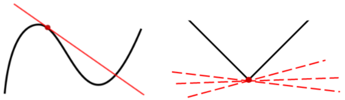
Optimization in Machine Learning

Mathematical Concepts: Differentiation and Derivatives



Learning goals

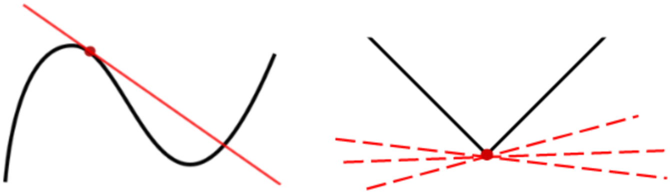
- Definition of smoothness
- Uni- & multivariate differentiation
- Gradient, partial derivatives
- Jacobian matrix
- Hessian matrix
- Lipschitz continuity

UNIVARIATE DIFFERENTIABILITY

Definition: A function $f : \mathcal{S} \subseteq \mathbb{R} \to \mathbb{R}$ is said to be **differentiable** for each inner point $x \in \mathcal{S}$ if the following limit exists:

$$f'(x) := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

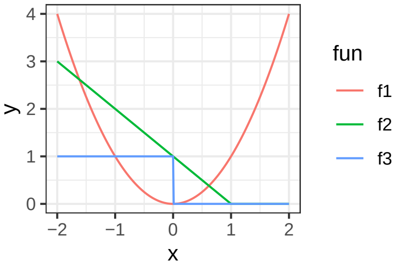
Intuitively: f can be approxed locally by a lin. fun. with slope $m = f'(x)$.



Left: Function is differentiable everywhere. **Right:** Not differentiable at the red point.

SMOOTH VS. NON-SMOOTH

- **Smoothness** of a function $f : \mathcal{S} \to \mathbb{R}$ is measured by the number of its continuous derivatives
- $\mathcal{C}^k$ is class of k-times continuously differentiable functions ($f \in \mathcal{C}^k$ means $f^{(k)}$ exists and is continuous)
- In this lecture, we call f "smooth", if at least $f \in \mathcal{C}^1$

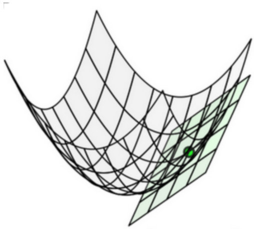


f_1 is smooth, f_2 is continuous but not differentiable, and f_3 is non-continuous.

MULTIVARIATE DIFFERENTIABILITY

Definition: $f : \mathcal{S} \subseteq \mathbb{R}^d \to \mathbb{R}$ is **differentiable** in $\mathbf{x} \in \mathcal{S}$ if there exists a (continuous) linear map $\nabla f(\mathbf{x}) : \mathcal{S} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ with

$$\lim_{\boldsymbol{h} \to 0} \frac{f(\mathbf{x} + \boldsymbol{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T \cdot \boldsymbol{h}}{||\boldsymbol{h}||} = 0$$



Geometrically: The function can be locally approximated by a tangent hyperplane.

Source: `https://github.com/jermwatt/machine_learning_refined`.

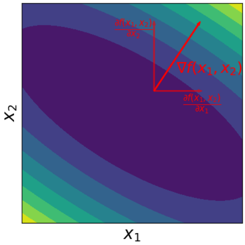
GRADIENT

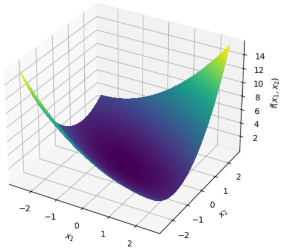
- Linear approximation is given by the **gradient**:

$$\nabla f = \frac{\partial f}{\partial x_1}\mathbf{e}_1 + \cdots + \frac{\partial f}{\partial x_d}\mathbf{e}_d = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d}\right)^T$$

- Elements of the gradient are called **partial derivatives**.
- To compute $\partial f/\partial x_j$, regard f as function of x_j only (others fixed)

Example: $f(\mathbf{x}) = x_1^2/2 + x_1 x_2 + x_2^2 \Rightarrow \nabla f(\mathbf{x}) = (x_1 + x_2, x_1 + 2x_2)^T$





DIRECTIONAL DERIVATIVE

The **directional derivative** tells how fast $f : \mathcal{S} \rightarrow \mathbb{R}$ is changing w.r.t. an arbitrary direction $\boldsymbol{v}$:

$$D_{\boldsymbol{v}} f(\mathbf{x}) := \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\boldsymbol{v}) - f(\mathbf{x})}{h} = \nabla f(\mathbf{x})^T \cdot \boldsymbol{v}.$$

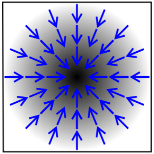
Example: The directional derivative for $\boldsymbol{v} = (1, 1)$ is:

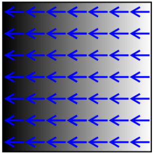
$$D_{\boldsymbol{v}} f(\mathbf{x}) = \nabla f(\mathbf{x})^T \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{\partial f}{\partial x_1} + \frac{\partial f}{\partial x_2}$$

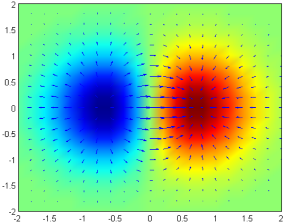
NB: Some people require that $||\boldsymbol{v}|| = 1$. Then, we can identify $D_{\boldsymbol{v}} f(\mathbf{x})$ with the instantaneous rate of change in direction $\boldsymbol{v}$ – and in our example we would have to divide by $\sqrt{2}$.

PROPERTIES OF THE GRADIENT

- **Orthogonal** to level curves/surfaces of a function
- Points in direction of **greatest increase** of f







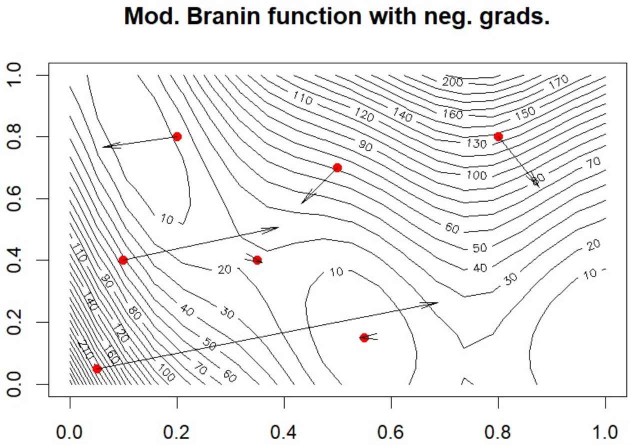
Proof: Let $\boldsymbol{v}$ be a vector with $\|\boldsymbol{v}\| = 1$ and θ the angle between $\boldsymbol{v}$ and $\nabla f(\mathbf{x})$.

$$D_{\boldsymbol{v}} f(\mathbf{x}) = \nabla f(\mathbf{x})^T \boldsymbol{v} = \|\nabla f(\mathbf{x})\| \, \|\boldsymbol{v}\| \cos(\theta) = \|\nabla f(\mathbf{x})\| \cos(\theta)$$

by the cosine formula for dot products and $\|\boldsymbol{v}\| = 1$. $\cos(\theta)$ is maximal if $\theta = 0$, hence if $\boldsymbol{v}$ and $\nabla f(\mathbf{x})$ point in the same direction.
(Alternative proof: Apply Cauchy-Schwarz to $\nabla f(\mathbf{x})^T \boldsymbol{v}$ and look for equality.)

Analogous: Negative gradient $-\nabla f(\mathbf{x})$ points in direction of greatest *de*crease

PROPERTIES OF THE GRADIENT



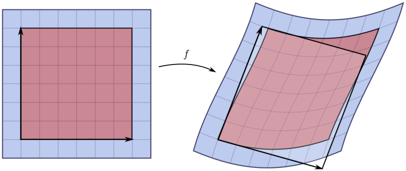
Length of arrows is norm of their gradient

JACOBIAN MATRIX

For vector-valued function $f = (f_1, \ldots, f_m)^T$, $f_j : \mathcal{S} \to \mathbb{R}$, the **Jacobian** matrix $J_f : \mathcal{S} \to \mathbb{R}^{m \times d}$ generalizes gradient by placing all ∇f_j in its rows:

$$J_f(\mathbf{x}) = \begin{pmatrix} \nabla f_1(\mathbf{x})^T \\ \vdots \\ \nabla f_m(\mathbf{x})^T \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_d} \end{pmatrix}$$

- Jacobian gives best linear approximation of distorted volumes



Source: Wikipedia

JACOBIAN DETERMINANT

Let $f \in \mathcal{C}^1$ and $\mathbf{x}_0 \in \mathcal{S}$.

Inverse function theorem: Let $\mathbf{y}_0 = f(\mathbf{x}_0)$. If $\det(J_f(\mathbf{x}_0)) \neq 0$, then

1. f is invertible in a neighborhood of $\mathbf{x}_0$,
2. $f^{-1} \in \mathcal{C}^1$ with $J_{f^{-1}}(\mathbf{y}_0) = J_f(\mathbf{x}_0)^{-1}$.

- $|\det(J_f(\mathbf{x}_0))|$: factor by which f expands/shrinks volumes near $\mathbf{x}_0$
- If $\det(J_f(\mathbf{x}_0)) > 0$, f preserves orientation near $\mathbf{x}_0$
- If $\det(J_f(\mathbf{x}_0)) < 0$, f reverses orientation near $\mathbf{x}_0$

HESSIAN MATRIX

For real-valued function $f : \mathcal{S} \to \mathbb{R}$, the **Hessian** matrix $H : \mathcal{S} \to \mathbb{R}^{d \times d}$ contains all their second derivatives (if they exist):

$$H(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \left(\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right)_{i,j=1,\ldots,d}$$

Note: Hessian of f is Jacobian of ∇f

Example: Let $f(\mathbf{x}) = \sin(x_1) \cdot \cos(2x_2)$. Then:

$$H(\mathbf{x}) = \begin{pmatrix} -\cos(2x_2) \cdot \sin(x_1) & -2\cos(x_1) \cdot \sin(2x_2) \\ -2\cos(x_1) \cdot \sin(2x_2) & -4\cos(2x_2) \cdot \sin(x_1) \end{pmatrix}$$

- If $f \in \mathcal{C}^2$, then H is symmetric
- Many local properties (geometry, convexity, critical points) are encoded by the Hessian and its spectrum ($\to$ later)

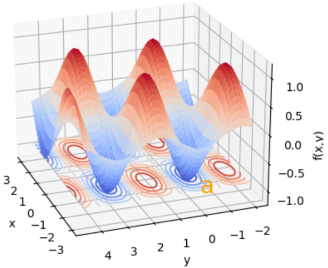
LOCAL CURVATURE BY HESSIAN

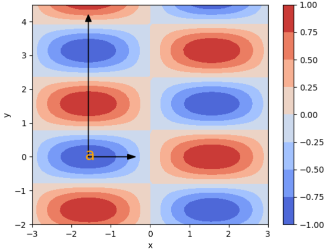
Eigenvector corresponding to largest (resp. smallest) **eigenvalue** of Hessian points in direction of largest (resp. smallest) **curvature**

Example (previous slide)**:** For $\boldsymbol{a} = (-\pi/2, 0)^T$, we have

$$H(\boldsymbol{a}) = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$$

and thus $\lambda_1 = 4, \lambda_2 = 1$, $\boldsymbol{v}_1 = (0, 1)^T$, and $\boldsymbol{v}_2 = (1, 0)^T$.





LIPSCHITZ CONTINUITY

Function $h : \mathcal{S} \to \mathbb{R}^m$ is **Lipschitz continuous** if slopes are bounded:

$$\|h(\mathbf{x}) - h(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \text{for each } \mathbf{x}, \mathbf{y} \in \mathcal{S} \text{ and some } L > 0$$

- **Examples** ($d = m = 1$)**:** $\sin(x)$, $|x|$
- **Not** examples: $1/x$ (but *locally* Lipschitz continuous), $\sqrt{x}$
- If $m = d$ and h **differentiable**:

$$h \text{ Lipschitz continuous with constant } L \iff J_h \preccurlyeq L \cdot \mathbf{I}_d$$

Note: $\mathbf{A} \preccurlyeq \mathbf{B} :\iff \mathbf{B} - \mathbf{A}$ is positive semidefinite, i.e., $\mathbf{v}^T(\mathbf{B} - \mathbf{A})\mathbf{v} \geq 0 \ \ \forall \mathbf{v} \neq 0$

Proof of "$\Rightarrow$" for $d = m = 1$**:**

$$h'(x) = \lim_{\epsilon \to 0} \frac{h(x+\epsilon) - h(x)}{\epsilon} \leq \lim_{\epsilon \to 0} \underbrace{\left| \frac{h(x+\epsilon) - h(x)}{\epsilon} \right|}_{\leq L} \leq \lim_{\epsilon \to 0} L = L$$

[**Proof** of "$\Leftarrow$" by mean value theorem: Show that $\lambda_{\max}(J_h) \leq L$.]

LIPSCHITZ GRADIENTS

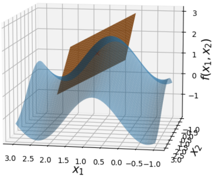
- Let $f \in \mathcal{C}^2$. Since $\nabla^2 f$ is Jacobian of $h = \nabla f$ $(m = d)$:

$$\nabla f \text{ Lipschitz continuous with constant } L \iff \nabla^2 f \preccurlyeq L \cdot \mathbf{I}_d$$

- Equivalently, eigenvalues of $\nabla^2 f$ are bounded by L
- **Interpretation:** Curvature in any direction is bounded by L
- Lipschitz gradients occur frequently in machine learning
  $\implies$ Fairly **weak assumption**
- Important for analysis of **gradient descent** optimization
  $\implies$ Descent lemma (later)

Optimization in Machine Learning

Mathematical Concepts: Taylor Approximations

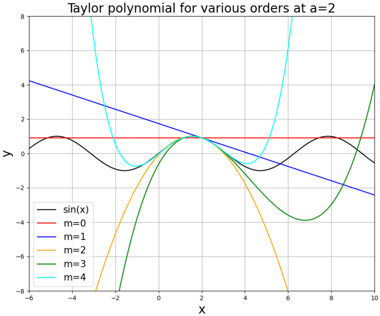


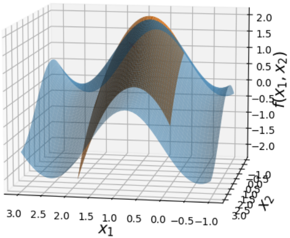
Learning goals

- Taylor's theorem (univariate)
- Taylor series (univariate)
- Taylor's theorem (multivariate)
- Taylor series (multivariate)

TAYLOR APPROXIMATIONS

- Mathematically fascinating: **Globally** approximate function by sum of polynomials determined by **local** properties
- Extremely important for **analyzing** optimization algorithms
- Geometry of **linear** and **quadratic** functions very well understood
  $\implies$ use them for **approximations**





TAYLOR'S THEOREM (UNIVARIATE)

Taylor's theorem: Let $I \subseteq \mathbb{R}$ be an open interval and $f \in \mathcal{C}^k(I, \mathbb{R})$. For each $a, x \in I$, it holds that

$$f(x) = \underbrace{\sum_{j=0}^{k} \frac{f^{(j)}(a)}{j!}(x-a)^j}_{T_k(x,a)} + R_k(x,a)$$

with the k-th **Taylor polynomial** T_k and a **remainder term**

$$R_k(x,a) = o(|x-a|^k) \quad \text{as } x \to a.$$

- There are explicit formulas for the remainder
- Wording: We "expand f via Taylor around a"

TAYLOR SERIES (UNIVARIATE)

- If $f \in C^\infty$, it *might* be expandable around $a \in I$ as a **Taylor series**

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k$$

- If Taylor series converges to f in an interval $I_0 \subseteq I$ centered at a (does not have to), we call f an *analytic function*
- Convergence if $R_k(x, a) \to 0$ as $k \to \infty$ for all $x \in I_0$
- Then, for all $x \in I_0$:

$$f(x) = \sum_{j=0}^{\infty} \frac{f^{(j)}(a)}{j!}(x-a)^j$$

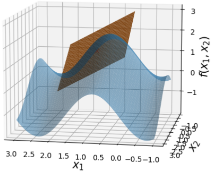
TAYLOR'S THEOREM (MULTIVARIATE)

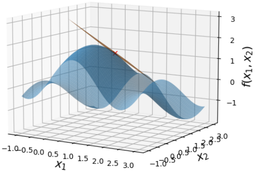
Taylor's theorem (1st order): For $f \in \mathcal{C}^1$, it holds that

$$f(\mathbf{x}) = \underbrace{f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})^T(\mathbf{x} - \boldsymbol{a})}_{T_1(\mathbf{x},\boldsymbol{a})} + R_1(\mathbf{x}, \boldsymbol{a}).$$

Example: $f(\mathbf{x}) = \sin(2x_1) + \cos(x_2)$, $\boldsymbol{a} = (1, 1)^T$. Since $\nabla f(\mathbf{x}) = \begin{pmatrix} 2\cos(2x_1) \\ -\sin(x_2) \end{pmatrix}$,

$$\begin{aligned} f(\mathbf{x}) &= T_1(\mathbf{x}) + R_1(\mathbf{x}, \boldsymbol{a}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})^T(\mathbf{x} - \boldsymbol{a}) + R_1(\mathbf{x}, \boldsymbol{a}) \\ &= \sin(2) + \cos(1) + (2\cos(2), -\sin(1))^T \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix} + R_1(\mathbf{x}, \boldsymbol{a}) \end{aligned}$$





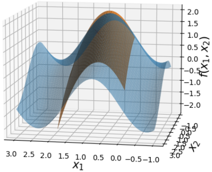
TAYLOR'S THEOREM (MULTIVARIATE)

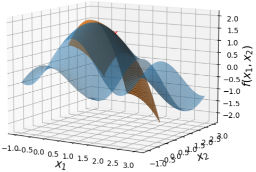
Taylor's theorem (2nd order): If $f \in \mathcal{C}^2$, it holds that

$$f(\mathbf{x}) = \underbrace{f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})^T(\mathbf{x} - \boldsymbol{a}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{a})^T \boldsymbol{H}(\boldsymbol{a})(\mathbf{x} - \boldsymbol{a})}_{T_2(\mathbf{x}, \boldsymbol{a})} + R_2(\mathbf{x}, \boldsymbol{a})$$

Example (continued): Since $H(\mathbf{x}) = \begin{pmatrix} -4\sin(2x_1) & 0 \\ 0 & -\cos(x_2) \end{pmatrix}$,

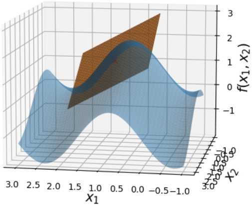
$$f(\mathbf{x}) = T_1(\mathbf{x}, \boldsymbol{a}) + \frac{1}{2}\begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix}^T \begin{pmatrix} -4\sin(2) & 0 \\ 0 & -\cos(1) \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix} + R_2(\mathbf{x}, \boldsymbol{a})$$

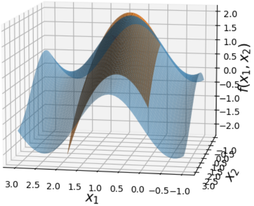




MULTIVARIATE TAYLOR APPROXIMATION

- Higher order k gives a better approximation
- $T_k(\mathbf{x}, \boldsymbol{a})$ is the best k-th order approximation to $f(\mathbf{x})$ near $\boldsymbol{a}$





Consider $T_2(\mathbf{x}, \boldsymbol{a}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})^T(\mathbf{x} - \boldsymbol{a}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{a})^T H(\boldsymbol{a})(\mathbf{x} - \boldsymbol{a})$. The first/second/third term ensures the values/slopes/curvatures of T_2 and f match at $\boldsymbol{a}$.

TAYLOR'S THEOREM (MULTIVARIATE)

The theorem for general order k requires a more involved notation.

Taylor's theorem (k-th order): If $f \in \mathcal{C}^k$, it holds that

$$f(\mathbf{x}) = \underbrace{\sum_{|\boldsymbol{\alpha}| \leq k} \frac{D^{\boldsymbol{\alpha}} f(\boldsymbol{a})}{\boldsymbol{\alpha}!} (\mathbf{x} - \boldsymbol{a})^{\boldsymbol{\alpha}}}_{T_k(\mathbf{x}, \boldsymbol{a})} + R_k(\mathbf{x}, \boldsymbol{a})$$

with $R_k(\mathbf{x}, \boldsymbol{a}) = o(\|\mathbf{x} - \boldsymbol{a}\|^k)$ as $\mathbf{x} \to \boldsymbol{a}$.

Notation: Multi-index $\boldsymbol{\alpha} \in \mathbb{N}^d$

- $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d$
- $\boldsymbol{\alpha}! = \alpha_1! \cdots \alpha_d!$
- $\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$
- $D^{\boldsymbol{\alpha}} f = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$

TAYLOR'S THEOREM (MULTIVARIATE)

Let us check for bivariate f ($d = 2$). For $|\boldsymbol{\alpha}| \leq 1$, we have

α_1	α_2	$\lvert\boldsymbol{\alpha}\rvert$	$D^{\boldsymbol{\alpha}} f$	$\boldsymbol{\alpha}!$	$(\mathbf{x} - \boldsymbol{a})^{\boldsymbol{\alpha}}$
0	0	0	f	1	1
1	0	1	$\partial f / \partial x_1$	1	$x_1 - a_1$
0	1	1	$\partial f / \partial x_2$	1	$x_2 - a_2$

and therefore

$$
\begin{aligned}
T_1(\mathbf{x}, \boldsymbol{a}) &= \frac{f(\boldsymbol{a})}{1} \cdot 1 + \frac{\partial f(\boldsymbol{a})}{\partial x_1}(x_1 - a_1) + \frac{\partial f(\boldsymbol{a})}{\partial x_2}(x_2 - a_2) \\
&= f(\boldsymbol{a}) + \begin{pmatrix} \frac{\partial f(\boldsymbol{a})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{a})}{\partial x_2} \end{pmatrix}^T \begin{pmatrix} x_1 - a_1 \\ x_2 - a_2 \end{pmatrix} \\
&= f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})^T (\mathbf{x} - \boldsymbol{a}).
\end{aligned}
$$

TAYLOR SERIES (MULTIVARIATE)

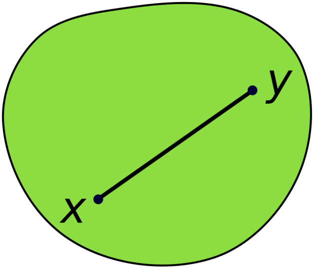
- Analogous to univariate case, if $f \in \mathcal{C}^\infty$, there *might* exist an open ball $B_r(\boldsymbol{a})$ with radius $r > 0$ around $\boldsymbol{a}$ such that the **Taylor series**

$$\sum_{|\boldsymbol{\alpha}| \geq 0} \frac{D^{\boldsymbol{\alpha}} f(\boldsymbol{a})}{\boldsymbol{\alpha}!} (\mathbf{x} - \boldsymbol{a})^{\boldsymbol{\alpha}}$$

  converges to f on $B_r(\boldsymbol{a})$
- Even if Taylor series converges, it might not converge to f
- Upper bound $R = \sup \{r \mid \text{Taylor series converges on } B_r(\boldsymbol{a})\}$ is called the **radius of convergence** of Taylor series around $\boldsymbol{a}$
- If $R > 0$ and f analytic, Taylor series converges *absolutely* and *uniformly* to f on *compact* sets inside $B_R(\boldsymbol{a})$
- No general convergence behaviour on boundary of $B_R(\boldsymbol{a})$

Optimization in Machine Learning

Mathematical Concepts: Convexity



Learning goals

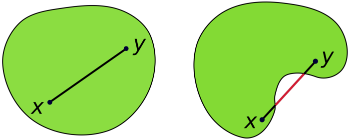
- Convex sets
- Convex functions

CONVEX SETS

A set of $\mathcal{S} \subseteq \mathbb{R}^d$ is **convex**, if for all $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ and all $t \in [0, 1]$ the following holds:

$$\mathbf{x} + t(\mathbf{y} - \mathbf{x}) \in \mathcal{S}$$

Intuitively: Connecting line between any $\mathbf{x}, \boldsymbol{y} \in \mathcal{S}$ lies completely in $\mathcal{S}$.



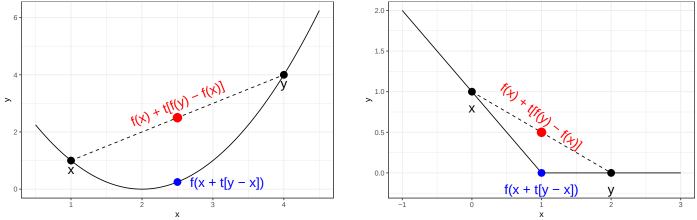
Left: convex set. **Right:** not convex. (Source: Wikipedia)

CONVEX FUNCTIONS

Let $f : \mathcal{S} \to \mathbb{R}$, $\mathcal{S}$ convex. f is **convex** if for all $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ and all $t \in [0, 1]$

$$f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) \leq f(\mathbf{x}) + t(f(\mathbf{y}) - f(\mathbf{x})).$$

Intuitively: Connecting line lies above function.



Left: Strictly convex function. **Right:** Convex, but not strictly.

Strictly convex if "$<$" instead of "$\leq$". **Concave** (strictly) if the inequality holds with "$\geq$" ("$>$"), respectively.

Note: f (strictly) concave $\Leftrightarrow$ $-f$ (strictly) convex.

EXAMPLES

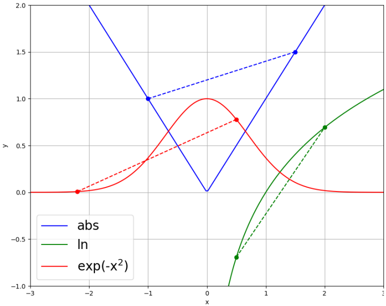
Convex function: $f(x) = |x|$

Proof:

$$\begin{aligned} f\,(x + t(y-x)) &= |x + t(y-x)| = |(1-t)x + t\cdot y| \\ &\le |(1-t)x| + |t\cdot y| = (1-t)|x| + t|y| \\ &= |x| + t\cdot(|y| - |x|) = f(x) + t\cdot(f(y) - f(x)) \end{aligned}$$

Concave function: $f(x) = \log(x)$

Neither nor: $f(x) = \exp(-x^2)$ (but log-concave)



OPERATIONS PRESERVING CONVEXITY

- **Nonnegatively weighted summation:** Weights $w_1, \dots, w_n \geq 0$, convex functions $f_1, \dots, f_n$: $w_1 f_1 + \cdots + w_n f_n$ also convex
  In particular: Sum of convex functions also convex
- **Composition:** g convex, f linear: $h = g \circ f$ also convex
  Proof:

$$\begin{aligned} h(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) &= g(f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))) \\ &= g(f(\mathbf{x}) + t(f(\mathbf{y}) - f(\mathbf{x}))) \\ &\leq g(f(\mathbf{x})) + t(g(f(\mathbf{y})) - g(f(\mathbf{x}))) \\ &= h(\mathbf{x}) + t(h(\mathbf{y}) - h(\mathbf{x})) \end{aligned}$$

- **Elementwise maximization:** $f_1, \dots, f_n$ convex functions:
  $g(\mathbf{x}) = \max \{f_1(\mathbf{x}), \dots, f_n(\mathbf{x})\}$ also convex

FIRST ORDER CONDITION

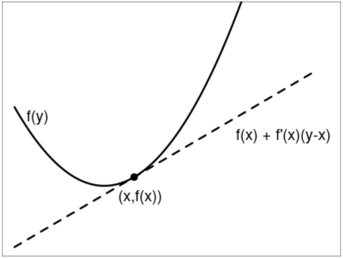
Prove convexity via **gradient**:

Let f be differentiable.

$$f \text{ (strictly) convex}$$

$$\iff$$

$$f(\mathbf{y}) \overset{(>)}{\geq} f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \text{ for all } \mathbf{x}, \mathbf{y} \in \mathcal{S} \text{ (s.t. } \mathbf{x} \neq \mathbf{y})$$



SECOND ORDER CONDITION

Matrix A is **positive (semi)definite** (p.(s.)d.) if $\boldsymbol{v}^T A \boldsymbol{v} \overset{(\geq)}{>} 0$ for all $\boldsymbol{v} \neq 0$.

Notation: $A \overset{(\succcurlyeq)}{\succ} 0$ for A p.(s.)d. and $B \overset{(\succcurlyeq)}{\succ} A$ if $B - A \overset{(\succcurlyeq)}{\succ} 0$

Prove convexity via **Hessian**:

Let $f \in \mathcal{C}^2$ and $H(\mathbf{x})$ be its Hessian.

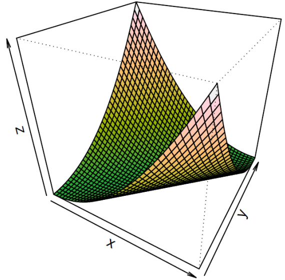
$$f \text{ (strictly) convex} \Longleftrightarrow H(\mathbf{x}) \overset{(\succ)}{\succcurlyeq} 0 \text{ for all } \mathbf{x} \in \mathcal{S}$$

Alternatively: Since $H(\mathbf{x})$ symmetric for $f \in \mathcal{C}^2$:

$$H(\mathbf{x}) \succcurlyeq 0 \Leftrightarrow \text{all eigenvalues of } H(\mathbf{x}) \geq 0$$

SECOND ORDER CONDITION

Example: $f(\mathbf{x}) = x_1^2 + x_2^2 - 2x_1x_2$, $\nabla f(\mathbf{x}) = \begin{pmatrix} 2x_1 - 2x_2 \\ 2x_2 - 2x_1 \end{pmatrix}$, $H(\mathbf{x}) = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$.

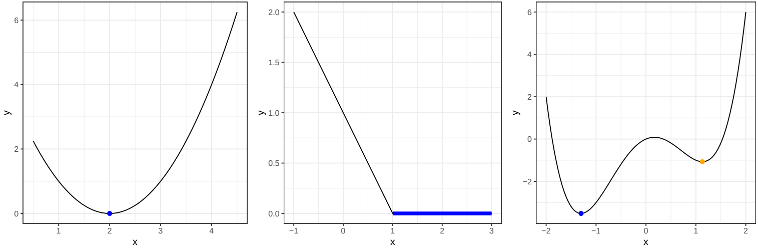


f is convex since $H(\mathbf{x})$ is p.s.d. for all $\mathbf{x} \in \mathcal{S}$:

$$\mathbf{v}^T \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \mathbf{v} = \mathbf{v}^T \begin{pmatrix} 2v_1 - 2v_2 \\ -2v_1 + 2v_2 \end{pmatrix} = 2v_1^2 - 2v_1v_2 - 2v_1v_2 + 2v_2^2$$
$$= 2v_1^2 - 4v_1v_2 + 2v_2^2 = 2(v_1 - v_2)^2 \geq 0.$$

CONVEX FUNCTIONS IN OPTIMIZATION

- For a convex function, every local optimum is also a global one
  $\Rightarrow$ No need for involved global optimizers, local ones are enough
- A strictly convex function has at most one optimal point
- Example for strictly convex function without optimum: $\exp$ on $\mathbb{R}$



Left: Strictly convex; exactly one local minimum, which is also global. **Middle:** Convex, but not strictly; all local optima are also global ones but not unique. **Right:** Not convex.

CONVEX FUNCTIONS IN OPTIMIZATION

"... in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

– R. Tyrrell Rockafellar. *SIAM Review*, 1993.

SIAM REVIEW
Vol. 35, No. 2, pp. 183–238, June 1993

LAGRANGE MULTIPLIERS AND OPTIMALITY*

R. TYRRELL ROCKAFELLAR†

Abstract. Lagrange multipliers used to be viewed as auxiliary variables introduced in a problem of constrained minimization in order to write first-order optimality conditions formally as a system of equations. Modern applications, with their emphasis on numerical methods and more complicated side conditions than equations, have demanded deeper understanding of the concept and how it fits into a larger theoretical picture.

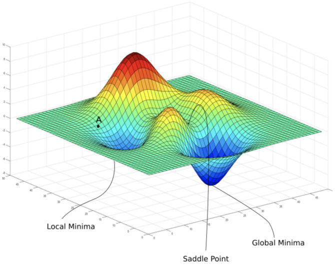
A major line of research has been the nonsmooth geometry of one-sided tangent and normal vectors to the set of points satisfying the given constraints. Another has been the game-theoretic role of multiplier vectors as solutions to a dual problem. Interpretations as generalized derivatives of the optimal value with respect to problem parameters have also been explored. Lagrange multipliers are now being seen as arising from a general rule for the subdifferentiation of a nonsmooth objective function which allows black-and-white constraints to be replaced by penalty expressions. This paper traces such themes in the current theory of Lagrange multipliers, providing along the way a free-standing exposition of basic nonsmooth analysis as motivated by and applied to this subject.

Key words. Lagrange multipliers, optimization, saddle points, dual problems, augmented Lagrangian, constraint qualifications, normal cones, subgradients, nonsmooth analysis

AMS subject classifications. 49K99, 58C20, 90C99, 49M29

Optimization in Machine Learning

Mathematical Concepts: Conditions for optimality



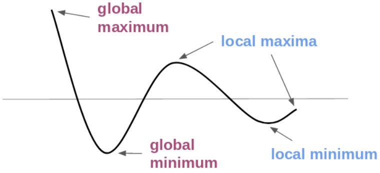
Learning goals

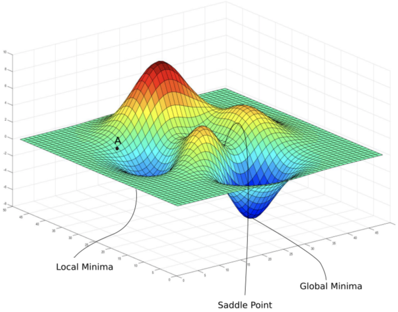
- Local and global optima
- First & second order conditions

DEFINITION LOCAL AND GLOBAL MINIMUM

Given $\mathcal{S} \subseteq \mathbb{R}^d$, $f : \mathcal{S} \to \mathbb{R}$:

- f has **global minimum** in $\mathbf{x}^* \in \mathcal{S}$, if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{S}$
- f has a **local minimum** in $\mathbf{x}^* \in \mathcal{S}$, if $\epsilon > 0$ exists s.t. $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in B_\epsilon(\mathbf{x}^*)$ ("ϵ"-ball around $\mathbf{x}^*$).





Source (**left**): `https://en.wikipedia.org/wiki/Maxima_and_minima`.

Source (**right**): `https://wngaw.github.io/linear-regression/`.

EXISTENCE OF OPTIMA

We regard the two main cases of $f : \mathcal{S} \to \mathbb{R}$:

- f **continuous**: If $\mathcal{S}$ is **compact**, f attains a minimum and a maximum (extreme value theorem).
- f **discontinuous**: **No general** statement possible about existence of optima.

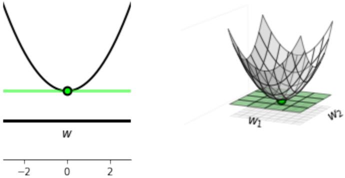
Example: $\mathcal{S} = [0, 1]$ compact, f discontinuous with

$$f(x) = \begin{cases} 1/x & \text{if } x > 0, \\ 0 & \text{if } x = 0. \end{cases}$$

FIRST ORDER CONDITION FOR OPTIMALITY

Observation: At an interior local optimum of $f \in \mathcal{C}^1$, first order Taylor approximation is flat, i.e., first order derivatives are zero.

This condition is therefore **necessary** and called **first order**.



Strictly convex functions (**left:** univariate, **right:** multivariate) with unique local minimum, which is the global one. Tangent (hyperplane) is perfectly flat at the optimum. (Source: Watt, *Machine Learning Refined*, 2020)

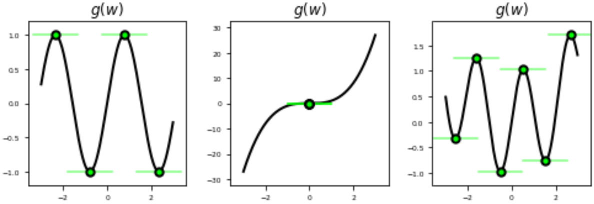
FIRST ORDER CONDITION FOR OPTIMALITY

First order condition: Gradient of f at local optimum $\mathbf{x}^* \in \mathcal{S}$ is zero:

$$\nabla f(\mathbf{x}^*) = (0, \dots, 0)^T$$

Points with zero first order derivative are called **stationary**.

Condition is **not sufficient**: Not all stationary points are local optima.



Left: Four points fulfill the necessary condition and are indeed optima.
Middle: One point fulfills the necessary condition but is not a local optimum.
Right: Multiple local minima and maxima.
(Source: Watt, 2020, Machine Learning Refined)

SECOND ORDER CONDITION FOR OPTIMALITY

Second order condition: Hessian of $f \in \mathcal{C}^2$ at stationary point $\mathbf{x}^* \in \mathcal{S}$ is positive or negative definite:

$$H(\mathbf{x}^*) \succ 0 \text{ or } H(\mathbf{x}^*) \prec 0$$

Interpretation: Curvature of f at local optimum is either positive in all directions or negative in all directions.

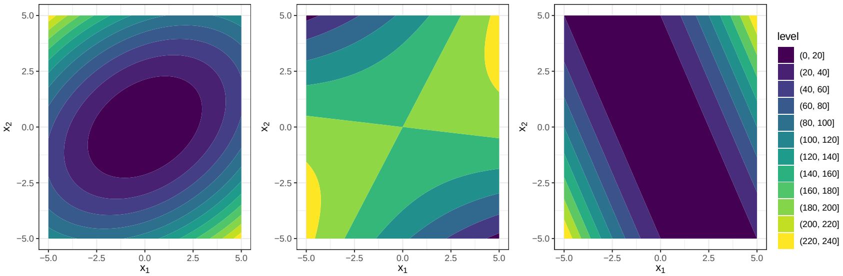
The second order condition is **sufficient** for a stationary point.

Proof: Later.

CONDITIONS FOR OPTIMALITY AND CONVEXITY

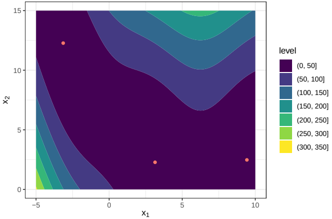
Let $f : \mathcal{S} \to \mathbb{R}$ be **convex**. Then:

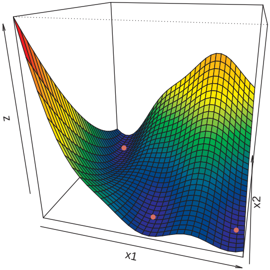
- Any local minimum is **also global** minimum
- If f **strictly convex**, f has **at most one** local minimum which would also be unique global minimum on $\mathcal{S}$



Three quadratic forms. **Left:** $H(\mathbf{x}^*)$ has two positive eigenvalues. **Middle:** $H(\mathbf{x}^*)$ has positive and negative eigenvalue. **Right:** $H(\mathbf{x}^*)$ has positive and a zero eigenvalue.

Example: Branin function





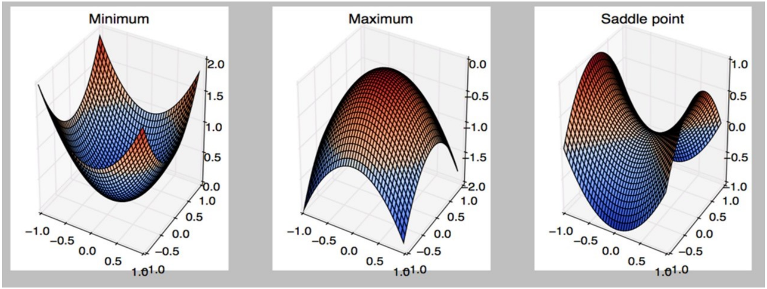
Spectra of Hessians (numerically computed):

	λ_1	λ_2
Left	22.29	0.96
Middle	11.07	1.73
Right	11.33	1.69

CONDITIONS FOR OPTIMALITY AND CONVEXITY

Definition: **Saddle point** at $\mathbf{x}$

- $\mathbf{x}$ stationary (necessary)
- $H(\mathbf{x})$ indefinite, i.e., positive and negative eigenvalues (sufficient)



CONDITIONS FOR OPTIMALITY AND CONVEXITY

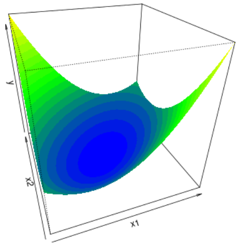
Examples:

- $f(x, y) = x^2 - y^2$, $\nabla f(x, y) = (2x, -2y)^T$,
  $H_f(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$
  $\implies$ Saddle point at $(0, 0)$ (sufficient condition met)

- $g(x, y) = x^4 - y^4$, $\nabla g(x, y) = (4x^3, -4y^3)^T$,
  $H_g(x, y) = \begin{pmatrix} 12x^2 & 0 \\ 0 & -12y^2 \end{pmatrix}$
  $\implies$ Saddle point at $(0, 0)$ (sufficient condition **not** met)

Optimization in Machine Learning

Mathematical Concepts: Quadratic forms I



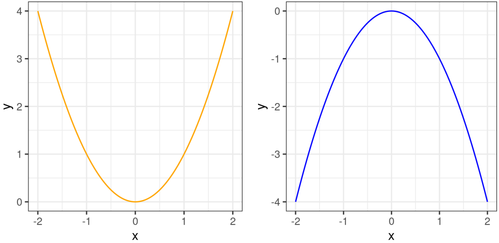
Learning goals

- Definition of quadratic forms
- Gradient, Hessian
- Optima

UNIVARIATE QUADRATIC FUNCTIONS

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

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

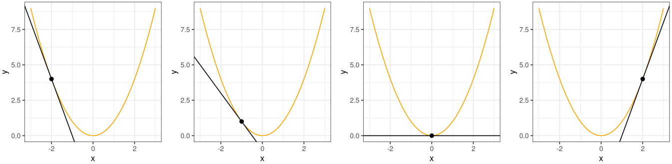


A quadratic function $q_1(x) = x^2$ (**left**) and $q_2(x) = -x^2$ (**right**).

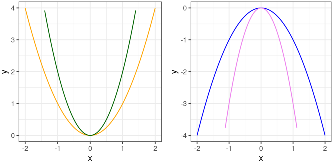
UNIVARIATE QUADRATIC FUNCTIONS

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 $q''(x) = 2 \cdot a$.

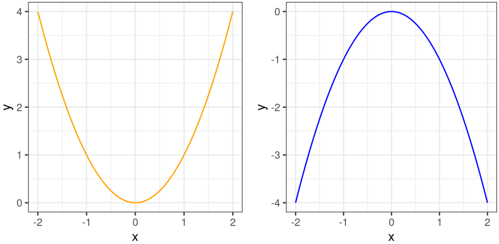


$q_1 = x^2$ (orange), $q_2 = 2x^2$ (green), $q_3(x) = -x^2$ (blue), $q_4 = -3x^2$ (magenta)

UNIVARIATE QUADRATIC FUNCTIONS

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

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



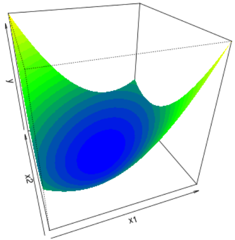
Left: $q_1(x) = x^2$ (convex). **Right:** $q_2(x) = -x^2$ (concave).

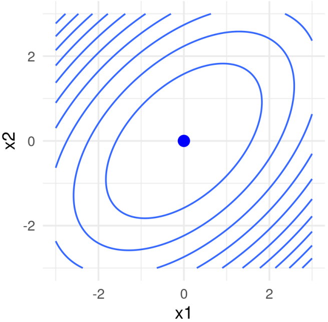
MULTIVARIATE QUADRATIC FUNCTIONS

A quadratic function $q : \mathbb{R}^d \rightarrow \mathbb{R}$ 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

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

If **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)

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

- The **Hessian** of q is

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

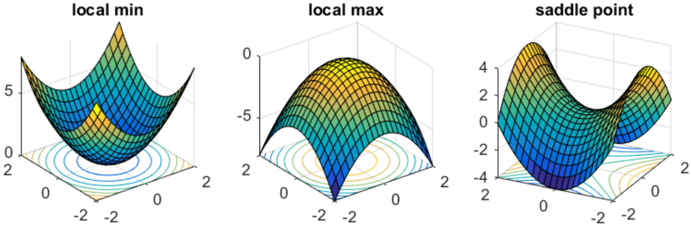
Curvature in direction of $\boldsymbol{v} \in \mathbb{R}^d$ is (by chain rule)

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

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.

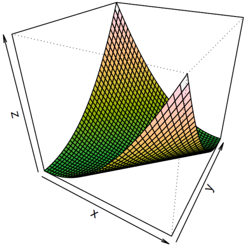
OPTIMA: RANK-DEFICIENT CASE

Example: Assume **A** is **not** full rank but has a zero eigenvalue with eigenvector $\boldsymbol{v}_0$.

- Recall: $\boldsymbol{v}_0$ spans null space of **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 $\nabla q(\mathbf{x}) = 2\mathbf{A}\mathbf{x} + \boldsymbol{b}$:

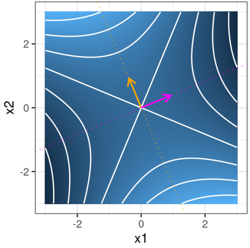
$$\nabla q(\mathbf{x} + \alpha \boldsymbol{v}_0) = 2\mathbf{A}(\mathbf{x} + \alpha \boldsymbol{v}_0) + \boldsymbol{b} = 2\mathbf{A}\mathbf{x} + \boldsymbol{b} = \nabla q(\mathbf{x})$$

- $\implies q$ 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 $\boldsymbol{v}_0$



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

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

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

- $q''(x) \overset{(>)}{\geq} 0$: q (strictly) convex. $q''(x) \overset{(<)}{\leq} 0$: 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 $q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$

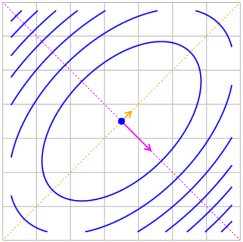
GEOMETRY OF QUADRATIC FUNCTIONS

Example: $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \implies \mathbf{H} = 2\mathbf{A} = \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix}$

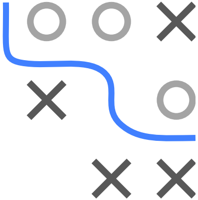
- Since $\mathbf{H}$ symmetric, eigendecomposition $\mathbf{H} = \mathbf{V}\Lambda\mathbf{V}^T$ with

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

$$\text{and } \Lambda = \begin{pmatrix} \lambda_{\max} & 0 \\ 0 & \lambda_{\min} \end{pmatrix} = \begin{pmatrix} 6 & 0 \\ 0 & 2 \end{pmatrix}.$$



GEOMETRY OF QUADRATIC FUNCTIONS



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

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

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

  Since $\|\boldsymbol{v}\| = \|\mathbf{x}\|$ ($\mathbf{V}$ orthogonal): $\max_{\|\mathbf{x}\|=1} \mathbf{x}^T\mathbf{H}\mathbf{x} \leq \lambda_{\max}$
  Additional: ${\boldsymbol{v}_{\max}}^T\mathbf{H}\boldsymbol{v}_{\max} = \mathbf{e}_1^T\Lambda\mathbf{e}_1 = \lambda_{\max}$
  Analogous: $\min_{\|\mathbf{x}\|=1} \mathbf{x}^T\mathbf{H}\mathbf{x} \geq \lambda_{\min}$ and ${\boldsymbol{v}_{\min}}^T\mathbf{H}\boldsymbol{v}_{\min} = \lambda_{\min}$

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

  Look at $q(\mathbf{x}) = \mathbf{x}^T\boldsymbol{A}\mathbf{x} + \boldsymbol{b}^T\mathbf{x} + c$
  Now use $\boldsymbol{y} = \mathbf{x} - \boldsymbol{w} = \mathbf{x} + \frac{1}{2}\boldsymbol{A}^{-1}\boldsymbol{b}$
  This already gives us the general form of an ellipse:
  $\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} = (\mathbf{x} - \boldsymbol{w})^T\boldsymbol{A}(\mathbf{x} - \boldsymbol{w}) = q(\mathbf{x}) + const$
  If we use $\boldsymbol{z} = \boldsymbol{V}^T y$ we obtain it in standard form
  $\sum_{i=1}^{n}\lambda_i z_i^2 = \boldsymbol{z}^T\boldsymbol{\Lambda}\boldsymbol{z} = y^T\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T y = \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} = q(\mathbf{x}) + const$

GEOMETRY OF QUADRATIC FUNCTIONS

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

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

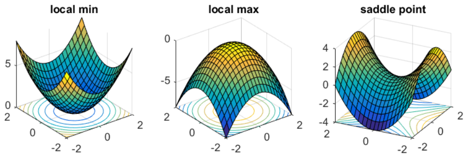
Choose $\epsilon > 0$ s.t. $|R_2(\mathbf{x}, \mathbf{x}^*)| < \frac{1}{2}\lambda_{\text{min}}\|\mathbf{x} - \mathbf{x}^*\|^2$ for each $\mathbf{x} \neq \mathbf{x}^*$ with $\|\mathbf{x} - \mathbf{x}^*\| < \epsilon$.
Then:

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

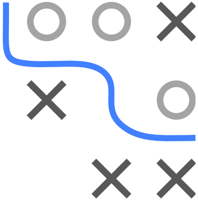
GEOMETRY OF QUADRATIC FUNCTIONS

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,
  - there is a (unique) global minimum.
- If **all** eigenvalues of $\mathbf{H} \overset{(<)}{\leq} 0$ ($\Leftrightarrow$ $\mathbf{H} \overset{(\prec)}{\preccurlyeq} 0$):
  - 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.



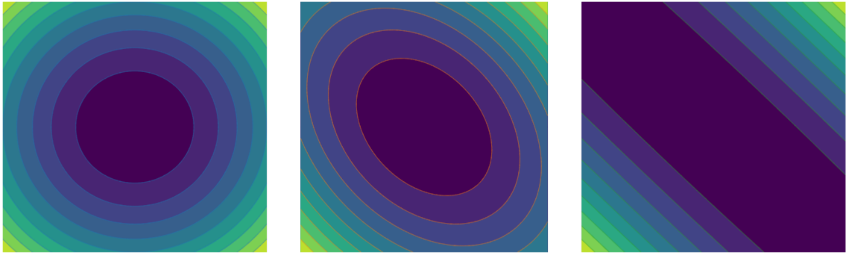
CONDITION AND CURVATURE



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

High condition means:

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

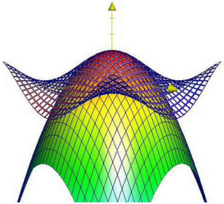


Left: Excellent condition. **Middle:** Good condition. **Right:** Bad condition.

APPROXIMATION OF SMOOTH FUNCTIONS

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

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



f and its second order approximation is shown by the dark and bright grid, respectively. (Source: `daniloroccatano.blog`)

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

Optimization in Machine Learning

Mathematical Concepts: Matrix Calculus

δ

Learning goals

- Rules of matrix calculus
- Connection of gradient, Jacobian and Hessian

SCOPE

- $\mathcal{X}/\mathcal{Y}$ denote space of **independent/dependent** variables
- Identify dependent variable with a **function** $y : \mathcal{X} \to \mathcal{Y}, x \mapsto y(x)$
- Assume y sufficiently smooth
- In matrix calculus, x and y can be **scalars**, **vectors**, or **matrices**:

Type	scalar x	vector $\mathbf{x}$	matrix $\mathbf{X}$
scalar y	$\partial y/\partial x$	$\partial y/\partial \mathbf{x}$	$\partial y/\partial \mathbf{X}$
vector $\mathbf{y}$	$\partial \mathbf{y}/\partial x$	$\partial \mathbf{y}/\partial \mathbf{x}$	–
matrix $\mathbf{Y}$	$\partial \mathbf{Y}/\partial x$	–	–

- We denote vectors/matrices in **bold** lowercase/uppercase letters

NUMERATOR LAYOUT

- **Matrix calculus:** collect derivative of each component of dependent variable w.r.t. each component of independent variable
- We use so-called **numerator layout** convention:

$$\frac{\partial y}{\partial \mathbf{x}} = \left(\frac{\partial y}{\partial x_1}, \cdots, \frac{\partial y}{\partial x_d} \right) = \nabla y^T \in \mathbb{R}^{1\times d}$$

$$\frac{\partial \mathbf{y}}{\partial x} = \left(\frac{\partial y_1}{\partial x}, \cdots, \frac{\partial y_m}{\partial x} \right)^T \in \mathbb{R}^{m}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial y_m}{\partial \mathbf{x}} \end{pmatrix} = \left(\frac{\partial \mathbf{y}}{\partial x_1} \cdots \frac{\partial \mathbf{y}}{\partial x_d} \right) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_d} \end{pmatrix} = J_{\mathbf{y}} \in \mathbb{R}^{m\times d}$$

SCALAR-BY-VECTOR

Let $\mathbf{x} \in \mathbb{R}^d$, $y, z : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{A}$ be a matrix.

- If y is a **constant** function: $\frac{\partial y}{\partial \mathbf{x}} = \mathbf{0}^T \in \mathbb{R}^{1\times d}$
- **Linearity**: $\frac{\partial (a\cdot y+z)}{\partial \mathbf{x}} = a\frac{\partial y}{\partial \mathbf{x}} + \frac{\partial z}{\partial \mathbf{x}}$ (a constant)
- **Product** rule: $\frac{\partial (y\cdot z)}{\partial \mathbf{x}} = y\frac{\partial z}{\partial \mathbf{x}} + \frac{\partial y}{\partial \mathbf{x}} z$
- **Chain** rule: $\frac{\partial g(y)}{\partial \mathbf{x}} = \frac{\partial g(y)}{\partial y}\frac{\partial y}{\partial \mathbf{x}}$ (g scalar-valued function)
- **Second** derivative: $\frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}^T} = \nabla^2 y^T$ ($= \nabla^2 y$ if $y \in \mathcal{C}^2$) (Hessian)
- $\frac{\partial (\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T)$
- $\frac{\partial (\mathbf{y}^T\mathbf{A}\mathbf{z})}{\partial \mathbf{x}} = \mathbf{y}^T\mathbf{A}\frac{\partial \mathbf{z}}{\partial \mathbf{x}} + \mathbf{z}^T\mathbf{A}^T\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ ($\mathbf{y}$, $\mathbf{z}$ vector-valued functions of $\mathbf{x}$)

VECTOR-BY-SCALAR

Let $x \in \mathbb{R}$ and $\mathbf{y}, \mathbf{z} : \mathbb{R} \to \mathbb{R}^m$.

- If $\mathbf{y}$ is a **constant** function: $\frac{\partial \mathbf{y}}{\partial x} = \mathbf{0} \in \mathbb{R}^m$
- **Linearity**: $\frac{\partial (a \cdot \mathbf{y} + \mathbf{z})}{\partial x} = a \frac{\partial \mathbf{y}}{\partial x} + \frac{\partial \mathbf{z}}{\partial x}$ (a constant)
- **Chain** rule: $\frac{\partial \mathbf{g}(\mathbf{y})}{\partial x} = \frac{\partial \mathbf{g}(\mathbf{y})}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial x}$ ($\mathbf{g}$ vector-valued function)
- $\frac{\partial (\mathbf{A}\mathbf{y})}{\partial x} = \mathbf{A} \frac{\partial \mathbf{y}}{\partial x}$ ($\mathbf{A}$ matrix)

VECTOR-BY-VECTOR

Let $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y}, \mathbf{z} : \mathbb{R}^d \to \mathbb{R}^m$.

- If $\mathbf{y}$ is a **constant** function: $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{0} \in \mathbb{R}^{m \times d}$
- $\frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I} \in \mathbb{R}^{d \times d}$
- **Linearity**: $\frac{\partial (a \cdot \mathbf{y} + \mathbf{z})}{\partial \mathbf{x}} = a \frac{\partial \mathbf{y}}{\partial \mathbf{x}} + \frac{\partial \mathbf{z}}{\partial \mathbf{x}}$ (a constant)
- **Chain** rule: $\frac{\partial \mathbf{g}(\mathbf{y})}{\partial \mathbf{x}} = \frac{\partial \mathbf{g}(\mathbf{y})}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ ($\mathbf{g}$ vector-valued function)
- $\frac{\partial (\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \mathbf{A}$, $\frac{\partial (\mathbf{x}^T \mathbf{B})}{\partial \mathbf{x}} = \mathbf{B}^T$ ($\mathbf{A}, \mathbf{B}$ matrices)

EXAMPLE

Consider $f : \mathbb{R}^2 \to \mathbb{R}$ with

$$f(\mathbf{x}) = \exp\left(-(\mathbf{x} - \mathbf{c})^T \mathbf{A}(\mathbf{x} - \mathbf{c})\right),$$

where $\mathbf{c} = (1, 1)^T$ and $\mathbf{A} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$.

Compute $\nabla f(\mathbf{x})$ at $\mathbf{x}^* = \mathbf{0}$:

1. Write $f(\mathbf{x}) = \exp(g(\mathbf{u}(\mathbf{x})))$ with $g(\mathbf{u}) = -\mathbf{u}^T \mathbf{A} \mathbf{u}$ and $\mathbf{u}(\mathbf{x}) = \mathbf{x} - \mathbf{c}$
2. **Chain** rule: $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \exp(g(\mathbf{u}(\mathbf{x}))) \frac{\partial g(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{x}}$
3. $\mathbf{u}^* := \mathbf{u}(\mathbf{x}^*) = (-1, -1)^T$, $g(\mathbf{u}^*) = -3$
4. $\frac{\partial g(\mathbf{u})}{\partial \mathbf{u}} = -2\mathbf{u}^T \mathbf{A}$, $\frac{\partial g(\mathbf{u}^*)}{\partial \mathbf{u}} = (3, 3)$
5. **Linearity**: $\frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial (\mathbf{x} - \mathbf{c})}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}}{\partial \mathbf{x}} - \frac{\partial \mathbf{c}}{\partial \mathbf{x}} = \mathbf{I}_2$
6. $\nabla f(\mathbf{x}^*) = \frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}}^T = (\exp(-3) \cdot (3, 3) \cdot \mathbf{I}_2)^T = \exp(-3) \begin{pmatrix} 3 \\ 3 \end{pmatrix}$