Descent Methods

CPSC 406 – Computational Optimization

Descent methods

- descent directions
- line search
- convergence

Descent directions

$$\min_x \; f(x) \quad \text{with} \quad f : \mathbb{R}^n \to \mathbb{R} \quad \text{continuously differentiable}$$

- **directional derivative** of f along ray $x + \alpha d$

$$f'(x; d) = \lim_{\alpha \to 0^+} \frac{f(x + \alpha d) - f(x)}{\alpha} = \nabla f(x)^T d$$

- d is a **descent direction** at x if

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

- by continuity, if d is a descent direction, then for some maximum step $\bar{\alpha}$

$$f(x + \alpha d) < f(x) \quad \forall \alpha \in (0, \bar{\alpha})$$

Question

Suppose we modify the standard gradient descent update by using a symmetric matrix B and setting

$$x^{k+1} = x^k - \alpha B \nabla f(x^k).$$

Under what condition on the eigenvalues of B is $-B \nabla f(x^k)$ guaranteed to be a descent direction for every nonzero $\nabla f(x^k)$?

a. All eigenvalues of B are strictly positive (i.e., B is positive definite).
b. All eigenvalues of B are non-positive.
c. All eigenvalues of B are less than -1.
d. B has at least one strictly positive eigenvalue.

Generic descent method

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

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

- compute descent direction $d^{(k)}$
- compute step size $\alpha^{(k)}$
- update $x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$
- stop if OPTIMAL or MAXITER

Questions:

- how to determine a starting point?
- what are advantages/disadvantages of different directions $d^{(k)}$?
- how to choose step size $\alpha^{(k)}$?
- reasonable stopping criteria?

Gradient descent

$$x^{k+1} = x^k + \alpha^k d, \qquad d = -\nabla f(x^k)$$

- if x^k is **not** stationary, ie, $\nabla f(x^k) \neq 0$, then negative gradient is a descent direction

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

- negative gradient is the **steepest descent direction** of f at x

$$-\frac{\nabla f(x)}{\|\nabla f(x)\|} = \underset{\|d\| \leq 1}{\operatorname{argmin}}\, f'(x; d) \quad \text{(most negative)}$$

Proof. Use Cauchy-Schwartz inequality: for any vectors $w, v \in \mathbb{R}^n$,

$$-\|w\| \cdot \|v\| \leq w^\mathsf{T} v \leq \|w\| \cdot \|v\|$$

and upper (or lower) bound achived if and only if w and v are parallel

Gradient method

Initialize: choose $x_0 \in \mathbb{R}^n$ and tolerance $\epsilon > 0$

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

1. choose step size α^k to approximately minimize

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

2. update $x^{k+1} = x^k - \alpha^k \nabla f(x^k)$
3. stop if $\|\nabla f(x^k)\| < \epsilon$

Step size selection

step size rules typically used in practice

- **exact** (generally not possible, except for quadratic f)

$$\alpha^k \in \operatorname*{argmin}_{\alpha \geq 0} \ \phi(\alpha), \qquad \phi(\alpha) := f(x^k + \alpha d^k)$$

- **constant** (cheap and easy, but requires analyzing f)

$$\alpha^k = \bar{\alpha} > 0 \quad \forall k$$

- **approximate** — backtracking linesearch, eg, Armijo (relatively cheap, no analysis required)
  - reduce α until sufficient decrease in f, ie, with $\mu \in (0, 1)$

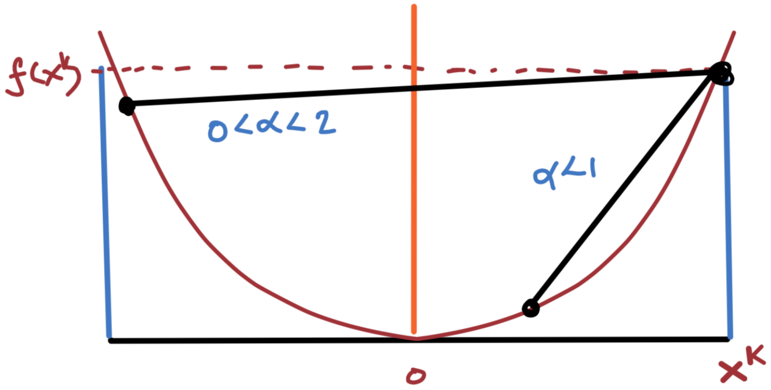
1. set $\alpha^k = \bar{\alpha} > 0$
2. until $f(x^k + \alpha^k d^k)$ "sufficiently less than" $f(x^k)$
   - $\alpha^k \leftarrow \alpha^k / 2$ (or some other divisor)
3. return $\alpha^{(k)}$

constant stepsize

Constant stepsize

- need to fix $\bar{\alpha} > 0$ small enough to ensure convergence
- sufficient condition: choose α small enough to guarantee

$$f(x^k + \bar{\alpha} d^k) < f(x^k) \quad \forall k$$



$f(x) = \frac{1}{2}x^2$ with x scalar and $d = -f'(x^k)$:

$$\begin{aligned} x^{k+1} &= x^k - \bar{\alpha} f'(x^k) \\ &= x^k - \bar{\alpha} x^k \\ &= (1 - \bar{\alpha}) x^k \\ &= (1 - \bar{\alpha})^{k+1} x^0 \end{aligned}$$

if $\bar{\alpha} \in (0, 2)$ then $|1 - \bar{\alpha}| < 1$ and

$$f(x^k) = \frac{1}{2}(1 - \bar{\alpha})^{2k} (x^{(0)})^2 \to 0 \quad \text{as} \quad k \to \infty$$

Constant stepsize — quadratic functions

$$f(x) = \tfrac{1}{2} x^\top H x + b^\top x + \gamma, \quad \text{with} \quad H \succ 0$$

- reliable constant stepsize $\bar{\alpha}$ depends on maximum eigenvalue

$$d^\top H d \leq \lambda_{\max}(H) \|d\|^2 \quad \forall d \in \mathbb{R}^n \qquad (1)$$

- behaviour of function value along steepest descent direction $d = -\nabla f(x)$

$$\begin{aligned} f(x + \alpha d) &= f(x) + \alpha d^\top \nabla f(x) + \tfrac{1}{2}\alpha^2 d^\top \nabla^2 f(x) d && \text{(exact because } f \text{ quadratic)} \\ &\leq f(x) - \alpha \|\nabla f(x)\|^2 + \tfrac{1}{2}\alpha^2 \lambda_{\max}(H) \|d\|^2 && \text{(by (1))} \\ &= f(x) - \underbrace{\left(\alpha - \tfrac{1}{2}\alpha^2 \lambda_{\max}(H)\right)}_{(\heartsuit)} \|\nabla f(x)\|^2 \end{aligned}$$

- if $\heartsuit > 0$ then $f(x + \alpha d) < f(x)$, as required, so choose

$$\alpha \in (0, 2/\lambda_{\max}(H))$$

Lipschitz smooth functions

for general smooth functions, constant stepsize depends on the Lipschitz constant of the gradient

Definition 1 (L-smooth functions) The function $f : \mathbb{R}^n \to \mathbb{R}$ is L-**Lipschitz smooth** if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^n$$

example — quadratic functions

$$f(x) = \frac{1}{2}x^\intercal Hx + b^\intercal x + \gamma, \quad \text{with} \quad H \succ 0$$

- f is $\lambda_{\max}(H)$-Lipschitz smooth because

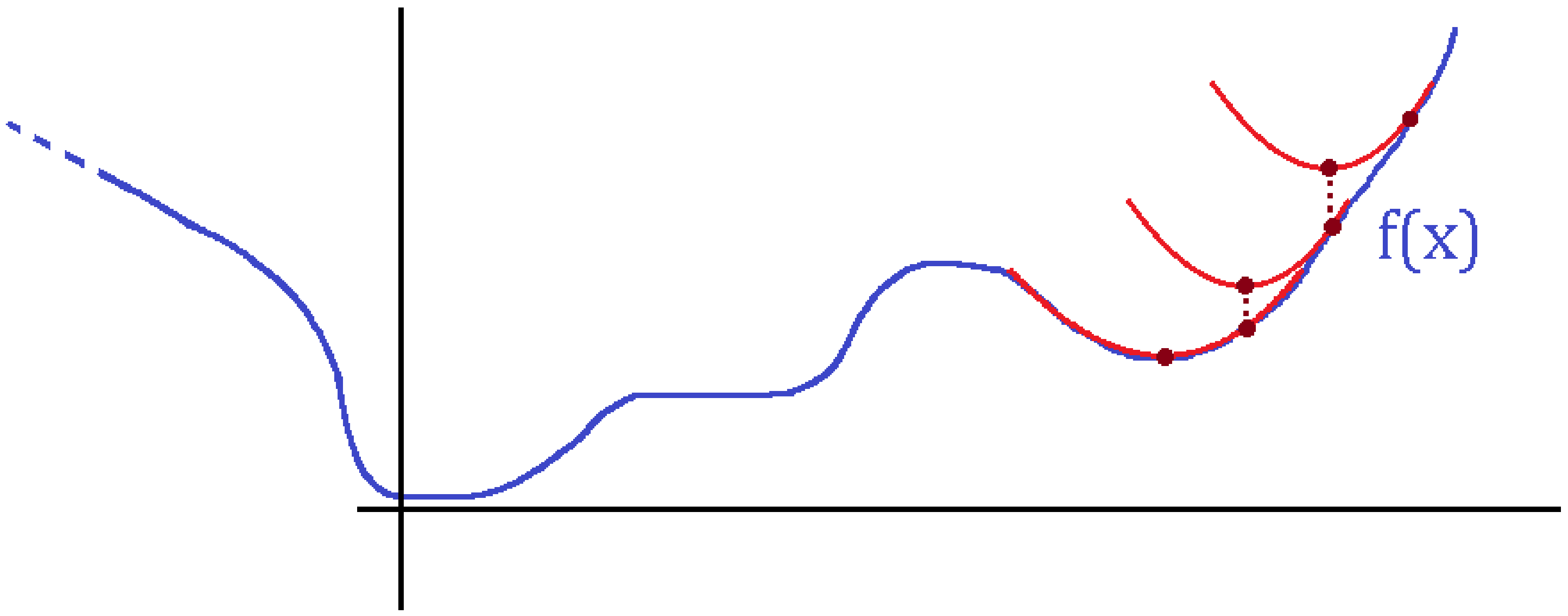
$$\begin{aligned}
\|\nabla f(x) - \nabla f(y)\| &= \|H(x - y)\| && (= \|(Hx + b) - (Hy + b)\|) \\
&= \|\Lambda U^\intercal (x - y)\| && (H = U\Lambda U^\intercal, \quad UU^\intercal = I) \\
&= \|\Lambda v\| && (v = U^\intercal (x - y)) \\
&= \sqrt{\textstyle\sum_{i=1}^n \lambda_i^2 v_i^2} \\
&\leq \lambda_{\max}(H)\|v\| \\
&= \lambda_{\max}(H)\|x - y\| && (\|v\| = \|x - y\|)
\end{aligned}$$

Second-order L-smooth characterization

If f is twice continuously differentiable, then f is L-Lipschitz smooth if and only if its Hessian is bounded by L, ie, for all $x \in \mathbb{R}^n$

$$\nabla^2 f(x) \preceq LI \quad \iff \quad LI - \nabla^2 f(x) \succeq 0$$

implies that quadratic approximation is a local upper bound



Question

Consider the nonlinear least-squares function

$$f(x) := \frac{1}{2}\|c(x)\|^2$$

where $c : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable with the $m \times n$ Jacobian $J(x) = \nabla c(x)^T$. Suppose the Jacobian's largest singular value is bounded by M for all x. Which of the following best describes the Lipschitz constant L for the gradient $\nabla f(x) = J(x)^T c(x)$?

a. $L = M$

b. $L = M^2$

c. $L = 2M$

d. $L = 2M^2$

(Recall that a function f is called L-smooth if $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all x, y.)

Example — logistic loss

- given feature/label pairs $(a_i, b_i) \in \mathbb{R}^n \times \{0,1\}, i = 1, \ldots, m$, find x to fit logistic model

$$\sigma(a_i^\intercal x) \approx b_i, \quad \text{where} \quad \sigma(t) = \frac{1}{1+e^{-t}}$$

- logistic loss problem, and objective gradient and Hessian

$$\min_x f(x) := -\sum_{i=1}^m b_i \log(\sigma(a_i^\intercal x)) + (1-b_i)\log(1-\sigma(a_i^\intercal x))$$

$$\nabla f(x) = A^\intercal r, \quad \nabla^2 f(x) = A^\intercal D A, \quad r = \sigma.(Ax) - b, \quad D = \mathbf{Diag}(r_i(1-r_i))_{i=1}^m$$

- because diagonals of D are in $(0, 1/4)$, for all unit-norm u,

$$u^\intercal \nabla^2 f(x) u = u^\intercal (A^\intercal D A) u \le \frac{1}{4} u^\intercal (A^\intercal A) u \le \frac{1}{4}\lambda_{\max}(A^\intercal A)$$

- so f is L-Lipschitz smooth with $L = \lambda_{\max}(A^\intercal A)/4$

exact linesearch

Exact linesearch

- exact linesearch typically only possible for quadratic functions

$$f(x) = \tfrac{1}{2} x^\top H x + b^\top x + \gamma, \quad \text{with} \quad H \succ 0$$

- exact linesearch solves the 1-dimensional optimization problem with d descent dir:

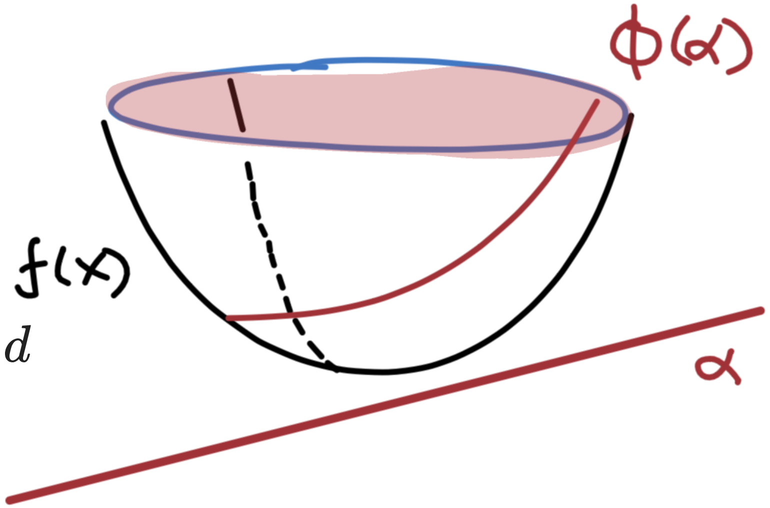
$$\min_{\alpha \geq 0} \; \phi(\alpha) := f(x + \alpha d)$$

- exact step computation:

$$\phi(\alpha) = \tfrac{1}{2}(x + \alpha d)^\top H (x + \alpha d) + b^\top (x + \alpha d) + \gamma$$

$$\phi'(\alpha) = \alpha d^\top H d + x^\top H d + b^\top d = \alpha d^\top H d + \nabla f(x)^\top d$$

$$\phi'(\alpha^*) = 0 \quad \Longleftrightarrow \quad \alpha^* = -\frac{\nabla f(x)^\top d}{d^\top H d}$$

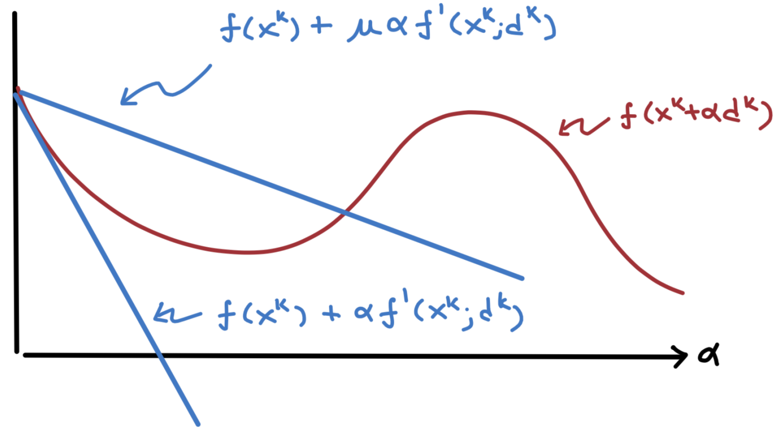


backtracking

Backtracking linesearch (Armijo)

pull back along descent direction d^k until **sufficient decrease** in f

- $f'(x^k; d^k) < 0$
- sufficient descent parameter $\mu \in (0, 1)$




```
function armijo(f, ∇f, x, d; μ=1e-4, α=1, ρ=0.5, maxits=10)
    for k in 1:maxits
       if f(x+α*d) < f(x) + μ*α*dot(∇f(x),d)
           return α
       end
       α *= ρ
    end
    error("backtracking linesearch failed")
end;
```

Convergence of gradient method

$f : \mathbb{R}^n \to \mathbb{R}$ L-smooth

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k)$$

with

- constant stepsize $\alpha^k = \bar{\alpha} \in (0, 2/L)$
- exact stepsize $\alpha^k = \operatorname{argmin}_{\alpha \geq 0} f(x^k + \alpha d^k)$
- backtracking stepsize α^k with $\mu \in (0, 1)$

guarantee – for all $k = 0, 1, 2, \ldots$

- **descent** (unless $\nabla f(x^k) = 0$)

$$f(x^{k+1}) < f(x^k)$$

- **convergence**

$$\|\nabla f(x^k)\| \to 0$$