Lecture 22: Quasi-Newton: The BFGS and SR1 Methods

Yudong Chen

1 The BFGS method

Closely related to DFP is the BFGS (Broyden-Fletcher-Goldfarb-Shanno) method, which is the most popular quasi-Newton method.

The high level idea of BFGS is similar to DFP, except that we switch the roles of B_k and H_k:

- works with a secant equation for H_{k+1} instead of B_{k+1};
- imposes a least change condition on H_{k+1} instead of B_{k+1}.

In particular recall the DFP secant equation:

$$\text{DFP:} \qquad y_k = B_{k+1} s_k. \tag{1}$$

Working with $H_{k+1} = B_{k+1}^{-1}$, BFGS considers the following secant equation:

$$\text{BFGS:} \qquad H_{k+1} y_k = s_k. \tag{2}$$

To find H_{k+1}, we solve the least-change problem

$$\begin{aligned} \min_H \; & \|H - H_k\|_W \\ \text{s.t.} \; & H = H^\top \\ & H y_k = s_k, \end{aligned} \tag{3}$$

where $\|\cdot\|_W$ is the weighted Frobenius norm with weight matrix $W = \bar{G}_k = \int_0^1 \nabla^2 f(x_k + t s_k) \mathrm{d}t$. The solution H_{k+1} and its inverse B_{k+1} are given in closed form by

$$\text{(BFGS)} \qquad \begin{aligned} H_{k+1} &= \left(I - \frac{s_k y_k^\top}{s_k^\top y_k} \right) H_k \left(I - \frac{y_k s_k^\top}{s_k^\top y_k} \right) + \frac{s_k s_k^\top}{s_k^\top y_k}, \\ B_{k+1} &= B_k - \underbrace{\frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k}}_{\text{rank-1}} + \underbrace{\frac{y_k y_k^\top}{y_k^\top s_k}}_{\text{rank-1}}. \end{aligned} \tag{4}$$

Similar to DFP, BFGS involves rank-2 updates and maintains positive definiteness (proof left as exercise).

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} computed using (4) are also positive definite.*

DFP and BFGS are duals of each other: one can be obtained from the other using the interchanges below.

DFP	B_{k+1}	s_k	y_k
BFGS	H_{k+1}	y_k	s_k

1.1 Implementation and performance

A direct implementation of BFGS stores the $d \times d$ matrix H_k explicitly. An alternative : store σ_0 for $H_0 = \sigma_0 I$ and the pairs $(s_0, y_0), (s_1, y_1), \ldots, (s_k, y_k)$, so H_{k+1} is stored implicitly. To form the search direction $-H_k \nabla f(x_k)$ from this implicit representation, it takes $O(d)$ operations for each step, so $O(dk)$ operations in total, and storage of $O(dk)$. For $k \leq d/5$, this is better than explicit storage with cost $O(d^2)$.

It is observed that BFGS tends to outperform DFP, as BFGS can more effectively recover from a bad Hessian approximation B_k.

Some numerical results on $f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$ (from Nocedal-Wright). To achieve $\|\nabla f(x_k)\| \leq 10^{-5}$, the steepest descent (i.e., GD) method required 5264 iterations, BFGS required 34, and Newton required 21. The table shows $\|x_k - x^*\|$ for the last few iterations.

steepest descent	BFGS	Newton
1.827e-04	1.70e-03	3.48e-02
1.826e-04	1.17e-03	1.44e-02
1.824e-04	1.34e-04	1.82e-04
1.823e-04	1.01e-06	1.17e-08

1.2 Convergence guarantees for BFGS

We consider the iteration $x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k)$, where B_k is updated according to BFGS (4), and α_k satisfies the Weak Wolfe Conditions with $c_1 \leq \frac{1}{2}$. Moreover, we will assume that the line search procedure will always try $\alpha_k = 1$ first and accept it when it satisfies the Wolfe Conditions.

We have global convergence guarantees for *convex* functions.

Theorem 1 (Global convergence; Thm 6.5 in Nocedal-Wright)**.** *Suppose that*

- *$f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, the sublevel set $\mathcal{L} := \left\{x \in \mathbb{R}^d \mid f(x) \leq f(x_0)\right\}$ is convex, and*

$$\forall x \in \mathcal{L}: \quad mI \preccurlyeq \nabla^2 f(x) \preccurlyeq MI$$

  for some $0 < m \leq M < \infty$. (Note that f has a unique minimizer x^* in $\mathcal{L}$.)

- *The initial B_0 is symmetric p.d.*

Then $\{x_k\}$ converges to the minimizer x^*.

Using Theorem 1, we can in fact show that the convergence is fast enough that

$$\sum_{k=1}^{\infty} \|x_k - x^*\| < \infty. \tag{5}$$

We have local superlinear convergence guarantees for general (possibly nonconvex) functions.

Theorem 2 (Local superlinear convergence; Thm 6.6 in Nocedal-Wright)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable. Suppose that the iterates of BFGS converge to a local minimizer x^* and satisfy (5), and the Hessian of f is positive definite and L-Lipschitz around x^*, i.e.,*

$$\left\|\nabla^2 f(x) - \nabla^2 f(x^*)\right\| \leq L \|x - x^*\|, \qquad \forall x \in \mathcal{N}_{x^*}.$$

Then $\{x_k\} \xrightarrow{k\to\infty} x^*$ at a superlinear rate.

The proof of Theorem 2 ends by showing that

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

In this case, Theorem 2 from Lecture 21 applies and guarantees superlinear convergence.

2 The SR1 (symmetric rank-1 update) method

Consider the rank-1 update

$$B_{k+1} = B_k + \sigma_k v_k v_k^\top,$$

where $\sigma_k \in \{-1, +1\}$ and $v_k \in \mathbb{R}^d$. We choose σ_k, B_k so that B_{k+1} satisfies the secant equation

$$B_{k+1} s_k = y_k, \tag{6}$$

where $s_k := x_{k+1} - x_k, y_k := \nabla f(x_{k+1}) - \nabla f(x_k)$. The secant equation is equivalent to

$$y_k - B_k s_k = \underbrace{\sigma_k (v_k^\top s_k)}_{\in\mathbb{R}} v_k. \tag{7}$$

Assume $v_k^\top s_k \neq 0$. Then v_k is parallel to $y_k - B_k s_k$, i.e., $v_k = \delta(y_k - B_k s_k)$ for some $\delta \in \mathbb{R}$. Substituting back, we get

$$y_k - B_k s_k = \underbrace{\sigma_k \delta^2 s_k^\top (y_k - B_k s_k)}_{\in\mathbb{R}} (y_k - B_k s_k).$$

For this equation to hold, we must have

$$\sigma_k = \operatorname{sign}\left(s_k^\top (y_k - B_k s_k)\right), \qquad \delta = \pm \frac{1}{\sqrt{\left|s_k^\top (y_k - B_k s_k)\right|}}$$

assuming that $\left|s_k^\top (y_k - B_k s_k)\right| \neq 0$.

The above choice of σ_k and δ are the only possible way of satisfying the secant equation with a symmetric rank-1 update. This gives the SR1 update rule for B_{k+1}:

$$\text{(SR1)} \qquad B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^\top}{s_k^\top (y_k - B_k s_k)}.$$

By Sherman-Morrison formula, we also have the update rule for $H_{k+1} = B_{k+1}^{-1}$:

$$\text{(SR1)} \qquad H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^\top}{y_k^\top (s_k - H_k y_k)}.$$

SR1 is very simple. However, even if B_k is p.d., B_{k+1} may not be. The same holds for H_k and H_{k+1}. Therefore, SR1 is in general not used with the update $x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k)$, as it need not give a descent direction. However, this B_k is quite useful in Trust-Region methods, which we will discuss later. The lack of positive definiteness may actually make B_k a better approximation to the true Hessian $\nabla^2 f(x_k)$ (which may be indefinite), compared to B_k generated by DFP/BFGS.

Another major issue of SR1: the numbers $s_k^\top(y_k - B_k s_k)$ and $y_k^\top(s_k - H_k y_k)$, which appear in the denominators of the update rules, may be zero (or very small). In this case, there is no symmetric rank-1 update that satisfies the secant equation. This may happen even when f is a convex quadratic.

Let us zoom in the above issue. Based on our derivation of SR1, there are three cases:

1. If $s_k^\top(y_k - B_k s_k) \neq 0$, then B_{k+1} is uniquely defined by the SR1 update rule above.

2. If $y_k = B_k s_k$, then by (7) the only way to satisfy the secant equation is with $B_{k+1} = B_k$.

3. If $y_k \neq B_k s_k$ and $s_k^\top(y_k - B_k s_k) = 0$, then there is no symmetric rank-1 update that satisfies the secant equation.

Due to the case 3, SR1 is numerically unstable. To have all the required properties of B_k, H_k, rank-2 updates (as in DFP/BFGS) are necessary.

Nevertheless, SR1 is still used, because:

1. there exists a simple safeguard that prevents numerical instability (see below);

2. there exist some setups (e.g., constrained optimization) where it is not possible to impose the curvature condition $y_k^\top s_k > 0$, which is necessary for DFP/BFGS, but not needed in SR1.

Safeguard for SR1: Apply SR1 update only if

$$\left| s_k^\top(y_k - B_k s_k) \right| \geq r \|s_k\| \|y_k - B_k s_k\|, \tag{8}$$

where r is some small constant (e.g., 10^{-8}). Otherwise, set $B_{k+1} = B_k$ (i.e., skip the update). Note that the skipping happens when B_k is already a good approximation of the true Hessian along the direction s_k.

Hessian approximation properties of SR1:

- (NW Theorem 6.1) For strongly convex quadratic function $f(x) = \frac{1}{2}x^\top A x + b^\top x$, if $s_k^\top(y_k - B_k s_k) \neq 0$ for all k, then SR1 iterates converges to the minimizer x^* in at most d step. Moreover, if its search directions $p_k = -B_k^{-1}\nabla f(x_k)$ are linearly independent, then $H_d = A^{-1}$.

- (NW Theorem 6.2) For general f with Lipschitz continuous Hessian, if $x_k \to x^*$, (8) holds for all k, and the steps $\{s_k\}$ uniformly linearly independent, then $B_k \to \nabla^2 f(x^*)$.

(Optional) Go through the proof of Theorem 6.1.