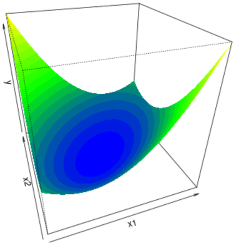
Optimization in Machine Learning

Mathematical Concepts
Quadratic forms I



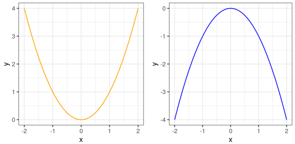
Learning goals

- Definition of quadratic functions
- Gradient, Hessian
- Optima

UNIVARIATE QUADRATIC

- Quadratic function $q : \mathbb{R} \to \mathbb{R}$

$$q(x) = ax^2 + bx + c, \quad a \neq 0$$

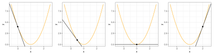


- Left: $q_1(x) = x^2$. Right: $q_2(x) = -x^2$

UNIVARIATE: BASIC PROPERTIES

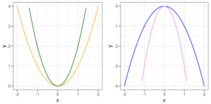
- Slope at $(x, q(x))$:

$$q'(x) = 2ax + b$$



- Curvature:

$$q''(x) = 2a$$



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

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

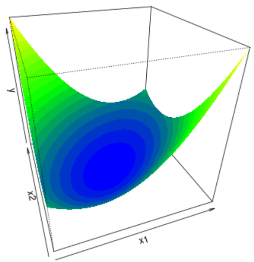
as 2nd derivative: $q''(x^*) = 2a \neq 0$

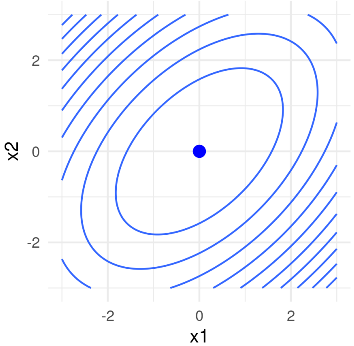
MULTIVARIATE QUADRATIC

- $q : \mathbb{R}^d \to \mathbb{R}$

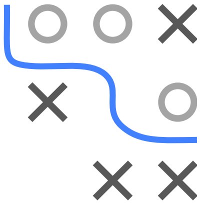
$$q(\boldsymbol{x}) = \boldsymbol{x}^T \mathbf{A} \boldsymbol{x} + \mathbf{b}^T \boldsymbol{x} + c$$

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





SYMMETRIZATION



- W.l.o.g. assume $\mathbf{A}$ symmetric, i.e., $\mathbf{A}^T = \mathbf{A}$
- If $\mathbf{A}$ not symmetric, there exists symmetric $\tilde{\mathbf{A}}$ with

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

- Justification

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

- $\tilde{\mathbf{A}}_1$ symmetric, $\tilde{\mathbf{A}}_2$ anti-symmetric ($\tilde{\mathbf{A}}_2^T = -\tilde{\mathbf{A}}_2$)
- Since $\boldsymbol{x}^T\mathbf{A}^T\boldsymbol{x}$ is a scalar, equal to its transpose

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

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

GRADIENT AND HESSIAN

- $q : \mathbb{R}^d \to \mathbb{R}$

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

- Gradient

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

- Under assumed symmetry: $\nabla q(\boldsymbol{x}) = (2\mathbf{A}\boldsymbol{x} + \mathbf{b})^T$
- Directional derivative: $\nabla q(\boldsymbol{x})\ \boldsymbol{v}$
- Hessian

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

- Under assumed symmetry: $\mathbf{H} = 2\mathbf{A}$
- Directional curvature: $\boldsymbol{v}^T \mathbf{H} \boldsymbol{v}$

OPTIMUM

- $q : \mathbb{R}^d \to \mathbb{R}$

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

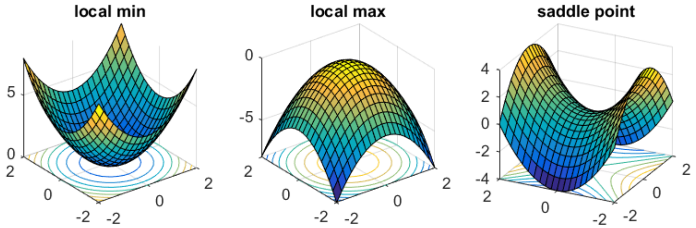
- Since **A** full rank, unique stationary point $\boldsymbol{x}^*$ (min, max, or saddle)

$$\nabla q(\boldsymbol{x}^*) = \mathbf{0}^T$$

$$(2\mathbf{A}\boldsymbol{x}^* + \mathbf{b})^T = \mathbf{0}^T$$

$$\boldsymbol{x}^* = -\tfrac{1}{2}\mathbf{A}^{-1}\mathbf{b}$$

- $q(\boldsymbol{x}^*) = c - \frac{1}{2}\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}$



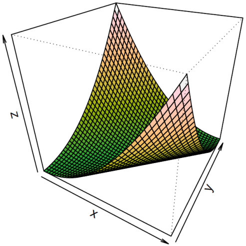
- Left: **A** pos. def. Middle: **A** neg. def. Right: **A** indefinite

OPTIMA: RANK-DEFICIENT CASE

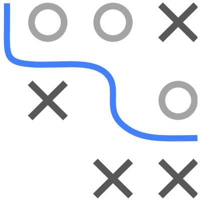
- $q : \mathbb{R}^d \to \mathbb{R}$

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

- Assume **A** symmetric now
- For stationary points to exist, we need : $\nabla q(\boldsymbol{x}) = 2\mathbf{A}\boldsymbol{x} + \mathbf{b} = 0$
- This implies we need $\mathbf{b} \in \mathit{range}(\mathbf{A})$, let's assume this is the case
- Let $\boldsymbol{x}_p$ be stationary, so $2\mathbf{A}\boldsymbol{x}_p = -\mathbf{b}$
- Then any point in affine space $\boldsymbol{x}_p + \mathit{ker}(\mathbf{A})$ is also stationary, with same function value and same Hessian (as it is constant)



0



OPTIMA: RANK-DEFICIENT CASE

- All affine spaces of form $\boldsymbol{x}_p + ker(\mathbf{A})$ for diff. valid $\boldsymbol{x}_p$ are the same
- Any stationary point must be in $\boldsymbol{x}_p + ker(\mathbf{A})$
- So $\boldsymbol{x}_p + ker(\mathbf{A})$ are all the stationary points, with same curvature
- If $\mathbf{A} \succeq 0$, they are all minima
- If $\mathbf{A} \preceq 0$, they are all maxima
- If $\mathbf{A}$ is indefinite, they are all saddle points