {g_{k+1}}^\top d_k = 0, \forall\, k = 0, \dots, n-1$
$x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$. Using telescopic sum until j, we obtain
$x_k = x_j + \sum_{i=j}^{k-1} \alpha_i d_i$, where $j \in \{0, \dots, k-1\}$

$$\therefore\ Ax_k - b = Ax_j - b + \sum_{i=j}^{k-1} \alpha_i A d_i$$
$$\implies g_k = g_j + \sum_{i=j}^{k-1} \alpha_i A d_i$$
$$\implies {g_k}^\top d_{j-1} = \underbrace{{g_j}^\top d_{j-1}}_{=0 \text{(first order necessary condition)}} + \sum_{i=j}^{k-1} \alpha_i \underbrace{{d_i}^\top A d_{j-1}}_{=0 \text{(A-conjugacy)}} = 0.$$

$\therefore\ {g_k}^\top d_j = 0\ \forall j = 0, \dots, k-1$
In other words, $g_k \perp \mathcal{B}_k$.

Convergence of Conjugate Descent

Convergence of Conjugate Descent

We need to show that $f(x_k) \leq f(x)$, $\forall x \in x_0 + \mathcal{B}_k$.

Convergence of Conjugate Descent

We need to show that $f(x_k) \leq f(x)$, $\forall x \in x_0 + \mathcal{B}_k$.

Or, equivalently

$$f(x_0 + \sum_{j=0}^{k-1} \alpha_j d_j) \leq f(x_0 + \sum_{j=0}^{k-1} \mu_j d_j), \quad \mu_j \in \mathbb{R}$$

Convergence of Conjugate Descent

We need to show that $f(x_k) \leq f(x)$, $\forall x \in x_0 + \mathcal{B}_k$.
Or, equivalently

$$f(x_0 + \sum_{j=0}^{k-1} \alpha_j d_j) \leq f(x_0 + \sum_{j=0}^{k-1} \mu_j d_j), \quad \mu_j \in \mathbb{R}$$

Using the Taylor series expansion of f around x_0,

Convergence of Conjugate Descent

We need to show that $f(x_k) \leq f(x)$, $\forall x \in x_0 + \mathcal{B}_k$.
Or, equivalently

$$f(x_0 + \sum_{j=0}^{k-1} \alpha_j d_j) \leq f(x_0 + \sum_{j=0}^{k-1} \mu_j d_j), \quad \mu_j \in \mathbb{R}$$

Using the Taylor series expansion of f around x_0,we get,

$$f(x_0) + \sum_{j=0}^{k-1} (\alpha_j {g_0}^\top d_j + \frac{1}{2} {\alpha_j}^2 {d_j}^\top A d_j) \leq f(x_0) + \sum_{j=0}^{k-1} (\mu_j {g_0}^\top d_j + \frac{1}{2} {\mu_j}^2 {d_j}^\top A d_j)$$

Remember, $\alpha_j = \arg\min_{\alpha} f(x_j + \alpha d_j)$.

Convergence of Conjugate Descent

We need to show that $f(x_k) \leq f(x), \forall x \in x_0 + \mathcal{B}_k$.
Or, equivalently

$$f(x_0 + \sum_{j=0}^{k-1} \alpha_j d_j) \leq f(x_0 + \sum_{j=0}^{k-1} \mu_j d_j), \quad \mu_j \in \mathbb{R}$$

Using the Taylor series expansion of f around x_0,we get,

$$f(x_0) + \sum_{j=0}^{k-1}(\alpha_j {g_0}^\top d_j + \frac{1}{2}{\alpha_j}^2 {d_j}^\top A d_j) \leq f(x_0) + \sum_{j=0}^{k-1}(\mu_j {g_0}^\top d_j + \frac{1}{2}{\mu_j}^2 {d_j}^\top A d_j)$$

Remember, $\alpha_j = \arg\min_\alpha f(x_j + \alpha d_j)$.Therefore,

$$f(x_j + \alpha_j d_j) \leq f(x_j + \mu_j d_j), \forall j \in \{0, \dots, n-1\} \quad (\because \alpha_j \text{ is the minimizer})$$

Convergence of Conjugate Descent

We need to show that $f(x_k) \leq f(x)$, $\forall x \in x_0 + \mathcal{B}_k$.
Or, equivalently

$$f(x_0 + \sum_{j=0}^{k-1} \alpha_j d_j) \leq f(x_0 + \sum_{j=0}^{k-1} \mu_j d_j), \quad \mu_j \in \mathbb{R}$$

Using the Taylor series expansion of f around x_0,we get,

$$f(x_0) + \sum_{j=0}^{k-1} (\alpha_j {g_0}^\top d_j + \frac{1}{2}{\alpha_j}^2 {d_j}^\top A d_j) \leq f(x_0) + \sum_{j=0}^{k-1} (\mu_j {g_0}^\top d_j + \frac{1}{2}{\mu_j}^2 {d_j}^\top A d_j)$$

Remember, $\alpha_j = \arg\min_\alpha f(x_j + \alpha d_j)$.Therefore,

$$f(x_j + \alpha_j d_j) \leq f(x_j + \mu_j d_j), \forall j \in \{0, \dots, n-1\} \quad (\because \alpha_j \text{ is the minimizer})$$

Again, using Taylor series expansion of f around x_j,

Convergence of Conjugate Descent

We need to show that $f(x_k) \leq f(x)$, $\forall x \in x_0 + \mathcal{B}_k$.
Or, equivalently

$$f(x_0 + \sum_{j=0}^{k-1} \alpha_j d_j) \leq f(x_0 + \sum_{j=0}^{k-1} \mu_j d_j), \quad \mu_j \in \mathbb{R}$$

Using the Taylor series expansion of f around x_0,we get,

$$f(x_0) + \sum_{j=0}^{k-1}(\alpha_j {g_0}^\top d_j + \frac{1}{2}{\alpha_j}^2 {d_j}^\top A d_j) \leq f(x_0) + \sum_{j=0}^{k-1}(\mu_j {g_0}^\top d_j + \frac{1}{2}{\mu_j}^2 {d_j}^\top A d_j)$$

Remember, $\alpha_j = \arg\min_\alpha f(x_j + \alpha d_j)$.Therefore,

$$f(x_j + \alpha_j d_j) \leq f(x_j + \mu_j d_j), \forall j \in \{0, \ldots, n-1\} \quad (\because \alpha_j \text{ is the minimizer})$$

Again, using Taylor series expansion of f around x_j,we get,

$$f(x_j) + \alpha_j {g_j}^\top d_j + \frac{1}{2}{\alpha_j}^2 {d_j}^T A d_j \leq f(x_j) + \mu_j {g_j}^\top d_j + \frac{1}{2}{\mu_j}^2 {d_j}^T A d_j$$

Convergence of Conjugate Descent

Convergence of Conjugate Descent

Claim

$g_j^\top d_j = g_0^\top d_j, \forall j$

Convergence of Conjugate Descent

Claim

${g_j}^\top d_j = {g_0}^\top d_j, \forall j$

Proof.

$x_j = x_0 + \sum_{i=0}^{j-1} \alpha_i d_i$

$\implies Ax_j - b = Ax_0 - b + \sum_{i=0}^{j-1} \alpha_i A d_i$

$\implies g_j = g_0 + \sum_{i=0}^{j-1} \alpha_i A d_i$

$\implies {g_j}^\top d_j = {g_0}^\top d_j + \sum_{i=0}^{j-1} \alpha_i {d_i}^\top A d_j$

$\therefore {g_j}^\top = {g_0}^\top d_j \, \forall j$ $\square$

Convergence of Conjugate Descent

Claim

${g_j}^\top d_j = {g_0}^\top d_j, \forall j$

Proof.

$x_j = x_0 + \sum_{i=0}^{j-1} \alpha_i d_i$
$\implies Ax_j - b = Ax_0 - b + \sum_{i=0}^{j-1} \alpha_i A d_i$
$\implies g_j = g_0 + \sum_{i=0}^{j-1} \alpha_i A d_i$
$\implies {g_j}^\top d_j = {g_0}^\top d_j + \sum_{i=0}^{j-1} \alpha_i {d_i}^\top A d_j$
$\therefore {g_j}^\top = {g_0}^\top d_j \, \forall j$ □

$\therefore \alpha_j {g_0}^\top d_j + \frac{1}{2}{\alpha_j}^2 {d_j}^T A d_j \leq \mu_j {g_0}^\top d_j + \frac{1}{2}{\mu_j}^2 {d_j}^T A d_j$

Convergence of Conjugate Descent

Claim

${g_j}^\top d_j = {g_0}^\top d_j, \forall j$

Proof.

$x_j = x_0 + \sum_{i=0}^{j-1} \alpha_i d_i$
$\implies Ax_j - b = Ax_0 - b + \sum_{i=0}^{j-1} \alpha_i A d_i$
$\implies g_j = g_0 + \sum_{i=0}^{j-1} \alpha_i A d_i$
$\implies {g_j}^\top d_j = {g_0}^\top d_j + \sum_{i=0}^{j-1} \alpha_i {d_i}^\top A d_j$
$\therefore {g_j}^\top = {g_0}^\top d_j \, \forall j$ □

$\therefore \alpha_j {g_0}^\top d_j + \frac{1}{2}{\alpha_j}^2 {d_j}^T A d_j \leq \mu_j {g_0}^\top d_j + \frac{1}{2}{\mu_j}^2 {d_j}^T A d_j$
Therefore, by summing over j we get,

$$f(x_0) + \sum_{j=0}^{k-1} (\alpha_j {g_0}^\top d_j + \frac{1}{2}{\alpha_j}^2 {d_j}^\top A d_j) \leq f(x_0) + \sum_{j=0}^{k-1} (\mu_j {g_0}^\top d_j + \frac{1}{2}{\mu_j}^2 {d_j}^\top A d_j)$$

Convergence of Conjugate Descent

Claim

$g_j{}^\top d_j = g_0{}^\top d_j, \forall j$

Proof.

$x_j = x_0 + \sum_{i=0}^{j-1} \alpha_i d_i$
$\implies Ax_j - b = Ax_0 - b + \sum_{i=0}^{j-1} \alpha_i A d_i$
$\implies g_j = g_0 + \sum_{i=0}^{j-1} \alpha_i A d_i$
$\implies g_j{}^\top d_j = g_0{}^\top d_j + \sum_{i=0}^{j-1} \alpha_i d_i{}^\top A d_j$
$\therefore g_j{}^\top = g_0{}^\top d_j \, \forall j$ $\square$

$\therefore \alpha_j g_0{}^\top d_j + \frac{1}{2}\alpha_j{}^2 d_j{}^T A d_j \leq \mu_j g_0{}^\top d_j + \frac{1}{2}\mu_j{}^2 d_j{}^T A d_j$
Therefore, by summing over j we get,

$$f(x_0) + \sum_{j=0}^{k-1}(\alpha_j g_0{}^\top d_j + \frac{1}{2}\alpha_j{}^2 d_j{}^\top A d_j) \leq f(x_0) + \sum_{j=0}^{k-1}(\mu_j g_0{}^\top d_j + \frac{1}{2}\mu_j{}^2 d_j{}^\top A d_j)$$

$$\implies f(x_0 + \textstyle\sum_{j=0}^{k-1} \alpha_j d_j) \leq f(x_0 + \sum_{j=0}^{k-1} \mu_j d_j), \quad \mu_j \in \mathbb{R}$$

Convergence of Conjugate Descent

Claim

$g_j{}^{\top} d_j = g_0{}^{\top} d_j, \forall j$

Proof.

$x_j = x_0 + \sum_{i=0}^{j-1} \alpha_i d_i$
$\implies Ax_j - b = Ax_0 - b + \sum_{i=0}^{j-1} \alpha_i A d_i$
$\implies g_j = g_0 + \sum_{i=0}^{j-1} \alpha_i A d_i$
$\implies g_j{}^{\top} d_j = g_0{}^{\top} d_j + \sum_{i=0}^{j-1} \alpha_i d_i{}^{\top} A d_j$
$\therefore\ g_j{}^{\top} = g_0{}^{\top} d_j \, \forall j$ $\square$

$\therefore\ \alpha_j g_0{}^{\top} d_j + \frac{1}{2}\alpha_j{}^2 d_j{}^T A d_j \leq \mu_j g_0{}^{\top} d_j + \frac{1}{2}\mu_j{}^2 d_j{}^T A d_j$
Therefore, by summing over j we get,

$$f(x_0) + \sum_{j=0}^{k-1}(\alpha_j g_0{}^{\top} d_j + \frac{1}{2}\alpha_j{}^2 d_j{}^{\top} A d_j) \leq f(x_0) + \sum_{j=0}^{k-1}(\mu_j g_0{}^{\top} d_j + \frac{1}{2}\mu_j{}^2 d_j{}^{\top} A d_j)$$

$$\implies f(x_0 + \textstyle\sum_{j=0}^{k-1} \alpha_j d_j) \leq f(x_0 + \sum_{j=0}^{k-1} \mu_j d_j), \quad \mu_j \in \mathbb{R}$$

$\therefore\ f(x_k) \leq f(x) \, \forall \, x \in x_0 + \mathbb{B}_k$

Convergence of Conjugate Descent

Claim

${g_j}^\top d_j = {g_0}^\top d_j, \forall j$

Proof.

$x_j = x_0 + \sum_{i=0}^{j-1} \alpha_i d_i$
$\implies Ax_j - b = Ax_0 - b + \sum_{i=0}^{j-1} \alpha_i A d_i$
$\implies g_j = g_0 + \sum_{i=0}^{j-1} \alpha_i A d_i$
$\implies {g_j}^\top d_j = {g_0}^\top d_j + \sum_{i=0}^{j-1} \alpha_i {d_i}^\top A d_j$
$\therefore\ {g_j}^\top = {g_0}^\top d_j \, \forall j$ □

$\therefore\ \alpha_j {g_0}^\top d_j + \frac{1}{2}{\alpha_j}^2 {d_j}^T A d_j \leq \mu_j {g_0}^\top d_j + \frac{1}{2}{\mu_j}^2 {d_j}^T A d_j$
Therefore, by summing over j we get,

$$f(x_0) + \sum_{j=0}^{k-1}(\alpha_j {g_0}^\top d_j + \frac{1}{2}{\alpha_j}^2 {d_j}^\top A d_j) \leq f(x_0) + \sum_{j=0}^{k-1}(\mu_j {g_0}^\top d_j + \frac{1}{2}{\mu_j}^2 {d_j}^\top A d_j)$$

$$\implies f(x_0 + \textstyle\sum_{j=0}^{k-1} \alpha_j d_j) \leq f(x_0 + \sum_{j=0}^{k-1} \mu_j d_j), \quad \mu_j \in \mathbb{R}$$

$\therefore\ f(x_k) \leq f(x) \, \forall\, x \in x_0 + \mathbb{B}_k$
So, at n^{th} iteration, $f(x_n) \leq f(x) \, \forall\, x \in x_0 + \mathcal{B}_n$.

Convergence of Conjugate Descent

Claim

${g_j}^\top d_j = {g_0}^\top d_j,\ \forall j$

Proof.

$x_j = x_0 + \sum_{i=0}^{j-1} \alpha_i d_i$

$\implies Ax_j - b = Ax_0 - b + \sum_{i=0}^{j-1} \alpha_i A d_i$

$\implies g_j = g_0 + \sum_{i=0}^{j-1} \alpha_i A d_i$

$\implies {g_j}^\top d_j = {g_0}^\top d_j + \sum_{i=0}^{j-1} \alpha_i {d_i}^\top A d_j$

$\therefore\ {g_j}^\top = {g_0}^\top d_j\ \forall j$ □

$\therefore\ \alpha_j {g_0}^\top d_j + \frac{1}{2}{\alpha_j}^2 {d_j}^T A d_j \le \mu_j {g_0}^\top d_j + \frac{1}{2}{\mu_j}^2 {d_j}^T A d_j$
Therefore, by summing over j we get,

$$f(x_0) + \sum_{j=0}^{k-1}(\alpha_j {g_0}^\top d_j + \frac{1}{2}{\alpha_j}^2 {d_j}^\top A d_j) \le f(x_0) + \sum_{j=0}^{k-1}(\mu_j {g_0}^\top d_j + \frac{1}{2}{\mu_j}^2 {d_j}^\top A d_j)$$

$$\implies f(x_0 + \textstyle\sum_{j=0}^{k-1} \alpha_j d_j) \le f(x_0 + \sum_{j=0}^{k-1} \mu_j d_j), \quad \mu_j \in \mathbb{R}$$

$\therefore\ f(x_k) \le f(x)\ \forall\, x \in x_0 + \mathbb{B}_k$
So, at n^{th} iteration, $f(x_n) \le f(x)\ \forall\, x \in x_0 + \mathcal{B}_n$. But, $x_0 + \mathcal{B}_n = \mathbb{R}^n$.

Convergence of Conjugate Descent

Claim

$g_j{}^\top d_j = g_0{}^\top d_j, \forall j$

Proof.

$x_j = x_0 + \sum_{i=0}^{j-1} \alpha_i d_i$

$\implies Ax_j - b = Ax_0 - b + \sum_{i=0}^{j-1} \alpha_i A d_i$

$\implies g_j = g_0 + \sum_{i=0}^{j-1} \alpha_i A d_i$

$\implies g_j{}^\top d_j = g_0{}^\top d_j + \sum_{i=0}^{j-1} \alpha_i d_i{}^\top A d_j$

$\therefore g_j{}^\top = g_0{}^\top d_j \forall j$ □

$\therefore \alpha_j g_0{}^\top d_j + \frac{1}{2}\alpha_j{}^2 d_j{}^T A d_j \leq \mu_j g_0{}^\top d_j + \frac{1}{2}\mu_j{}^2 d_j{}^T A d_j$

Therefore, by summing over j we get,

$$f(x_0) + \sum_{j=0}^{k-1}(\alpha_j g_0{}^\top d_j + \frac{1}{2}\alpha_j{}^2 d_j{}^\top A d_j) \leq f(x_0) + \sum_{j=0}^{k-1}(\mu_j g_0{}^\top d_j + \frac{1}{2}\mu_j{}^2 d_j{}^\top A d_j)$$

$$\implies f(x_0 + \textstyle\sum_{j=0}^{k-1} \alpha_j d_j) \leq f(x_0 + \sum_{j=0}^{k-1} \mu_j d_j), \quad \mu_j \in \mathbb{R}$$

$\therefore f(x_k) \leq f(x) \forall x \in x_0 + \mathbb{B}_k$

So, at n^{th} iteration, $f(x_n) \leq f(x) \forall x \in x_0 + \mathcal{B}_n$. But, $x_0 + \mathcal{B}_n = \mathbb{R}^n$.

Hence, $x_n = x^*$

Procedure to Obtain Conjugate Directions: Gram-Schmidt Procedure

Procedure to Obtain Conjugate Directions

Procedure to Obtain Conjugate Directions

We use Gram-Schmidt procedure to obtain conjugate directions $d_0, \ldots, d_{n-1}$.

Procedure to Obtain Conjugate Directions

We use Gram-Schmidt procedure to obtain conjugate directions $d_0, \ldots, d_{n-1}$. . To this end, we need to start with a linearly independent set.

Procedure to Obtain Conjugate Directions

We use Gram-Schmidt procedure to obtain conjugate directions $d_0, \ldots, d_{n-1}$. . To this end, we need to start with a linearly independent set.
Suppose $\{-g_0, \ldots, -g_{n-1}\}$ is a linearly independent set (We will show this shortly).

Procedure to Obtain Conjugate Directions

We use Gram-Schmidt procedure to obtain conjugate directions $d_0, \ldots, d_{n-1}$. . To this end, we need to start with a linearly independent set.
Suppose $\{-g_0, \ldots, -g_{n-1}\}$ is a linearly independent set (We will show this shortly).

- Let $d_0 = -g_0$

Procedure to Obtain Conjugate Directions

We use Gram-Schmidt procedure to obtain conjugate directions $d_0, \ldots, d_{n-1}$. . To this end, we need to start with a linearly independent set.

Suppose $\{-g_0, \ldots, -g_{n-1}\}$ is a linearly independent set (We will show this shortly).

- Let $d_0 = -g_0$
- $d_k = -g_k + \sum_{j=0}^{k-1} \beta_j d_j, \quad k = 1, \ldots, n-1$

Procedure to Obtain Conjugate Directions

We use Gram-Schmidt procedure to obtain conjugate directions $d_0, \ldots, d_{n-1}$. . To this end, we need to start with a linearly independent set.
Suppose $\{-g_0, \ldots, -g_{n-1}\}$ is a linearly independent set (We will show this shortly).

- Let $d_0 = -g_0$
- $d_k = -g_k + \sum_{j=0}^{k-1} \beta_j d_j, \quad k = 1, \ldots, n-1$

But, we want $d_0, \ldots, d_{n-1}$ to be $A - \textit{conjugate}$ vectors.

Procedure to Obtain Conjugate Directions

We use Gram-Schmidt procedure to obtain conjugate directions $d_0, \ldots, d_{n-1}$. . To this end, we need to start with a linearly independent set.
Suppose $\{-g_0, \ldots, -g_{n-1}\}$ is a linearly independent set (We will show this shortly).

- Let $d_0 = -g_0$
- $d_k = -g_k + \sum_{j=0}^{k-1} \beta_j d_j, \quad k = 1, \ldots, n-1$

But, we want $d_0, \ldots, d_{n-1}$ to be $A - \mathit{conjugate}$ vectors. Therefore,

$$d_i{}^\top A d_k = -d_i{}^\top A g_k + \sum_{j=0}^{k-1} \beta_i d_i{}^\top A d_j, \quad i = 0, \ldots, k-1$$

$$\therefore\ 0 = -d_i{}^\top A g_k + \beta_i d_i{}^\top A d_i, \quad i = 0, \ldots, k-1$$

Procedure to Obtain Conjugate Directions

We use Gram-Schmidt procedure to obtain conjugate directions $d_0, \dots, d_{n-1}$. . To this end, we need to start with a linearly independent set.
Suppose $\{-g_0, \dots, -g_{n-1}\}$ is a linearly independent set (We will show this shortly).

- Let $d_0 = -g_0$
- $d_k = -g_k + \sum_{j=0}^{k-1} \beta_j d_j, \quad k = 1, \dots, n-1$

But, we want $d_0, \dots, d_{n-1}$ to be $A - \textit{conjugate}$ vectors. Therefore,

$$d_i{}^\top A d_k = -d_i{}^\top A g_k + \sum_{j=0}^{k-1} \beta_i d_i{}^\top A d_j, \quad i = 0, \dots, k-1$$

$$\therefore\ 0 = -d_i{}^\top A g_k + \beta_i d_i{}^\top A d_i, \quad i = 0, \dots, k-1$$

$$\implies \beta_i = \frac{g_k{}^\top A d_i}{d_i{}^\top A d_i},\ \therefore\ d_k = -g_k + \textstyle\sum_{j=0}^{k-1} \left(\dfrac{g_k{}^\top A d_j}{d_j{}^T A d_j} \right) d_j$$

Procedure to Obtain Conjugate Directions

Procedure to Obtain Conjugate Directions

Now, we show that $\{-g_0, \ldots, -g_{n-1}\}$ is a linearly independent set.

Procedure to Obtain Conjugate Directions

Now, we show that $\{-g_0, \ldots, -g_{n-1}\}$ is a linearly independent set.
Note that $\text{span}\{d_0, \ldots, d_{k-1}\} = \text{span}\{-g_0, \ldots, -g_{k-1}\}$.

Procedure to Obtain Conjugate Directions

Now, we show that $\{-g_0, \ldots, -g_{n-1}\}$ is a linearly independent set.
Note that $\text{span}\{d_0, \ldots, d_{k-1}\} = \text{span}\{-g_0, \ldots, -g_{k-1}\}$.
We already established that $g_k \perp \mathcal{B}_k$.

Procedure to Obtain Conjugate Directions

Now, we show that $\{-g_0, \dots, -g_{n-1}\}$ is a linearly independent set.
Note that $\text{span}\{d_0, \dots, d_{k-1}\} = \text{span}\{-g_0, \dots, -g_{k-1}\}$.
We already established that $g_k \perp \mathcal{B}_k$.
$\implies -g_k \perp \text{span}\{d_0, \dots, d_{k-1}\}$.

Procedure to Obtain Conjugate Directions

Now, we show that $\{-g_0, \ldots, -g_{n-1}\}$ is a linearly independent set.
Note that $\text{span}\{d_0, \ldots, d_{k-1}\} = \text{span}\{-g_0, \ldots, -g_{k-1}\}$.
We already established that $g_k \perp \mathcal{B}_k$.
$\implies -g_k \perp \text{span}\{d_0, \ldots, d_{k-1}\}$.
$\therefore -g_k \perp \text{span}\{-g_0, \ldots, -g_{k-1}\}$

Procedure to Obtain Conjugate Directions

Now, we show that $\{-g_0, \ldots, -g_{n-1}\}$ is a linearly independent set.
Note that span$\{d_0, \ldots, d_{k-1}\}$ = span$\{-g_0, \ldots, -g_{k-1}\}$.
We already established that $g_k \perp \mathcal{B}_k$.
$\implies -g_k \perp$ span$\{d_0, \ldots, d_{k-1}\}$.
$\therefore -g_k \perp$ span$\{-g_0, \ldots, -g_{k-1}\}$
$\therefore \{-g_0, \ldots, -g_{n-1}\}$ is a linearly independent set.

Procedure to Obtain Conjugate Directions

Procedure to Obtain Conjugate Directions

Thus, we have

$$d_0 = -g_0$$

$$d_k = -g_k + \sum_{j=0}^{k-1} \left(\frac{{g_k}^\top A d_j}{{d_j}^T A d_j} \right) d_j \quad \forall\, k = 1, \ldots, n-1$$

Procedure to Obtain Conjugate Directions

Thus, we have

$$d_0 = -g_0$$

$$d_k = -g_k + \sum_{j=0}^{k-1} \left(\frac{{g_k}^\top A d_j}{{d_j}^T A d_j} \right) d_j \quad \forall\, k = 1, \ldots, n-1$$

But, the update still depends on A and we need to get rid of that.

Procedure to Obtain Conjugate Directions

Thus, we have

$$d_0 = -g_0$$

$$d_k = -g_k + \sum_{j=0}^{k-1} \left(\frac{{g_k}^\top A d_j}{{d_j}^T A d_j} \right) d_j \quad \forall\, k = 1, \ldots, n-1$$

But, the update still depends on A and we need to get rid of that. To this end, note that $x_{j+1} = x_j + \alpha_j d_j \implies g_{j+1} = g_j + \alpha_j A d_j$.

Procedure to Obtain Conjugate Directions

Thus, we have

$$d_0 = -g_0$$
$$d_k = -g_k + \sum_{j=0}^{k-1} \left(\frac{{g_k}^\top A d_j}{{d_j}^T A d_j} \right) d_j \quad \forall\, k = 1, \ldots, n-1$$

But, the update still depends on A and we need to get rid of that. To this end, note that $x_{j+1} = x_j + \alpha_j d_j \implies g_{j+1} = g_j + \alpha_j A d_j$. Therefore,

$$A d_j = \frac{1}{\alpha_j}(g_{j+1} - g_j)$$

Procedure to Obtain Conjugate Directions

Thus, we have

$$d_0 = -g_0$$

$$d_k = -g_k + \sum_{j=0}^{k-1} \left(\frac{{g_k}^\top A d_j}{{d_j}^T A d_j} \right) d_j \quad \forall\, k = 1, \ldots, n-1$$

But, the update still depends on A and we need to get rid of that. To this end, note that $x_{j+1} = x_j + \alpha_j d_j \implies g_{j+1} = g_j + \alpha_j A d_j$. Therefore,

$$A d_j = \frac{1}{\alpha_j}(g_{j+1} - g_j)$$

Hence,

$$d_k = -g_k + \sum_{j=0}^{k-1} \left(\frac{{g_k}^T (g_{j+1} - g_j)}{{d_j}^T (g_{j+1} - g_j)} \right) d_j$$

Procedure to Obtain Conjugate Directions

Thus, we have

$$d_0 = -g_0$$

$$d_k = -g_k + \sum_{j=0}^{k-1} \left(\frac{{g_k}^\top A d_j}{{d_j}^T A d_j} \right) d_j \quad \forall\, k = 1, \ldots, n-1$$

But, the update still depends on A and we need to get rid of that. To this end, note that $x_{j+1} = x_j + \alpha_j d_j \implies g_{j+1} = g_j + \alpha_j A d_j$. Therefore,

$$A d_j = \frac{1}{\alpha_j}(g_{j+1} - g_j)$$

Hence,

$$d_k = -g_k + \sum_{j=0}^{k-1} \left(\frac{{g_k}^T (g_{j+1} - g_j)}{{d_j}^T (g_{j+1} - g_j)} \right) d_j$$

$$d_k = -g_k + \left(\frac{{g_k}^T g_k}{{d_{k-1}}^T (g_k - g_{k-1})} \right) d_{k-1}$$

Procedure to Obtain Conjugate Directions

Procedure to Obtain Conjugate Directions

Due to exact line search, ${g_k}^T d_{k-1} = 0$.

Procedure to Obtain Conjugate Directions

Due to exact line search, ${g_k}^T d_{k-1} = 0$.Note that,

$$d_{k-1} = -g_{k-1} + \beta_{k-2} d_{k-2}$$

Procedure to Obtain Conjugate Directions

Due to exact line search, ${g_k}^T d_{k-1} = 0$.Note that,

$$d_{k-1} = -g_{k-1} + \beta_{k-2} d_{k-2}$$

$\implies$

$$-{d_{k-1}}^T g_{k-1} = {g_{k-1}}^T g_{k-1} + \beta_{k-2} {g_{k-1}}^T d_{k-2}$$

Procedure to Obtain Conjugate Directions

Due to exact line search, ${g_k}^T d_{k-1} = 0$.Note that,

$$d_{k-1} = -g_{k-1} + \beta_{k-2} d_{k-2}$$

$\implies$

$$-{d_{k-1}}^T g_{k-1} = {g_{k-1}}^T g_{k-1} + \beta_{k-2} {g_{k-1}}^T d_{k-2}$$

Therefore,

$$d_k = -g_k + \left(\frac{{g_k}^T g_k}{{g_{k-1}}^T g_{k-1}} \right) d_{k-1}, \quad k = 1, \ldots, n-1$$

The above update is called Fletcher-Reeves update

Conclusion

Conclusion

Conclusion

- Solution to convex quadratic problem $\iff$ solution to system of linear equations

Conclusion

- Solution to convex quadratic problem $\iff$ solution to system of linear equations
- The curse of inverse computation can be avoided if the Hessian matrix is positive definite

Conclusion

- Solution to convex quadratic problem $\iff$ solution to system of linear equations
- The curse of inverse computation can be avoided if the Hessian matrix is positive definite
- Conjugate gradient (descent) method finds the optimal solution in at most n iterations

References

[1] David G Luenberger, Yinyu Ye, et al. *Linear and nonlinear programming*. Vol. 2. Springer, 1984.

[2] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. 2e. New York, NY, USA: Springer, 2006.