Optimization

Pt. 3 Quasi-Newton Methods

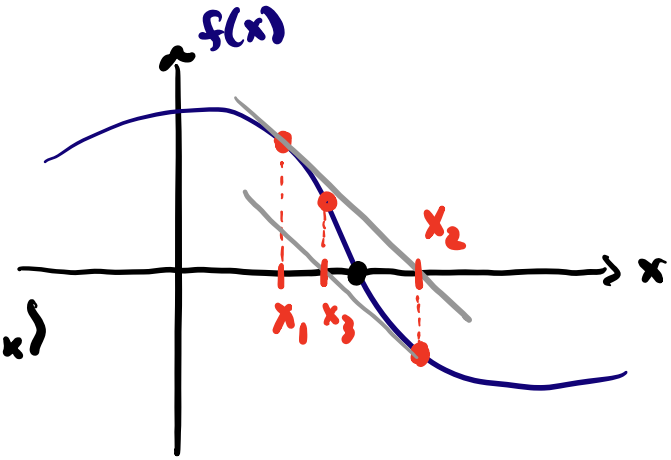
Recap Newton's Method (in $d=1$)

Solve $f(x)=0$

Given x_0

for $k=1,2,3,\ldots$

$$x_{k+1} = x_k - f(x_k)/f'(x_k)$$



For optimization (critical pts of $f(x)$), run Newton to find $f'(x)=0$:

$$x_{k+1} = x_k - f'(x_k)/f''(x_k)$$

$\Rightarrow$ "Second Order" Method

Typically converges quadratically (for $f(x)=0$)

(1) f'' continuous

(2) $f'(x) \neq 0$

(3) x_0 suff. close to root x

Newton's Method in $d \geq 1$

$f: \mathbb{R}^d \to \mathbb{R}^d$, find $f(x) = 0$

$$f(x + \Delta x) = f(x) + J_f(x)\Delta x + o(\|\Delta x\|)$$

$$\Rightarrow \quad \Delta x \approx J_f^{-1}(x)\left[f(x+\Delta x) - f(x)\right]$$

$= 0$ when $x + \Delta x$ is root of linear approx

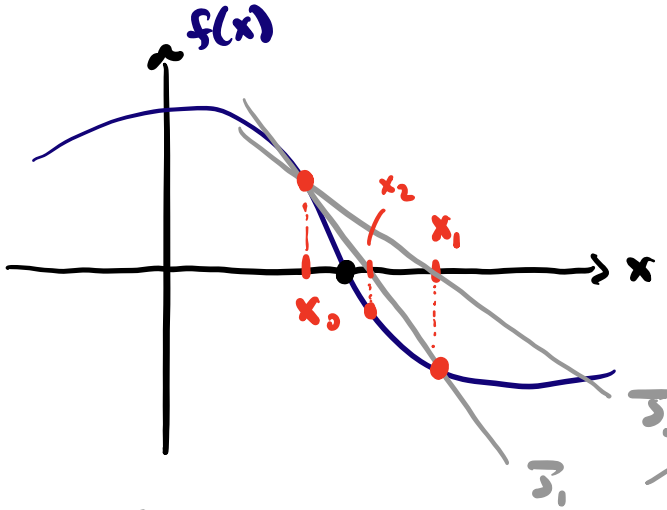
Newton step $\Rightarrow \quad x_{k+1} = x_k + J_f^{-1}(x_k) f(x_k)$

Note: adapts to non-square systems via pseudoinverse or least-squares solve.

Quasi-Newton Methods

Computing the Jacobian and solving $J_f(x)\Delta x_k = f(x_k)$ at each step is not always computationally tractable, e.g. when d is very large or computing $F(x)$ is very expensive and J_f is not known analytically.

Idea: approximate $J_k \approx J_f(x_k)$ on the fly, using cheap updates at each iteration.



Given x_0, J_0

for $k = 1, 2, 3, \ldots$

x_{k+1} = Newton update

J_{k+1} = secant between $f(x_k)$ and $f(x_{k+1})$.

- If $x_k \to x$, secant line becomes better and better approx to tangent line

- Secant line requires no explicit knowledge of derivatives of f.

Secant eq. $$\underbrace{f(x_{k+1}) - f(x_k)}_{\Delta f_k} = J_{k+1}\underbrace{(x_{k+1} - x_k)}_{\Delta x_k}$$

for $d > 1$, this does not uniquely determine J_{k+1}, so we need to impose more constraints.

"No-change" eq. $J_{k+1} q = J_k q$ when $q^T \Delta x_k = 0$.

These uniquely determine J_{k+1} from J_k:

$$(J_{k+1} - J_k)q = u \Delta x^T q \quad \text{for some } u$$

$\Rightarrow$ No-change satisfied

$$\Delta f_k = [J_k + u \Delta x^T] \Delta x = J_k \Delta x + u \Delta x^T \Delta x$$

$$u = (\Delta f_k - J_k \Delta x) / \Delta x^T \Delta x$$

$\Rightarrow$ Secant satisfied

$$J_{k+1} = J_k + \underbrace{(\Delta f_k - J_k \Delta x) \frac{\Delta x^T}{\Delta x^T \Delta x}}_{\text{rank-1 update to } J_k}$$

Can show that this choice of J_k satisfies

$$\text{argmin} \; \| J_{k+1} - J_k \|_F$$

$$\text{s.t. secant cond.}$$

"Minimal info update" = minimal norm soln to $\Delta f_k = J_{k+1} \Delta x_k$

Pseudocode (simplest form)

Given x_0, J_0, $f_0 = f(x_0)$

for $k = 1, 2, 3, \ldots$

$$x_{k+1} = x_k - J_k^{-1} f_k$$

$$J_{k+1} = J_k + (\Delta f_k - J_k \Delta x_k) \frac{\Delta x_k^T}{\Delta x_k^T \Delta x_k}$$

Since we actually need J_k^{-1} at each step, we can use the Sherman-Morrison formula to update J_{k+1}^{-1} from J_k^{-1} fast.

$$J_{k+1}^{-1} = J_k^{-1} + (\Delta x_k - J_k^{-1} \Delta f_k) \frac{\Delta x_k^T J_k^{-1}}{\Delta x_k^T J_k^{-1} \Delta f_k}$$

This is called Broyden's first update.

Broyden's second update comes from applying secant and no-change conditions directly to $G_{k+1} = J_{k+1}^{-1}$.

Secant: $\Delta x_k = G_{k+1} \Delta f_k$

No-Change: $G_{k+1} q_k = G_k q_k$ when $q_k^T \Delta f_k = 0$

$$\Rightarrow G_{k+1} = G_k + (\Delta x_k - G_k \Delta f_k) \frac{\Delta f_k^T}{\Delta f_k^T \Delta f_k}$$

The 2nd update minimizes

$$\text{argmin} \|G_{k+1} - G_k\|_F$$

s.t. Secant condition

BFGS Update

In optimization, consider minimizing

$$f: \mathbb{R}^d \to \mathbb{R}.$$

Run Newton on ∇f using Hessian H:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix} \qquad H = \begin{bmatrix} \frac{\partial^2 f}{\partial^2 x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & & \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial^2 x_d} \end{bmatrix}$$

$\Rightarrow$ H is symmetric when f is 2x cont. diff.

$\Rightarrow$ At smooth local min, H is SPD
(analogous to $f''(x) > 0$ in 1D)

Notice that Broyden updates are <u>not</u> symmetric. Since we want approximations $H_k \to H$ when $x_k \to x$ (local min), can we adapt Broyden updates to keep H_k PSD?

<u>Idea:</u> Use symmetric low-rank update to enforce symmetry while satisfying secant condition.

$$\Delta g_k = \nabla f(x_{k+1}) - \nabla f(x_k), \quad \Delta x_k = x_{k+1} - x_k$$

Secant eq. $\quad H_{k+1} \Delta x_k = \Delta g_k$

Symm. Rank 2 update: $\quad H_{k+1} = H_k + \alpha u u^T + \beta v v^T$

with $u = \Delta g_k$ and $v = H_k \Delta x_k$, i.e.

$$\Delta H_k = H_{k+1} - H_k = \alpha \underbrace{\Delta g_k (\Delta g_k)^T}_{\text{symm. type 2 update}} + \beta \underbrace{H_k \Delta x_k (H_k \Delta x_k)^T}_{\text{symm. type 2 update}}$$

Choose α, β to satisfy secant eqn.

$$\Delta g_k = H_{k+1} \Delta x_k = H_k \Delta x_k + \alpha \Delta g_k (\Delta g_k)^T \Delta x_k$$

$$+ \beta H_k \Delta x_k \Delta x_k^T H_k \Delta x_k$$

$\Rightarrow$ need $\alpha = [(\Delta g_k)^T \Delta x_k]^{-1}$, $\beta = -[\Delta x_k^T H_k \Delta x_k]^{-1}$

to get $\Delta g_k = H_k \Delta x_k + \Delta g_k - H_k \Delta x_k$ ✓

$$H_{k+1} = H_k + \frac{\Delta g_k \Delta g_k^T}{\Delta g_k^T \Delta g_k} - \frac{H_k \Delta x_k \Delta x_k^T H_k}{\Delta x_k^T H_k \Delta x_k}$$

We can also use the Sherman-Morrison-Woodbury formula to get a fast rank-2 update formula directly for $H_k^{-1} \to H_{k+1}^{-1}$.

One can also show that BFGS update is PSD under appropriate restrictions on Δx_k.