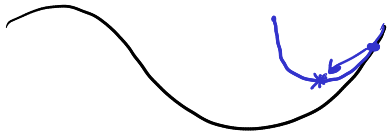
## Last time

Newton's Method

Quasi-Newton  Approximate $H^{-1}$



## Conjugate Gradient and Momentum

### Line Search

Sat. Wolfe $\Rightarrow$ Global Convergence (Thm 3.2)

Direction Affects Convergence Rate

### Trust Region

### Momentum

- Not in N+W (2006)
- Training of Neural Networks
- Availability of Auto-diff

### Directions for Line Search

Gradient (Steepest)

Conjugate Gradient

L-BFGS

Quasi Newton

BFGS

Newton

### Convergence Rate

Linear

No theorem in N.W. slower (for quadratic, LCG converges in $n$ steps)

Super linear

Quadratic

Corolla

Computational Cost sensitive to $n$

Cadillac (Luxury)

## Conjugate Gradient

$$f(\vec{x}) = \frac{1}{2}\vec{x}^T A\vec{x} + \vec{b}^T\vec{x} + c$$

$$\nabla f(\vec{x}) = A\vec{x} + \vec{b} = 0$$

Newton: Invert

$$x^* = -A^{-1}b$$

### Conjugate Direction

Def: $\vec{d}^i$ and $\vec{d}^j$ are mutually conjugate w.r.t. $A$ if

$$\vec{d}^{iT} A \vec{d}^j = 0$$

(not generally orthogonal)

# Linear Conjugate Gradient (LCG)

Start with $\vec{d}^1 = -\vec{g}^1$

loop

$\vec{x}^{k+1} \leftarrow \vec{x}^k + \alpha^k \vec{d}^k$

$d^{k+1} \leftarrow -\vec{g}^{k+1} + \beta^{k+1} \vec{d}^k$

$\alpha^k$ is minimizer along $\vec{d}^k$

$$\alpha^k = -\frac{\vec{d}^{k\,T}(A\vec{x}^k + \vec{b})}{\vec{d}^{k\,T} A \vec{d}^k}$$

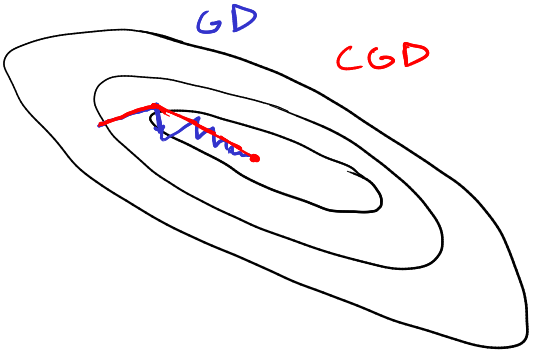
$$\vec{d}^{k\,T} A \vec{d}^{k-1} = 0$$

$$(-\vec{g}^k + \beta^k \vec{d}^{k-1}) A \vec{d}^{k-1} = 0$$

$$-\vec{g}^k A \vec{d}^{k-1} + \beta^k \vec{d}^{k-1} A \vec{d}^{k-1} = 0$$

$$\beta^k = \frac{\vec{g}^k A \vec{d}^{k-1}}{\vec{d}^{k-1\,T} A \vec{d}^{k-1}}$$

LCG finds minimum of $f$ in $n$ steps (N+W Thm 5.1)




Large Nonlinear Problems: Unknown $A$ (or expensive)

- Approximate $\alpha$ with line search
- Approximate $\beta$ (below)

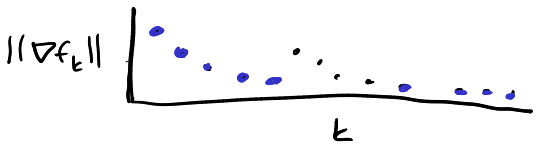
↳ $\alpha$? $\beta$?

Fletcher Reeves

$$\beta^k = \frac{\vec{g}^{k\,T}\vec{g}^k}{\vec{g}^{k-1\,T}\vec{g}^{k-1}}$$

→ N.W. Theorem 5.7

Strong Wolfe line search $\Rightarrow \liminf_{k\to\infty} \|\nabla f_k\| = 0$




Polak-Ribière

$$\beta^k = \max\left\{\frac{\vec{g}^{k\,T}(\vec{g}^k - \vec{g}^{k-1})}{\vec{g}^{k-1\,T}\vec{g}^{k-1}}, 0\right\}$$

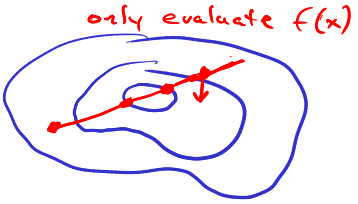
Relationship between CG and Quasi-Newton

p. 180 of N+W

"Memoriless" BFGS (at each step, set $Q^{k-1} = I$)

then equivalent to Hestenes-Stiefel CG

Break: If you have R.M. Autodiff why would you prefer fixed step size gradient descent over ~~line search?~~ backtracking line search




# Momentum / Adaptive Scaling

("Vanilla") Momentum

$$\vec{v}^{k+1} \leftarrow \beta \vec{v} - \alpha \vec{g}^k$$
$$\vec{x}^{k+1} \leftarrow \vec{x}^k + \vec{v}^{k+1}$$

Tendency to overshoot

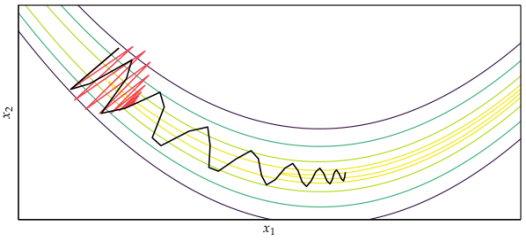



Figure 5.5. Gradient descent and the momentum method compared on the Rosenbrock function with $b = 100$; see appendix B.6.

Momentum

Nesterov Momentum

$$\vec{v}^{k+1} \leftarrow \beta \vec{v}^k - \alpha \nabla f\left(\vec{x}^k + \beta \vec{v}^k\right)$$
$$\vec{x}^{k+1} \leftarrow \vec{x}^k + \vec{v}^{k+1}$$

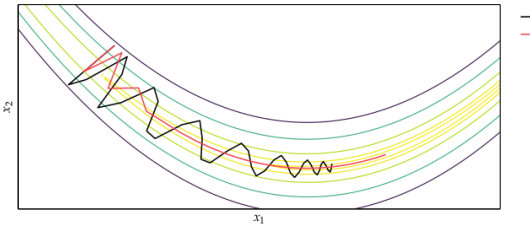



Figure 5.6. The momentum and Nesterov momentum methods compared on the Rosenbrock function with $b = 100$; see appendix B.6.

Step size adaptation

Adagrad

$$x_i^{k+1} \leftarrow x_i^k - \frac{\alpha}{\varepsilon + \sqrt{s_i^k}} g_i^k$$
$$s_i^k = \sum_{j=1}^{k} \left(g_i^j\right)^2$$

$\varepsilon$: small ($10^{-8}$)

monotonically increases

monotonic decrease of learning rate

element-wise

RMS Prop

$$\hat{s}_i^k = \gamma \hat{s}_i^{k-1} + (1-\gamma)\left(g_i^k\right)^2 \qquad \gamma = 0.9$$
$$\mathrm{RMS}(g_i) \equiv \sqrt{\hat{s}_i}$$

Ada delta

$$\vec{x}_i^{k+1} \leftarrow \vec{x}_i^k - \frac{\mathrm{RMS}(\Delta x_i)}{\varepsilon + \mathrm{RMS}(g_i)} g_i^k$$

No learning rate

Ada m

$$\vec{v}^{k+1} \leftarrow \gamma_v \vec{v}^k + (1-\gamma_v)\vec{g}^k$$
$$\vec{s}^{k+1} \leftarrow \gamma_s \vec{s}^k + (1-\gamma_s)\left(\vec{g}^k \odot \vec{g}^k\right)$$
$$\hat{\vec{v}}^{k+1} \leftarrow \frac{\vec{v}^{k+1}}{1-\gamma_v}$$
$$\hat{\vec{s}}^{k+1} \leftarrow \frac{\vec{s}^{k+1}}{1-\gamma_s}$$
$$\vec{x}^{k+1} = \vec{x}^k - \frac{\alpha \hat{\vec{v}}^{k+1}}{\varepsilon + \sqrt{\hat{\vec{s}}^{k+1}}}$$