QUASI-NEWTON METHODS

Rank one and rank two updates

Newton method

- For unconstraint minimization
- To minimize $f(\theta)$ which is convex and twice differentiable
- Iterate by

$$\theta_{n+1} = \theta_n - H^{-1} \nabla f(\theta)$$

- advantages: simple to apply, fast convergence
- disadvantages: local convergence, requires second derivatives, solution of linear equation

Quasi-Newton

- Instead of the true Hessian, an initial matrix H_0 is chosen (usually $H_0 = I$) which is subsequently updated by an update formula: $H_{n+1} = H_n + H_n^u$
- This updating can also be done with the inverse of the Hessian H^{-1} as follows: Let $B = H^{-1}$; then the updating formula for the inverse is also of the form $B_{n+1} = B_n + B_n^u$

- *Big question: What is the update matrix?*

Secant Condition

- Quasi-Newton updates satisfy

$$H_{n+1}(\theta_{n+1}-\theta_n) = \nabla f(\theta_{n+1})-\nabla f(\theta_n)$$

- Interpretation
- define second-order approximation at θ_{n+1}

$$f_{quad}(z) = f(\theta_{n+1}) + df(\theta_{n+1})(z-\theta_{n+1}) + \frac{1}{2}(z-\theta_{n+1})^t H_{n+1}(z-\theta_{n+1})$$

- secant condition implies that gradient of f_{quad} agrees with the gradient of f at θ_n
- Let $B = H^{-1}$, then the secant condition becomes

$$\theta_{n+1}-\theta_n = B_{n+1}(\nabla f(\theta_{n+1})-\nabla f(\theta_n))$$

Rank one and rank two updates

- Let $B_{k+1} = B_k + B_k^u$, $g_n = \nabla f(\theta_{n+1}) - \nabla f(\theta_n)$, and $d_n = \theta_{n+1} - \theta_n$, the condition becomes

$$d_n = B_n g_n + B_n^u g_n \ (*)$$

- A general form of solution is $B_n^u = a\ uu^t + b\ vv^t$, where a and b are scalars, and u and v are vectors satisfying (*)
- b = 0: rank one updates
- b ≠ 0: rank two updates – BFGS, DFP

Rank-One Quasi-Newton Method

- Secant condition: $\nabla f(\theta_{n+1}) - \nabla f(\theta_n) = H_{n+1}(\theta_{n+1} - \theta_n)$
- Update to H_n:

$$H_{n+1} = H_n + a_n u_n u_n^t,$$

where constant c_n and vector v_n are determined by

$$a_n = -\frac{1}{(H_n d_n - g_n)^t d_n}, u_n = H_n d_n - g_n.$$

- When $(H_n d_n - g_n)^t d_n$ is too close to 0,
  - Either H_n is retained for H_{n+1},
  - Or use trust region strategy:
    - Minimize quadratic approxiamtion to $f(\theta)$ subject to spherical constraint $\|\theta - \theta_n\|^2 \leq r^2$ for a fixed radius r.
    - Has a solution regardless of whether H_n is positive definite.
    - Prevent absurdly large steps in the early stages of minimization.

Backtrack

- Hereditary positive definiteness: positive definiteness is guaranteed to be transferred from one iteration to the next.

$$H_{n+1} = H_n + a_n u_n u_n^t,$$

- If H_n is positive definite and $a_n \geq 0$, then H_{n+1} will be positive definite.
- If $a_n < 0$, then it may be necessary to backtrack
  - Shrink a_n towards 0 until positive definiteness is achieved.

Broyden-Fletcher-Goldfarb-Shanno (BFGS) update

- BFGS update

$$H_{n+1} = H_n + \frac{g_n g_n^t}{g_n^t d_n} - \frac{H_n d_n d_n^t H_n}{d_n^t H_n d_n}$$

- where $g_n = \nabla f(\theta_{n+1}) - \nabla f(\theta_n)$, $d_n = \theta_{n+1} - \theta_n$
- Inverse update

$$B_{n+1} = (I - \frac{d_n g_n^t}{g_n^t d_n}) B_n (I - \frac{g_n d_n^t}{g_n^t d_n}) + \frac{d_n d_n^t}{g_n^t d_n}$$

- Note that $g_n{}^t d_n > 0$ for strictly convex f

Positive Definiteness

- If $g_n{}^t d_n > 0$, BFGS update preserves positive definiteness of H_n
- proof: from inverse update formula

$$x^t H_{n+1}^{-1} x = (x - \frac{d_n^t x}{d_n^t g_n} g_n)^t H_n^{-1} (x - \frac{d_n^t x}{d_n^t g_n} g_n) + \frac{(d_n^t x)^2}{g_n^t d_n}$$

- If $H_n^{-1} \succ 0$, both terms are nonnegative for all x
- Second term is zero only if $d_n^t x = 0$; the first term is zero only if $x = 0$
- This ensures that $\Delta\theta = -H_n^{-1} \nabla f(\theta_n)$ is a descent direction

Convergence

global result

- if f is strongly convex, BFGS with backtracking line search converges from any θ_0 and $H_0 \succ 0$

Local convergence

- If *f* is strongly convex and $df^2(\theta)$ is Lipschitz continuous, local convergence is **superlinear:** for sufficiently large *n*,

$$\|\theta_{n+1} - \theta^*\|_2 \leqslant c_n \|\theta_n - \theta^*\|_2 \to 0$$

- where $c_n \to 0$

Quasi-Newton Algorithm

given starting point θ_0 and $H_0 \succ 0$

For n = *1, 2, . . .,* until a stopping criterion is satisfied

1. compute quasi-Newton direction $\Delta\theta = -H_n^{-1}\nabla f(\theta_n)$
2. determine step size *t* (e.g., by backtracking line search)
3. Compute $\theta_{n+1} = \theta_n + t\Delta\theta$
4. Compute update matrix according to a given formula, and update H_n or H_n^{-1}

Comments

Initialization

- True Hessian
- aI, where a is in the range of the eigenvalues of the true Hessian

Pros and Cons

- Avoid calculation of second derivatives
- Simplify computation of search direction
- Global convergence even with inexact line searches
- Quadratic convergence of Newton's Method is lost
- Can get stuck on a saddle point