

1 Outline

In this lecture, we study

- Newton's method.
- Convergence of Newton's method.

2 Quasi-Newton method

Remember our comparison of gradient descent and Newton's method. Gradient descent has a cheaper iteration cost of $O(d)$, while Newton's method has a lower iteration complexity of $O(\log \log(1/\epsilon))$. Quasi-Newton methods are designed to achieve the best of both worlds. We will study methods that achieve an iteration cost of $O(d^2)$ and an iteration complexity of $o(\log(1/\epsilon))$.

The basic outline of a quasi-Newton method is as follows.

Algorithm 1 Quasi-Newton method

```
Initialize  $x_1$  and a positive definite matrix  $B_1$ .  
for  $t = 1, \dots, T - 1$  do  
    Solve  $B_t g_t = -\nabla f(x_t)$ .  
    Update  $x_{t+1} = x_t + \eta_t g_t$ .  
    Compute  $B_{t+1}$  from  $B_t$ .  
end for  
Return  $x_T$ .
```

Here, a candidate for B_t is $\nabla^2 f(x_t)$, in which case, $g_t = -\nabla^2 f(x_t)^{-1} \nabla f(x_t)$ corresponds to a Newton iteration. Note that given x_t and B_t , we can obtain g_t and x_{t+1} . For the next iteration, we need to design B_{t+1} . There are several desired properties when selecting B_{t+1} based on B_t .

- We want B_{t+1} to be symmetric and positive definite.
- We want B_{t+1} to be close to B_t , or we want to compute B_{t+1} from B_t easily.
- We want B_{t+1} to satisfy

$$\nabla f(x_{t+1}) - \nabla f(x_t) = B_{t+1} g_t.$$

A motivation for this is that for one dimensional problem, we have

$$h'(x_{t+1}) = \frac{h(x_{t+1}) - h(x_t)}{x_{t+1} - x_t}$$

and thus $h(x_{t+1}) - h(x_t) = h'(x_{t+1})(x_{t+1} - x_t)$.

Hereinafter, we stick to the following notations for ease of exposition.

$$B^+ = B_{t+1}, \quad B = B_t, \quad s^+ = g_{t+1}, \quad s = g_t, \quad y = \nabla f(x_{t+1}) - \nabla f(x_t).$$

Hence, the goal is to compute B^+ satisfying the desired properties. In particular, we want B^+ to be positive definite and satisfy

$$y = B^+ s,$$

which is called the *secant equation*.

2.1 Symmetric rank-one (SR1) update

Remember that we want B^+ to be something that is “close” to B . One way is to add a rank-one matrix to B to obtain B^+ . To be precise, let $a \in \mathbb{R}$ and $u \in \mathbb{R}^d$. Then we update

$$B^+ = B + auu^\top.$$

Then the secant equation requires that

$$y - Bs = a(u^\top s)u,$$

in which case $y - Bs$ and u are scalar multiples of each other. Hence, we can set $u = y - Bs$ and $a = 1/((y - Bs)^\top s)$. Then B^+ is given by

$$B^+ = B + \frac{(y - Bs)(y - Bs)^\top}{(y - Bs)^\top s}.$$

Next, to compute s^+ satisfying $B^+ s^+ = -\nabla f(x^+)$, which corresponds to $B_{t+1}g_{t+1} = -\nabla f(x_{t+1})$, we need to obtain the inverse of B^+ . In fact, the inverse of B^+ based on the SR1 update is given easily.

Lemma 22.1 (Sherman-Morrison formula). *Let $B \in \mathbb{R}^{d \times d}$ be invertible, and let $u, v \in \mathbb{R}^d$.*

$$(B + uv^\top)^{-1} = B^{-1} - \frac{B^{-1}uv^\top B^{-1}}{1 + v^\top B^{-1}u}.$$

Based on this lemma,

$$\begin{aligned} (B^+)^{-1} &= B^{-1} - \frac{B^{-1}(y - Bs)(y - Bs)^\top B^{-1}}{(y - Bs)^\top s + (y - Bs)^\top B^{-1}(y - Bs)} \\ &= B^{-1} + \frac{(s - B^{-1}y)(s - B^{-1}y)^\top}{(s - B^{-1}y)^\top y}. \end{aligned}$$

However, B^+ is not necessarily positive definite, even if B is.

2.2 Broyden-Fletcher-Goldfarb-Shanno (BFGS) update

Our next attempt is to add a rank-two matrix, which is the sum of two rank-one matrices. To be specific, let $a, b \in \mathbb{R}$ and $u, v \in \mathbb{R}^d$. Then we update

$$B^+ = B + auu^\top + bvv^\top.$$

Then the secant equation requires that

$$y - Bs = a(u^\top s)u + b(v^\top s)v,$$

in which case, we can set $u = -Bs$ and $v = y$. Then

$$B^+ = B - \frac{Bss^\top B}{s^\top Bs} + \frac{yy^\top}{y^\top s}.$$

This update rule is called the Broyden-Fletcher-Goldfarb-Shanno (BFGS) update.

Lemma 22.2 (Woodbury formula). *Let B, D be invertible matrices and U, V be matrices of appropriate dimensions. Then*

$$(B + UDV)^{-1} = B^{-1} - B^{-1}U(D^{-1} + VB^{-1}U)^{-1}VB^{-1}.$$

Then

$$\begin{aligned} (B^+)^{-1} &= \left(B + \underbrace{\begin{bmatrix} Bs & y \end{bmatrix}}_U \underbrace{I}_D \underbrace{\begin{bmatrix} -1/s^\top Bs & 0 \\ 0 & 1/y^\top s \end{bmatrix}}_V \underbrace{\begin{bmatrix} s^\top B \\ y^\top \end{bmatrix}}_V \right)^{-1} \\ &= \left(I - \frac{sy^\top}{y^\top s} \right) B^{-1} \left(I - \frac{ys^\top}{y^\top s} \right) + \frac{ss^\top}{y^\top s}. \end{aligned}$$

Once we have obtained the inverse of B , computing the inverse of B^+ boils down to rank-one matrix multiplications, which costs $O(d^2)$ time steps.

Moreover, the resulting matrix B^+ is positive definite.

$$x^\top (B^+)^{-1} x = \left(x - \frac{x^\top s}{y^\top s} y \right)^\top B^{-1} \left(x - \frac{s^\top x}{y^\top s} y \right) + \frac{(x^\top s)^2}{y^\top s}.$$

Here, since B is positive definite, so is its inverse. Hence, the first term is strictly positive. Moreover,

$$y^\top s = \frac{1}{\eta_t} (\nabla f(x_{t+1}) - \nabla f(x_t))^\top (x_{t+1} - x_t) \geq 0$$

due to the convexity of f . Therefore, $x^\top (B^+)^{-1} x > 0$ for any nonzero x , and thus B^+ is positive definite.

2.3 Davidon-Fletcher-Powell (DFP) update

The DFGS update adds a rank-two matrix to the current matrix B . We may add a rank-two matrix to the inverse, and the corresponding update rule is called the Davidon-Fletcher-Powell (DFP) update. To be specific,

$$(B^+)^{-1} = B^{-1} + auu^\top + bvv^\top,$$

which is equivalent to

$$B^+ = \left(B^{-1} + auu^\top + bvv^\top \right)^{-1}$$

Then the secant equation requires that

$$s - B^{-1}y = a(u^\top y)u + b(v^\top y)v.$$

Following the same argument from the previous part, we have

$$(B^+)^{-1} = B^{-1} - \frac{B^{-1}yy^\top B^{-1}}{y^\top B^{-1}y} + \frac{ss^\top}{s^\top y}.$$

Moreover,

$$B^+ = \left(I - \frac{ys^\top}{s^\top y}\right) B \left(I - \frac{sy^\top}{s^\top y}\right) + \frac{yy^\top}{s^\top y}.$$

2.4 Broyden class

We have discussed the BFGS and DFP update rules to run quasi-Newton iterations. We can interpolate them by taking

$$B^+ = (1 - \phi)B_{\text{BFGS}}^+ + \phi B_{\text{DFP}}^+ \quad (22.1)$$

for some fixed ϕ where B_{BFGS}^+ and B_{DFP}^+ denote the matrices obtained by the BFGS and DFP update rules, respectively. Then

$$\begin{aligned} B^+ &= B_{\text{BFGS}}^+ + \phi(B_{\text{DFP}}^+ - B_{\text{BFGS}}^+) \\ &= B_{\text{BFGS}}^+ + \phi\left(B - \frac{Bss^\top B}{s^\top Bs} - \left(I - \frac{ys^\top}{s^\top y}\right) B \left(I - \frac{sy^\top}{s^\top y}\right)\right) \\ &= B_{\text{BFGS}}^+ + \phi(s^\top Bs) \left(\frac{y}{y^\top s} - \frac{Bs}{s^\top Bs}\right) \left(\frac{y}{y^\top s} - \frac{Bs}{s^\top Bs}\right)^\top. \end{aligned}$$

The Broyden class is the family of up update rules given by (22.1) for any ϕ . Of course, the BFGS and DFP updates belong to the Broyden class, corresponding to $\phi = 0$ and $\phi = 1$, respectively. In fact, the SR1 update is also in the Broyden class, as it corresponds to

$$\phi = \frac{y^\top s}{y^\top s - s^\top Bs}.$$

2.5 Convergence of quasi-Newton methods

As for Newton's method, we assume the following conditions to guarantee convergence of quasi-Newton methods.

- f is twice continuously differentiable.
- f is m -strongly convex and M -smooth in the ℓ_2 norm.
- The Hessian of f is L -Lipschitz continuous in the ℓ_2 norm.

It is proved that both BFGS and DFP with backtracking line search guarantee

$$\|x_{t+1} - x^*\|_2 \leq c_t \|x_t - x^*\|_2$$

where $c_t \rightarrow 0$ as $t \rightarrow \infty$. Remember that gradient descent guarantees

$$\|x_{t+1} - x^*\|_2 \leq \gamma \|x_t - x^*\|_2$$

for some “fixed” γ . Hence, quasi-Newton methods result in faster convergence. At the same time, each iteration requires $O(d^2)$ time, which is smaller than $O(d^3)$.