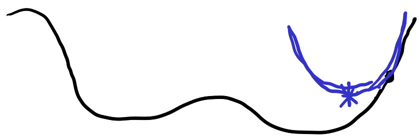
→ Newton + Quasi-Newton

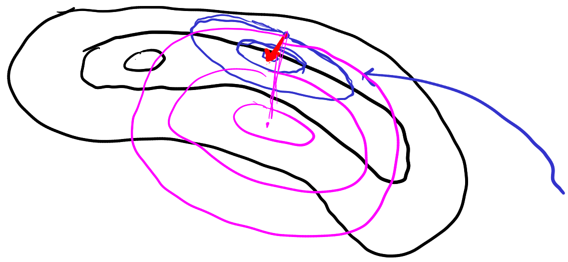
Last time:

- Grad. Descent w/ fixed step size is slow
- Line Search + Wolfe Cond ⇒ convergence to a fixed pt. for a wide range of descent directions
- Backtracking

Newton's Method



$$x^{k+1} \leftarrow x^k - \frac{f'(x^k)}{f''(x^k)}$$



$$f^k \equiv f(\vec{x}^k)$$
$$\vec{g}^k \equiv \nabla f(\vec{x}^k)$$
$$H^k = \nabla^2 f(\vec{x}^k)$$

$$q(\vec{x}) = f^k + (\vec{g}^k)^T(\vec{x}-\vec{x}^k) + \tfrac{1}{2}(\vec{x}-\vec{x}^k)^T H^k (\vec{x}-\vec{x}^k)$$

$$\nabla q(\vec{x}) = \vec{g}^k + H^k(\vec{x}^* - \vec{x}^k) = 0$$

$$H^k \vec{x}^* = -\vec{g}^k + H^k \vec{x}^k$$

$$\boxed{\vec{x}^{k+1} = \vec{x}^* = \vec{x}^k - (H^k)^{-1}\vec{g}^k}$$ Newton Step

$$\boxed{\vec{d}^k = -(H^k)^{-1}\vec{g}^k}$$ Newton Direction

$\cancel{\vec{x}^0}$

loop

$$\vec{d}^k \leftarrow -(H^k)^{-1}\vec{g}^k$$
$$\alpha^k \leftarrow \underset{\alpha}{\operatorname{argmin}}\left(f(\vec{x} + \alpha \vec{d}^k)\right) \leftarrow \text{backtracking}$$
$$\vec{x}^{k+1} = \vec{x} + \alpha^k \vec{d}^k$$

Advantages

	Gradient	Newton
Ease of Implementation	✓	(auto-diff makes easier)
Practical Efficiency (#steps)		✓
" " (computation)	✓	Evaluating + Inverting expensive
Natural Step Size		✓ (=1)
Theory	Thm 3.3	✓ Thm 3.5

Rates of Convergence

"quotient"

Q-linear

$\exists \; r \in (0,1)$ such that

Grad

$$\frac{\|\vec{x}^{k+1} - \vec{x}^*\|}{\|\vec{x}^k - \vec{x}^*\|} \le r \qquad \text{for all } k \text{ sufficiently large}$$

Q-superlinear

Quasi Newton

$$\lim_{k \to \infty} \frac{\|\vec{x}^{k+1} - \vec{x}^*\|}{\|\vec{x}^k - \vec{x}^*\|} = 0$$

Q-quadratic

$\exists \; M > 0$ such that

Newton

$$\frac{\|\vec{x}^{k+1} - \vec{x}^*\|}{\|\vec{x}^k - \vec{x}^*\|^2} \le M \qquad \text{for all } k \text{ sufficiently large}$$

Theorem 3.3.

grad desc-

When the steepest descent method with exact line searches (3.26) is applied to the strongly convex quadratic ~~function~~ (3.24), the error norm (3.27) satisfies

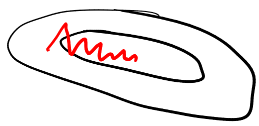
$$\|x_{k+1} - x^*\|_Q^2 \le \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 \|x_k - x^*\|_Q^2, \tag{3.29}$$

Linear

where $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ *are the eigenvalues of* Q.

$\vec{x}^T Q \vec{x}$

$$\|\vec{x}\|_Q = \sqrt{x^T Q x}$$



Theorem 3.5.

Suppose that f *is twice differentiable and that the Hessian* $\nabla^2 f(x)$ *is Lipschitz continuous (see (A.42)) in a neighborhood of a solution* x^* *at which the sufficient conditions (Theorem 2.4) are satisfied. Consider the iteration* $x_{k+1} = x_k + p_k$*, where* p_k *is given by (3.30). Then*

Newton Step

(i) *if the starting point* x_0 *is sufficiently close to* x^**, the sequence of iterates converges to* x^**;*

(ii) *the rate of convergence of* $\{x_k\}$ *is quadratic; and*

(iii) *the sequence of gradient norms* $\{\|\nabla f_k\|\}$ *converges quadratically to zero.*

Want: Newton-Like convergence
Only gradient evaluations

Break: In 1-D case how do we approximate Newton's method with only f, f' evaluations?

Secant Method

$$f''(x^k) \approx \frac{f'(x^k) - f'(x^{k-1})}{x^k - x^{k-1}}$$

$$x^{k+1} \leftarrow x^k - \frac{x^k - x^{k-1}}{f'(x^k) - f'(x^{k-1})} f'(x^k)$$

Quasi-Newton methods

$$\vec{x}^{k+1} \leftarrow \vec{x}^k - \alpha^k Q^k \vec{g}^k \qquad Q^k \text{ approximates } (H^k)^{-1}$$

α^k: backtracking

Typically $Q^{(1)} = I$

$$\vec{\gamma}^{k+1} \equiv \vec{g}^{k+1} - \vec{g}^k$$

$$\vec{\delta}^{k+1} \equiv \vec{x}^{k+1} - \vec{x}^k$$

DFP (Davidon - Fletcher - Powell)

Argonne Natl. Lab Tech Rep. 1959

1991 - first article in SIAM journal

$$Q \leftarrow Q - \frac{Q\vec{\gamma}\vec{\gamma}^T Q}{\vec{\gamma}^T Q \vec{\gamma}} + \frac{\vec{\delta}\vec{\delta}^T}{\vec{\delta}^T \vec{\gamma}}$$

BFGS (Broyden - Fletcher - Goldfarb - Shanno)

$$Q \leftarrow Q - \left(\frac{\vec{\delta}\vec{\gamma}^T Q + Q\vec{\gamma}\vec{\delta}^T}{\vec{\delta}^T \vec{\gamma}}\right) + \left(1 + \frac{\vec{\gamma}^T Q \vec{\gamma}}{\vec{\delta}^T \vec{\gamma}}\right)\left(\frac{\vec{\delta}\vec{\delta}^T}{\vec{\delta}^T \vec{\gamma}}\right)$$

Theorem 6.6.

Suppose that f is twice continuously differentiable and that the iterates generated by the BFGS algorithm converge to a minimizer x^* at which Assumption 6.2 holds. Suppose also that (6.52) holds. Then x_k converges to x^* at a superlinear rate.

$$(6.52) \rightarrow \sum_{k=0}^{\infty} \|\vec{x}^k - \vec{x}^*\| < \infty$$

BFGS (+ other Quasi-Newton converge superlinearly)

Problem: Maintains large, dense Q

L-BFGS : store m values of $\vec{\delta}, \vec{\gamma}$

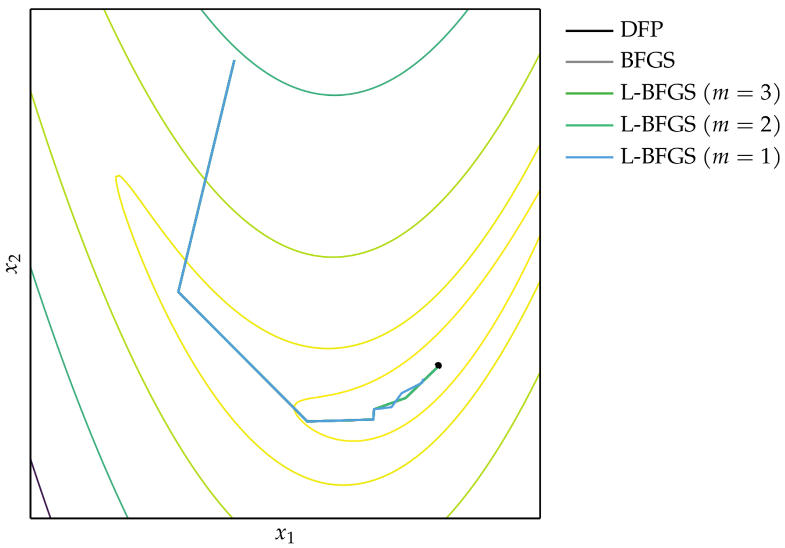
$\vec{q}^{(m)} = \nabla f(\vec{x}^{k})$ then

$$\vec{q}^{(i)} = q^{(i+1)} - \frac{(\vec{\delta}^{(i+1)})^T \vec{q}^{(i+1)}}{(\vec{\gamma}^{(i+1)})^T \vec{\delta}^{(i+1)}} \vec{\gamma}^{(i+1)}$$

$$\vec{z}^{(0)} = \frac{\vec{\gamma}^{(m)} \odot \vec{\delta}^{(m)} \odot \vec{q}^{(m)}}{(\vec{\gamma}^{(m)})^T \vec{\gamma}^{(m)}}$$

$$\vec{z}^{(i)} = \vec{z}^{(i-1)} + \vec{\delta}^{(i-1)} \left(\frac{(\vec{\delta}^{(i-1)})^T \vec{q}^{(i-1)}}{(\vec{\gamma}^{(i-1)})^T \vec{\delta}^{(i-1)}} - \frac{(\vec{\gamma}^{(i-1)})^T \vec{z}^{(i-1)}}{(\vec{\gamma}^{(i-1)})^T \vec{\delta}^{(i-1)}} \right)$$

$$\underline{\vec{d} = -\vec{z}^{(m)}}$$



Gradient	Q-N	Newton
Linearly	Super Linearly	Quad
	estimating Q or keeping history	inverting Hessian