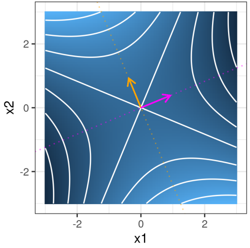
Optimization in Machine Learning

Mathematical Concepts
Quadratic functions II



Learning goals

- Geometry of quadratic functions
- Spectrum of Hessian

PROPERTIES OF QUADRATIC FUNCTIONS

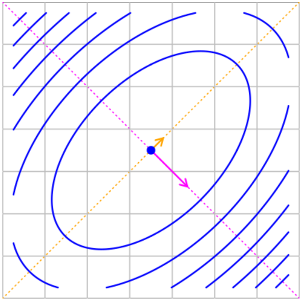
- $q(\boldsymbol{x}) = \boldsymbol{x}^T \mathbf{A}\boldsymbol{x} + \mathbf{b}^T \boldsymbol{x} + c$
- Under symmetry: $\mathbf{H} = 2\mathbf{A}$
- Convexity/concavity of q depend on eigenvalues of $\mathbf{H}$

GEOMETRY OF QUADRATIC FUNCTIONS

- Example: $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \Rightarrow \mathbf{H} = 2\mathbf{A} = \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix}$
- Since $\mathbf{H}$ symmetric: eigendecomposition $\mathbf{H} = \mathbf{V}\Lambda\mathbf{V}^T$

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

$$\Lambda = \begin{pmatrix} {\color{magenta}\lambda_{\text{max}}} & 0 \\ 0 & {\color{orange}\lambda_{\text{min}}} \end{pmatrix} = \begin{pmatrix} 6 & 0 \\ 0 & 2 \end{pmatrix}$$



SPECTRUM AND CURVATURE

- $\boldsymbol{v}_{\text{max}}$ direction of highest curvature, with curvature value λ_{max}

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

- Since $||\boldsymbol{v}|| = ||\boldsymbol{x}||$ ($\mathbf{V}$ orthogonal): $\max_{||\boldsymbol{v}||=1} \boldsymbol{v}^T \mathbf{H} \boldsymbol{v} \leq \lambda_{\text{max}}$
- For $\boldsymbol{v}_{\text{max}}$ we obtain this upper bound: $\boldsymbol{v}_{\text{max}}^T \mathbf{H} \boldsymbol{v}_{\text{max}} = \mathbf{e}_1^T \Lambda \mathbf{e}_1 = \lambda_{\text{max}}$
- Analogously, $\boldsymbol{v}_{\text{min}}$ direction of lowest curvature, with curvature value λ_{min}
- Contour lines of any quadratic function are ellipses

SECOND ORDER CONDITION

- Recall: Second order condition for optimality is sufficient
- If $H(\boldsymbol{x}^*) \succ 0$ at stationary point $\boldsymbol{x}^*$, then $\boldsymbol{x}^*$ local minimum ($\prec$ for maximum)

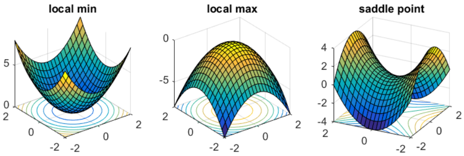
$$f(\boldsymbol{x}) = f(\boldsymbol{x}^*) + \underbrace{\nabla f(\boldsymbol{x}^*)}_{=0}(\boldsymbol{x} - \boldsymbol{x}^*) + \frac{1}{2}\underbrace{(\boldsymbol{x} - \boldsymbol{x}^*)^T H(\boldsymbol{x}^*)(\boldsymbol{x} - \boldsymbol{x}^*)}_{\geq \lambda_{\min} \|\boldsymbol{x} - \boldsymbol{x}^*\|^2}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

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

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) + \underbrace{\frac{1}{2}\lambda_{\min}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 + R_2(\boldsymbol{x}, \boldsymbol{x}^*)}_{>0} > f(\boldsymbol{x}^*)$$

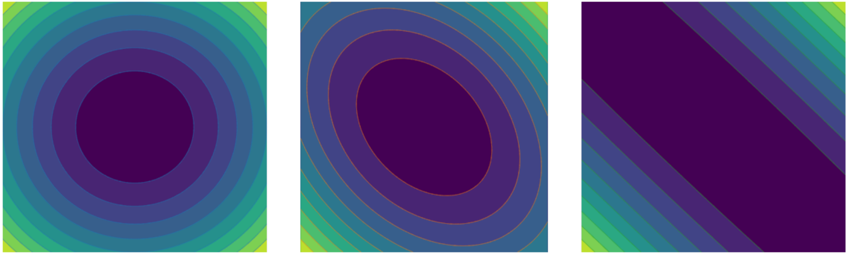
EIGENVALUES AND SHAPE

- 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
  - (Unique) global minimum
- If all eigenvalues of $\mathbf{H} \overset{(<)}{\leq} 0$ ($\Leftrightarrow \mathbf{H} \overset{(\prec)}{\preccurlyeq} 0$):
  - q (strictly) concave
  - (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

- $\kappa(\mathbf{H}) = \kappa(\mathbf{A}) = |\lambda_{\text{max}}| / |\lambda_{\text{min}}|$
- High condition means
  - $|\lambda_{\text{max}}| \gg |\lambda_{\text{min}}|$
  - Curvature along $\boldsymbol{v}_{\text{max}} \gg$ along $\boldsymbol{v}_{\text{min}}$
  - Problem for algorithms like gradient descent

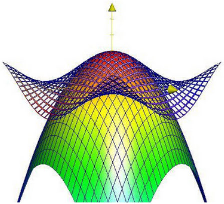


Left: Excellent condition. Middle: Good condition. Right: Bad condition.

APPROXIMATION OF SMOOTH FUNCTIONS

- Any $f \in \mathcal{C}^2$ can be locally approximated by quadratic function (second order Taylor)

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



f and second order approximation: dark vs bright grid. (Source: `daniloroccatano.blog`)

- $\implies$ Hessians provide information about **local** geometry of a function