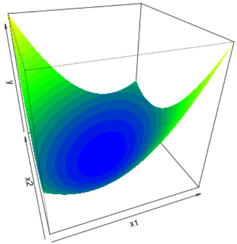
Optimization in Machine Learning

Mathematical Concepts
Quadratic forms I



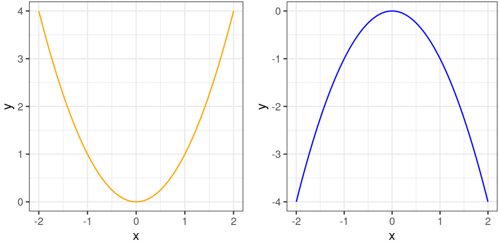
Learning goals

- Definition of quadratic forms
- Gradient, Hessian
- Optima

UNIVARIATE QUADRATIC FUNCTIONS

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

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

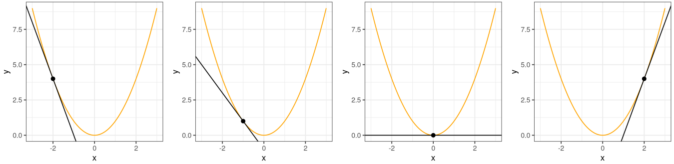


A quadratic function $q_1(x) = x^2$ (**left**) and $q_2(x) = -x^2$ (**right**).

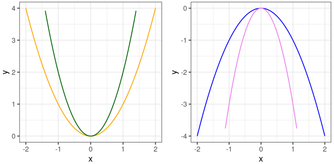
UNIVARIATE QUADRATIC FUNCTIONS / 2

Basic properties:

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

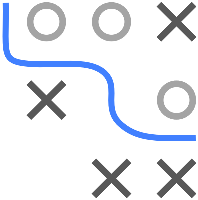


- **Curvature** of q is given by $q''(x) = 2 \cdot a$.



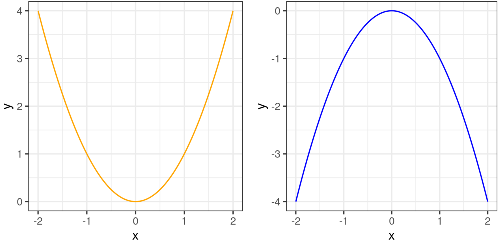
$q_1 = x^2$ (orange), $q_2 = 2x^2$ (green), $q_3(x) = -x^2$ (blue), $q_4 = -3x^2$ (magenta)

UNIVARIATE QUADRATIC FUNCTIONS / 3



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

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



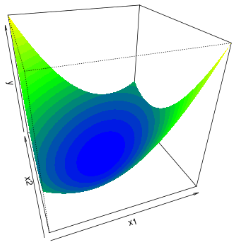
Left: $q_1(x) = x^2$ (convex). **Right:** $q_2(x) = -x^2$ (concave).

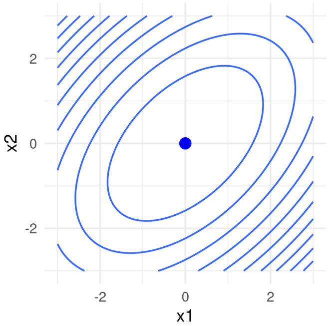
MULTIVARIATE QUADRATIC FUNCTIONS

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 / 2

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

If $\mathbf{A}$ 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

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

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)

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

- The **Hessian** of q is

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

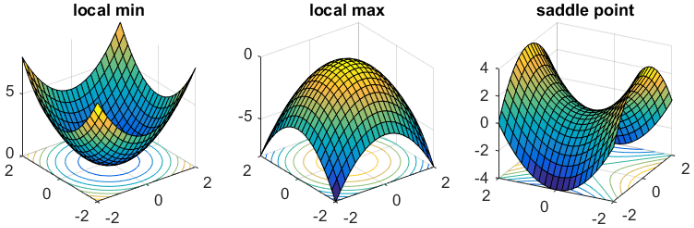
  Curvature in direction of $\boldsymbol{v} \in \mathbb{R}^d$ is (by chain rule)

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

OPTIMUM

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

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



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 $\boldsymbol{v}_0$