Optimization Methods (CS1.404), Spring 2024
Lecture 17

Naresh Manwani

Machine Learning Lab, IIIT-H

March 14th, 2024

Quasi-Newton Method (Rank One Correction)

1: **Initialize:** The starting point $\mathbf{x}_0$, Symmetric positive definite matrix B_0 and the tolerance parameter $\epsilon > 0$, Set $k = 0$

2: **while** $\|\mathbf{g}_k\| > \epsilon$ **do**

3: $\quad \mathbf{d}_k = -B_k \mathbf{g}_k$

4: $\quad$ Find α_k along $\mathbf{d}_k$ such that

- $f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) < f(\mathbf{x}_k)$
- α_k satisfies Armijo-Wolfe condition

5: $\quad \mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$

6: $\quad$ Find B_{k+1} as

$$B_{k+1} = B_k + \frac{(\boldsymbol{\delta}_k - B_k \boldsymbol{\gamma}_k)(\boldsymbol{\delta}_k - B_k \boldsymbol{\gamma}_k)^T)}{(\boldsymbol{\delta}_k - B_k \boldsymbol{\gamma}_k)^T \boldsymbol{\gamma}_k}$$

7: $\quad k = k + 1$

8: **end while**

9: **Output:** $\mathbf{x}^* = \mathbf{x}_k$, a stationary point of f.

Convergence of SR1 Applied on Quadratic Functions

Theorem1: For Quadratic Functions

Consider SR1 quasi-Newton algorithm applied to a quadratic function with positive definite Hessian H. Then, for any starting point $\mathbf{x}_0$ and any symmetric starting matrix B_0, the sequence of iterates $\mathbf{x}_k$ generated by SR1 converges to the minimizer in n-steps, provided $(\boldsymbol{\delta}_k - B_k\boldsymbol{\gamma}_k)^T\boldsymbol{\gamma}_k \neq 0, \forall k$. Moreover, if n-steps are performed and $\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{n-1}$ are linearly independent, then $B_n = H^{-1}$.

Convergence of SR1 Applied on General Functions

- For general nonlinear functions, the SR1 update continues to generate good Hessian approximations under certain conditions.

Theorem

Suppose that f is twice continuously differentiable, and that its Hessian is bounded and Lipschitz continuous in a neighborhood of a point x^*. Let $\mathbf{x}_k, \ k = 0, 1, \ldots$ be any sequence of iterates such that $\mathbf{x}_k \to \mathbf{x}^*$ for some $\mathbf{x}^* \in \mathbb{R}^n$. Suppose in addition that the $|(\boldsymbol{\delta}_k - B_k\boldsymbol{\gamma}_k)^T\boldsymbol{\gamma}_k| \geq r\|\boldsymbol{\delta}_k - B_k\boldsymbol{\gamma}_k\|.\|\boldsymbol{\gamma}_k\|$ holds for all k, for some $r \in (0, 1)$, and that $\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{n-1}$ are linearly independent. Then, the matrices B_k generated by the SR1 updating formula satisfy

$$\lim_{k\to\infty} \|B_k - \nabla^2 f(\mathbf{x}^*)\| = 0$$

SR1: Some Remarks

- SR1 is a simple and elegant way to use the information gathered during two consecutive iterations to update B_k.
- B_{k+1} is positive definite if $(\boldsymbol{\delta}_k - B_k\boldsymbol{\gamma}_k)^T\boldsymbol{\gamma}_k > 0$ which can not be guaranteed for every k.
- Numerical difficulties happen if $(\boldsymbol{\delta}_k - B_k\boldsymbol{\gamma}_k)^T\boldsymbol{\gamma}_k$ is close to 0.

Rank Two Correction Quasi newton Method

- Given that B_k is symmetric and positive definite matrix, let

$$B_{k+1} = B_k + \alpha \mathbf{u}\mathbf{u}^T + \beta \mathbf{v}\mathbf{v}^T$$

- B_{k+1} is required to satisfy Quasi-Newton condition. Thus,

$$\alpha \mathbf{u}^T \gamma_k \mathbf{u} + \beta \mathbf{v}^T \gamma_k \mathbf{v} = \boldsymbol{\delta}_k - B_k \gamma_k$$

- Letting $\alpha \mathbf{u}^T \gamma_k = \beta \mathbf{v}^T \gamma_k = 1$, we get $\mathbf{u} + \mathbf{v} = \boldsymbol{\delta}_k - B_k \gamma_k$. Taking $\mathbf{u} = \boldsymbol{\delta}_k$ and $\mathbf{v} = -B_k \gamma_k$, we get

$$\alpha^{-1} = \mathbf{u}^T \gamma_k = \boldsymbol{\delta}_k^T \gamma_k$$
$$\beta^{-1} = \mathbf{v}^T \gamma_k = -\gamma_k^T B_k \gamma_k$$

- Therefore, we get the following update for B_{k+1}:

$$B_{k+1} = B_k + \frac{\boldsymbol{\delta}_k^T \boldsymbol{\delta}_k}{\boldsymbol{\delta}_k^T \gamma_k} - \frac{B_k \gamma_k \gamma_k^T B_k}{\gamma_k^T B_k \gamma_k}$$

- This update is called DFP named after Davidson, Fletcher and Powell.

Rank Two Correction Quasi newton Method

Theorem

Given that B_k is symmetric and positive definite, B_{k+1} generated by DFP is symmetric and positive definite.

DFP Quasi-Newton Algorithm

1: **Initialize:** The starting point $\mathbf{x}_0$, Symmetric positive definite matrix B_0 and the tolerance parameter $\epsilon > 0$, Set $k = 0$

2: **while** $\|\mathbf{g}_k\| > \epsilon$ **do**

3: $\mathbf{d}_k = -B_k\mathbf{g}_k$

4: Find α_k along $\mathbf{d}_k$ such that

- $f(\mathbf{x}_k + \alpha_k\mathbf{d}_k) < f(\mathbf{x}_k)$
- α_k satisfies Armijo-Wolfe condition

5: $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$

6: Find B_{k+1} as

$$B_{k+1} = B_k + \frac{\boldsymbol{\delta}_k^T\boldsymbol{\delta}_k}{\boldsymbol{\delta}_k^T\boldsymbol{\gamma}_k} - \frac{B_k\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^T B_k}{\boldsymbol{\gamma}_k^T B_k\boldsymbol{\gamma}_k}$$

7: $k = k + 1$

8: **end while**

9: **Output:** $\mathbf{x}^* = \mathbf{x}_k$, a stationary point of f.

BFGS Algorithm

- Quasi Newton condition requires $B_{k+1}\boldsymbol{\gamma}_k = \boldsymbol{\delta}_k$ to hold for all k.
- Assume that we want to approximate the Hessian H_{k+1} rather than its inverse. Let $G_{k+1} = B_{k+1}^{-1}$ which approximates Hessian H_{k+1}.
- Then, Quasi-Newton condition would result into $G_{k+1}\boldsymbol{\delta}_k = \boldsymbol{\gamma}_k$.
- Rank two update of G_{k+1} will have the form

$$G_{k+1} = G_k + \alpha \mathbf{u}\mathbf{u}^T + \beta \mathbf{v}\mathbf{v}^T$$

- G_{k+1} is required to satisfy Quasi-Newton condition. Thus,

$$\alpha \mathbf{u}^T \boldsymbol{\delta}_k \mathbf{u} + \beta \mathbf{v}^T \boldsymbol{\delta}_k \mathbf{v} = \boldsymbol{\gamma}_k - G_k \boldsymbol{\delta}_k$$

- Letting $\alpha \mathbf{u}^T \boldsymbol{\delta}_k = \beta \mathbf{v}^T \boldsymbol{\delta}_k = 1$, we get $\mathbf{u} + \mathbf{v} = \boldsymbol{\gamma}_k - G_k \boldsymbol{\delta}_k$. Taking $\mathbf{u} = \boldsymbol{\gamma}_k$ and $\mathbf{v} = -G_k \boldsymbol{\delta}_k$, we get

$$\alpha^{-1} = \mathbf{u}^T \boldsymbol{\delta}_k = \boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k$$
$$\beta^{-1} = \mathbf{v}^T \boldsymbol{\delta}_k = -\boldsymbol{\delta}_k^T G_k \boldsymbol{\delta}_k$$

- Therefore, we get the following update for B_{k+1}:

$$G_{k+1} = G_k + \frac{\boldsymbol{\gamma}_k^T \boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k} - \frac{G_k \boldsymbol{\delta}_k \boldsymbol{\delta}_k^T G_k}{\boldsymbol{\delta}_k^T G_k \boldsymbol{\delta}_k}$$

- Next step is to find B_{k+1} as G_{k+1}^{-1}.
- We use Sherman-Morrison Formula to find G_{k+1}^{-1}.
  $(A + \mathbf{u}\mathbf{v}^T)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^T A^{-1}}{1+\mathbf{v}^T A^{-1}\mathbf{u}}$
- Applying this formula twice to G_{k+1}, we get

$$
\begin{aligned}
B_{k+1}^{BFGS} = B_k^{BFGS} &+ \left(1 + \frac{\gamma_k^T B_k^{BFGS} \gamma_k}{\boldsymbol{\delta}_k^T \gamma_k}\right) \frac{\boldsymbol{\delta}_k \boldsymbol{\delta}_k^T}{\boldsymbol{\delta}_k^T \gamma_k} \\
&- \left(\frac{\boldsymbol{\delta}_k \gamma_k^T B_k^{BFGS} + B_k^{BFGS} \gamma_k \boldsymbol{\delta}_k^T}{\boldsymbol{\delta}_k^T \gamma_k}\right)
\end{aligned}
$$

Broyden Family

- $B_{k+1}(\phi) = \phi B_{k+1}^{BFGS} + (1-\phi) B_{k+1}^{DFP}$, where $\phi \in [0, 1]$

Broyden Familty Quasi-Newton Algorithm

1: **Initialize:** The starting point $\mathbf{x}_0$, Symmetric positive definite matrix B_0 and the tolerance parameter $\epsilon > 0$, Set $k = 0$
2: **while** $\|\mathbf{g}_k\| > \epsilon$ **do**
3: $\quad \mathbf{d}_k = -B_k \mathbf{g}_k$
4: $\quad$ Find α_k along $\mathbf{d}_k$ such that
- $f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) < f(\mathbf{x}_k)$
- α_k satisfies Armijo-Wolfe condition

5: $\quad \mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$
6: $\quad$ Find B_{k+1} as

$$B_{k+1}(\phi) = \phi B_{k+1}^{BFGS} + (1-\phi) B_{k+1}^{DFP}$$

$\quad$ where $\phi \in [0, 1]$.
7: $\quad k = k + 1$
8: **end while**
9: **Output:** $\mathbf{x}^* = \mathbf{x}_k$, a stationary point of f.