In applications it's usually not possible to calculate a min explicitely ⇒ use iterative algorithms to construct a sequence of points $\vec{w}_0, \vec{w}_1, \dots \in \mathbb{R}^d$ and hope that sequence converges to min $\vec{x}^*$ of a function $f(\vec{x})$

Iterative algorithm:

Initialise: choose starting value $\vec{w}_0 \in \mathbb{R}^d$

1. Calculate search direction $\vec{p}_k$ for $k = 0, \dots$
2. Update $\vec{w}_{k+1} = \vec{w}_k + \alpha_k \vec{p}_k$ $k = 0, \dots$ $\alpha_k \dots$ step length, learning rate
   such that $f(\vec{w}_{k+1}) < f(\vec{w}_k)$
3. Check stopping criterion: if satisfied return $\vec{w}_{k+1}$, otherwise go to step 1.

Line search (steepest descent) (SD)

$$\vec{p}_k = -\nabla f(\vec{w}_k)$$ ⇐ go in direction of max descent

Taylor expansion: $\alpha_k = \alpha = \text{const}$ and let $\vec{p}$ be a direction

$$f(\vec{w}_k + \alpha \vec{p}) = f(\vec{w}_k) + \alpha \vec{p}^T \nabla f(\vec{w}_k) + \mathcal{O}(\alpha^2) \quad | - f(\vec{w}_k) + \lim \text{in } \alpha$$

$$\left.\frac{df(\vec{w}_k + \alpha p)}{d\alpha}\right|_{\alpha=0} = \underbrace{\langle \vec{p}, \nabla f(\vec{w}_k) \rangle}_{<0 \text{ then descent}}$$ ⇐ directional derivative in direction $\vec{p}$

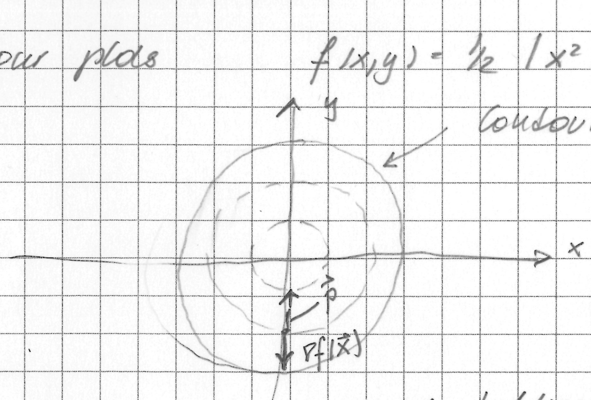
$$\Rightarrow \vec{p} = -\alpha \nabla f(\vec{w}_k), \ \alpha > 0$$

Cauchy Schwarz gives bounds

$$-\|\vec{p}\|_2 \, \|\nabla f(\vec{w}_k)\|_2 \le \langle \vec{p}, \nabla f(\vec{w}_k) \rangle \le \|\vec{p}\|_2 \, \|\nabla f(\vec{w}_k)\|_2$$

equality if we choose $\vec{p} = -\alpha \|\nabla f\|$

Visualisation: contour plots $f(x,y) = \frac{1}{2}(x^2 + y^2)$



contour lines where $f(x,y) = \text{const}$

angle between $\vec{p}$ and $\nabla f(\vec{w})$

$$\langle \vec{p}, f(\vec{w}) \rangle = \|\vec{p}\| \, \|\nabla f(\vec{w})\| \cos\theta$$ ⇐ max descent when $\cos\theta = -1$

How do choose step size α / α_k

α too big $\Rightarrow$ over shoot
α too small $\Rightarrow$ takes forever

Finding the right α ≙ step size control see Chapter 6 for a more general discussion

In case of (SD) consider

$$f(\vec{w}) = \tfrac{1}{2} \|\vec{A}\vec{w} - \vec{b}\|_2^2 \qquad A \in \mathbb{R}^{m\times n} \quad b \in \mathbb{R}^m$$

$$\Rightarrow \nabla f(\vec{w}) = \vec{A}^T(\vec{A}\vec{w} - \vec{b}) \Rightarrow \vec{w}_{k+1} = \vec{w}_k - \alpha_k \vec{A}^T(\vec{A}\vec{w}_k - \vec{b})$$

To find optimal we seek to minimise the following function

$$\alpha \mapsto \varphi(\alpha) = f(\vec{w}_k - \alpha \underbrace{\vec{A}^T(\vec{A}\vec{w}_k - \vec{b})}_{=: -\vec{r} = -\nabla f(\vec{w}_k)})$$

$$\varphi'(\alpha) = \frac{d}{d\alpha} f(\vec{w}_k + \alpha \vec{r}_k) = \langle \nabla f(\vec{w}_k + \alpha \vec{r}_k), \vec{r}_k \rangle$$

$$\nabla f(\vec{w}_k + \alpha \vec{r}_k) = -\vec{A}^T(\vec{b} - \vec{A}(\vec{w}_k + \alpha \vec{r}_k)) = +\vec{A}^T(\vec{A}\vec{w}_k + \alpha \vec{r}_k) - \vec{A}^T\vec{b} = \vec{A}^T[\vec{A}(\vec{w}_k + \alpha \vec{r}_k - \vec{b})]$$

$$= \langle A^T[\vec{A}(\vec{w}_k + \alpha \vec{r}_k) - \vec{b}], \vec{r}_k \rangle$$

$$= \langle \underbrace{A^T(\vec{A}\vec{w}_k - b)}_{-\vec{r}_k}, \vec{r}_k \rangle + \alpha \langle \vec{A}^T\vec{A}\vec{r}_k, \vec{r}_k \rangle$$

$$= -\langle \vec{r}_k, \vec{r}_k \rangle + \alpha \langle \vec{A}^T\vec{A}\vec{r}_k, \vec{r}_k \rangle = -\vec{r}_k^T\vec{r}_k + \alpha \vec{r}_k^T\vec{A}^T\vec{A}\vec{r}_k = 0$$

since we want to find min w.r.t. α

$$\Rightarrow \alpha_k = \frac{\vec{r}_k^T\vec{r}_k}{\vec{r}_k^T\vec{A}^T\vec{A}\vec{r}_k} = \frac{\|\vec{r}_k\|_2^2}{\|\vec{A}\vec{r}_k\|_2^2}$$

Update of $\vec{r}_k$:

$$\vec{r}_{k+1} = \vec{A}^T(\vec{b} - \vec{A}\vec{w}_{k+1})$$

$$= \vec{A}^T(\vec{b} - \vec{A}(\vec{w}_k + \alpha_k \vec{r}_k)) = \underbrace{\vec{A}^T(\vec{b} - \vec{A}\vec{w}_k}_{\vec{r}_k} - \alpha_k \vec{A}\vec{r}_k)$$

$$= \vec{r}_k - \alpha_k \vec{A}\vec{r}_k$$

Gradient descent for linear least square

$$\alpha_k = \frac{\vec{r}_k^T\vec{r}_k}{\vec{r}_k^T\vec{A}^T\vec{A}\vec{r}_k}$$

$$\vec{w}_{k+1} = \vec{w}_k + \alpha_k \vec{r}_k$$

$$\vec{r}_{k+1} = \vec{r}_k - \alpha_k \vec{A}^T\vec{A}\vec{r}_k$$

Stopping criterion: ◦) $\vec{r} = \vec{0} \Leftrightarrow \vec{w}$ is a stationary point $\Rightarrow \|\vec{r}_k\| < \varepsilon \quad \varepsilon > 0, \ \varepsilon \ll 1$

$\Rightarrow$ not good if fct. is flat

◦) $\|\vec{w}_{k+1} - \vec{w}_k\|_2 < \varepsilon$