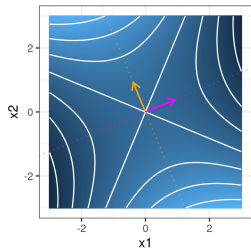


Optimization in Machine Learning

Mathematical Concepts

Quadratic forms II



Learning goals

- Geometry of quadratic forms
- Spectrum of Hessian

PROPERTIES OF QUADRATIC FUNCTIONS

Recall: Quadratic form q

- Univariate: $q(x) = ax^2 + bx + c$
- Multivariate: $q(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c$

General observation: If $q \geq 0$ ($q \leq 0$), q is convex (concave)

Univariate function: Second derivative is $q''(x) = 2a$

- $q''(x) \stackrel{(>)}{\geq} 0$: q (strictly) convex. $q''(x) \stackrel{(<)}{\leq} 0$: q (strictly) concave.
- High (low) absolute values of $q''(x)$: high (low) curvature

Multivariate function: Second derivative is $\mathbf{H} = 2\mathbf{A}$

- Convexity/concavity of q depend on eigenvalues of \mathbf{H}
- Let us look at an example of the form $q(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$



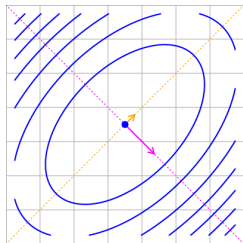
GEOMETRY OF QUADRATIC FUNCTIONS

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 $\mathbf{H} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ with

$$\mathbf{V} = \begin{pmatrix} | & | \\ \textcolor{violet}{v}_{\max} & \textcolor{brown}{v}_{\min} \\ | & | \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \text{ orthogonal}$$

$$\text{and } \mathbf{\Lambda} = \begin{pmatrix} \textcolor{violet}{\lambda}_{\max} & 0 \\ 0 & \textcolor{brown}{\lambda}_{\min} \end{pmatrix} = \begin{pmatrix} 6 & 0 \\ 0 & 2 \end{pmatrix}.$$



GEOMETRY OF QUADRATIC FUNCTIONS / 2

- \mathbf{v}_{\max} (\mathbf{v}_{\min}) direction of highest (lowest) curvature

Proof: With $\mathbf{v} = \mathbf{V}^T \mathbf{x}$:

$$\mathbf{x}^T \mathbf{H} \mathbf{x} = \mathbf{x}^T \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \mathbf{x} = \mathbf{v}^T \mathbf{\Lambda} \mathbf{v} = \sum_{i=1}^d \lambda_i v_i^2 \leq \lambda_{\max} \sum_{i=1}^d v_i^2 = \lambda_{\max} \|\mathbf{v}\|^2$$

Since $\|\mathbf{v}\| = \|\mathbf{x}\|$ (\mathbf{V} orthogonal): $\max_{\|\mathbf{x}\|=1} \mathbf{x}^T \mathbf{H} \mathbf{x} \leq \lambda_{\max}$

Additional: $\mathbf{v}_{\max}^T \mathbf{H} \mathbf{v}_{\max} = \mathbf{e}_1^T \mathbf{\Lambda} \mathbf{e}_1 = \lambda_{\max}$

Analogous: $\min_{\|\mathbf{x}\|=1} \mathbf{x}^T \mathbf{H} \mathbf{x} \geq \lambda_{\min}$ and $\mathbf{v}_{\min}^T \mathbf{H} \mathbf{v}_{\min} = \lambda_{\min}$

- Contour lines of any quadratic form are ellipses
(with eigenvectors of \mathbf{A} as principal axes, principal axis theorem)

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

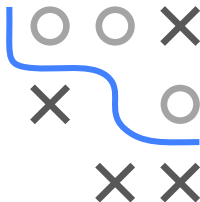
Now use $\mathbf{y} = \mathbf{x} - \mathbf{w} = \mathbf{x} + \frac{1}{2} \mathbf{A}^{-1} \mathbf{b}$

This already gives us the general form of an ellipse:

$$\mathbf{y}^T \mathbf{A} \mathbf{y} = (\mathbf{x} - \mathbf{w})^T \mathbf{A} (\mathbf{x} - \mathbf{w}) = q(\mathbf{x}) + \text{const}$$

If we use $\mathbf{z} = \mathbf{V}^T \mathbf{y}$ we obtain it in standard form

$$\sum_{i=1}^n \lambda_i z_i^2 = \mathbf{z}^T \mathbf{\Lambda} \mathbf{z} = \mathbf{y}^T \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \mathbf{y} = \mathbf{y}^T \mathbf{A} \mathbf{y} = q(\mathbf{x}) + \text{const}$$



GEOMETRY OF QUADRATIC FUNCTIONS / 3

Recall: **Second order condition for optimality** is **sufficient**.

We skipped the **proof** at first, but can now catch up on it.

If $H(\mathbf{x}^*) \succ 0$ at stationary point \mathbf{x}^* , then \mathbf{x}^* is local minimum (\prec for maximum).

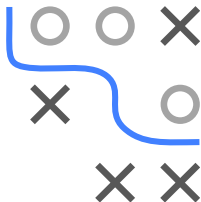
Proof: Let $\lambda_{\min} > 0$ denote the smallest eigenvalue of $H(\mathbf{x}^*)$. Then:

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

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

Then:

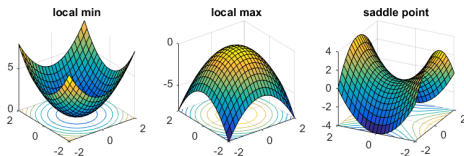
$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \underbrace{\frac{1}{2} \lambda_{\min} \|\mathbf{x} - \mathbf{x}^*\|^2 + R_2(\mathbf{x}, \mathbf{x}^*)}_{>0} > f(\mathbf{x}^*) \quad \text{for each } \mathbf{x} \neq \mathbf{x}^* \text{ with } \|\mathbf{x} - \mathbf{x}^*\| < \epsilon.$$



GEOMETRY OF QUADRATIC FUNCTIONS / 4

If spectrum of \mathbf{A} is known, also that of $\mathbf{H} = 2\mathbf{A}$ is known.

- If **all** eigenvalues of $\mathbf{H} \stackrel{(>)}{\geq} 0$ ($\Leftrightarrow \mathbf{H} \stackrel{(>)}{\succcurlyeq} 0$):
 - q (strictly) convex,
 - there is a (unique) global minimum.
- If **all** eigenvalues of $\mathbf{H} \stackrel{(<)}{\leq} 0$ ($\Leftrightarrow \mathbf{H} \stackrel{(<)}{\preccurlyeq} 0$):
 - q (strictly) concave,
 - there is a (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.

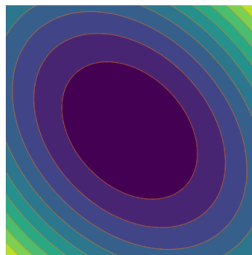
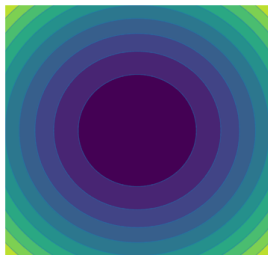
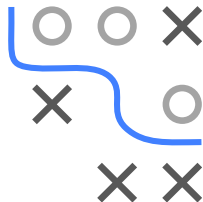


CONDITION AND CURVATURE

Condition of $\mathbf{H} = 2\mathbf{A}$ is given by $\kappa(\mathbf{H}) = \kappa(\mathbf{A}) = |\lambda_{\max}|/|\lambda_{\min}|$.

High condition means:

- $|\lambda_{\max}| \gg |\lambda_{\min}|$
- Curvature along $\mathbf{v}_{\max} \gg$ curvature along \mathbf{v}_{\min}
- **Problem** for optimization algorithms like **gradient descent** (later)

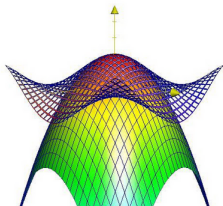


Left: Excellent condition. **Middle:** Good condition. **Right:** Bad condition.

APPROXIMATION OF SMOOTH FUNCTIONS

Any function $f \in \mathcal{C}^2$ can be locally approximated by a quadratic function via second order Taylor approximation:

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



f and its second order approximation is shown by the dark and bright grid, respectively.
(Source: [danioloroccatano.blog](#))

\Rightarrow Hessians provide information about **local** geometry of a function.

