Optimization Methods (CS1.404), Spring 2024
Lecture 14

Naresh Manwani

Machine Learning Lab, IIIT-H

March 4th, 2024

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Need for Cholesky Factorization

- In Levenberg Marquardt algorithm, it is required to validate the positive definiteness of the matrix $\nabla^2 f(\mathbf{x}_k) + \mu_k \mathbf{I}$.
- Another issue is to solve the equation $\nabla^2 f(\mathbf{x}_k)\mathbf{d} = -\nabla f(\mathbf{x}_k)$ in general for Newton method.
- These two issues are resolved using Cholesky factorization.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Solving $\mathbf{Ax} = \mathbf{b}$ using Cholesky Factorization

- Let $\mathbf{A}$ be $n \times n$ positive definite matrix. Cholesky factorization of $\mathbf{A}$ has the form $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower triangular $n \times n$ matrix whose diagonal is positive
- Given the Cholesky factorization, equation $\mathbf{Ax} = \mathbf{b}$ can be solved in following two steps.
  - Find the solution $\mathbf{u}$ of $\mathbf{Lu} = \mathbf{b}$
  - Find the solution $\mathbf{x}$ of $\mathbf{L}^T\mathbf{x} = \mathbf{u}$.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Cholesky Factorization Algorithm

- The computation of Cholesky factorization is done using a simple recursive approach.
- Consider the following block matrix partitioning of the matrices A and L.

$$\mathbf{A} = \begin{pmatrix} A_{11} & \mathbf{A}_{21} \\ \mathbf{A}_{12} & \mathbf{A}_{22} \end{pmatrix} \qquad \mathbf{L} = \begin{pmatrix} L_{11} & \mathbf{0} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{pmatrix}$$

  where $A_{11} \in \mathbb{R}, \mathbf{A}_{21} \in \mathbb{R}^{(n-1)\times 1}, \mathbf{A}_{22} \in \mathbb{R}^{(n-1)\times(n-1)}, L_{11} \in \mathbb{R}$, $\mathbf{L}_{21} \in \mathbb{R}^{(n-1)\times 1}, \mathbf{L}_{22} \in \mathbb{R}^{(n-1)\times(n-1)}$.
- Since $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, we have

$$\begin{pmatrix} A_{11} & \mathbf{A}_{21} \\ \mathbf{A}_{12} & \mathbf{A}_{22} \end{pmatrix} = \begin{pmatrix} L_{11}^2 & L_{11}\mathbf{L}_{21}^T \\ L_{11}\mathbf{L}_{21} & \mathbf{L}_{21}\mathbf{L}_{21}^T + \mathbf{L}_{22}\mathbf{L}_{22}^T \end{pmatrix}$$

Cholesky Factorization Algorithm: Continue

- Therefore, in particular $L_{11} = \sqrt{A_{11}}$, $\mathbf{L}_{21} = \frac{1}{\sqrt{A_{11}}}\mathbf{A}_{12}^T$.
- $\mathbf{L}_{22}\mathbf{L}_{22}^T = \mathbf{A}_{22} - \mathbf{L}_{21}\mathbf{L}_{21}^T = \mathbf{A}_{22} - \frac{1}{A_{11}}\mathbf{A}_{12}^T\mathbf{A}_{12}$.
- We are left with the task of Cholesky factorization of $(n-1) \times (n-1)$ matrix $\mathbf{A}_{22} - \frac{1}{A_{11}}\mathbf{A}_{12}^T\mathbf{A}_{12}$.
- We keep following the above procedure and we can get the complete Cholesky factorization.
- The algorithm for Cholesky factorization will find a solution only if all the diagonal elements l_{ii} that are computed during the process are positive, so that computing their square root is possible.
- The positiveness of these elements is equivalent to the property that the matrix to be factored is positive definite.
- Therefore, the Cholesky factorization process can be viewed as a criteria for positive definiteness.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Coordinate Descent Method

- Consider the problem $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function.
- Coordinate descent method works as follows.
  1. For every coordinate variable x_i, $i \in \{1, \ldots, n\}$, minimize $f(\mathbf{x})$ with respect to x_i, keeping other variables x_j, $j \neq i$ constant.
  2. Repeat the above process in step 1 until some stopping condition is satisfied.

Coordinate Descent Method

Algorithm

- **Input:** $\epsilon > 0$ (tolerance parameter)
- **Initialize:** $\mathbf{x}_1$, $k = 1$
- **Step 1:** Set $\mathbf{d}_k = \mathbf{e}_k$, where $\mathbf{e}_k$ is k^{th} basis vector of standard basis of $\mathbb{R}^n$
- **Step 2:** Set $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$, where $\alpha_k = \arg\min_{\alpha \in \mathbb{R}} f(\mathbf{x}_k + \alpha \mathbf{d}_k)$.
- **Step 3:**
  - If $(\|\nabla f(\mathbf{x}_k)\| \leq \epsilon)$
    - then output $\mathbf{x}^* = \mathbf{x}_k$
  - Else if $(k = n)$
    - Set $\mathbf{x}_1 = \mathbf{x}_{k+1}$ and repeat from Step 2.
  - Else,
    - Set $k = k + 1$ and repeat from Step 2.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Coordinate Descent Method on Quadratic Functions

- **For convex quadratic functions of n variables, above algorithm converges in n-steps.**
  - Example 1: $\min_{\mathbf{x}\in\mathbb{R}^2}\ 4x_1^2 + x_2^2$ (spherical contours). Take $\mathbf{x}_0 = (-1, -1)$. Coordinate descent method finds minimizer in two steps.
  - Example 2: $\min_{\mathbf{x}\in\mathbb{R}^2}\ 4x_1^2 + x_2^2 - 2x_1x_2$ (elliptical contours) . Take $\mathbf{x}_0 = (-1, -1)$. Coordinate descent method does not converge in two steps.
- In other words, when the objective function is separable in terms of variables (hessian is diagonal), then coordinate descent method will find $\mathbf{x}^*$ in n-steps if there are n-variables.
- When objective function is not separable in variables, then Hessian is not diagonal. Coordinate descent method will not find minimizer in n-steps.
- Can we choose $\mathbf{d}_1, \ldots, \mathbf{d}_n$ in such a way that it converges in n-steps?

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Conjugate Directions

Definition

Let Q be a real symmetric $n \times n$ matrix. The directions $\mathbf{d}_0, \ldots, \mathbf{d}_{n-1}$ are Q-conjugate if, for all $i \neq j$, we have $\mathbf{d}_i^T Q \mathbf{d}_j = 0$.

Lemma

Let Q be a symmetric positive definite $n \times n$ matrix. Let directions $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k \in \mathbb{R}^n$ $(k \leq n-1)$ are Q-conjugate, then they are linearly independent.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Conjugate Directions: Example 1

- Let $f(x_1, x_2) = 4x_1^2 + x_2^2 - 2x_1x_2$
- Hessian $H = \begin{pmatrix} 8 & -2 \\ -2 & 2 \end{pmatrix}$
- Let $\mathbf{d}_0 = (1, 0)^T$
- Then the conjugate direction $\mathbf{d}_1 = (a, b)^T$ would satisfy $\mathbf{d}_0^T H \mathbf{d}_1 = 0$.
- This results in relation $8a - 2b = 0$. Thus, we can take $\mathbf{d}_1 = (1, 4)^T$.

Conjugate Directions: Example 2

- Let $Q = \begin{pmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}$
- Let $\mathbf{d}_0 = (1, 0, 0)^T$.
- Now, we want to find $\mathbf{d}_1 = (a, b, c)^T$ which is Q-conjugate to $\mathbf{d}_0$. We require $\mathbf{d}_0^T H \mathbf{d}_1 = 0$. Which results in relation $3a + c = 0$. So, we can choose $\mathbf{d}_1 = (1, 0, -3)^T$.
- Now, we want to find $\mathbf{d}_2 = (e, f, g)^T$ which is Q-conjugate to $\mathbf{d}_0$ and $\mathbf{d}_1$. So, we get the conditions $3e + g = 0$ and $-6f - 8g = 0$. We can choose $\mathbf{d}_2 = (1, 4, -3)^T$.

Choosing Conjugate Directions

- A systematic procedure for finding Q-conjugate directions can be developed using the idea of Gram-Schmidt algorithm of transforming a given basis of $\mathbb{R}^n$ into an orthogonal basis of $\mathbb{R}^n$.
- For a symmetric matrix matrix H, orthogonal eigenvectors of H itself are H-conjugate.
  - Let $\mathbf{v}_1$ and $\mathbf{v}_2$ are mutually orthonormal eigenvectors of H corresponding to eigenvalues λ_1 and λ_2.
  - Then $\mathbf{v}_1^T H \mathbf{v}_2 = \lambda_2 \mathbf{v}_1^T \mathbf{v}_2 = 0$.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Conjugate Direction Method

- Consider minimization problem $\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\mathbf{x}^T H\mathbf{x} + \mathbf{c}^T\mathbf{x}$, where H is symmetric positive definite matrix.
- Let $\mathbf{x}_0$ be the initial parameters.
- Let $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1} \in \mathbb{R}^n$ be H-conjugate directions.
- As we know that these conjugate directions are linearly independent. We can write any $\mathbf{x} - \mathbf{x}_0 \in \mathbb{R}^n$ as a linear combination of these conjugate directions. Thus,

$$\mathbf{x} - \mathbf{x}_0 = \sum_{i=0}^{n-1} \alpha_i \mathbf{d}_i$$

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Conjugate Direction Method - Continue

- Given $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1} \in \mathbb{R}^n$ and $\mathbf{x}_0 \in \mathbb{R}^n$, the above minimization problem becomes

$$\begin{aligned}
\phi(\boldsymbol{\alpha}) &= \frac{1}{2}\left(\mathbf{x}_0 + \sum_{i=0}^{n-1}\alpha_i\mathbf{d}_i\right)^T H\left(\mathbf{x}_0 + \sum_{i=0}^{n-1}\alpha_i\mathbf{d}_i\right) + \mathbf{c}^T\left(\mathbf{x}_0 + \sum_{i=0}^{n-1}\alpha_i\mathbf{d}_i\right) \\
&= \frac{1}{2}\mathbf{x}_0^T H\mathbf{x}_0 + \frac{1}{2}\left(\sum_{i=0}^{n-1}\alpha_i\mathbf{d}_i\right)^T H\left(\sum_{i=0}^{n-1}\alpha_i\mathbf{d}_i\right) + \left(\sum_{i=0}^{n-1}\alpha_i\mathbf{d}_i\right) H\mathbf{x}_0 \\
&\quad + \mathbf{c}^T\left(\mathbf{x}_0 + \sum_{i=0}^{n-1}\alpha_i\mathbf{d}_i\right) \\
&= \frac{1}{2}\mathbf{x}_0^T H\mathbf{x}_0 + \frac{1}{2}\sum_{i=0}^{n-1}\alpha_i^2\mathbf{d}_i^T H\mathbf{d}_i + \sum_{i=0}^{n-1}\alpha_i\mathbf{d}_i^T H\mathbf{x}_0 + \mathbf{c}^T\left(\mathbf{x}_0 + \sum_{i=0}^{n-1}\alpha_i\mathbf{d}_i\right) \\
&= \sum_{i=0}^{n-1}\left(\frac{1}{2}(\mathbf{x}_0 + \alpha_i\mathbf{d}_i)^T H(\mathbf{x}_0 + \alpha_i\mathbf{d}_i) + \mathbf{c}^T(\mathbf{x}_0 + \alpha_i\mathbf{d}_i)\right) \\
&\quad - (n-1)\left(\frac{1}{2}\mathbf{x}_0^T H\mathbf{x}_0 + \mathbf{c}^T\mathbf{x}_0\right)
\end{aligned}$$

Conjugate Direction Method - Continue

- Ignoring the constant term, we define a new function

$$\psi(\boldsymbol{\alpha}) = \sum_{i=0}^{n-1} \left(\frac{1}{2}(\mathbf{x}_0 + \alpha_i \mathbf{d}_i)^T H(\mathbf{x}_0 + \alpha_i \mathbf{d}_i) + \mathbf{c}^T(\mathbf{x}_0 + \alpha_i \mathbf{d}_i) \right)$$

- ψ is separable in terms of $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$, which are our new optimization variables.
- Minimizing ψ with respect to α_i, we get

$$\alpha_i^* = -\frac{\mathbf{d}_i^T(H\mathbf{x}_0 + \mathbf{c})}{\mathbf{d}_i^T H \mathbf{d}_i}$$

- $\mathbf{x}^* = \mathbf{x}_0 + \sum_{i=0}^{n-1} \alpha_i^* \mathbf{d}_i$

Basic Conjugate Direction Algorithm

Given starting point $\mathbf{x}_0$ and H conjugate directions $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1} \in \mathbb{R}^n$, the Conjugate Direction Algorithm works as follows:

- For $(k = 0, 1, \ldots, n-1)$
  - $\nabla f(\mathbf{x}_k) = H\mathbf{x}_k + \mathbf{c}$
  - $\alpha_k = -\frac{\nabla f(\mathbf{x}_k)^T \mathbf{d}_k}{\mathbf{d}_k^T H \mathbf{d}_k}$
  - $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$

Theorem

Consider minimization problem $\min_{\mathbf{x} \in \mathbb{R}^n} \ \frac{1}{2}\mathbf{x}^T H \mathbf{x} + \mathbf{c}^T \mathbf{x}$, where H is symmetric positive definite matrix. For any starting point $\mathbf{x}_0$ and H conjugate directions $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1} \in \mathbb{R}^n$, the **Basic Conjugate Direction Algorithm** converges to the unique $\mathbf{x}^*$ (that solves $H\mathbf{x}^* + \mathbf{c} = \mathbf{0}$) in n-steps; that is $\mathbf{x}_n = \mathbf{x}^*$.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD