Lecture 21: Quasi-Newton Methods

Yudong Chen

1 Generic quasi-Newton method

A generic quasi-Newton (QN) method takes the form

$$x_{k+1} = x_k - \alpha_k \underbrace{(B_k)^{-1} \nabla f(x_k)}_{-p_k}, \tag{QN}$$

where $B_k \succ 0$. We assume that the stepsize α_k is chosen by a linear procedure to satisfy the weak/strong Wolfe conditions (both sufficient decrease and curvature).[1] [2]

We want a B_k that is easier to compute than the Hessian $\nabla^2 f(x_k)$ but has the same "effect" as $\nabla^2 f(x_k)$. In particular, B_k should be such that the search direction $p_k = -B_k^{-1}\nabla f(x_k)$ approximates the Newton direction $p_k^{\mathrm{N}} = -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$. The goal is to achieve superlinear convergence, i.e., faster than first-order methods.

1.1 General results

The theorem below is general and applies to any search direction p_k. We will later apply this theorem to the quasi-newton method (QN).

Theorem 1 (Theorem 3.6 in Nocedal-Wright)**.** *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable. Consider the iteration $x_{k+1} = x_k + \alpha_k p_k$, where p_k is a descent direction and α_k satisfies Weak Wolfe Conditions (WWC) with $c_1 \leq \frac{1}{2}$. If the sequence $\{x_k\}$ converges to a point x^* such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$, and if the search direction p_k satisfies*

$$\lim_{k\to\infty} \frac{\left\|\nabla f(x_k) + \nabla^2 f(x_k) p_k\right\|_2}{\|p_k\|_2} = 0, \tag{1}$$

then

1. *the unit stepsize $\alpha_k = 1$ is admissible (i.e., satisfies WWC) for all sufficient large k;*
2. *if $\alpha_k = 1$ for all $k > k_0$, where $k_0 < \infty$, then $\{x_k\}$ converges to x^k superlinearly.*

As a general result, Theorem 1 can be applied to the damped Newton's method, which uses $p_k = -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$ and trivially satisfies (1). Therefore, the theorem guarantees that damped Newton's method with backtracking line search accepts the stepszie $\alpha_k = 1$ for k sufficiently large, in which case it reduces to basic Newton's method and converges quadratically.

[1] For reasons to become clear later, it is important that the curvature condition (not just sufficient decrease) holds. Therefore, backtracking line search is less appropriate for Quasi-Newton methods.

[2] It is often assumed that the line search procedure will try $\alpha_k = 1$ first and accept this stepsize if it satisfies the Wolfe Conditions.

For a QN search direction $p_k = -B_k^{-1}\nabla f(x_k)$, it can be shown that the condition (1) is equivalent to

$$\lim_{k\to\infty} \frac{\left\|\left(B_k - \nabla^2 f(x_k)\right) p_k\right\|_2}{\|p_k\|_2} = 0. \tag{2}$$

The above equation can be written as $\left\|\left(B_k - \nabla^2 f(x_k)\right) p_k\right\| = o(\|p_k\|)$. Note that this condition may hold even if B_k does not converge to $\nabla^2 f(x^*)$. It suffices that B_k approximates $\nabla^2 f(x_k)$ well *along the search directions* p_k. This is a general guideline for choosing B_k.

Below is another equivalent form of the above condition.

Claim 1. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable. Assume that $\{x_k\}$ converges to a point x^* such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$. Then, condition (2) is equivalent to

$$\left\|p_k - p_k^{\mathrm{N}}\right\|_2 = o(\|p_k\|_2), \tag{3}$$

where $p_k := -B_k^{-1}\nabla f(x_k)$ be the Quasi-Newton search direction, and $p_k^{\mathrm{N}} := -\left(\nabla^2 f(x_k)\right)^{-1}\nabla f(x_k)$ is the Newton direction.

Proof of Claim 1. We first show (2) $\implies$ $\left\|p_k - p_k^{\mathrm{N}}\right\| = o(\|p_k\|)$. Since $p_k = -B_k^{-1}\nabla f(x_k)$, we can write

$$p_k^{\mathrm{N}} = -\left(\nabla^2 f(x_k)\right)^{-1}\nabla f(x_k) = \left(\nabla^2 f(x_k)\right)^{-1} B_k p_k.$$

Hence

$$\begin{aligned}
\left\|p_k - p_k^{\mathrm{N}}\right\| &= \left\|p_k - \left(\nabla^2 f(x_k)\right)^{-1} B_k p_k\right\| \\
&= \left\|\left(\nabla^2 f(x_k)\right)^{-1}\left(\nabla^2 f(x_k) - B_k\right) p_k\right\| \\
&\le \left\|\left(\nabla^2 f(x_k)\right)^{-1}\right\| \cdot \left\|\left(\nabla^2 f(x_k) - B_k\right) p_k\right\| \\
&\le 2\left\|\left(\nabla^2 f(x^*)\right)^{-1}\right\| \cdot o\left(\|p_k\|\right) \\
&\qquad \text{because } \left\|\left(\nabla^2 f(x_k)\right)^{-1}\right\| \le 2\left\|\left(\nabla^2 f(x^*)\right)^{-1}\right\| \text{ for } k \text{ sufficient large by Hessian continuity,} \\
&\qquad \text{and by (2)} \\
&= o(\|p_k\|).
\end{aligned}$$

We next show $\left\|p_k - p_k^{\mathrm{N}}\right\| = o(\|p_k\|) \implies$ (2). From what we have derived above:

$$p_k - p_k^{\mathrm{N}} = \left(\nabla^2 f(x_k)\right)^{-1}\left(\nabla^2 f(x_k) - B_k\right) p_k,$$

hence

$$\left(\nabla^2 f(x_k) - B_k\right) p_k = \nabla^2 f(x_k)\left(p_k - p_k^{\mathrm{N}}\right).$$

It follows that

$$\begin{aligned}
\left\|\left(\nabla^2 f(x_k) - B_k\right) p_k\right\| &= \left\|\nabla^2 f(x_k)\left(p_k - p_k^{\mathrm{N}}\right)\right\| \\
&\le \left\|\nabla^2 f(x_k)\right\| \left\|p_k - p_k^{\mathrm{N}}\right\| \\
&= O(1) \cdot o(\|p_k\|),
\end{aligned}$$

where the last step holds since $\left\|\nabla^2 f(x_k)\right\| \le 2\left\|\nabla^2 f(x^*)\right\| = O(1)$. □

In fact, the equivalent conditions (2) and (3) are both necessary and sufficient for superlinear convergence of QN method, as shown in the following theorem.

Theorem 2 (Theorem 3.7 in Nocedal-Wright)**.** *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable. Consider the iteration (QN) with $\alpha_k = 1$. Assume that $\{x_k\}$ converges to a point x^* such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$. Then the convergence is superlinear if and only if (2) holds.*

Proof of Theorem 2. We only prove the "if" part; "only if" part is left as exercise.

Assume $\left\|p_k - p_k^{\mathrm{N}}\right\| = o(\|p_k\|)$. Want to show superlinear convergence, i.e., $\|x_{k+1} - x^*\| = o\left(\|x_k - x^*\|\right)$. We have

$$\begin{aligned}
\|x_{k+1} - x^*\| &= \|x_k + p_k - x^*\| \\
&= \left\|x_k + p_k^{\mathrm{N}} - x^* + p_k - p_k^{\mathrm{N}}\right\| \\
&\le \left\|x_k + p_k^{\mathrm{N}} - x^*\right\| + \left\|p_k - p_k^{\mathrm{N}}\right\| \\
&= O\left(\|x_k - x^*\|^2\right) + o\left(\|p_k\|\right) \\
&= o\left(\|x_k - x^*\|\right) + o\left(\|p_k\|\right).
\end{aligned}$$

It remains to show $\|p_k\| = O\left(\|x_k - x^*\|\right)$. Note that $\left\|p_k - p_k^{\mathrm{N}}\right\| = o\left(\|p_k\|\right)$ implies

$$\begin{aligned}
\|p_k\| &= O\left(\left\|p_k^{\mathrm{N}}\right\|\right) \\
&= O\left(\left\|x_k + p_k^{\mathrm{N}} - x^* - (x_k - x^*)\right\|\right) \\
&\le O\Big(\underbrace{\left\|x_k + p_k^{\mathrm{N}} - x^*\right\|}_{=o(\|x_k - x^*\|)} + \|x_k - x^*\|\Big) \\
&= O\left(\|x_k - x^*\|\right).
\end{aligned}$$

□

1.2 Basic ideas of quasi-Newton

We want to choose B_k such that

1. B_k is a good estimate of $\nabla^2 f(x_k)$ in the sense of (2), which guarantees superlinear convergence;
2. B_k can be formed by "cheap" operations, without actually computing the Hessian $\nabla^2 f(x_k)$.

We consider Quasi-Newton methods that only use *gradient* evaluation to compute B_k. To get information about $\nabla^2 f$ from ∇f, we make use of one form of Taylor's Theorem:

$$\nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f\left(x + t(y - x)\right)(y - x)\mathrm{d}t.$$

The first idea is to take finite differences $\nabla f(x + e_i) - \nabla f(x)$ along n directions $e_i, i = 1, \ldots, n$. This is too expensive.

Instead, we only use the gradients we will evaluate anyway, namely $\nabla f(x_k)$.

In the sequel, we discuss four popular Quasi-Newton methods: DFP, BFGS, SR1, and L-BFGS.

2 The DFP method

The DFP (Davidon-Fletcher-Powell) is one of the earliest efficient quasi-Newton methods.

Quadratic model

To derive the DFP method, we begin with the following local quadratic model of f:

$$f(x_k + p) \approx m_k(p) := f(x_k) + \langle \nabla f(x_k), p \rangle + \frac{1}{2} p^\top B_k p.$$

Note that $f(x_k) = m_k(0)$ and $\nabla f(x_k) = \nabla m_k(0)$. The QN search direction is given by

$$p_k = -B_k^{-1} \nabla f(x_k) = \operatorname*{argmin}_{p \in \mathbb{R}^d} m_k(p).$$

We then compute $x_{k+1} = x_k + \alpha_k p_k$, where α_k is stepsize determined using a line search procedure.

Suppose B_k has been computed. We move on to the next iteration, where the quadratic model is

$$m_{k+1}(p) = f(x_{k+1}) + \langle \nabla f(x_{k+1}), p \rangle + \frac{1}{2} p^\top B_{k+1} p.$$

Instead of computing B_{k+1} from scratch, we will use B_k to compute B_{k+1}.

Secant equation

We want to choose B_{k+1} so that m_{k+1} is a good quadratic model of f. A reasonable condition is that the gradient of m_{k+1} agrees with the gradient of f at both x_k and x_{k+1}. By construction, we automatically have $\nabla m_{k+1}(0) = \nabla f(x_{k+1})$.

What about $\nabla f(x_k)$? Note that

$$\nabla m_{k+1}(-\alpha_k p_k) = \nabla f(x_{k+1}) - \alpha_k B_{k+1} p_k,$$

and we want the RHS to agree with $\nabla f(x_k)$. That is, we want B_{k+1} to satisfy the equation

$$\alpha_k B_{k+1} p_k = \nabla f(x_{k+1}) - \nabla f(x_k).$$

Let us introduce the shorthands

$$\begin{aligned} s_k &:= \alpha_k p_k = x_{k+1} - x_k, && \text{displacement} \\ y_k &:= \nabla f(x_{k+1}) - \nabla f(x_k). && \text{change in gradients} \end{aligned}$$

Then the above equation can be written compactly as

$$B_{k+1} s_k = y_k, \tag{4}$$

which is called the *secant equation.*

Curvature condition

If $B_{k+1} \succ 0$, then left multiplying both sides of (4) by $s_k^\top$ gives

$$s_k^\top y_k > 0, \tag{5}$$

which is called the *curvature condition*. This is a necessary for the existence of a p.d. B_{k+1} satisfying the secant equation (4).

- The curvature condition will be automatically satisfied if f is strongly convex, since

$$s_k^\top y_k = \langle \nabla f(x_{k+1}) - \nabla f(x_k), x_{k+1} - x_k \rangle > 0,$$

  which is the strong monotonicity/coercivity property of the gradient.

- The curvature condition does not automatically hold for nonconvex functions. It holds if α_k (the stepsize for the *previous* iteration k) satisfies the second Wolfe condition. In particular, by WW2 (curvature condition), we have

$$\langle \nabla f(x_{k+1}), s_k \rangle \geq c_2 \langle \nabla f(x_k), s_k \rangle, \quad \text{where } c_2 \in (0,1),$$

  hence

$$\begin{aligned} \langle y_k, s_k \rangle &= \langle \nabla f(x_{k+1}) - \nabla f(x_k), s_k \rangle \\ &\geq \underbrace{(c_2 - 1)}_{<0} \underbrace{\langle \nabla f(x_k), s_k \rangle}_{<0} > 0. \end{aligned}$$

When the curvature condition holds, the secant equation $B_{k+1}s_k = y_k$ has a solution. In fact, it has infinitely many solutions.

Choosing B_{k+1}

To uniquely specify B_{k+1}, we can enforce that it is the "closest" matrix to B_k that satisfies the above conditions. In particular, we compute B_{k+1} by solving

$$\begin{aligned} \min_B \ & \|B - B_k\| \\ \text{s.t. } & B = B^\top \\ & Bs_k = y_k, \end{aligned} \tag{6}$$

where $\|\cdot\|$ is some matrix norm.

A norm that gives an easy-to-compute (and affine-invariant) solution is the weighted Frobenius norm

$$\|A\|_W := \left\| W^{1/2} A W^{1/2} \right\|_F,$$

where W is a p.d. weight matrix, $W^{1/2}$ denotes the matrix square root of W (HW1 Q6), and $\|C\|_F^2 := \sum_{i=1}^d \sum_{j=1}^d C_{ij}^2$ denotes the Frobenius norm of a matrix C. Here the weight W can be any matrix that satisfies $Wy_k = s_k$. For example, we can take $W = \bar{G}_k^{-1}$, where $\bar{G}_k = \int_0^1 \nabla^2 f(x_k + t\alpha_k p_k) \mathrm{d}t$ is the average Hessian. Then $Wy_k = s_k$ holds by Taylor's Theorem:

$$\int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k)) \underbrace{(x_{k+1} - x_k)}_{s_k} \mathrm{d}t = \underbrace{\nabla f(x_k + 1) - \nabla f(x_k)}_{y_k}.$$

The DFP update rules

With the above choice of the norm and weigh matrix, the unique solution to the problem (6) is given by

$$\text{(DFP)} \qquad B_{k+1} = \left(I - \frac{y_k s_k^\top}{y_k^\top s_k} \right) B_k \left(I - \frac{s_k y_k^\top}{y_k^\top s_k} \right) + \frac{y_k y_k^\top}{y_k^\top s_k}. \tag{7}$$

The inverse $H_{k+1} = B_{k+1}^{-1}$ can also be computed recursively and efficiently, using the Sherman-Morrison-Woodbury formula (exercise):

$$\text{(DFP)} \qquad H_{k+1} = H_k - \underbrace{\frac{H_k y_k y_k^\top H_k}{y_k^\top H_k y_k}}_{\text{rank-1}} + \underbrace{\frac{s_k s_k^\top}{y_k^\top s_k}}_{\text{rank-1}}. \tag{8}$$

The above two equations involve rank-2 modifications (exercise: show that $B_{k+1} - B_k$ has rank at most 2). This structure can be exploited for efficient storage and computation.

In the least-change problem (6), we do not explicit enforce positive definiteness of B. This property holds automatically.

Fact 1. *If B_k and H_k are positive definite and $y_k^\top s_k > 0$, then B_{k+1} and H_{k+1} are also positive definite.*

Proof. Take any vector $z \neq 0$. From (7) we have

$$z^\top B_{k+1} z = \left(z - s_k \cdot \frac{y_k^\top z}{y_k^\top s_k} \right)^\top B_k \left(z - s_k \cdot \frac{y_k^\top z}{y_k^\top s_k} \right) + \frac{(y_k^\top z)^2}{y_k^\top s_k}.$$

If $y_k^\top z \neq 0$, the second RHS term is positive. If $y_k^\top z = 0$, then $z - s_k \cdot \frac{y_k^\top z}{y_k^\top s_k} = z \neq 0$ and hence the first RHS term is positive (since $B_k \succ 0$). We conclude that $B_{k+1} \succ 0$ and consequently $H_{k+1} = B_{k+1}^{-1} \succ 0$. □

DFP is a precursor of the BFGS (Broyden-Fletcher-Goldfarb-Shanno) method, the most popular quasi-Newton method.

Appendices

Sherman-Morrison-Woodbury formula:

$$\left(A + UV^\top \right)^{-1} = A^{-1} - A^{-1} U \left(I + V^\top A U \right)^{-1} V^\top A^{-1},$$

which is valid when the matrix dimensions are compatible and all inverses on the RHS are well-defined.