10. GRADIENT DESCENT

- Gradient Descent
- step size selection
- Scaled gradient descent.

Gradient method.

Input: $\varepsilon > 0$ (tolerance)

$x_0 \in \mathbb{R}^n$ (starting iterate).

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

- evaluate gradient $g_k = \nabla f(x_k)$.
- choose step length α_k based on decreasing $f(x_k + \alpha g_k)$.
- $x_{k+1} = x_k - \alpha_k g_k$.
- stop if $\|\nabla f(x_{k+1})\| < \varepsilon$.

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

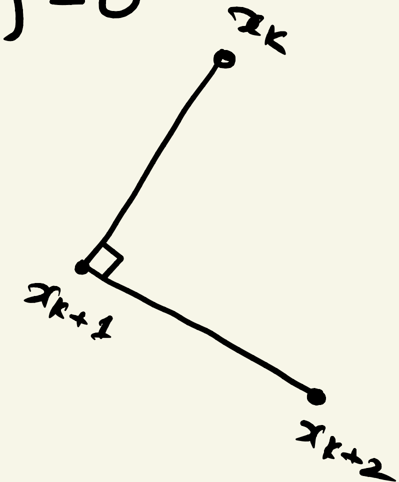
$$\alpha_{exact} = - \frac{\nabla f(x_k)^T d_k}{d_k^T A d_k} > 0 \qquad d_k = -g_k.$$

$$f(x,y) = x^2 + y^2 \Rightarrow A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \; b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \; c = 0.$$

"Zig-Zag" of Gradient descent with exact line search.

Let $x_1, x_2, \ldots, x_n$ be iterates of gradient descent with exact line search. Then

$$(x_{k+2} - x_{k+1})^T (x_{k+1} - x_k) = 0$$



Gradient descent method with constant stepsize.

- Constant stepsize i.e. $\alpha^k = \bar{\alpha}$ $k = 0, 1, 2, \ldots$
- $\bar{\alpha}$ is too small $\Rightarrow$ convergence is slow

  $\bar{\alpha}$ is too large $\Rightarrow$ gradient method diverges

  How to choose $\bar{\alpha}$?
- $\bar{\alpha}$ has to satisfy $\bar{\alpha} \in (0, \alpha_{max})$ for method to converge
- α_{max} is determined by Lipschitz constant of ∇f.

Lipschitz continuity of Gradient

A continuously differentiable function $f: \mathbb{R}^n \to \mathbb{R}$ has a Lipschitz continuous gradient with parameter L if

$$\|\nabla f(x) - \nabla f(y)\| \le L \|x-y\| \quad \text{(2-norm)}.$$

for all x, y and some $L \in \mathbb{R}$.

Example: $f(x) = \frac{1}{2} x^T A x + b^T x + c \qquad A \succ 0$

$$\nabla f(x) = Ax + b.$$

$$\|\nabla f(x) - \nabla f(y)\| = \|(Ax+b) - (Ay-b)\|$$

$$= \|A(x-y)\| = \frac{\|A(x-y)\|}{\|x-y\|} \|x-y\|$$

$$\le \|A\| \|x-y\| \qquad \|A\| = \lambda_{max}(A).$$

$$\|A\| = \sup_{\|x\|=1} \|Ax\|$$

Example:

$A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, $\|A\| = 2$

Constant stepsize threshold

- If $f: \mathbb{R}^n \to \mathbb{R}$ has a L-Lipschitz continuous gradient and a minimizer exists, then the gradient method with constant stepsize $\bar{\alpha}$ converges if

$$\bar{\alpha} \in (0, 2/L).$$

For example: Quadratic function.

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

- $f(x) = \frac{1}{2} x^T A x + b^T x + c$
- $L = \|A\| = \lambda_{max}(A) = 2$
- Assume minimizer exists ($b \in R(A)$)
- Gradient method converges for $\bar{\alpha} \in (0, 1)$.

Convergence of Gradient method.

For the minimization of $f: R^n \to R$ bounded below with L- Lipschitz gradient and one of the line search:

① constant stepsize $\bar{\alpha} \in (0, 2/L)$

② exact line search

③ back tracking line search $\mu \in (0,1)$.

Then

ⓐ $f(x_{k+1}) < f(x_k)$ for all $k = 0, 1, \ldots$ unless $\nabla f(x_k) = 0$ [decreasing].

ⓑ $\|\nabla f(x_k)\| \to 0$ [stationary point].

Condition number of a matrix.

The condition number of a $n \times n$ positive definite matrix A is defined by

$$\kappa(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} \geq 1 \qquad \left[\kappa(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)} \right]$$

- ill-conditioned if $\kappa(A)$ is large.
- condition number of Hessian at solution influences the speed of convergence of gradient method.

$$H = \nabla^2 f(x^*).$$

$\kappa(H)$ small implies fast convergence.

Rosenbrock Function.

$$f(x_1,x_2) = 100\,(x_2 - x_1^2)^2 + (1-x_1)^2$$

$$\nabla f(x_1,x_2) = \begin{bmatrix} -400x_1(x_2-x_1^2) - 2(1-x_1) \\ 200\,(x_2 - x_1^2) \end{bmatrix}$$

$$\nabla^2 f(x_1,x_2) = \begin{bmatrix} -400x_2 + 1200x_1^2 + 2 & -400x_1 \\ -400x_1 & 200 \end{bmatrix}$$

Solution $(x_1,x_2) = (1,1)$ (check $\nabla f(1,1) = 0$)

$$\nabla^2 f(1,1) = \begin{bmatrix} 802 & -400 \\ -400 & 200 \end{bmatrix}$$

unique global min.

back tracking: fix $\mu \in (0,1)$. reduce α until

$$f(x_k) - f(x_k + \alpha d_k) \geq -\mu\alpha\nabla f(x_k)^T d_k.$$