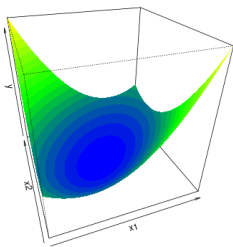


Optimization in Machine Learning

Mathematical Concepts: Quadratic forms I



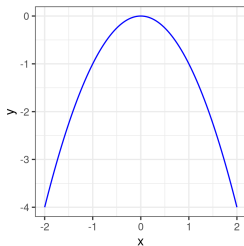
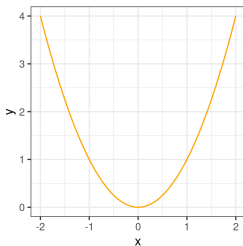
Learning goals

- Definition of quadratic forms
- Gradient, Hessian
- Optima

UNIVARIATE QUADRATIC FUNCTIONS

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

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

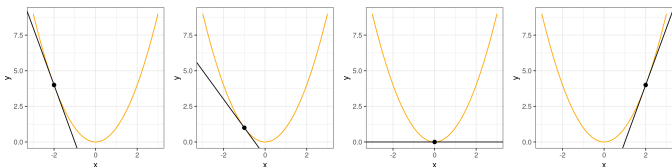


A quadratic function $q_1(x) = x^2$ (**left**) and $q_2(x) = -x^2$ (**right**).

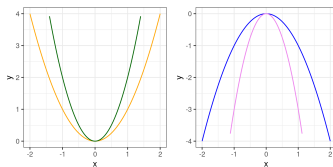
UNIVARIATE QUADRATIC FUNCTIONS

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)

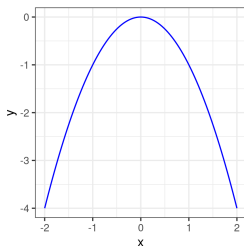
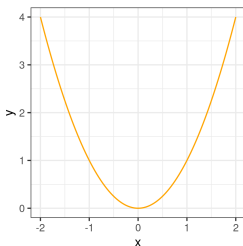
UNIVARIATE QUADRATIC FUNCTIONS

- **Convexity/Concavity:**

- $a > 0$: q convex, bounded from below, unique global **minimum**
- $a < 0$: q concave, bounded from above, unique global **maximum**

- **Optimum x^* :**

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



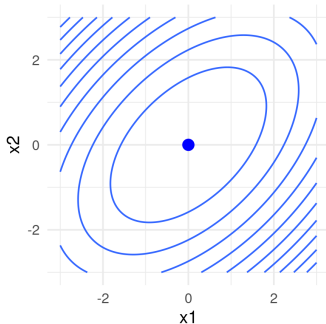
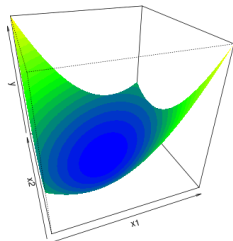
Left: $q_1(x) = x^2$ (convex). **Right:** $q_2(x) = -x^2$ (concave).

MULTIVARIATE QUADRATIC FUNCTIONS

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

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

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



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:

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

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

$$\nabla q(\mathbf{x}) = (\mathbf{A}^T + \mathbf{A}) \mathbf{x} + \mathbf{b} = 2\mathbf{A}\mathbf{x} + \mathbf{b} \in \mathbb{R}^d$$

Derivative in direction $\mathbf{v} \in \mathbb{R}^d$ is (by chain rule)

$$\left. \frac{dq(\mathbf{x} + h \cdot \mathbf{v})}{dh} \right|_{h=0} = \nabla q(\mathbf{x} + h\mathbf{v})^T \mathbf{v} \Big|_{h=0} = \nabla q(\mathbf{x})^T \mathbf{v}.$$

- The **Hessian** of q is

$$\nabla^2 q(\mathbf{x}) = (\mathbf{A}^T + \mathbf{A}) = 2\mathbf{A} =: \mathbf{H} \in \mathbb{R}^{d \times d}$$

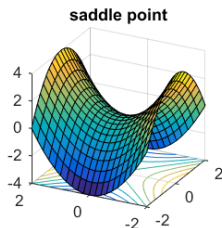
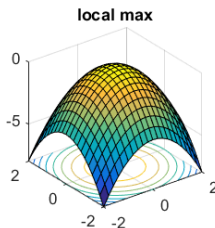
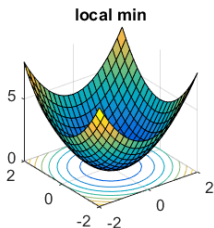
Curvature in direction of $\mathbf{v} \in \mathbb{R}^d$ is (by chain rule)

$$\left. \frac{d^2 q(\mathbf{x} + h \cdot \mathbf{v})}{dh^2} \right|_{h=0} = \mathbf{v}^T \nabla^2 q(\mathbf{x} + h\mathbf{v}) \mathbf{v} \Big|_{h=0} = \mathbf{v}^T \mathbf{H} \mathbf{v}.$$

OPTIMUM

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

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



Left: \mathbf{A} positive definite. **Middle:** \mathbf{A} negative definite. **Right:** \mathbf{A} indefinite.

OPTIMA: RANK-DEFICIENT CASE

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

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

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

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

