Need approximation B_k of $\nabla^2 f(x_k)$ based on the past $(x_i, \nabla f(x_i))$.

(1) B_k is symmetric

(2) $m_k(x_k) = f(x_k)$, $\nabla m_k(x_k) = \nabla f(x_k)$

$B_k s_k = y_k \iff$ (3) $\nabla m_k(x_{k-1}) = \nabla f(x_{k-1})$ ← Model should capture curvature we observed.

(4) $B_k \succ 0$

(5) Want "cheap updates" for B_k from B_{k-1} (namely $O(d^2)$)

Note $\nabla m_k(x) = \overset{f(x_k)}{\downarrow} g_k^T(x - x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k)$

By (2) $= f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(\quad)^T B_k (\quad)$

$\Rightarrow \nabla m_k(x_{k-1}) = \nabla f(x_k) + B_k(x_{k-1} - x_k) = \nabla f(x_{k-1})$ by (3)

$\Rightarrow B_k(x_{k-1} - x_k) = \nabla f(x_{k-1}) - \nabla f(x_k)$

$\Leftrightarrow B_k s_k = y_k$ **The Secant Equation**

$\frac{d(d+1)}{2}$ unknowns, d constraints

where $s_k = x_k - x_{k-1}$ "run"

$y_k = \nabla f(x_k) - \nabla f(x_{k-1})$. "rise"

We pick B_k having seen s_k, y_k.

Our steps need $-B_k^{-1}\nabla f(x_k)$, really, we want cheap updates B_k^{-1} from B_{k-1}^{-1}.

Lemma (Sherman-Morrison) For any invertible $A \in \mathbb{R}^{d\times d}$ and vectors $u, v \in \mathbb{R}^d$, If $1+v^TA^{-1}u \neq 0$, then $(A+uv^T)$ is invertible with

$$(A+uv^T)^{-1} = A^{-1} - \frac{(A^{-1}u)(A^{-T}v)^T}{1+v^TA^{-1}u}$$

Rank k updates work the same (Woodbury Identity).

(5a) $B_k - B_{k-1}$ to be rank one (meaning $= uv^T$)

If B_k comes out positive definite, then

$$y_k^T s_k > 0 \qquad \text{(why? } B_k s_k = y_k \Rightarrow s_k^T B_k^T s_k = y_k^T s_k > 0\text{)}$$

$$\| \\ (\nabla f(x_k) - \nabla f(x_{k-1}))^T (x_k - x_{k-1}) > 0$$

"monotone gradient"

$\Uparrow$

f being strongly convex.

If f is nonconvex, we need algorithmic tricks to get $y_k^T s_k > 0$ (linesearches).

6. Update Formulas for Quasi Newton

Assume (1) - (4), (5a).

$$B_{k+1} = B_k + uv^T \qquad \text{(by (5a))}$$

B_{k+1}, B_k are symmetric (by (1))

$\Rightarrow B_{k+1} - B_k$ is symmetric

$\Rightarrow uv^T$ is symmetric.

$\Rightarrow uv^T = \alpha \cdot ww^T$ for some $w \in \mathbb{R}^d$, $\alpha \in \mathbb{R}$

$\Rightarrow B_{k+1} = B_k + \alpha ww^T$.

(3) $\Rightarrow B_{k+1} s_{k+1} = y_{k+1}$

Case 1. If $B_k s_{k+1} = y_{k+1}$, then $w = 0$.

Case 2. If $B_k s_{k+1} \neq y_{k+1}$,

then $(B_k + \alpha ww^T) s_{k+1} = y_{k+1}$

$\Rightarrow B_k s_{k+1} - y_{k+1} = -\alpha w(w^T s_{k+1})$

$\Rightarrow w = \dfrac{B_k s_{k+1} - y_{k+1}}{-\alpha w^T s_{k+1}} = \beta \cdot (B_k s_{k+1} - y_{k+1})$

for some β

$$B_{k+1} = B_k + \underbrace{\beta_k^2}_{\text{LW}} \underbrace{=}_{\gamma} (B_k s_{k+1} - y_{k+1})(B_k s_{k+1} - y_{k+1})^T$$

(3) $\Rightarrow$ $B_k s_{k+1} + \gamma (B_k s_{k+1} - y_{k+1})(B_k s_{k+1} - y_{k+1})^T s_{k+1} = y_{k+1}$

$\Rightarrow (1 + \gamma (B_k s_{k+1} - y_{k+1})^T s_{k+1}) B_k s_{k+1}$

$$- (\gamma (B_k s_{k+1} - y_{k+1})^T s_{k+1} + 1) y_{k+1} = 0$$

$$\Rightarrow \gamma = \frac{-1}{(B_k s_{k+1} - y_{k+1})^T s_{k+1}}$$

$$\Rightarrow B_{k+1} = B_k - \frac{(B_k s_{k+1} - y_{k+1})(B_k s_{k+1} - y_{k+1})^T}{(B_k s_{k+1} - y_{k+1})^T s_{k+1}}$$

$-$ ↑ $+?$; $(B_k s_{k+1} - y_{k+1}) \neq 0$; $s_{k+1} \neq 0$

Symmetric Rank One update (SR1).

This is the unique way to satisfy (1) - (5a). ↑ (if a way exists)

B_{k+1} might become negative definite

(or worse we divide by zero).

$\Rightarrow$ We don't have (4) necessarily.

For example, suppose $B_0 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, and observe

$$s_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad y_1 = \begin{pmatrix} -3 \\ 2 \end{pmatrix}$$

$s_1^T y_1 > 0$ "monotone curvature" "look convex"

SR1 update $\Rightarrow B_1 = \begin{pmatrix} 2 & 1 \\ 1 & -3 \end{pmatrix}$ which is indefinite.

Next step might not descend.

Theorem Suppose f is C^2, $x_k \to x^*$, $\nabla^2 f$ is bounded and Lips.

Further, suppose s_k are uniformly linearly independent.

Then if $\left|(y_{k+1} - B_k s_{k+1})^T s_{k+1}\right| > \gamma \left| y_{k+1} - B_k s_{k+1} \right| \left| s_{k+1} \right|$ for some $\gamma \in (0,1)$,

$$\lim B_k = \nabla^2 f(x^*) \quad \text{(under SR1)}.$$

We are stuck on rank one updates.

Look for rank two updates.

Lemma U is a symmetric rank two matrix iff

$$U = \beta v v^T + \delta w w^T$$

for some $\beta, \delta \neq 0$, v, w linearly independent.

Want $B_{k+1} - B_k = U$ ← rank two, symmetric (5b)

(previously for rank one, $\omega = B_k s_{k+1} - y_{k+1}$)

Guess a good path

$u = y_{k+1}$

$\omega = B_k s_{k+1}$

Then (3) $\Rightarrow B_{k+1} s_{k+1} = y_{k+1}$

$$\Rightarrow \left(B_k + \beta y_{k+1} y_{k+1}^T + \delta (B_k s_{k+1})(B_k s_{k+1})^T\right) s_{k+1} = y_{k+1}$$

for some β, δ

$$\Rightarrow B_k s_{k+1} \underbrace{\left(1 + \delta\, s_{k+1}^T B_k^T s_{k+1}\right)}_{=0} + y_{k+1} \underbrace{\left(\beta y_{k+1}^T s_{k+1} - 1\right)}_{=0} = 0$$

$$\Rightarrow \delta = \frac{-1}{s_{k+1}^T B_k s_{k+1}} \qquad \Rightarrow \beta = \frac{1}{y_{k+1}^T s_{k+1}}$$

>0 if B_k is p.s.d. $\Leftrightarrow$ (4)

>0 if enantone condition.

$$\Rightarrow B_{k+1} = B_k + \frac{y_{k+1} y_{k+1}^T}{y_{k+1}^T s_{k+1}} - \frac{B_k s_{k+1} (B_k s_{k+1})^T}{s_{k+1}^T B_k s_{k+1}}$$

BFGS Update (1970, 4 way invented)

B_k is positive definite, $B_k \succeq 0 \Leftrightarrow$ positive semidefinite.

Lemma If $B_k \succ 0$ and $s_{k+1}^T y_{k+1} > 0$, then

$B_{k+1} \succ 0$ (under BFGS update)

Proof. HW5 Q2(a).

$\Rightarrow$ BFGS satisfies (1)-(5b)

Not unique, For example, instead of y_{k+1}, $B_k s_{k+1}$,

pick $B_k^{-1} y_k$, s_{k+1} (another fine choice for building a rank two solution)

(1)-(3) $\Rightarrow$ Davidon-Fletcher-Powell Update (DFP)

$$B_{k+1} = \left(I - \frac{y_{k+1} s_{k+1}^T}{y_{k+1}^T s_{k+1}}\right) B_k \left(I - \frac{s_{k+1} y_{k+1}^T}{s_{k+1}^T y_{k+1}}\right) + \frac{y_{k+1} y_{k+1}^T}{y_{k+1}^T s_{k+1}}.$$

Infinitely many valid rank two updates.

What is the "best" rank two update?

Pick optimally.

Lemma. This also preserves positive definiteness

Proof. HW5 Q2(b).

One approach. Pick B_{k+1} that keeps information from B_k

min the relative entropy between $N(0, B_{k+1})$ and $N(0, B_k)$.

$$\min \quad \underbrace{\operatorname{tr}(B_k^{-1} X)}_{=\sum \lambda_i} - \underbrace{\log \det (B_k^{-1} X)}_{=\prod \lambda_i} - n$$

$$\approx \sum (\lambda_i - \log(\lambda_i) - 1)$$

$$\text{s.t.} \quad X s_{k+1} = y_{k+1}.$$

Obj minimizes at $X = B_k$ (although not feasible)

with value zero

This convex in X, minimizers are just the BFGS update.