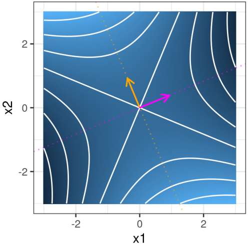
Optimization in Machine Learning

Mathematical Concepts: Quadratic forms II



Learning goals

- Geometry of quadratic forms
- Eigenspectrum

PROPERTIES OF QUADRATIC FUNCTIONS

Univariate function

2nd derivative is a $q''(x) = 2 \cdot a$. Basic properties of q can be read-off:

- $q''(x) > 0$: q convex; $q''(x) < 0$: q concave
- High (lows) absolute values of $q''(x)$: high (low) curvature

Multivariate function

2nd derivative is a symmetric matrix of values **H** (called Hessian).

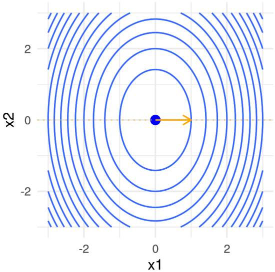
Now: See how Eigenspectrum of **H** encodes the basic properties of q.

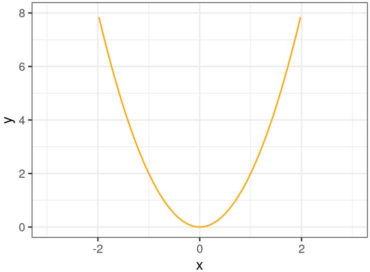
PROPERTIES OF QUADRATIC FUNCTIONS (DIAG)

Example 1: Function composed of two univariate quadratic terms

$$q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{x} = 2 \cdot x_1^2 + x_2^2$$

$$\text{with } \nabla q(\mathbf{x}) = 2 \cdot \mathbf{A} \cdot \mathbf{x} = 4 \cdot x_1 + 2 \cdot x_2, \quad \boldsymbol{H} = 2 \cdot \mathbf{A} = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$$





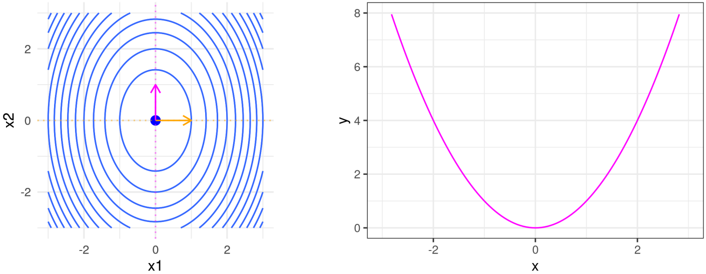
q has a high positive curvature of 4 in the direction of $v = (1, 0)^\top$,

PROPERTIES OF QUADRATIC FUNCTIONS (DIAG)

Example 1: Function composed of two univariate quadratic terms

$$q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{x} = 2 \cdot x_1^2 + x_2^2$$

$$\text{with } \nabla q(\mathbf{x}) = 2 \cdot \mathbf{A} \cdot \mathbf{x} = 4 \cdot x_1 + 2 \cdot x_2, \quad \boldsymbol{H} = 2 \cdot \mathbf{A} = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$$



q has a high positive curvature of 4 in the direction of $v = (1, 0)^\top$, and a lower (positive) curvature of 2 in direction of $v = (0, 1)^\top$.

PROPERTIES OF QUADRATIC FUNCTIONS (DIAG)

Takeaway I:

- Hessian encodes curvature
- If the Hessian **H** is diagonal, the diagonal elements encode the curvature of the function:
  - *i*-th diagonal element gives us the curvature in the direction of $\boldsymbol{v} = \boldsymbol{e}_i$ because
    $$\boldsymbol{v}^\top \boldsymbol{H}\boldsymbol{v} = \boldsymbol{e}_i^\top \boldsymbol{H}\boldsymbol{e}_i = h_{ii}.$$
  - The curvature in an arbitrary direction $\boldsymbol{v} \in \mathbb{R}^d$, $\|\boldsymbol{v}\| = 1$, is
    $$\boldsymbol{v}^\top \mathbf{H}\boldsymbol{v} = h_{11}v_1^2 + h_{22}v_2^2 + ... + h_{dd}v_d^2.$$

PROPERTIES OF QUADRATIC FUNCTIONS (DIAG)

Takeaway I:

- Hessian encodes curvature
- If the Hessian **H** is diagonal, the diagonal elements encode the curvature of the function:
  - i-th diagonal element gives us the curvature in the direction of $\boldsymbol{v} = \boldsymbol{e}_i$ because
    $$\boldsymbol{v}^\top \boldsymbol{H}\boldsymbol{v} = \boldsymbol{e}_i^\top \boldsymbol{H}\boldsymbol{e}_i = h_{ii}.$$
  - The curvature in an arbitrary direction $\boldsymbol{v} \in \mathbb{R}^d$, $\|\boldsymbol{v}\| = 1$, is
    $$\boldsymbol{v}^\top \mathbf{H}\boldsymbol{v} = h_{11}v_1^2 + h_{22}v_2^2 + ... + h_{dd}v_d^2.$$
- For general (non-diagonal) matrices we analyze the **eigenspectrum** of **H**

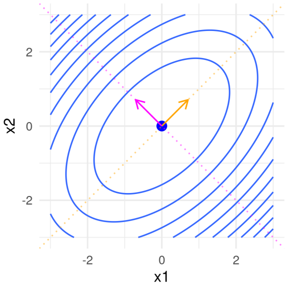
Note: For diagonal matrices the eigenspectrum is is to read-off: Diagonal elements of **H eigenvalues**, unit vectors **eigenvectors**

$$\mathbf{H}\boldsymbol{e}_1 = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} = 4 \cdot \boldsymbol{e}_1; \qquad \mathbf{H}\boldsymbol{e}_2 = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} = 2 \cdot \boldsymbol{e}_2$$

PROPERTIES OF QUADRATIC FUNCTIONS

Example 2:

$$
\begin{aligned}
q(\mathbf{x}) &= \mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \mathbf{x}, \\
\text{with } \nabla q(\mathbf{x}) &= 2 \cdot \mathbf{A} \cdot \mathbf{x}, \qquad \nabla^2 q(\mathbf{x}) = \boldsymbol{H} = 2 \cdot \mathbf{A} = \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix}
\end{aligned}
$$

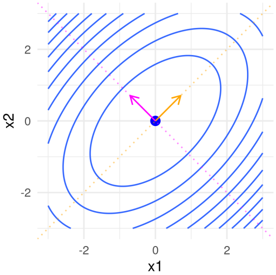


In the general case, the curvature is determined by the Eigenspectrum of $\boldsymbol{H}$.

PROPERTIES OF QUADRATIC FUNCTIONS

Takeaway II:

- Geometrically, directions of highest / lowest curvature along main axes of ellipses representing the contour lines of q.
- Mathematically, the direction with the highest (lowest) curvature is the direction of the eigenvector $\boldsymbol{v}_{\text{max}}$ ($\boldsymbol{v}_{\text{min}}$) belonging to largest (smallest) eigenvalue λ_{max} (λ_{min}) of $\boldsymbol{H}$.



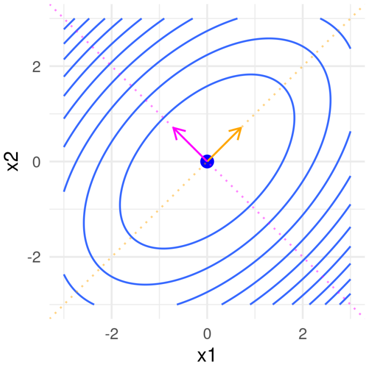
The eigenvectors and eigenvalues of $\boldsymbol{H} = \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix}$ are:

$$\boldsymbol{v}_{\text{min}} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda_{\text{min}} = 2$$

$$\boldsymbol{v}_{\text{max}} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \lambda_{\text{max}} = 3$$

PROPERTIES OF QUADRATIC FUNCTIONS

Direction $\boldsymbol{v}_{\max}$ is also direction in which the function increases fastest.

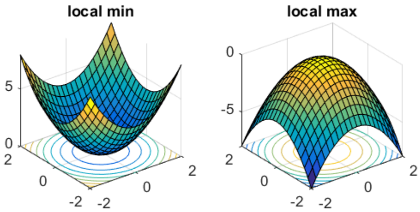


"Walking" the same distance along $\boldsymbol{v}_{\max}$ (magenta) makes us pass more level curves than walking along any other direction.

PROPERTIES OF QUADRATIC FUNCTIONS

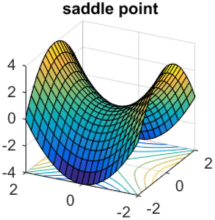
If eigenspectrum of **A** is known, i.e. the set of its eigenvalues $\{\lambda_1, \lambda_2, ..., \lambda_d\}$, also eigenspectrum of $\boldsymbol{H} = 2 \cdot \mathbf{A}$ is known and we can read off:

- If **all** eigenvalues of the $\boldsymbol{H}$ are > 0 (we call $\boldsymbol{H}$ positive definite):
  - the function q is convex,
  - there is a unique global minimum.
- If **all** eigenvalues of the $\boldsymbol{H}$ are < 0 (we call $\boldsymbol{H}$ negative definite):
  - the function q is concave,
  - there is a unique global maximum.



PROPERTIES OF QUADRATIC FUNCTIONS

- If there are both positive and negative eigenvalues (we call $\boldsymbol{H}$ indefinite):
  - the function q is neither concave nor convex,
  - there is a saddle point.



PROPERTIES OF QUADRATIC FUNCTIONS

Example: Sketch the following function

$$q(\mathbf{x}) \quad = \quad \mathbf{x}^\top \begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix} \mathbf{x}$$

Step 1: Compute the Hessian

$$\boldsymbol{H} \quad = \quad 2 \cdot \mathbf{A} = \begin{pmatrix} -2 & -2 \\ -2 & 2 \end{pmatrix}$$

PROPERTIES OF QUADRATIC FUNCTIONS

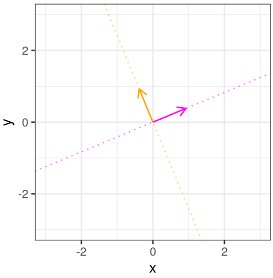
Example: Sketch the following function

$$q(\mathbf{x}) = \mathbf{x}^\top \begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix} \mathbf{x}$$

Step 2: Compute eigenvectors / -values:

$$\boldsymbol{v}_1 = \begin{pmatrix} 1-\sqrt{2} \\ 1 \end{pmatrix}, \qquad \lambda_1 = 2\sqrt{2}$$

$$\boldsymbol{v}_2 = \begin{pmatrix} 1+\sqrt{2} \\ 1 \end{pmatrix}, \qquad \lambda_2 = -2\sqrt{2}.$$



PROPERTIES OF QUADRATIC FUNCTIONS

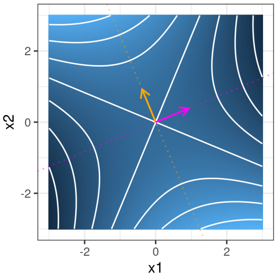
Example: Sketch the following function

$$q(\mathbf{x}) = \mathbf{x}^\top \begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix} \mathbf{x}$$

Step 2: Compute eigenvectors / -values:

$$\boldsymbol{v}_1 = \begin{pmatrix} 1-\sqrt{2} \\ 1 \end{pmatrix}, \qquad \lambda_1 = 2\sqrt{2}$$

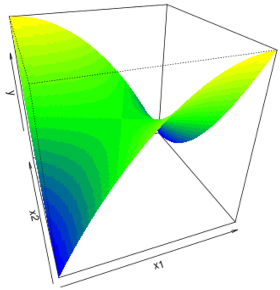
$$\boldsymbol{v}_2 = \begin{pmatrix} 1+\sqrt{2} \\ 1 \end{pmatrix}, \qquad \lambda_2 = -2\sqrt{2}.$$

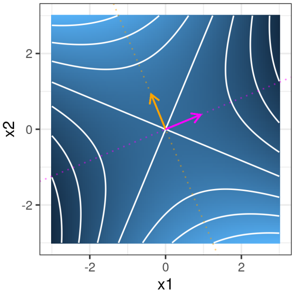


PROPERTIES OF QUADRATIC FUNCTIONS

Example: Sketch the following function

$$q(\mathbf{x}) = \mathbf{x}^\top \begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix} \mathbf{x}$$

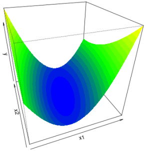


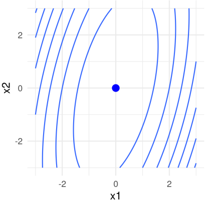


EIGENSPECTRUM AND CONDITION

Also the condition can be read off from Eigenspectrum: $\kappa(\mathbf{A}) = \frac{|\lambda_{\text{max}}|}{|\lambda_{\text{min}}|}$.
A high condition means:

- The absolute value of the biggest eigenvalue λ_{max} is much larger than the absolute value of the lowest eigenvalue λ_{min}.
- The curvature in the direction of minimum curvature ($\mathbf{v}_{\text{max}}$) is much lower than the one in the direction of maximum curvature ($\mathbf{v}_{\text{min}}$).
- We will see later: optimization algorithms like gradient descent will have difficulties optimizing such functions.

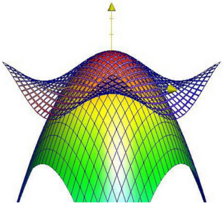




INTERPRETATION OF GENERAL FUNCTIONS

Every function can be locally approximated by a quadratic function via 2nd order Taylor approximation:

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



f is shown as the hollow grid and its second-order approximation at $(0, 0)$ as a continuous surface. Source: `daniloroccatano.blog`.

By analyzing $\nabla^2 f(\tilde{\boldsymbol{x}})$ we can gain a local understanding of a function's geometry.