Optimization Prerequisites

HESSIAN MATRIX

For real-valued function $f : \mathcal{S} \to \mathbb{R}$, the **Hessian** matrix $H : \mathcal{S} \to \mathbb{R}^{d \times d}$ contains all their second derivatives (if they exist):

$$H(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \left(\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right)_{i,j=1,\dots,d}$$

Note: Hessian of f is Jacobian of ∇f

Example: Let $f(\mathbf{x}) = \sin(x_1) \cdot \cos(2x_2)$. Then:

$$H(\mathbf{x}) = \begin{pmatrix} -\cos(2x_2) \cdot \sin(x_1) & -2\cos(x_1) \cdot \sin(2x_2) \\ -2\cos(x_1) \cdot \sin(2x_2) & -4\cos(2x_2) \cdot \sin(x_1) \end{pmatrix}$$

- If $f \in \mathcal{C}^2$, then H is symmetric
- Many local properties (geometry, convexity, critical points) are encoded by the Hessian and its spectrum ($\to$ later)

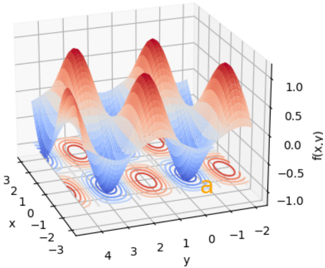
LOCAL CURVATURE BY HESSIAN

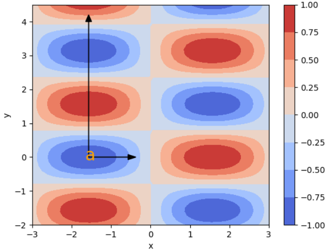
Eigenvector corresponding to largest (resp. smallest) **eigenvalue** of Hessian points in direction of largest (resp. smallest) **curvature**

Example (previous slide)**:** For $\boldsymbol{a} = (-\pi/2, 0)^T$, we have

$$H(\boldsymbol{a}) = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$$

and thus $\lambda_1 = 4$, $\lambda_2 = 1$, $\boldsymbol{v}_1 = (0, 1)^T$, and $\boldsymbol{v}_2 = (1, 0)^T$.

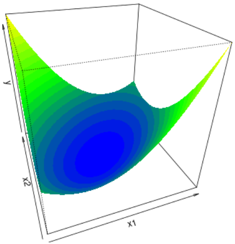




https://www.desmos.com/3d

Optimization in Machine Learning

Mathematical Concepts: Quadratic forms I



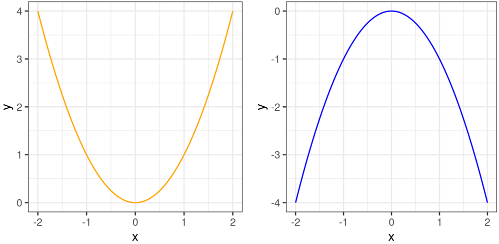
Learning goals

- Definition of quadratic forms
- Gradient, Hessian
- Optima

UNIVARIATE QUADRATIC FUNCTIONS

Consider a **quadratic function** $q : \mathbb{R} \to \mathbb{R}$

$$q(x) = a \cdot x^2 + b \cdot x + c, \qquad a \neq 0.$$

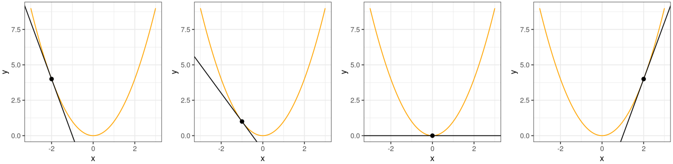


A quadratic function $q_1(x) = x^2$ (**left**) and $q_2(x) = -x^2$ (**right**).

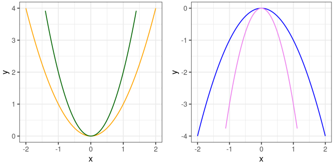
UNIVARIATE QUADRATIC FUNCTIONS

Basic properties:

- **Slope** of tangent at point $(x, q(x))$ is given by $q'(x) = 2 \cdot a \cdot x + b$



- **Curvature** of q is given by $q''(x) = 2 \cdot a$.

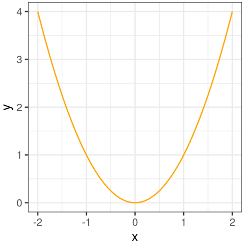


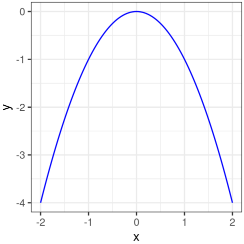
$q_1 = x^2$ (orange), $q_2 = 2x^2$ (green), $q_3(x) = -x^2$ (blue), $q_4 = -3x^2$ (magenta)

UNIVARIATE QUADRATIC FUNCTIONS

- **Convexity/Concavity**:
  - $a > 0$: q convex, bounded from below, unique global **minimum**
  - $a < 0$: q concave, bounded from above, unique global **maximum**
- **Optimum** x^*:

$$q'(x^*) = 0 \quad \Leftrightarrow \quad 2ax^* + b = 0 \quad \Leftrightarrow \quad x^* = \frac{-b}{2a}$$





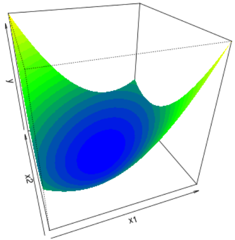
Left: $q_1(x) = x^2$ (convex). **Right:** $q_2(x) = -x^2$ (concave).

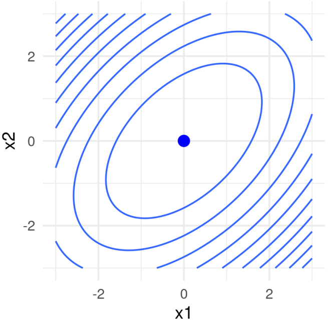
MULTIVARIATE QUADRATIC FUNCTIONS

A quadratic function $q : \mathbb{R}^d \to \mathbb{R}$ has the following form:

$$q(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \boldsymbol{b}^T \mathbf{x} + c$$

with $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ full-rank matrix, $\boldsymbol{b} \in \mathbb{R}^d$, $c \in \mathbb{R}$.





MULTIVARIATE QUADRATIC FUNCTIONS

W.l.o.g., assume **A symmetric**, i.e., $\mathbf{A}^T = \mathbf{A}$.

If **A** not symmetric, there is always a symmetric matrix $\tilde{\mathbf{A}}$ s.t.

$$q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\tilde{\mathbf{A}}\mathbf{x} = \tilde{q}(\mathbf{x}).$$

Justification: We write

$$q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} = \frac{1}{2}\mathbf{x}^T\underbrace{(\mathbf{A}+\mathbf{A}^T)}_{\tilde{\mathbf{A}}_1}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\underbrace{(\mathbf{A}-\mathbf{A}^T)}_{\tilde{\mathbf{A}}_2}\mathbf{x}$$

with $\tilde{\mathbf{A}}_1$ symmetric, $\tilde{\mathbf{A}}_2$ anti-symmetric (i.e., $\tilde{\mathbf{A}}_2^T = -\tilde{\mathbf{A}}_2$). Since $\mathbf{x}^T\mathbf{A}^T\mathbf{x}$ is a scalar, it is equal to its transpose:

$$\begin{aligned}\mathbf{x}^T(\mathbf{A}-\mathbf{A}^T)\mathbf{x} &= \mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{A}^T\mathbf{x} = \mathbf{x}^T\mathbf{A}\mathbf{x} - \left(\mathbf{x}^T\mathbf{A}^T\mathbf{x}\right)^T \\ &= \mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{A}\mathbf{x} = 0.\end{aligned}$$

Therefore, $q(\mathbf{x}) = \tilde{q}(\mathbf{x})$ with $\tilde{q}(\mathbf{x}) = \mathbf{x}^T\tilde{\mathbf{A}}\mathbf{x}$ with $\tilde{\mathbf{A}} = \tilde{\mathbf{A}}_1/2$.

GRADIENT AND HESSIAN

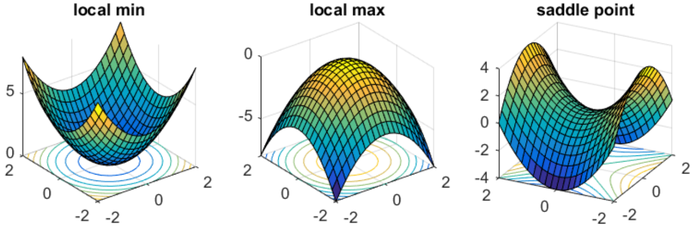
- The **gradient** of q is
$$\nabla q(\mathbf{x}) = \left(\mathbf{A}^T + \mathbf{A}\right)\mathbf{x} + \boldsymbol{b} = 2\mathbf{A}\mathbf{x} + \boldsymbol{b} \in \mathbb{R}^d$$
Derivative in direction $\boldsymbol{v} \in \mathbb{R}^d$ is (by chain rule)
$$\frac{\mathrm{d}q(\mathbf{x} + h \cdot \boldsymbol{v})}{\mathrm{d}h}\Bigg|_{h=0} = \nabla q(\mathbf{x} + h\boldsymbol{v})^T \boldsymbol{v}\Bigg|_{h=0} = \nabla q(\mathbf{x})^T \boldsymbol{v}.$$

- The **Hessian** of q is
$$\nabla^2 q(\mathbf{x}) = \left(\boldsymbol{A}^T + \boldsymbol{A}\right) = 2\mathbf{A} =: \mathbf{H} \in \mathbb{R}^{d\times d}$$
Curvature in direction of $\boldsymbol{v} \in \mathbb{R}^d$ is (by chain rule)
$$\frac{\mathrm{d}^2 q(\mathbf{x} + h \cdot \boldsymbol{v})}{\mathrm{d}h^2}\Bigg|_{h=0} = \boldsymbol{v}^T \nabla^2 q(\mathbf{x} + h\boldsymbol{v})\boldsymbol{v}\Bigg|_{h=0} = \boldsymbol{v}^T \mathbf{H}\boldsymbol{v}.$$

OPTIMUM

Since **A** has full rank, there exists a *unique* stationary point $\mathbf{x}^*$ (minimum, maximum, or saddle point):

$$\nabla q(\mathbf{x}^*) = 0$$
$$2\mathbf{A}\mathbf{x}^* + \boldsymbol{b} = 0$$
$$\mathbf{x}^* = -\frac{1}{2}\mathbf{A}^{-1}\boldsymbol{b}.$$



Left: **A** positive definite. **Middle:** **A** negative definite. **Right:** **A** indefinite.

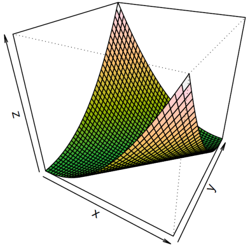
OPTIMA: RANK-DEFICIENT CASE

Example: Assume $\mathbf{A}$ is **not** full rank but has a zero eigenvalue with eigenvector $\boldsymbol{v}_0$.

- Recall: $\boldsymbol{v}_0$ spans null space of $\mathbf{A}$, i.e., $\mathbf{A}(\alpha \boldsymbol{v}_0) = 0$ for each $\alpha \in \mathbb{R}$
- $\implies \mathbf{A}(\mathbf{x} + \alpha \boldsymbol{v}_0) = \mathbf{A}\mathbf{x}$
- Since $\nabla q(\mathbf{x}) = 2\mathbf{A}\mathbf{x} + \boldsymbol{b}$:

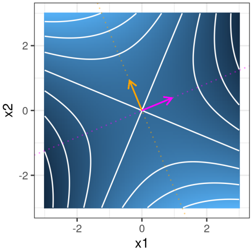
$$\nabla q(\mathbf{x} + \alpha \boldsymbol{v}_0) = 2\mathbf{A}(\mathbf{x} + \alpha \boldsymbol{v}_0) + \boldsymbol{b} = 2\mathbf{A}\mathbf{x} + \boldsymbol{b} = \nabla q(\mathbf{x})$$

- $\implies$ q has infinitely many stationary points along line $\mathbf{x}^* + \alpha \boldsymbol{v}_0$
- Since $\mathbf{H} = 2\mathbf{A}$, kind of stationary point not changing along $\boldsymbol{v}_0$



Optimization in Machine Learning

Mathematical Concepts
Quadratic forms II



Learning goals

- Geometry of quadratic forms
- Spectrum of Hessian

PROPERTIES OF QUADRATIC FUNCTIONS

Recall: Quadratic form q

- Univariate: $q(x) = ax^2 + bx + c$
- Multivariate: $q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + \boldsymbol{b}^T\mathbf{x} + c$

General observation: If $q \geq 0$ ($q \leq 0$), q is convex (concave)

Univariate function: Second derivative is $q''(x) = 2a$

- $q''(x) \overset{(>)}{\geq} 0$: q (strictly) convex. $q''(x) \overset{(<)}{\leq} 0$: q (strictly) concave.
- High (low) absolute values of $q''(x)$: high (low) curvature

Multivariate function: Second derivative is $\mathbf{H} = 2\mathbf{A}$

- Convexity/concavity of q depend on eigenvalues of $\mathbf{H}$
- Let us look at an example of the form $q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$

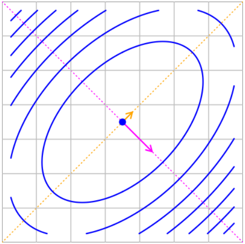
GEOMETRY OF QUADRATIC FUNCTIONS

Example: $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \implies \mathbf{H} = 2\mathbf{A} = \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix}$

- Since $\mathbf{H}$ symmetric, eigendecomposition $\mathbf{H} = \mathbf{V}\Lambda\mathbf{V}^T$ with

$$\mathbf{V} = \begin{pmatrix} | & | \\ v_{\text{max}} & v_{\text{min}} \\ | & | \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \text{ orthogonal}$$

$$\text{and } \Lambda = \begin{pmatrix} \lambda_{\text{max}} & 0 \\ 0 & \lambda_{\text{min}} \end{pmatrix} = \begin{pmatrix} 6 & 0 \\ 0 & 2 \end{pmatrix}.$$



GEOMETRY OF QUADRATIC FUNCTIONS

- $\boldsymbol{v}_{\max}$ ($\boldsymbol{v}_{\min}$) direction of highest (lowest) curvature

  Proof: With $\boldsymbol{v} = \mathbf{V}^T\mathbf{x}$:

  $$\mathbf{x}^T\mathbf{H}\mathbf{x} = \mathbf{x}^T\mathbf{V}\Lambda\mathbf{V}^T\mathbf{x} = \boldsymbol{v}^T\Lambda\boldsymbol{v} = \sum_{i=1}^{d}\lambda_i v_i^2 \leq \lambda_{\max}\sum_{i=1}^{d} v_i^2 = \lambda_{\max}\|\boldsymbol{v}\|^2$$

  Since $\|\boldsymbol{v}\| = \|\mathbf{x}\|$ ($\mathbf{V}$ orthogonal): $\max_{\|\mathbf{x}\|=1}\mathbf{x}^T\mathbf{H}\mathbf{x} \leq \lambda_{\max}$
  Additional: ${\boldsymbol{v}_{\max}}^T\mathbf{H}\boldsymbol{v}_{\max} = \mathbf{e}_1^T\Lambda\mathbf{e}_1 = \lambda_{\max}$

  Analogous: $\min_{\|\mathbf{x}\|=1}\mathbf{x}^T\mathbf{H}\mathbf{x} \geq \lambda_{\min}$ and ${\boldsymbol{v}_{\min}}^T\mathbf{H}\boldsymbol{v}_{\min} = \lambda_{\min}$

- Contour lines of any quadratic form are ellipses
  (with eigenvectors of A as principal axes, principal axis theorem)

  Look at $q(\mathbf{x}) = \mathbf{x}^T\boldsymbol{A}\mathbf{x} + \boldsymbol{b}^T\mathbf{x} + c$
  Now use $\boldsymbol{y} = \mathbf{x} - \boldsymbol{w} = \mathbf{x} + \frac{1}{2}\boldsymbol{A}^{-1}\boldsymbol{b}$
  This already gives us the general form of an ellipse:
  $\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} = (\mathbf{x} - \boldsymbol{w})^T\boldsymbol{A}(\mathbf{x} - \boldsymbol{w}) = q(\mathbf{x}) + const$
  If we use $\boldsymbol{z} = \boldsymbol{V}^T y$ we obtain it in standard form
  $\sum_{i=1}^{n}\lambda_i z_i^2 = \boldsymbol{z}^T\boldsymbol{\Lambda}\boldsymbol{z} = y^T\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T y = \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} = q(\mathbf{x}) + const$

GEOMETRY OF QUADRATIC FUNCTIONS

Recall: **Second order condition for optimality** is **sufficient**.

We skipped the **proof** at first, but can now catch up on it.
If $H(\mathbf{x}^*) \succ 0$ at stationary point $\mathbf{x}^*$, then $\mathbf{x}^*$ is local minimum ($\prec$ for maximum).

Proof: Let $\lambda_{\min} > 0$ denote the smallest eigenvalue of $H(\mathbf{x}^*)$. Then:

$$f(\mathbf{x}) = f(\mathbf{x}^*) + \underbrace{\nabla f(\mathbf{x}^*)}_{=0}{}^T(\mathbf{x} - \mathbf{x}^*) + \frac{1}{2}\underbrace{(\mathbf{x} - \mathbf{x}^*)^T H(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)}_{\geq \lambda_{\min} \|\mathbf{x} - \mathbf{x}^*\|^2 \text{ (see above)}} + \underbrace{R_2(\mathbf{x}, \mathbf{x}^*)}_{= o(\|\mathbf{x} - \mathbf{x}^*\|^2)} .$$

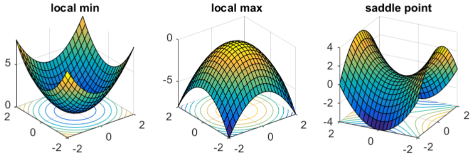
Choose $\epsilon > 0$ s.t. $|R_2(\mathbf{x}, \mathbf{x}^*)| < \frac{1}{2}\lambda_{\min}\|\mathbf{x} - \mathbf{x}^*\|^2$ for each $\mathbf{x} \neq \mathbf{x}^*$ with $\|\mathbf{x} - \mathbf{x}^*\| < \epsilon$.
Then:

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \underbrace{\frac{1}{2}\lambda_{\min}\|\mathbf{x} - \mathbf{x}^*\|^2 + R_2(\mathbf{x}, \mathbf{x}^*)}_{>0} > f(\mathbf{x}^*) \quad \text{for each } \mathbf{x} \neq \mathbf{x}^* \text{ with } \|\mathbf{x} - \mathbf{x}^*\| < \epsilon.$$

GEOMETRY OF QUADRATIC FUNCTIONS

If spectrum of $\mathbf{A}$ is known, also that of $\mathbf{H} = 2\mathbf{A}$ is known.

- If **all** eigenvalues of $\mathbf{H} \overset{(>)}{\geq} 0$ ($\Leftrightarrow \mathbf{H} \overset{(\succ)}{\succcurlyeq} 0$):
  - q (strictly) convex,
  - there is a (unique) global minimum.
- If **all** eigenvalues of $\mathbf{H} \overset{(<)}{\leq} 0$ ($\Leftrightarrow \mathbf{H} \overset{(\prec)}{\preccurlyeq} 0$):
  - q (strictly) concave,
  - there is a (unique) global maximum.
- If $\mathbf{H}$ has both positive and negative eigenvalues ($\Leftrightarrow \mathbf{H}$ indefinite):
  - q neither convex nor concave,
  - there is a saddle point.

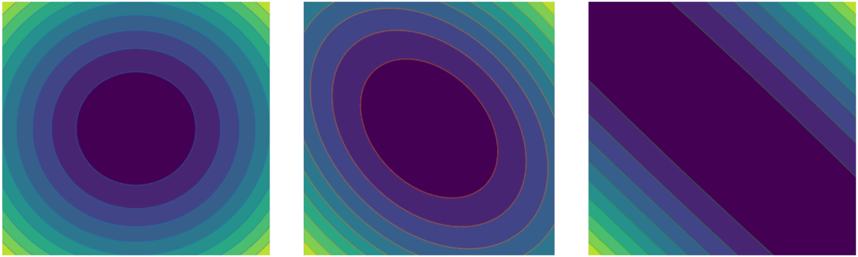


CONDITION AND CURVATURE

Condition of $\mathbf{H} = 2\mathbf{A}$ is given by $\kappa(\mathbf{H}) = \kappa(\mathbf{A}) = |\lambda_{\text{max}}|/|\lambda_{\text{min}}|$.

High condition means:

- $|\lambda_{\text{max}}| \gg |\lambda_{\text{min}}|$
- Curvature along $\mathbf{v}_{\text{max}} \gg$ curvature along $\mathbf{v}_{\text{min}}$
- **Problem** for optimization algorithms like **gradient descent** (later)

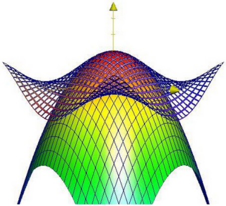


Left: Excellent condition. **Middle:** Good condition. **Right:** Bad condition.

APPROXIMATION OF SMOOTH FUNCTIONS

Any function $f \in \mathcal{C}^2$ can be locally approximated by a quadratic function via second order Taylor approximation:

$$f(\mathbf{x}) \approx f(\tilde{\boldsymbol{x}}) + \nabla f(\tilde{\boldsymbol{x}})^T(\mathbf{x} - \tilde{\boldsymbol{x}}) + \frac{1}{2}(\mathbf{x} - \tilde{\boldsymbol{x}})^T \nabla^2 f(\tilde{\boldsymbol{x}})(\mathbf{x} - \tilde{\boldsymbol{x}})$$



f and its second order approximation is shown by the dark and bright grid, respectively. (Source: `daniloroccatano.blog`)

$\Longrightarrow$ Hessians provide information about **local** geometry of a function.

https://www.geogebra.org/m/M2P4KsRe

See common_functions.ipynb

FIRST ORDER CONDITION

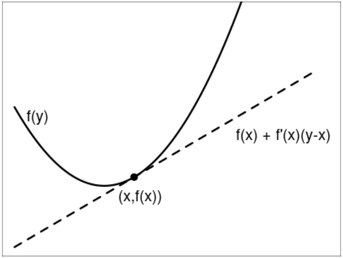
Prove convexity via **gradient**:

Let f be differentiable.

$$f \text{ (strictly) convex}$$

$$\Longleftrightarrow$$

$$f(\mathbf{y}) \overset{(>)}{\geq} f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \text{ for all } \mathbf{x}, \mathbf{y} \in \mathcal{S} \text{ (s.t. } \mathbf{x} \neq \mathbf{y})$$



SECOND ORDER CONDITION

Matrix A is **positive (semi)definite** (p.(s.)d.) if $\boldsymbol{v}^T A \boldsymbol{v} \overset{(\geq)}{>} 0$ for all $\boldsymbol{v} \neq 0$.

Notation: $A \overset{(\succcurlyeq)}{\succ} 0$ for A p.(s.)d. and $B \overset{(\succcurlyeq)}{\succ} A$ if $B - A \overset{(\succcurlyeq)}{\succ} 0$

Prove convexity via **Hessian**:

Let $f \in \mathcal{C}^2$ and $H(\mathbf{x})$ be its Hessian.

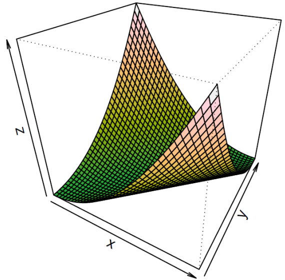
$$f \text{ (strictly) convex} \iff H(\mathbf{x}) \overset{(\succ)}{\succcurlyeq} 0 \text{ for all } \mathbf{x} \in \mathcal{S}$$

Alternatively: Since $H(\mathbf{x})$ symmetric for $f \in \mathcal{C}^2$:

$$H(\mathbf{x}) \succcurlyeq 0 \Leftrightarrow \text{all eigenvalues of } H(\mathbf{x}) \geq 0$$

SECOND ORDER CONDITION

Example: $f(\mathbf{x}) = x_1^2 + x_2^2 - 2x_1x_2$, $\nabla f(\mathbf{x}) = \begin{pmatrix} 2x_1 - 2x_2 \\ 2x_2 - 2x_1 \end{pmatrix}$, $H(\mathbf{x}) = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$.



f is convex since $H(\mathbf{x})$ is p.s.d. for all $\mathbf{x} \in \mathcal{S}$:

$$\mathbf{v}^T \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \mathbf{v} = \mathbf{v}^T \begin{pmatrix} 2v_1 - 2v_2 \\ -2v_1 + 2v_2 \end{pmatrix} = 2v_1^2 - 2v_1v_2 - 2v_1v_2 + 2v_2^2$$

$$= 2v_1^2 - 4v_1v_2 + 2v_2^2 = 2(v_1 - v_2)^2 \geq 0.$$

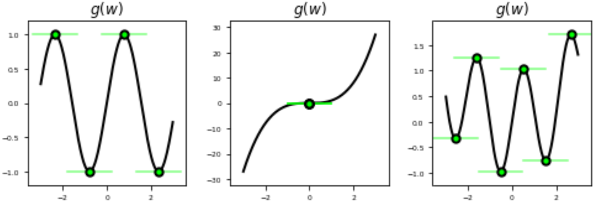
FIRST ORDER CONDITION FOR OPTIMALITY

First order condition: Gradient of f at local optimum $\mathbf{x}^* \in \mathcal{S}$ is zero:

$$\nabla f(\mathbf{x}^*) = (0, \ldots, 0)^T$$

Points with zero first order derivative are called **stationary**.

Condition is **not sufficient**: Not all stationary points are local optima.



Left: Four points fulfill the necessary condition and are indeed optima.
Middle: One point fulfills the necessary condition but is not a local optimum.
Right: Multiple local minima and maxima.
(Source: Watt, 2020, Machine Learning Refined)

SECOND ORDER CONDITION FOR OPTIMALITY

Second order condition: Hessian of $f \in \mathcal{C}^2$ at stationary point $\mathbf{x}^* \in \mathcal{S}$ is positive or negative definite:

$$H(\mathbf{x}^*) \succ 0 \text{ or } H(\mathbf{x}^*) \prec 0$$

Interpretation: Curvature of f at local optimum is either positive in all directions or negative in all directions.

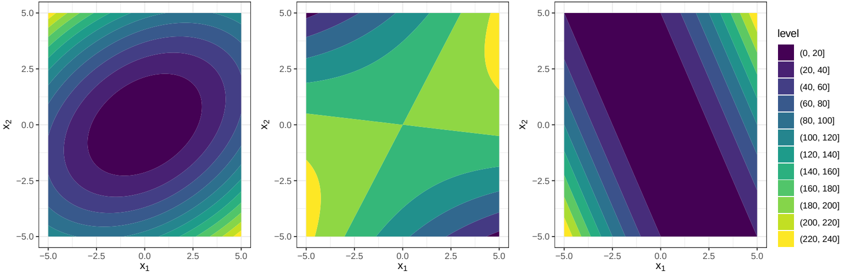
The second order condition is **sufficient** for a stationary point.

Proof: Later.

CONDITIONS FOR OPTIMALITY AND CONVEXITY

Let $f : \mathcal{S} \to \mathbb{R}$ be **convex**. Then:

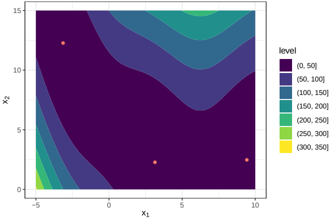
- Any local minimum is **also global** minimum
- If f **strictly convex**, f has **at most one** local minimum which would also be unique global minimum on $\mathcal{S}$

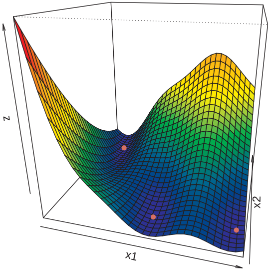


Three quadratic forms. **Left:** $H(\mathbf{x}^*)$ has two positive eigenvalues. **Middle:** $H(\mathbf{x}^*)$ has positive and negative eigenvalue. **Right:** $H(\mathbf{x}^*)$ has positive and a zero eigenvalue.

CONDITIONS FOR OPTIMALITY AND CONVEXITY

Example: Branin function





Spectra of Hessians (numerically computed):

	λ_1	λ_2
Left	22.29	0.96
Middle	11.07	1.73
Right	11.33	1.69

CONDITIONS FOR OPTIMALITY AND CONVEXITY

Definition: **Saddle point** at $\mathbf{x}$

- $\mathbf{x}$ stationary (necessary)
- $H(\mathbf{x})$ indefinite, i.e., positive and negative eigenvalues (sufficient)