Optimization Theory and Algorithm **Lecture 15 - 06/15/2021**

Lecture 15

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

1 Quasi-Newton Method

To overcome the deficiency of computing Newton Equation for large-scale problems and the non-decreasing property of $\{f(\mathbf{x}^t)\}$ in NR-algorithms, Quasi-Newton method is proposed.

1.1 General Quasi-Newton Method

Consider the Taylor expansion of $\nabla f(\mathbf{x}^t)$ at $\mathbf{x}^{t+1}$, then

$$\nabla f(\mathbf{x}^t) = \nabla f(\mathbf{x}^{t=1}) + \nabla^2 f(\mathbf{x}^{t+1})(\mathbf{x}^t - \mathbf{x}^{t+1})(\mathbf{x}^t - \mathbf{x}^{t+1}) + o(\|\mathbf{x}^t - \mathbf{x}^{t+1}\|).$$

Let $\mathbf{y}^t = \nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^t)$ and $\mathbf{d}^t = \mathbf{x}^{t+1} - \mathbf{x}^t$, then it should be

$$\nabla^2 f(\mathbf{x}^{t+1})\mathbf{d}^t + o(\|\mathbf{d}^t\|) = \mathbf{y}^t. \tag{1}$$

The Quasi-Newton method uses the approximated Hessian matrix B^{t+1} that satisfies

$$\mathbf{y}^t = B^{t+1}\mathbf{d}^t$$

to replace $\nabla^2 f(\mathbf{x}^{t+1})$ in the NR algorithm. Alternatively, we construct an approximated inverse Hessian matrix H^{t+1} that satisfies

$$H^{t+1}\mathbf{y}^t = \mathbf{d}^t$$

to replace $(\nabla^2 f(\mathbf{x}^{t+1}))^{-1}$ in the NR algorithm.

Algorithm 1 General Quasi-Newton Algorithm with Line Search

1: **Input:** Given a initial starting point $\mathbf{x}^0 \in dom(f)$, $B^0 \in^{n\times n}$ (or H^0) and $t = 0$.
2: **while** Stop condition is false **do**
3: $\mathbf{d}^t = -(B^t)^{-1}\nabla f(\mathbf{x}^t)$ or $\mathbf{d}^t = -H^t\nabla f(\mathbf{x}^t)$,
4: Line search a proper step size s_t,
5: $\mathbf{x}^{t+1} = \mathbf{x}^t + s_t\mathbf{d}^t$,
6: Update $B^{t+1} \leftarrow B^t$ or $H^{t+1} \leftarrow H^t$
7: $t := t + 1$.
8: **end while**
9: **Output:** $\mathbf{x}^T$, where T is the last step index.

The main steps of Quasi-Newton algorithm is the step 4 and 6. Next, we will discuss how to select a proper step size s_t.

Actually, we hope that any B^t is a positive and definite matrix. So,

$$(\mathbf{d}^t)^\top B^t\mathbf{d}^t = (\mathbf{d}^t)^\top\mathbf{y}^t > 0 \text{ (Curvature Condition)}. \tag{2}$$

Definition 1 *Let* $\mathbf{d}^t$ *be a descent direction of* $\mathbf{x}^t$*. In addition, if*

$$f(\mathbf{x}^t + s\mathbf{d}^t) \leq f(\mathbf{x}^t) + c_1 s \nabla f(\mathbf{x}^t)^\top \mathbf{d}^t, \tag{3}$$

$$\nabla f(\mathbf{x}^t + s\mathbf{d}^t)\mathbf{d}^t \geq c_2 \nabla f(\mathbf{x}^t)^\top \mathbf{d}^t, \tag{4}$$

then s *is said to satisfy the Wolfe condition where* $c_1, c_2 \in (0,1), c_1 < c_2$*.*

In step 4, we use the Wolfe condition to search a step size. Assume that $\mathbf{d}^t$ is the direction we focused, then it satisfies $\nabla f(\mathbf{x}^{t+1})\mathbf{d}^t \geq c_2 \nabla f(\mathbf{x}^t)^\top \mathbf{d}^t$. So,

$$(\nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^{t+1}))^\top \mathbf{d}^t = (\mathbf{y}^t)^\top \mathbf{d}^t \geq (c_2 - 1)\nabla f(\mathbf{x}^t)^\top \mathbf{d}^t > 0. \tag{5}$$

This means that the step size that selected by Wolfe conditions can achieve the "Curvature Condition."

Finally, let us show three commonly used methods to update the approximated Hessian in Step 6.

1.2 SR1

The first method is called "Symmetric Rank One" (SR1) method. Suppose that we have B^t, then using a rank one matrix $uu^\top, u \in \mathbb{R}^n$ to update B^t. That is

$$B^{t+1} = B^t + \alpha_t \mathbf{u}^t (\mathbf{u}^t)^\top. \tag{6}$$

We hope that

$$\begin{aligned} B^{t+1}\mathbf{d}^t &= \mathbf{y}^t \\ (B^t + \alpha_t \mathbf{u}^t (\mathbf{u}^t)^\top)\mathbf{d}^t &= \mathbf{y}^t \\ (\alpha_t (\mathbf{u}^t)^\top \mathbf{d}^t)\mathbf{u}^t &= \mathbf{y}^t - B^t \mathbf{d}^t. \end{aligned}$$

We guess that

$$\begin{aligned} \mathbf{u}^t &= \mathbf{y}^t - B^t \mathbf{d}^t, \\ \alpha_t &= \frac{1}{(\mathbf{u}^t)^\top \mathbf{d}^t} = \frac{1}{(\mathbf{y}^t - B^t \mathbf{d}^t)^\top \mathbf{d}^t}. \end{aligned}$$

Then we have that

$$B^{t+1} = B^t + \frac{(\mathbf{y}^t - B^t \mathbf{d}^t)(\mathbf{y}^t - B^t \mathbf{d}^t)^\top}{(\mathbf{y}^t - B^t \mathbf{d}^t)^\top \mathbf{d}^t}. \tag{7}$$

The similar result can be obtained for H^t as:

$$H^{t+1} = H^t + \frac{(\mathbf{d}^t - H^t \mathbf{y}^t)(\mathbf{d}^t - H^t \mathbf{y}^t)^\top}{(\mathbf{d}^t - H^t \mathbf{y}^t)^\top \mathbf{y}^t}. \tag{8}$$

The biggest drawback of SR1 is that cannot guarantee B^t or H^t is positive and definite. In practice, SR1 is not commonly used.

1.3 BFGS

The second method of updating H^t or B^t is called BFGS. The name comes from Figure 1.

The basic idea is to use Rank 2 matrix! Namely,

$$B^{t+1} = B^t + \alpha_t \mathbf{u}^t (\mathbf{u}^t)^\top + \beta_t \mathbf{v}^t (\mathbf{v}^t)^\top. \tag{9}$$

Figure 1: BFGS

The similar with the derivatives of SR1, we have that

$$\begin{aligned}
&B^{t+1}\mathbf{d}^t = \mathbf{y}^t \\
&(B^t + \alpha_t \mathbf{u}^t(\mathbf{u}^t)^\top + \beta_t \mathbf{v}^t(\mathbf{v}^t)^\top)\mathbf{d}^t = \mathbf{y}^t \\
&(\alpha_t(\mathbf{u}^t)^\top \mathbf{d}^t)\mathbf{u}^t + (\beta_t(\mathbf{v}^t)^\top \mathbf{d}^t)\mathbf{v}^t = \mathbf{y}^t - B^t\mathbf{d}^t.
\end{aligned}$$

We guess that

$$\begin{aligned}
&\mathbf{u}^t = \mathbf{y}^t \\
&\alpha_t = \frac{1}{(\mathbf{u}^t)^\top \mathbf{d}^t} = \frac{1}{(\mathbf{y}^t)^\top \mathbf{d}^t} \\
&\mathbf{v}^t = B^t\mathbf{d}^t \\
&\beta_t = -\frac{1}{(\mathbf{v}^t)^\top \mathbf{y}^t} = -\frac{1}{(\mathbf{d}^t)^\top B^t \mathbf{d}^t}.
\end{aligned}$$

So,

$$B^{t+1} = B^t + \frac{\mathbf{y}^t(\mathbf{y}^t)^\top}{(\mathbf{y}^t)^\top \mathbf{d}^t} - \frac{B^t\mathbf{d}^t(B^t\mathbf{d}^t)^\top}{(\mathbf{d}^t)^\top B^t \mathbf{d}^t}. \tag{10}$$

Theorem 1 *(Sherman- Morrison-Woodbury Formula)*

Suppose that $A \in \mathbb{R}^{n^2}$ and $C \in \mathbb{R}^{k^2}$ are invertible, and $U \in \mathbb{R}^{n\times k}, V \in \mathbb{R}^{k\times n}$, then $A + UCV$ is invertible if and only if $C^{-1} + VA^{-1}U$ is ivertible, and

$$(A + UCV)^{-1} = A^{-1} - A^{-1}UV(C^{-1} + VA^{-1}U)^{-1}A^{-1}. \tag{11}$$

Remark 1 *In Theorem 1, if we set $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, then*

$$(A + \mathbf{u}\mathbf{v}^\top)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^\top A^{-1}}{1 + \mathbf{v}^\top A^{-1}\mathbf{u}}. \tag{12}$$

Then, based on SMW Theorem 1

$$H^{t+1} = (I - \rho_t \mathbf{d}^t(\mathbf{y}^t)^\top)^\top H^t(I - \rho_t \mathbf{d}^t(\mathbf{y}^t)^\top) + \rho_t \mathbf{d}^t(\mathbf{d}^t)^\top, \tag{13}$$

where $\rho_t = \frac{1}{(\mathbf{d}^t)^\top \mathbf{y}^t}$.

Based on this formula, we can find that $H^t \succ 0$ implies $H^{t+1} \succ 0$.

For large-scale optimization problems, L-BFGS algorithm is designed. Please See 6.5.4 in the text book.

1.4 DFP

Comparing with BFGS, DFP focus on updating H^t in advance. So, it has

$$H^{t+1} = H^t + \alpha_t \mathbf{u}^t (\mathbf{u}^t)^\top + \beta_t \mathbf{v}^t (\mathbf{v}^t)^\top. \tag{14}$$

Based on $H^{t+1}\mathbf{y}^t = \mathbf{d}^t$, we can obtain that

$$H^{t+1} = H^t + \frac{\mathbf{d}^t (\mathbf{d}^t)^\top}{(\mathbf{y}^t)^\top \mathbf{d}^t} - \frac{H^t \mathbf{y}^t (H^t \mathbf{y}^t)^\top}{(\mathbf{y}^t)^\top H^t \mathbf{y}^t}. \tag{15}$$

So, based on SMW Theorem 1

$$B^{t+1} = (I - \rho_t \mathbf{y}^t (\mathbf{d}^t)^\top)^\top B^t (I - \rho_t \mathbf{y}^t (\mathbf{d}^t)^\top) + \rho_t \mathbf{y}^t (\mathbf{y}^t)^\top. \tag{16}$$

1.5 Theory

Please see 6.5.3 of the text book.

References