CPSC 406

Computational Optimization

Dept of Computer Science

University of British Columbia

GRADIENT DESCENT

Descent Directions

- Unconstrained nonlinear optimization:

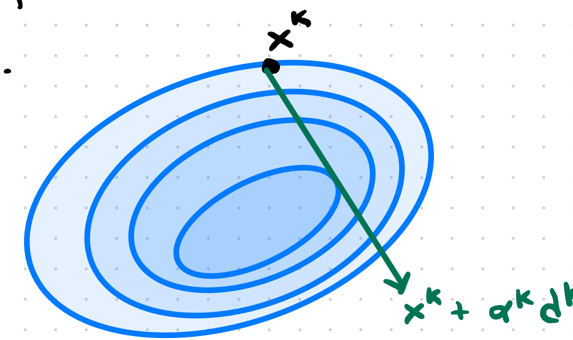
$$\min_{x \in \mathbb{R}^n} f(x) \quad , \quad f: \mathbb{R}^n \to \mathbb{R} \quad , \quad \text{continuosly differentiable}$$

- We will consider iterative algorithms of the form

$$x^{k+1} = x^k + \alpha^k d^k \quad , \quad k = 0, 1, 2, \ldots$$

where

- d^k ≡ search direction
- α^k ≡ step length



- A search direction $d \neq 0$ is a <u>descent dir</u> for f at x if the directional derivative is negative, ie,

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

Descent Property

If $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, and $d \in \mathbb{R}^n$ is a descent direction at x, then for some $\varepsilon > 0$,

$$f(x + \alpha d) < f(x) \qquad \forall \alpha \in (0, \varepsilon] \qquad \text{(Descent)}$$

Proof :

- Because $f'(x; d) < 0$

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

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

$$\frac{f(x+\alpha d) - f(x)}{\alpha} < 0 \qquad \forall \ \alpha \in (0, \varepsilon]$$

which implies (Descent)

GENERIC DESCENT METHOD — conceptual algorithm

Initialization: choose $x_0 \in \mathbb{R}^n$

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

(a) compute descent direction d^k

(b) compute stepsize α^k st $f(x^k + \alpha^k d^k) < f(x^k)$

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

(d) check stopping criteria

Questions

- How to determine a starting point?
- what are advantages / disadvantages of different dirs d^k?
- How to compute a step length α^k?
- when to stop?

STEP SIZE SELECTION α^k

[descent direction computation later]

STEPSIZE SELECTION (α^k)

These are the selection rules most used in practice:

1. **Constant stepsize**: $\alpha^k = \bar{\alpha} \quad \forall k$ (needs additional conditions on f to be reliable.)

2. **Exact Linesearch**: choose α^k to minimize f along ray $x^k + \alpha d^k$:

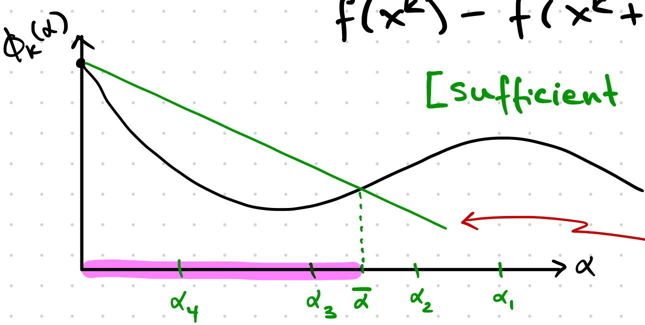
$$\alpha^k \in \underset{\alpha \geq 0}{\operatorname{argmin}} \; f(x^k + \alpha d^k)$$

3. **Backtracking "Armijo" Linesearch**: for some parameter $\mu \in (0,1)$ reduce α (eg, $\alpha \leftarrow \alpha/2$ beginning with $\alpha = 1$) until

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

[sufficient decrease property - see α_3 & α_4 below]

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



$f(x^k) + \mu \alpha \underbrace{\nabla f(x^k)^T d^k}_{\leq 0 \text{ by def'n of } d^k}$

Exact Linesearch for Quadratic Functions

An exact linesearch is typically only possible for quadratic func's:

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

Exact linesearch solves the 1-dimensional optimization problem

$$\min_{\alpha \geq 0} f(x + \alpha d) \qquad \text{(where } d \text{ is descent dir)}$$

Derivation of solution [details in class]:

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

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

$$\frac{d}{d\alpha} f(x+\alpha d) = 0 \iff \left\{ \alpha = \frac{-\nabla f(x)^T d}{d^T A d} > 0 \right\}$$

$$\iff \left\{ \begin{array}{ll} d^T A d > 0 & (A \succ 0) \\ \nabla f(x)^T d < 0 & (d \text{ is descent dir}) \end{array} \right\}$$

SEARCH DIRECTIONS d^k

GRADIENT DESCENT

$$d^k := -g^k \quad \text{where} \quad g^k := \nabla f(x^k)$$

- The negative gradient direction $(-g_k)$ provides descent:

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

if $g_k \neq 0$, ie, x^k is not already stationary.

- The negative gradient $g \equiv -\nabla f(x)$ is the steepest descent direction of f at x, ie, it solves

$$\min \{ f'(x;d) \mid \|d\| = 1 \} \qquad [\text{set } g \equiv \nabla f(x)]$$

Proof:

$$f'(x;d) \equiv g^T d \geq -\|g\| \cdot \|d\| \qquad [\text{Cauchy-Schwartz Ineq}]$$

$$\geq -\|g\| \qquad [\|d\| = 1]$$

Lower bound is achieved by setting $d = -g/\|g\|$

GRADIENT METHOD

Input : $\varepsilon > 0$ (tolerance)
x_0 (starting iterate)

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

- evaluate gradient $g^k = \nabla f(x^k)$
- choose step length α^k based on reducing the function

$$\phi(\alpha) = f(x^k - \alpha g^k)$$

[see stepsize selection slide

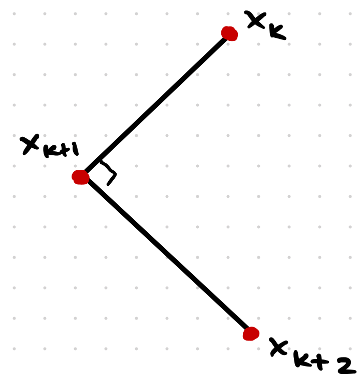
- $x^{k+1} = x^k - \alpha^k g^k$
- STOP if $\|\nabla f(x^{k+1})\| < \varepsilon$

[DEMO on function $f(x,y) = x^2 + 2y^2$]

"Zig-Zag" of Gradient Method

Let $x_1, x_2, x_3, \dots$ be the iterates generated by the gradient method. Then

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



<u>Proof</u> By definition of the gradient update:

$$\left.\begin{aligned} x_{k+1} &= x_k - \alpha_k \nabla f_k \\ x_{k+2} &= x_{k+1} - \alpha_{k+1} \nabla f_{k+1} \end{aligned}\right\} \Rightarrow \begin{aligned} x_{k+1} - x_k &= -\alpha_k \nabla f_k \\ x_{k+2} - x_{k+1} &= -\alpha_{k+1} \nabla f_{k+1} \end{aligned}$$

$$\Rightarrow (x_{k+2} - x_{k+1})^T (x_{k+1} - x_k) = 0 \iff \nabla f_k^T \nabla f_{k+1} = 0.$$

Because $\alpha^k \in \operatorname{argmin} \{ \phi(\alpha) := f(x_k - \alpha \nabla f_k) \}$

$$0 = \phi'(\alpha_k) = -\nabla f^T \underbrace{x_k - \alpha \nabla f_k)}_{\equiv x_{k+1}} = -\nabla f_k^T \nabla f(x_{k+1})$$

$$\iff \nabla f(x_k)^T \nabla f(x_{k+1}) = 0 \quad //$$

"Zig-Zag" behavior often the reason why the gradient method is slow.

Gradient Method with Constant Stepsize

- Constant stepsize sets $\alpha^k = \bar{\alpha}$ for all k.
- How to choose $\bar{\alpha}$?
  - $\bar{\alpha}$ too small $\Rightarrow$ gradient method slow
  - $\bar{\alpha}$ too large $\Rightarrow$ gradient method diverges

[Demo on min $x^2 + 2y^2$ with different values of constant stepsize]

- Must choose steplength $\bar{\alpha} \in (0, \alpha_{max})$ for method to converge.
- α_{max} depends on a property of $\nabla f(x)$ called Lipschitz continuity.

LIPSCHITZ CONTINUITY OF GRADIENT

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

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \qquad \text{[2-norm throughout]}$$

for all vectors x, y and some $L > 0$ constant.

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

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

$$\begin{aligned} \|\nabla f(x) - \nabla f(y)\| &= \|(Ax-b) - (Ay-b)\| \\ &= \|Ax - Ay\| \\ &= \|A(x-y)\| \leq \|A\| \cdot \|x-y\| \end{aligned}$$

$\|A\|_2 = \lambda_{max}(A)$

EXAMPLE: Take $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ from previous slide. $\|A\|_2 = 2$.

Constant Step-Size Threshold

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

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

- Previous quadratic example
  - $f(x) = \frac{1}{2}x^T A x + b^T x + c$ with $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$
  - $L = \lambda_{max}(A) = 2$
  - gradient method converges for all $\bar{\alpha} \in (0, 1)$

Convergence of the Gradient Method

For the minimization of $f: \mathbb{R}^n \to \mathbb{R}$ bnd below with L-Lipschitz gradient and one of the linesearches

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

(2) exact line search

(3) backtracking linesearch with $\mu \in (0,1)$

Then

(a) $f(x_{k+1}) < f(x_k)$ for all $k = 0, 1, 2, \ldots$ unless $\nabla f(x_k) = 0$

(Decreasing)

(b) $\|\nabla f(x_k)\| \to 0$ as $k \to \infty$.

(convergence to stationary point)

(Backtracking linesearch demo)

CONDITION NUMBER OF A MATRIX

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

$$K(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} \geq 1$$

- ill-conditioned matrices have $K(A)$ large
- Condition number of the Hessian at the solution influences the speed at which the gradient method converges

$$H \equiv \nabla^2 f(x^*)$$

$K(H)$ small $\Rightarrow$ GM typically converges quickly

$K(H)$ large $\Rightarrow$ GM " " slowly

Example: 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 $(x_1,x_2) = (1,1)$ is unique. [Verify $\nabla f(1,1) = 0$]

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

[Backtracking Demo]

Scaled Gradient Method

$$(P) \quad \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \qquad f : \mathbb{R}^n \to \mathbb{R}$$

- Make a linear change of variables with

  S nonsingular $n \times n$, $\quad x = Sy \quad$ ie $\quad y := S^{-1}x$

$$(P_{\text{scaled}}) \quad \underset{y \in \mathbb{R}^n}{\text{minimize}} \quad g(y) := f(Sy)$$

- Apply gradient method to scaled problem:

$$y_{k+1} = y_k - \alpha_k \nabla g(y_k) \quad \text{with} \quad \nabla g(y) = S^T \nabla f(Sy)$$

- Multiply on left by S:

$$x_{k+1} = x_k - \alpha_k S S^T \nabla f(x_k)$$

- Scaled gradient method : with $D = SS^T$,

$$x_{k+1} = x_k - \alpha_k D \nabla f(x_k)$$

Scaled Descent

The scaled gradient $-D\nabla f(x)$ is a descent direction:

$$f'(x; -D\nabla f(x)) = -\nabla f(x)^T D \nabla f(x) < 0$$

because $D = SS^T \succ 0$ (S nonsingular)

Scaled Gradient Method

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

- choose scaling matrix D_k
- compute scaled gradient $d_k = D_k \nabla f(x_k)$
- Compute steplength α_k by linesearch on the func'n
  $$\phi(\alpha) = f(x_k - \alpha d_k)$$
- $x_{k+1} = x_k - \alpha_k d_k$
- STOP if $\|\nabla f(x_{k+1})\| \le tol$

Choosing the Scaling Matrix

- Scaled gradient method is just the gradient method applied to g:

$$g(y) = f(D^{\frac{1}{2}}y) \qquad = f(x) \qquad [D = SS^T = D^{\frac{1}{2}}D^{\frac{1}{2}}]$$

$$\nabla g(y) = D^{\frac{1}{2}}\nabla f(D^{\frac{1}{2}}y) \qquad = D^{\frac{1}{2}}\nabla f(x)$$

$$\nabla^2 g(y) = D^{\frac{1}{2}}\nabla^2 f(D^{\frac{1}{2}}y)D^{\frac{1}{2}} \qquad = D^{\frac{1}{2}}\nabla^2 f(x) D^{\frac{1}{2}}$$

- Choose D_k so to make $D_k^{\frac{1}{2}}\nabla^2 f_k D_k^{\frac{1}{2}}$ as well conditioned as possible. Take $H_k \equiv \nabla^2 f(x_k)$:

$$D_k = \begin{cases} H_k^{-1} \succ 0 & [\text{Newton}] \quad D_k^{\frac{1}{2}} H_k D_k^{\frac{1}{2}} = I \quad \text{cond}(I) = 1 \\[2em] (H_k + \lambda I)^{-1} & \begin{bmatrix}\text{Damped} \\ \text{Newton}\end{bmatrix} \quad D_k^{\frac{1}{2}} H_k D_k^{\frac{1}{2}} \to I \text{ as } \lambda \to 0 \\[2em] \text{diag}\left(\dfrac{\partial^2 f(x_k)}{\partial x_i^2}\right)^{-1} & \begin{bmatrix}\text{Diagonal} \\ \text{scaling}\end{bmatrix} \\ \quad \succ 0 & \end{cases}$$

GAUSS-NEWTON METHOD

$$\text{(NLS)} \quad \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) = \frac{1}{2}\|r(x)\|^2$$

$r_i : \mathbb{R}^n \to \mathbb{R}$
cont. diff'l
$i = 1, \ldots, m$

Gauss-Newton Method:

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \ \frac{1}{2}\|\underbrace{r_k + A_k(x - x_k)}_{\left[\begin{array}{l}\text{linearization of} \\ r \text{ at } x_k\end{array}\right]}\|^2$$

$r_k = r(x_k)$

$$A_k = \begin{bmatrix} \nabla r_1(x_k)^T \\ \vdots \\ \nabla r_m(x_k)^T \end{bmatrix}$$

$$= (A_k^T A_k)^{-1} A_k^T (A_k x_k - r_k)$$

$$= x_k - (A_k^T A_k)^{-1} A_k^T r_k$$

$$= x_k - (A_k^T A_k)^{-1} \nabla f(x_k)$$

$\nabla f(x_k) = A_k^T r_k$

Thus, we see that the Gauss-Newton method is a scaled gradient method with the scaling matrix

$$D_k = (A_k^T A_k)^{-1}$$