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

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

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

$$\text{s.t.} \quad X s_{k+1} = y_{k+1}.$$
$$X \succ 0$$

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

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

[KKT Conditions give BFGS Update]

Entropy is not symmetric. Instead minimize the entropy of previous $N(0, B_k)$ from $N(0, B_{k+1})$

$$\min \quad tr(X^{-1} B_k) - \log\det(X^{-1} B_k) - n$$
$$\text{s.t.} \quad X s_{k+1} = y_{k+1} \quad (\Leftrightarrow X^{-1} y_{k+1} = s_{k+1})$$
$$X \succ 0 \quad (\Leftrightarrow X^{-1} \succ 0)$$

Convex in X^{-1}. Optimal solution DFP Update. (Dual to BFGS).

Entropies are trickier, lets pick based on some matrix norm.

$$\min \quad \|B - B_k\|$$
$$\text{s.t.} \quad B = B^T, \quad B \succ 0, \quad B s_{k+1} = y_{k+1}$$

Pick $\|A\| = \|W^{1/2} A W^{1/2}\|_F$ for any $W y_{k+1} = s_{k+1}$.

inverse of something solving secant equation.

Then DFP update minimizes this.

We mainly need B_k^{-1}, $p_k = -B_k^{-1} \nabla f(x_k)$.

$$\min \quad \|B^{-1} - B_k^{-1}\|$$
$$\text{s.t.} \quad B = B^T, \quad B \succ 0, \quad B s_{k+1} = y_{k+1}.$$

Then minimized by BFGS (for any W).

$B^{-1} = (\text{BFGS})^{-1}$

We need $y_{k+1}^T s_{k+1} > 0$ every step.

$$\Updownarrow$$

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^T \underbrace{(x_{k+1} - x_k)}_{\alpha_k p_k} > 0$$

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

$$\alpha_k \Big(\text{directional der at } x_{k+1} \text{ in } p_k - \underbrace{\text{directional der at } x_k \text{ in } p_k}_{\substack{\text{negative since } p_k \\ \text{is a descent direction.}}}\Big) > 0$$

If $\alpha_k = \min_{\alpha} f(x_k + \alpha p_k)$ "exact linesearch",

then (1st order optimality) $\frac{d}{d\alpha} f(x_k + \alpha p_k) = 0$

$$\Leftrightarrow \nabla f(x_k + \alpha_k p_k)^T \alpha p_k = 0$$

$$\Leftrightarrow \nabla f(x_{k+1})^T p_k = 0.$$

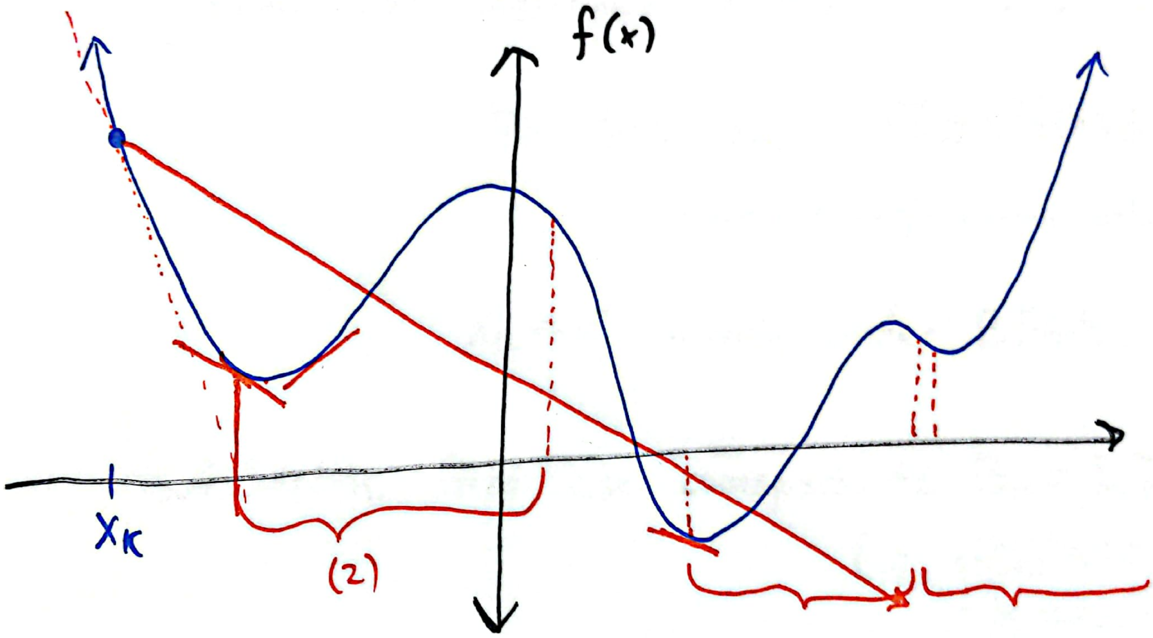
$$\Rightarrow y_{k+1}^T s_{k+1} > 0 \quad (\text{in particular } = p_k^T \nabla f(x_k))$$

<u>Theorem</u> (Thm 5.28 of Rusczynski...)

If use exact linesearches, then BFGS and DFP will produce the same sequence of iterates.

This motivates a new linesearch criteria : Wolfe Conditions

$$
\begin{cases}
(1) & f(x_k + \alpha p_k) \le f(x_k) - \eta \alpha \, p_k^T \nabla f(x_k) \, , \quad \eta \in (0,1) \\
(2) & \nabla f(x_k + \alpha p_k)^T p_k \ge c \cdot \underbrace{\nabla f(x_k)^T p_k}_{<0} \, , \quad c \in (\eta, 1)
\end{cases}
$$



<u>Lemma</u>. There exist intervals satisfying both Wolfe Conditions for any C^1 function, bounded below.

$\Rightarrow$ We have a well-defined Quasi-Newton.

7. Quasi-Newton Convergence Rates

Full details/proofs Nocedal+Wright Section 6.4

We are like gradient descent...

Theorem (6.5) Assuming $B_0 \succ 0$, f is C^2,

$\forall x, f(x) \le f(x_0), \ \mu I \preceq \nabla^2 f(x) \preceq LI$

(smooth/strongly convex below x_0),

$x_k \to x^*$ linearly under BFGS.

Proof Sketch. Grad Desc coorsponds to $B_k = LI$.

$(q_k = -B_k^{-1} \nabla f(x_k)$
$= 1/L \nabla f(x_k))$

Show B_k from BFGS stays similar to LI

by examing

$$\psi((LI)^{-1}B_k) = tr((LI)^{-1}B_k) - \log\det((LI)^{-1}B_k) - n.$$

Inductively look at the change in this

we see the angles $\theta_k = \angle p_k, q_k$

$p_k = -B_k^{-1}\nabla f(x_k), \ q_k = \frac{-1}{L}\nabla f(x_k)$

have $\cos\theta_k$ bounded away from zero.

$\Rightarrow$ Follow similar linear rate as GD. $\square$

We are like Newton's Method...

<u>Theorem (6.6)</u> Assuming $B_0 \succ 0$, and $x_k \to x^*$
$\nabla f(x^*) \succ 0$
$\nabla^2 f(x)$ is Lipschitz near x^*,

then $x_k \to x^*$ superLinearly (under BFGS)

Proof Sketch. We need B_k to really be acting similar to $\nabla^2 f(x^*)$ as we converge.

In particular, we need $p_k = -B_k^{-1} \nabla f(x_k)$ to "converge" to $\underbrace{-\nabla^2 f(x^*)^{-1} \nabla f(x_k)}_{= q_k}$

Inductively examining the change in

$$\psi(\nabla^2 f(x^*)^{-1} B_k) = tr(\cdot) - \log\det(\,) - n.$$

shows $\theta_k = \langle p_k, q_k \to 0$. $\cos\theta_k \to 1$.

$\Rightarrow$ Our steps converge to Newton's trajectory. $\square$

8. Practical Improvements

+ Only need $\nabla f()$ at each step

+ Superlinear convergence

- Only work up to $10^4 \sim 10^5$ for d
  (need to store B_k^{-1}, size $O(d^2)$).
  (1^{st} order methods work $10^6 \sim 10^9$)

Solution: Limited Memory Quasi-Newton / BFGS
(LBFGS)

Restarted Method. Pick $m \in [2, 30]$. $B_0 = I$
$B_0^{-1} = I$

$$B_k^{-1} = (B_0^{-1} + w_1 w_1^T + w_2 w_2^T + \ldots + w_{2k} w_{2k}^T)$$

(works $10^6 - 10^8$)

$$B_k^{-1} \nabla f(x_k) = \nabla f(x_k) + w_1 (w_1^T \nabla f) + w_2 (w_2^T \nabla f) \ldots$$

looks like conjugate directions up next.

After m steps, restart method. $O(dm)$ memory

Alternatively, track last m updates, drop old ones.