09. GRADIENT DESCENT

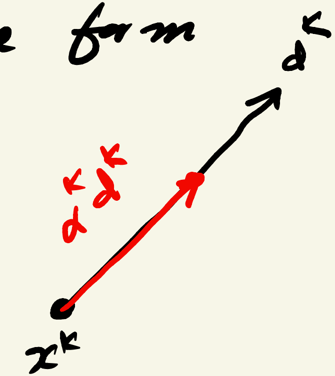
- Descent methods
- Descent direction and step size
- Gradient descent.

Descent direction

- Unconstrained optimization (non-linear)

$$\min_{x \in \mathbb{R}^n} f(x). \qquad f: \mathbb{R}^n \to \mathbb{R} \quad \text{cont. diff.}$$

- We will consider iterative algorithms of the form



$$x^{k+1} = x^k + \alpha^k d^k$$

$$\alpha^k := \text{step size} \qquad d^k := \text{descent direction}$$

- (Defintion) A search direction is a descent direction for f at $x \in \mathbb{R}^n$ if the directional derivative of f at x is negative, i.e:

$$f'(x; d) := \nabla f(x)^T d < 0.$$

Descent property

If f is cont. diff. and d is a descent direction,

then $\exists$ $\varepsilon > 0$ s.t.

$$f(x+\alpha d) < f(x)$$

$$\nabla f(x)^T d < 0$$

for $\alpha \in [0, \varepsilon]$.

proof: • Because $f'(x; d) < 0$ we have:

$$\lim_{\alpha \to 0^+} \frac{f(x+\alpha d) - f(x)}{\alpha} < 0$$

• Using definition of limits, $\exists$ $\varepsilon > 0$ s.t.

$$\frac{f(x+\alpha d) - f(x)}{\alpha} < 0 \quad \text{for } \alpha \in [0, \varepsilon].$$

$\lim_{x \to a} f(x) = L$ use (ε, δ) definition of limit.

Gen : Method for Descent direction.

Initialization : choose $x \in \mathbb{R}^n$.

For $k = 0, 1, 2, \dots$

(a) compute descent direction $d^k \in \mathbb{R}^n$.

(b) compute a stepsize α^k s.t. $f(x^k + \alpha^k d^k) < f(x^k)$

(c) update $x^{k+1} = x^k + \alpha^k d^k$

(d) Stoping criteria.

<u>Question:</u>

- How to choose d^k?
- what are advantages/disadvantages of different d^k?
- How do we compute α^k?
- stoping criteria?

Stepsize selection α^k.

- Constant stepsize: $\alpha^k = \bar{\alpha} \in \mathbb{R}$ for all k.

- Exact line search: choose α^k that minimizes the 1-dimensional optimization problem

$$\alpha^k := \arg\min_{\alpha \geq 0} f(x^k + \alpha d^k)$$

- Diminishing stepsize: choose α^k that satisfies

$$\alpha^k \to 0 \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha^k = \infty$$

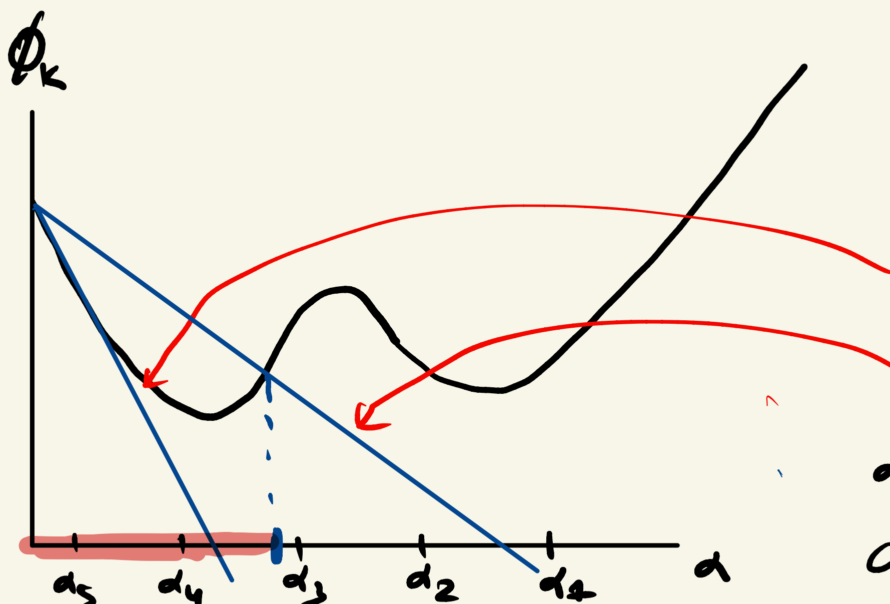
eg: $\alpha^k = \frac{1}{k}, \frac{1}{\sqrt{k}}$

- Backtracking "Armijo" linesearch: For some parameter $\mu \in (0,1)$ reduce stepsize α (eg: $\alpha \leftarrow \alpha/2$ starting at $\alpha = 1$) until

$$f(x^k) - f(x^k + \alpha d^k) \geq -\mu \alpha \nabla f(x^k)^T d^k.$$

$$\phi_k(\alpha) = f(x^k + \alpha d^k)$$

$$\phi_k(0) = \nabla f(x^k)^T d^k.$$



$$f(x^k) + \alpha \nabla f(x^k)^T d^k.$$

$$f(x^k) + \alpha \mu \nabla f(x^k)^T d^k$$

α_4, α_5 satisfies the sufficient decrease conditions.

Sufficient decrease condition.

- The sufficient decrease condition is satisfied for small enough α^k.

- suppose $d \neq 0$, $d \in \mathbb{R}^n$ is a descent direction of f at x. Let $\mu \in (0,1)$. There exists $\varepsilon > 0$ s.t

$$f(x) - f(x+\alpha d) \geq -\mu \alpha \nabla f(x)^T d.$$

for some $\alpha \in [0, \varepsilon]$

Exact line search for quadratic function.

An exact line search may not be possible for general function but is possible for quadratic function.

$$f(x) = \frac{1}{2}x^T A x + b^T x + c, \quad A \succ 0, \; A = A^T$$

Exact linesearch solve 1-d optimization problem

$$\min_{\alpha \geq 0} f(x + \alpha d). \qquad \nabla f(x) = Ax + b.$$

Derivation: $f(x+\alpha d) = \frac{1}{2}(x+\alpha d)^T A(x+\alpha d) + b^T(x+\alpha d) + c$

$$= \frac{1}{2}x^T A x + \alpha x^T A d + \frac{\alpha^2}{2} d^T A d + b^T x + \alpha b^T d + c$$

so, $\frac{d}{d\alpha} f(x+\alpha d) = x^T A d + \alpha d^T A d + b^T d = \alpha d^T A d + d^T \nabla f(x)$

so, $\frac{d}{d\alpha} f(x+\alpha d) = 0 \Rightarrow \alpha = -\frac{d^T \nabla f(x)}{d^T A d} > 0$

Gradient descent

$$d^k = -g_k \qquad g_k \equiv \nabla f(x^k)$$

- The negative gradient direction $-g_k$ provides a descent direction:

$$f'(x^k; -g^k) = -\nabla f(x^k)^T g_k = -\|g_k\|_2^2 < 0$$

if x^k is not a stationary point.

- The negative gradient $d \equiv -\nabla f(x)$ is the steepest direction of descent i.e.

$$\min_d \{ f'(x; d) \mid \|d\|_2 = 1 \}.$$

proof: $f'(x;d) = \nabla f(x)^T d$

$\geq -\|\nabla f(x)\| \, \|d\|$ Cauchy Schwartz

$= -\|\nabla f(x)\|$ $|\nabla f^T d| \leq \|\nabla f\| \, \|d\|$

The lower is achieved by $d = -\dfrac{\nabla f(x)}{\|\nabla f(x)\|}$

Input : $\varepsilon > 0$ tolerance

x_0 starting point

For $k = 0, 1, 2, \dots$

- evaluate gradient $g^k = \nabla f(x^k)$
- chose a step size α^k that satisfies $\phi_k(x^k - \alpha g_k)$

$$f(x^k - \alpha^k g^k) \leq f(x^k).$$

- $x^{k+1} = x^k - \alpha^k g^k$
- stop if $\|\nabla f(x^{k+1})\| \leq \varepsilon.$