Second-Order Optimality

CPSC 406 – Computational Optimization

Hessians and second-order optimality

- sufficient optimality conditions in $\mathbb{R}$
- positive definite matrices
- Hessians
- quadratic functions
- sufficient optimality conditions in $\mathbb{R}^n$

Necessary conditions (1-D)

Suppose that $f : \mathbb{R} \to \mathbb{R}$ is twice continuosly differentiable

necessary optimality conditions

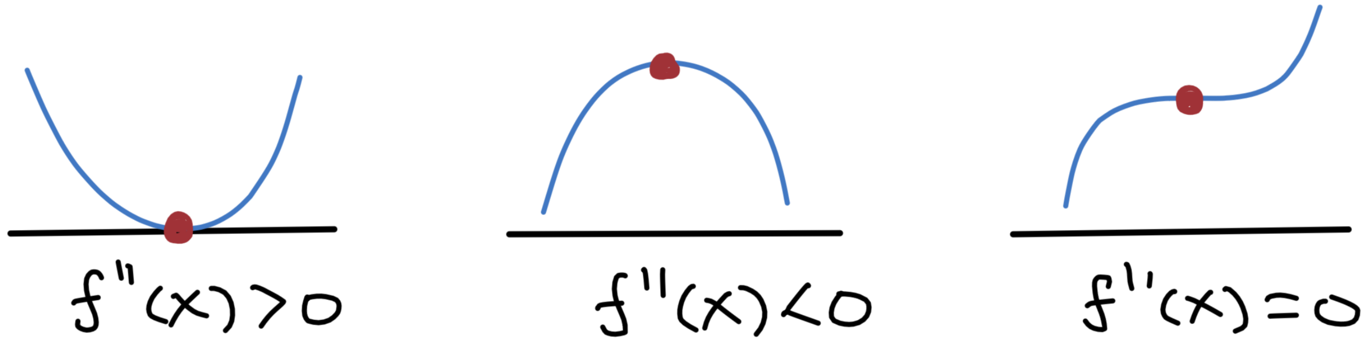
x^* is a local minimizer **only if**

- (**first-order**) $f'(x^*) = 0$
- (**second-order**) $f''(x^*) \geq 0$

sufficient optimality conditions

x^* is a local minimizer **if**

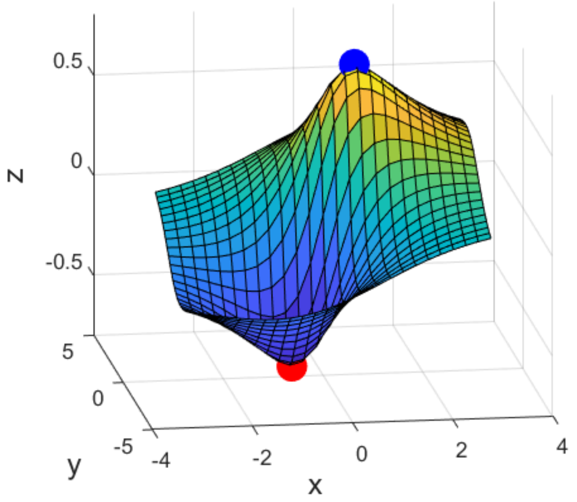
- (**first-order**) $f'(x^*) = 0$
- (**second-order**) $f''(x^*) > 0$

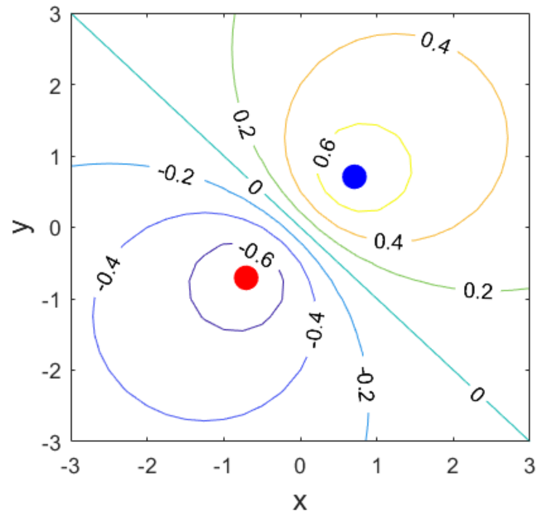


- generalize second-order conditions to $\mathbb{R}^n$

Example

$$\min_{x \in \mathbb{R}^2} \frac{x_1 + x_2}{x_1^2 + x_2^2 + 1}$$






```julia
using ForwardDiff
f(x) = (x[1]+x[2])/(x[1]^2+x[2]^2+1)
∇f(x) = ForwardDiff.gradient(f, x)

x =  [1, 1]/sqrt(2);
@show ∇f(+x)
@show ∇f(-x);
```

```
∇f(+x) = [1.1102230246251565e-16, 1.1102230246251565e-16]
∇f(-x) = [1.1102230246251565e-16, 1.1102230246251565e-16]
```

- Both x and $-x$ are stationary. Which is minimial/maximal?

positive definite matrices

Positive (semi)definite matrices

Let H by n-by-n matrix with $H = H^\intercal$ (symmetric)

- H is **positive semidefinite** ($H \succeq 0$) if

$$x^\intercal H x \geq 0 \quad \text{for all} \quad x \in \mathbb{R}^n$$

- H is **positive definite** ($H \succ 0$) if

$$x^\intercal H x > 0 \quad \text{for all} \quad 0 \neq x \in \mathbb{R}^n$$

- H is **negative semidefinite** if $-H$ is positive semidefinite, ie, $H \preceq 0 \iff -H \succeq 0$
- H is **negative definite** if $-H$ is positive definite, ie, $H \prec 0 \iff -H \succ 0$

- H is **indefinite** if it is neither positive nor negative semidefinite, ie,

$$\exists\, x \neq y \in \mathbb{R}^n \quad \text{such that} \quad x^\intercal H x > 0 \quad \text{and} \quad y^\intercal H y < 0$$

Question

The matrix $H = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$ is

a. positive definite

b. positive semidefinite

c. negative definite

d. negative semidefinite

e. indefinite

Diagonal matrices

$$D = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix} = \mathbf{Diag}(d_1, d_2, \ldots, d_n)$$

- $D \succ 0 \quad \iff \quad d_i > 0$ for all i
- $D \succeq 0 \quad \iff \quad d_i \geq 0$ for all i

Eigenpairs of symmetric matrices

Let H be a n-by-n matrix. Then $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}$ is an **eigenvector**/**eigenvalue** pair of H if

$$Hx = \lambda x$$

Theorem 1 (Eigenvalues of symmetric matrices) If H is n-by-n and symmetric, then there exists n orthogonal eigenvectors and all eigenvalues are real.

$$\begin{Bmatrix} Hx_1 = \lambda_1 x_1 \\ Hx_2 = \lambda_2 x_2 \\ \vdots \\ Hx_n = \lambda_n x_n \end{Bmatrix} \quad \text{or} \quad HX = X\Lambda$$

where $X^\intercal = X^{-1}$ (orthogonal) and Λ is a diagonal matrix of eigenvalues:

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

Eigenvalues and definiteness

The matrix H is positive (semi) definite if and only if all of its eigenvalues are (nonnegative) positive.

proof (positive definite)

- by spectral theorem,

$$X^\top H X = \Lambda \quad \text{where} \quad X^\top = X^{-1} \quad \text{and} \quad \Lambda = \mathbf{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

- for any $x \in \mathbb{R}^n$ there exists $y = (y_1, \ldots, y_n)$ such that $x = Xy$ and

$$x^\top H x = y^\top X^\top H X y = y^\top \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2$$

- thus, $x^\top H x > 0$ for all $x \neq 0$ (ie, H positive definite) if and only if

$$\sum_{i=1}^{n} \lambda_i y_i^2 > 0 \quad \text{for all} \quad y \neq 0 \quad \iff \quad \lambda_i > 0 \quad \text{for all} \quad i = 1 : n$$

Example

```julia
using LinearAlgebra
```

$$H = \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}$$

```julia
H = [4 1; 1 3]
@show eigvals(H);
```

```
eigvals(H) = [2.381966011250105, 4.618033988749895]
```

$$H = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1/10 \end{bmatrix}$$

```julia
H = ones(3,3)
H[3, 3] = 1/10
@show eigvals(H);
```

```
eigvals(H) = [-0.6536725037400826, -2.3721342664653315e-17, 2.7536725037400815]
```

Equivalent conditions

Let H be a n-by-n symmetric matrix.

positive definite equivalences:

1. all eigenvalues of H are positive
2. $x^\intercal H x > 0$ for all $0 \neq x \in \mathbb{R}^n$
3. $H = R^\intercal R$ for some nonsingular n-by-n matrix R
4. H is symmetric and all of its leading principal minors are positive

positive semidefinite equivalences:

1. all eigenvalues of H are nonnegative
2. $x^\intercal H x \geq 0$ for all $x \in \mathbb{R}^n$
3. $H = R^\intercal R$ for some n-by-n matrix R
4. H is symmetric and all of its principal minors are nonnegative

Hessians

For $f : \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable, the **Hessian** of f at $x \in \mathbb{R}^n$ is the n-by-n symmetric matrix

$$H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{bmatrix} \qquad \frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

example

$$f(x) = x_1^2 + 8x_1x_2 - 2x_3^3, \quad \nabla f(x) = \begin{bmatrix} 2x_1 + 8x_2 \\ 8x_1 \\ -6x_3^2 \end{bmatrix}, \quad H(x) = \begin{bmatrix} 2 & 8 & 0 \\ 8 & 0 & 0 \\ 0 & 0 & -12x_3 \end{bmatrix}$$

Quadratic Functions

Quadratic functions

Quadratic functions over $\mathbb{R}^n$ have the form

$$f(x) = \tfrac{1}{2} x^\intercal H x + b^\intercal x + \gamma$$

where H is symmetric and $b \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$.

- $n = 1$

$$f(x) = \tfrac{1}{2} h x^2 + bx + \gamma, \quad H = [h]$$

- $n = 2$

$$\begin{aligned} f(x) &= \tfrac{1}{2} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_1 & b_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \gamma \\ &= \tfrac{1}{2} h_{11} x_1^2 + h_{12} x_1 x_2 + \tfrac{1}{2} h_{22} x_2^2 + b_1 x_1 + b_2 x_2 + \gamma \end{aligned}$$

Quadratic functions and symmetry

$$f(x) = \frac{1}{2}x^\top H x + b^\top x + \gamma, \quad \nabla f(x) = Hx + b, \quad \nabla^2 f(x) = H$$

We can always assume without loss of generality that

$$H = H^\top \quad \text{(symmetric)}$$

Suppose that $H \neq H^\top$:

$$x^\top H x = \frac{1}{2}x^\top H x + \frac{1}{2}x^\top H^\top x = x^\top \left[\frac{1}{2}(H + H^\top)\right] x$$

Thus we can replace H with $\frac{1}{2}(H + H^\top)$ and not change the function value

Optimality for quadratic functions

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}x^\intercal Hx + b^\intercal x + \gamma$$

$$\nabla f(x) = Hx + b, \quad \nabla^2 f(x) = H$$

optimality conditions

- (necessary) x^* is optimal **only if** $\nabla f(x^*) = Hx^* + b = 0$ (stationary)
- (sufficient) if stationary and $H \succeq 0$, then x^* is a global minimizer
- (sufficient) if stationary and $H \succ 0$, then x^* is the unique global minimizer

proof

for all $d \neq 0$,

$$f(x^* + d) - f(x^*) = d^\intercal \underbrace{\nabla f(x^*)}_{=0} + \underbrace{\frac{1}{2}d^\intercal \overbrace{\nabla^2 f(x^*)}^{=H} d}_{\geq 0} \begin{cases} \geq 0 & \text{if } H \succeq 0 \\ > 0 & \text{if } H \succ 0 \end{cases}$$

Nonlinear functions

Directional second derivatives

Given $f : \mathbb{R}^n \to \mathbb{R}$, recall the directional derivative

$$f'(x; d) = \lim_{\alpha \to 0^+} \frac{f(x + \alpha d) - f(x)}{\alpha} = d^\intercal \nabla f(x)$$

Definition 1 The **directional second derivative** of f at x in the direction d is

$$f''(x; d) = \lim_{\alpha \to 0^+} \frac{f'(x + \alpha d; d) - f'(x; d)}{\alpha} = d^\intercal \nabla^2 f(x) d$$

partial 2nd derivatives are the directional 2nd derivatives along each canonical basis vector e_i:

$$\frac{\partial^2 f}{\partial x_i^2}(x) = f''(x; e_i) \quad \text{with} \quad e_i(j) = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases}$$

Linear and quadratic approximations

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable.

Theorem 2 (Linear approximation) For all $x \in \mathbb{R}^n$ and $\epsilon > 0$, for each $y \in \epsilon\mathbf{B}(x)$ there exists $z \in [x, y]$ such that

$$f(y) = f(x) + \nabla f(x)^\intercal (y - x) + \frac{1}{2}(y - x)^\intercal \nabla^2 f(z)(y - x)$$

Theorem 3 (Quadratic approximation) For all x and d in $\mathbb{R}^n$,

$$f(x + d) = f(x) + \nabla f(x)^\intercal d + \frac{1}{2} d^\intercal \nabla^2 f(x) d + o(\|d\|^2)$$

Second-order necessary conditions

For $f : \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable and $\bar{x} \in \mathbb{R}^n$ *stationary* (ie, $\nabla f(\bar{x}) = 0$)

- $\bar{x}$ is a **local min** $\implies \nabla^2 f(\bar{x}) \succeq 0$
- $\bar{x}$ is a **local max** $\implies \nabla^2 f(\bar{x}) \preceq 0$

proof sketch for local min (analogous for local max). If $\bar{x}$ is a local min, then for all $d \neq 0$

$$0 \leq f(\bar{x} + \alpha d) - f(\bar{x}) = d^\intercal \underbrace{\nabla f(\bar{x})}_{=0} + \tfrac{1}{2}\alpha^2 d^\intercal \nabla^2 f(\bar{x}) d + o(\alpha^2 \|d\|^2)$$

Divide both sides by α^2 and take the limit as $\alpha \to 0^+$. Because $o(\alpha^2\|d\|^2)/\alpha^2 \to 0$,

$$0 \leq d^\intercal \nabla^2 f(\bar{x}) d$$

Because this hold for all $d \neq 0$,

$$\nabla^2 f(\bar{x}) \succeq 0$$

Sufficient conditions for optimality

For $f : \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable and $\bar{x} \in \mathbb{R}^n$ *stationary*,

- $\nabla f(\bar{x}) = 0$
- $\nabla^2 f(\bar{x}) \succ 0 \quad \Longrightarrow \quad \bar{x}$ is a local min
- $\nabla^2 f(\bar{x}) \prec 0 \quad \Longrightarrow \quad \bar{x}$ is a local max

proof sketch for local min (analogous for local max). By linear approximation theorem and continuity of $\nabla^2 f$, for any x close enough to $\bar{x}$ there exists $z \in [\bar{x}, x]$ such that

$$f(x) - f(\bar{x}) = (x - \bar{x})^\mathsf{T} \underbrace{\nabla f(\bar{x})}_{=0} + \frac{1}{2}(x - \bar{x})^\mathsf{T} \underbrace{\nabla^2 f(z)}_{\succ 0}(x - \bar{x}) > 0$$

Question

$$f(x) = x_1^2 + 8x_1x_2 - 2x_3^3, \quad \nabla f(x) = \begin{bmatrix} 2x_1 + 8x_2 \\ 8x_1 \\ -6x_3^2 \end{bmatrix}, \quad H(x) = \begin{bmatrix} 2 & 8 & 0 \\ 8 & 0 & 0 \\ 0 & 0 & -12x_3 \end{bmatrix}$$

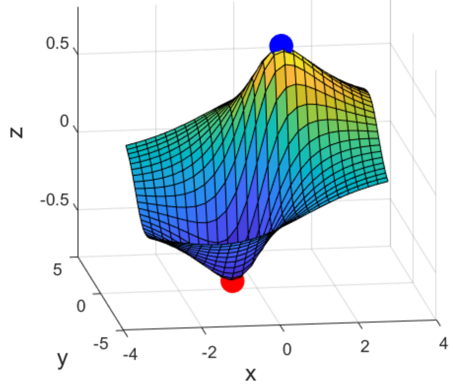
The stationary point $x^* = (0, 0, 0)$ is a

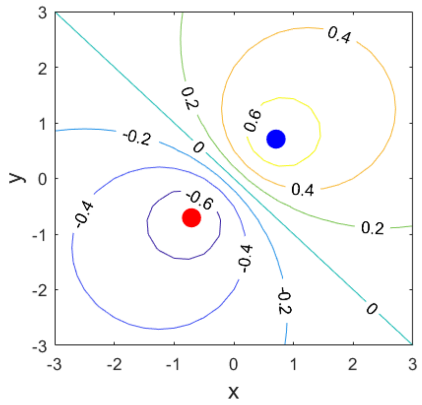
a. minimizer

b. maximizer

c. saddle point

Example





$$\min_{x,y} \; f(x,y) = \frac{x+y}{x^2+y^2+1}$$

$$\nabla f(x,y) = \frac{1}{(x^2+y^2+1)^2} \begin{bmatrix} y^2 - 2xy - x^2 + 1 \\ x^2 - 2xy - y^2 + 1 \end{bmatrix}$$

Stationary points $\nabla f(x^*, y^*) = 0$:

$$\underbrace{(x_1^*, y_1^*) = -\frac{1}{\sqrt{2}}(1,1)}_{\text{minimizer}} \qquad \underbrace{(x_2^*, y_2^*) = +\frac{1}{\sqrt{2}}(1,1)}_{\text{maximizer}}$$

Hessian of f at these points:

$$\nabla^2 f(x_1^*, y_1^*) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \succ 0 \qquad \nabla^2 f(x^*)2, y_2^*) = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \prec 0$$