$\boxed{\text{Option 2}}$ Move small values to ε, Keep large negative eigenvalues, but make them positive.

Compute $\nabla f(x_k)$, $\nabla^2 f(x_k) = V \Lambda V^T$

Pick $\varepsilon > 0$

$$\bar{\Lambda} = \operatorname{diag}(\bar{\lambda}), \quad \text{where } \bar{\lambda}_i = \begin{cases} \lambda_i & \text{if } \lambda_i \geq \varepsilon \\ \varepsilon & \text{if } -\varepsilon \leq \lambda_i \leq \varepsilon \\ -\lambda_i & \text{if } \lambda_i \leq -\varepsilon \end{cases}$$

$$B_k = V \bar{\Lambda} V^T \succ 0.$$

$$\Rightarrow p = -B_k^{-1} \nabla f(x_k)$$

$$= -\left(\begin{pmatrix} V_+ & V_\varepsilon & V_- \end{pmatrix} \begin{pmatrix} \Lambda_+ & & \\ & \varepsilon I & \\ & & -\Lambda_- \end{pmatrix} \begin{pmatrix} V_+^T \\ V_\varepsilon^T \\ V_-^T \end{pmatrix} \right)^{-1} \nabla f(x_k)$$

$$= -\begin{pmatrix} V_+ & V_\varepsilon & V_- \end{pmatrix} \begin{pmatrix} \Lambda_+^{-1} & & \\ & \frac{1}{\varepsilon} I & \\ & & -\Lambda_-^{-1} \end{pmatrix} \begin{pmatrix} V_+^T \nabla f(x_k) \\ V_\varepsilon^T \nabla f(x_k) \\ V_-^T \nabla f(x_k) \end{pmatrix}$$

$$= \underbrace{-V_+ \Lambda_+^{-1} V_+^T \nabla f(x_k)}_{\substack{\text{descent from Newton} \\ \text{in pos eigendirections} \\ \text{of } \nabla^2 f(x_k)}} \underbrace{- \frac{1}{\varepsilon} V_\varepsilon V_\varepsilon^T \nabla f(x_k)}_{\substack{\text{grad descent} \\ \text{in "null space"} \\ \text{of } \nabla^2 f(x_k)}} + \underbrace{V_- \Lambda_-^{-1} V_-^T \nabla f(x_k)}_{\substack{\text{negative of the} \\ \text{ascent dir of Newton} \\ \text{in neg eigendirections} \\ \Rightarrow \text{descent}}}$$

$\boxed{\text{Option 2.1 (Annapurna)}}$

$$B_k = \underbrace{V \bar{\Lambda} V^T}$$

$$\Lambda = \text{diag}(\bar{\lambda}), \quad \bar{\lambda}_i = \begin{cases} \lambda_i & \text{if } \lambda_i \geq \varepsilon \\ M & \text{if } \lambda_i < \varepsilon \end{cases}$$

for $M >> 0$.

As $M \to \infty$, $p = -B_k^{-1} \nabla f(x_k)$

$$\to -V_+ \Lambda_+^{-1} V_+^T \nabla f(x_k)$$

$\boxed{\text{Option 3}}$ Avoid Spectral Decomposition, fix smallest eigenvalue.

Compute $\lambda_{min}(\nabla^2 f(x_k))$

Pick $\varepsilon > 0$

If $\lambda_{min} > \varepsilon$, then $B_k = \nabla^2 f(x_k)$ (since $\nabla^2 f(x_k) \succeq \varepsilon I$)

else $B_k = \nabla^2 f(x_k) + \gamma I$

$\gamma = \varepsilon - \lambda_{min}(\nabla^2 f(x_k))$.

$\Rightarrow B_k \succeq \varepsilon I \succ 0$.

$\Rightarrow$ Descent from lemma in $p = -(\nabla^2 f(x_k) + \gamma I)^{-1} \nabla f(x_k)$

As $\gamma \to 0$, $p \to p^N$ (Newton step)

As $\gamma \to \infty$, $\frac{p}{\|p\|} \to \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$. (Grad dir)

3. Convergence Gaurontees

When $\nabla^2 f(x_k) \succeq \varepsilon I$, all of these have $B_k = \nabla^2 f(x_k)$

$\Rightarrow$ Newton's Method.

So local quadratic convergence still holds.
(assuming strong convexity)

For global guarantees, we need a descent lemma.

Lemma (HW5) Suppose ∇f is L-Lipschitz and
$x_{k+1} = x_k - \alpha B_k^{-1} \nabla f(x_k)$.

If $B_k \succ 0$, then

$$f(x_{k+1}) \leq f(x_k) - \left(\frac{\alpha}{\lambda_{max}(B_k)} - \frac{L\alpha^2}{2\lambda_{min}^2(B_k)} \right) \|\nabla f(x_k)\|_2^2$$

[Recovers old lemma when $B_k = I$]

$\Rightarrow$ Backtracking works. Some exponential works.

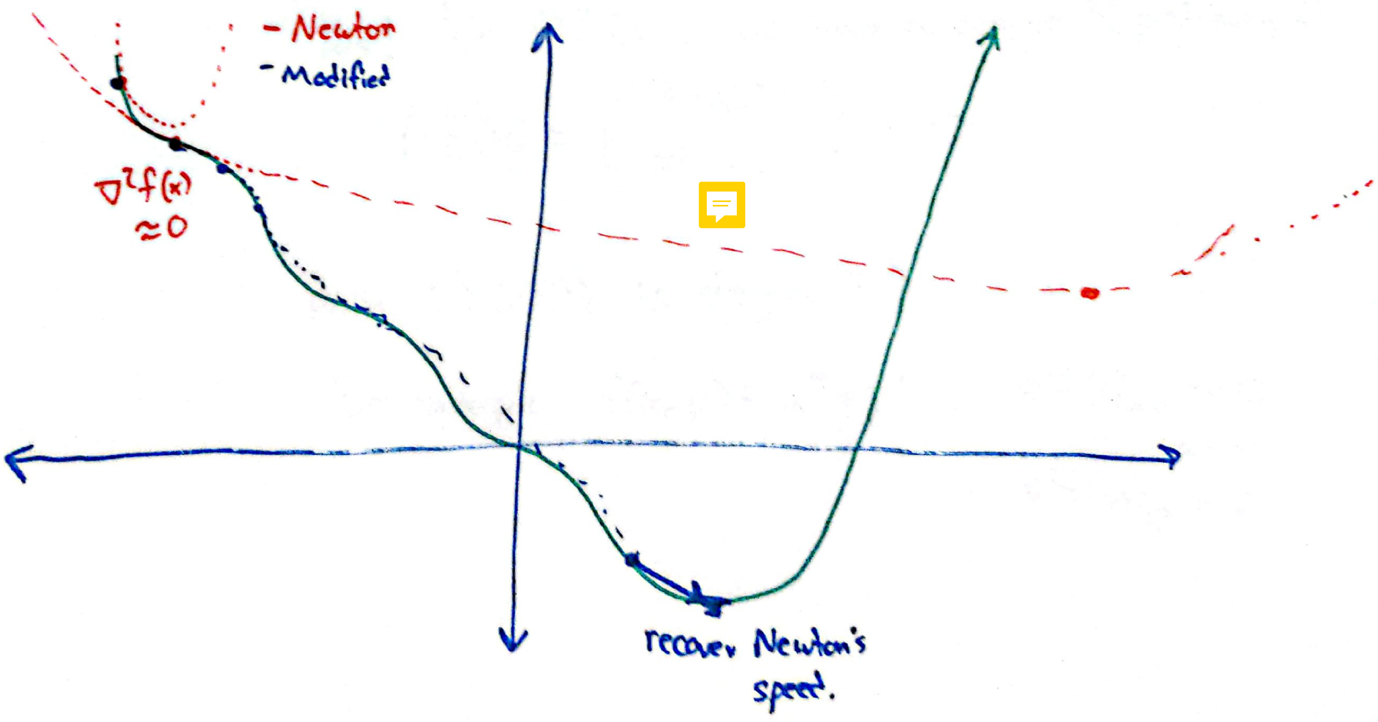
$\Rightarrow f(x_{k+1}) \leq f(x_k) - C \|\nabla f(x_k)\|_2^2 \leq f(x_0) - \sum_{i=0}^{k} C \|\nabla f(x_i)\|_2^2$

<u>Theorem</u> If f is C^1 with L-Lipschitz gradient, $\min f > \infty$, and B_k has eigenvalues bounded away from 0 and ∞, then there exists a constant M s.t.

$$\min_{i \le k} \|\nabla f(x_i)\| \le \frac{M}{\sqrt{k}} .$$

[Matchs Gradient Descent, essentially some proof].

$\Rightarrow$ Modified Methods converge globally, slowly, but if we approach some strict local min ($\nabla^2 f(x^*) \succ 0$), then we get Newton's fast quadratic convergence.



4. Computational Concerns (Again)

Still need to compute $\nabla^2 f(x)$

Still need linear system solves: $B_k p = -\nabla f(x_k)$

(or worse inverses

(or worse diagonalizations))

cost $O(d^3)$

$\Rightarrow$ At most $d \approx 1000$

Worried about bad conditioning

(singular $\nabla^2 f(x_k)$ $\Rightarrow$ B_k with eigenvalues $O(\varepsilon)$

$\Rightarrow$ $p_k = O(\frac{1}{\varepsilon})$).

Does bad conditioning occur?

Yes! HW4Q3(b), we have a degree 4 polynomial

$$\left(F(x,\lambda) = \begin{pmatrix} (A-\lambda I)x \\ x^T x - 1 \end{pmatrix} = 0 \right.$$

$\min \|F(x)\|_2^2$ is degree 4)

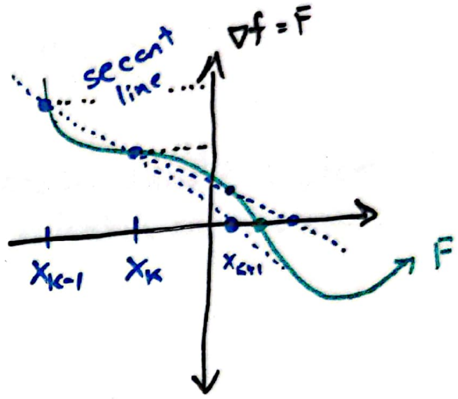
For example $f(x,y) = x^4 + y^2$, $\nabla^2 f(x,y)_{xx} \to 0$ as $x \to 0$

$\nabla^2 f(x,y)_{yy} = 2$.

5. Approximant Hessians and Secant Equations

Recall the Secant Method for $F: \mathbb{R} \to \mathbb{R}$

$$\nabla F(x_k) \approx \underbrace{\frac{F(x_k) - F(x_{k-1})}{x_k - x_{k-1}}}_{\overset{!!}{B_k}}$$



$$x_{k+1} = x_k - \frac{F(x_k)}{B_k}$$

Avoids Jacobian/Hessian Computations

Still superlinear convergence $\quad x_k \to x^*$

$$\Rightarrow x_k - x_{k-1} \to 0$$

$$\Rightarrow B_k \to \nabla f(x_k) \to \nabla F(x^*)$$

Goal: Get these two improvements for $\mathbb{R}^d$

(iteration cost $O(d^2)$, avoid inverses linear systems)

$\Rightarrow 10^4$, or $10^5 \approx d$ sized

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) $m_K(x_K) = f(x_K)$, $\nabla m_K(x_K) = \nabla f(x_K)$

(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 $\cancel{\nabla} m_K(x) = \overset{f(x_K)}{\downarrow} g_K^T(x-x_K) + \frac{1}{2}(x-x_K)^T B_K (x-x_K)$

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

$$\Rightarrow \nabla m_K(x_{K-1}) = \nabla f(x_K) + B_K(x_{K-1} - x_K) = \nabla f(x_{K-1}) \text{ by (3)}$$

$$\Rightarrow B_K(x_{K-1} - x_K) = \nabla f(x_{K-1}) - \nabla f(x_K)$$

$$\Leftrightarrow B_K s_K = y_K$$ The Secant Equation

where $s_K = x_K - x_{K-1}$ "run"

$y_K = \nabla f(x_K) - \nabla f(x_{K-1})$.
"rise"