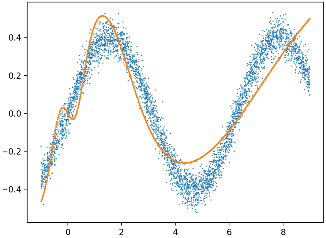
# Basics of Numerical Optimization: Preliminaries

**Ju Sun**
Computer Science & Engineering
University of Minnesota, Twin Cities

February 11, 2020

# Supervised learning as function approximation




- Underlying true function: $f_0$
- Training data: $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ with $\boldsymbol{y}_i \approx f_0(\boldsymbol{x}_i)$
- Choose a family of functions $\mathcal{H}$, so that $\exists f \in \mathcal{H}$ and $f$ and $f_0$ are close
- Find $f$, i.e., optimization

$$\min_{f \in \mathcal{H}} \sum_i \ell\left(\boldsymbol{y}_i, f\left(\boldsymbol{x}_i\right)\right) + \Omega\left(f\right)$$

- **Approximation capacity: Univeral approximation theorems (UAT)**
  $\Longrightarrow$ replace $\mathcal{H}$ by $\mathrm{DNN}_{\boldsymbol{W}}$, i.e., a deep neural network with weights $\boldsymbol{W}$
- **Optimization:**

$$\min_{\boldsymbol{W}} \sum_i \ell\left(\boldsymbol{y}_i, \mathrm{DNN}_{\boldsymbol{W}}\left(\boldsymbol{x}_i\right)\right) + \Omega\left(\boldsymbol{W}\right)$$

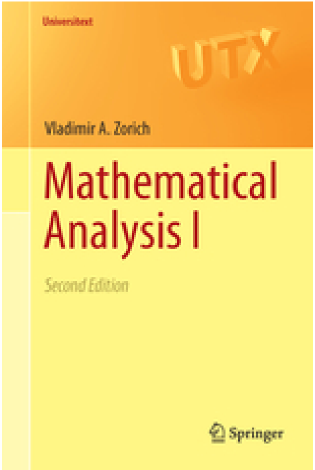
- **Generalization:** how to avoid over-complicated $\mathrm{DNN}_{\boldsymbol{W}}$ in view of UAT

Now we start to focus on **optimization**.

# Outline

Elements of multivatiate calculus

Optimality conditions of unconstrained optimization


[Munkres, 1997, Zorich, 2015, Coleman, 2012]

## Our notation

- scalars: $x$, vectors: $\boldsymbol{x}$, matrices: $\boldsymbol{X}$, tensors: $\mathcal{X}$, sets: $S$
- vectors are always **column vectors**, unless stated otherwise
- $x_i$: $i$-th element of $\boldsymbol{x}$, $x_{ij}$: $(i, j)$-th element of $\boldsymbol{X}$, $\boldsymbol{x}^i$: $i$-th row of $\boldsymbol{X}$ as a **row vector**, $\boldsymbol{x}_j$: $j$-th column of $\boldsymbol{X}$ as a **column vector**
- $\mathbb{R}$: real numbers, $\mathbb{R}_+$: positive reals, $\mathbb{R}^n$: space of $n$-dimensional vectors, $\mathbb{R}^{m\times n}$: space of $m \times n$ matrices, $\mathbb{R}^{m\times n\times k}$: space of $m \times n \times k$ tensors, etc
- $[n] \doteq \{1, \ldots, n\}$

# Differentiability — first order

Consider $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^m$

- Definition: **First-order differentiable** at a point $\boldsymbol{x}$ if there exists a matrix $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ such that

$$\frac{f(\boldsymbol{x}+\boldsymbol{\delta}) - f(\boldsymbol{x}) - \boldsymbol{B}\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2} \to \boldsymbol{0} \quad \text{as} \quad \boldsymbol{\delta} \to \boldsymbol{0}.$$

$$i.e., \quad f(\boldsymbol{x}+\boldsymbol{\delta}) = f(\boldsymbol{x}) + \boldsymbol{B}\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2) \quad \text{as} \quad \boldsymbol{\delta} \to \boldsymbol{0}.$$

- $\boldsymbol{B}$ is called the (Fréchet) derivative. When $m = 1$, $\boldsymbol{b}^\intercal$ (i.e., $\boldsymbol{B}^\intercal$) called **gradient**, denoted as $\nabla f(\boldsymbol{x})$. For general $m$, also called **Jacobian** matrix, denoted as $\boldsymbol{J}_f(\boldsymbol{x})$.
- Calculation: $b_{ij} = \frac{\partial f_i}{\partial x_j}(\boldsymbol{x})$
- **Sufficient condition**: if all partial derivatives exist and are **continuous** at $\boldsymbol{x}$, then $f(\boldsymbol{x})$ is differentiable at $\boldsymbol{x}$.

# Calculus rules

Assume $f, g : \mathbb{R}^n \to \mathbb{R}^m$ are differentiable at a point $\boldsymbol{x} \in \mathbb{R}^n$.

- **linearity**: $\lambda_1 f + \lambda_2 g$ is differentiable at $\boldsymbol{x}$ and $\nabla \left[\lambda_1 f + \lambda_2 g\right](\boldsymbol{x}) = \lambda_1 \nabla f(\boldsymbol{x}) + \lambda_2 \nabla g(\boldsymbol{x})$
- **product**: assume $m = 1$, $fg$ is differentiable at $\boldsymbol{x}$ and $\nabla \left[fg\right](\boldsymbol{x}) = f(\boldsymbol{x}) \nabla g(\boldsymbol{x}) + g(\boldsymbol{x}) \nabla f(\boldsymbol{x})$
- **quotient**: assume $m = 1$ and $g(\boldsymbol{x}) \neq 0$, $\frac{f}{g}$ is differentiable at $\boldsymbol{x}$ and $\nabla \left[\frac{f}{g}\right](\boldsymbol{x}) = \frac{g(\boldsymbol{x}) \nabla f(\boldsymbol{x}) - f(\boldsymbol{x}) \nabla g(\boldsymbol{x})}{g^2(\boldsymbol{x})}$
- **Chain rule**: Let $f : \mathbb{R}^m \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}^k$, and $f$ is differentiable at $\boldsymbol{x}$ and $\boldsymbol{y} = f(\boldsymbol{x})$ and $h$ is differentiable at $\boldsymbol{y}$. Then, $h \circ f : \mathbb{R}^n \to \mathbb{R}^k$ is differentiable at $\boldsymbol{x}$, and

$$\boldsymbol{J}_{[h \circ f]}(\boldsymbol{x}) = \boldsymbol{J}_h(f(\boldsymbol{x})) \boldsymbol{J}_f(\boldsymbol{x}).$$

**When** $k = 1$**,**

$$\nabla \left[h \circ f\right](\boldsymbol{x}) = \boldsymbol{J}_f^\top(\boldsymbol{x}) \nabla h(f(\boldsymbol{x})).$$

# Differentiability — second order

Consider $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ and assume $f$ is 1st-order differentiable in a small ball around $\boldsymbol{x}$

- Write $\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x}) \doteq \left[\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right)\right](\boldsymbol{x})$ provided the right side well defined
- **Symmetry**: If both $\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})$ and $\frac{\partial f^2}{\partial x_i \partial x_j}(\boldsymbol{x})$ exist and both are continuous at $\boldsymbol{x}$, then they are equal.
- **Hessian (matrix)**:

$$\nabla^2 f(\boldsymbol{x}) \doteq \left[\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})\right]_{j,i}, \tag{1}$$

  where $\left[\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})\right]_{j,i} \in \mathbb{R}^{n \times n}$ has its $(j,i)$-th element as $\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})$.
- $\nabla^2 f$ is symmetric.
- **Sufficient condition**: if all $\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})$ exist and are **continuous**, $f$ is 2nd-order differentiable at $\boldsymbol{x}$ (not converse; we omit the definition due to its technicality).

# Taylor's theorem

**Vector version**: consider $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$

- If $f$ is 1st-order differentiable at $\boldsymbol{x}$, then

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle + o(\|\boldsymbol{\delta}\|_2) \text{ as } \boldsymbol{\delta} \to \mathbf{0}.$$

- If $f$ is 2nd-order differentiable at $\boldsymbol{x}$, then

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle + \frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{\delta} \right\rangle + o(\|\boldsymbol{\delta}\|_2^2) \text{ as } \boldsymbol{\delta} \to \mathbf{0}.$$

**Matrix version**: consider $f(\boldsymbol{X}) : \mathbb{R}^{m \times n} \to \mathbb{R}$

- If $f$ is 1st-order differentiable at $\boldsymbol{X}$, then

$$f(\boldsymbol{X}+\boldsymbol{\Delta}) = f(\boldsymbol{X}) + \langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta} \rangle + o(\|\boldsymbol{\Delta}\|_F) \text{ as } \boldsymbol{\Delta} \to \mathbf{0}.$$

- If $f$ is 2nd-order differentiable at $\boldsymbol{X}$, then

$$f(\boldsymbol{X}+\boldsymbol{\Delta}) = f(\boldsymbol{X}) + \langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta} \rangle + \frac{1}{2} \left\langle \boldsymbol{\Delta}, \nabla^2 f(\boldsymbol{X}) \boldsymbol{\Delta} \right\rangle + o(\|\boldsymbol{\Delta}\|_F^2)$$

$$\text{as } \boldsymbol{\Delta} \to \mathbf{0}.$$

# Taylor approximation — asymptotic uniqueness

Let $f : \mathbb{R} \to \mathbb{R}$ be $k$ ($k \geq 1$ integer) times differentiable at a point $x$. If $P(\delta)$ is a $k$-th order polynomial satisfying $f(x + \delta) - P(\delta) = o(\delta^k)$ as $\delta \to 0$, then $P(\delta) = P_k(\delta) \doteq f(x) + \sum_{i=1}^{k} \frac{1}{k!} f^{(k)}(x) \delta^k$.

**Generalization to the vector version**

- Assume $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is 1-order differentiable at $\boldsymbol{x}$. If $P(\boldsymbol{\delta}) \doteq f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{\delta} \rangle$ satisties that

$$f(\boldsymbol{x} + \boldsymbol{\delta}) - P(\boldsymbol{\delta}) = o(\|\boldsymbol{\delta}\|_2) \quad \text{as } \boldsymbol{\delta} \to \mathbf{0},$$

  then $P(\boldsymbol{\delta}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle$, i.e., the 1st-order Taylor expansion.
- Assume $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is 2-order differentiable at $\boldsymbol{x}$. If $P(\boldsymbol{\delta}) \doteq f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{\delta} \rangle + \frac{1}{2} \langle \boldsymbol{\delta}, \boldsymbol{H}\boldsymbol{\delta} \rangle$ with $\boldsymbol{H}$ symmetric satisties that

$$f(\boldsymbol{x} + \boldsymbol{\delta}) - P(\boldsymbol{\delta}) = o(\|\boldsymbol{\delta}\|_2^2) \quad \text{as } \boldsymbol{\delta} \to \mathbf{0},$$

  then $P(\boldsymbol{\delta}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle + \frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{\delta} \right\rangle$, i.e., the 2nd-order Taylor expansion. We can read off $\nabla f$ and $\nabla^2 f$ if we know the expansion!

**Similarly for the matrix version**. See Chap 5 of [Coleman, 2012] for other forms of Taylor theorems and proofs of the asymptotic uniqueness.

# Asymptotic uniqueness — why interesting?

## Two ways of deriving gradients and Hessians (Recall HW0!)

(a) Derive the gradient and Hessian of the linear least-squares function $f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$. Please include your calculation details.

(b) Let $\sigma = \frac{1}{1+e^{-x}}$, i.e., the *logistic function*. Derive the gradient of the matrix-variable function $g(\boldsymbol{W}) = \|\boldsymbol{y} - \sigma(\boldsymbol{W}\boldsymbol{x})\|_2^2$, where $\sigma$ is applied to the vector $\boldsymbol{W}\boldsymbol{x}$ elementwise. This is regression based on a simplified one-neuron network. Please include your calculation details.

(a) Consider the least-squares objective $f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ again. Recall that for any two vectors $\boldsymbol{a}, \boldsymbol{b}$, $\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 = \|\boldsymbol{a}\|_2^2 - 2\boldsymbol{a}^\intercal \boldsymbol{b} + \|\boldsymbol{b}\|_2^2$. Now $f(\boldsymbol{x} + \boldsymbol{\delta}) = \|(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) - \boldsymbol{A}\boldsymbol{\delta}\|_2^2$. Expand this square by the previous formula, and compare it to the 2nd order Taylor expansion by plugging your results from **Problem 1(a)**. Are they equal or not? Why? (Hint: You may find this fact useful: for any two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ and any matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$, $\langle \boldsymbol{u}, \boldsymbol{M}\boldsymbol{v} \rangle = \langle \boldsymbol{M}^\intercal \boldsymbol{u}, \boldsymbol{v} \rangle$. This can be derived from the trace cyclic property above. )

(b) Consider the one-neuron network regression again: $g(\boldsymbol{W}) = \|\boldsymbol{y} - \sigma(\boldsymbol{W}\boldsymbol{x})\|_2^2$ with $\sigma = \frac{1}{1+e^{-x}}$, i.e., the *logistic function*. Let's try to work out its 1st order Taylor expansion by direct expansion as follows.

- Show that $\sigma((\boldsymbol{W} + \boldsymbol{\Delta})\boldsymbol{x}) = \sigma(\boldsymbol{W}\boldsymbol{x}) + \sigma'(\boldsymbol{W}\boldsymbol{x}) \odot (\boldsymbol{\Delta}\boldsymbol{x}) + o(\|\boldsymbol{\Delta}\|_F)$ when $\boldsymbol{\Delta} \to \boldsymbol{0}$. Here, both $\sigma$ and $\sigma'$ are applied elementwise, and $\odot$ denotes the elementwise (Hadamard) product.
- So $\boldsymbol{y} - \sigma((\boldsymbol{W} + \boldsymbol{\Delta})\boldsymbol{x}) = (\boldsymbol{y} - \sigma(\boldsymbol{W}\boldsymbol{x})) - \sigma'(\boldsymbol{W}\boldsymbol{x}) \odot (\boldsymbol{\Delta}\boldsymbol{x}) - o(\|\boldsymbol{\Delta}\|_F)$ when $\boldsymbol{\Delta} \to \boldsymbol{0}$. Substitute this back into the square and use the identity $\|\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}\|_2^2 = \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2 + \|\boldsymbol{c}\|_2^2 + 2\boldsymbol{a}^\intercal \boldsymbol{b} + 2\boldsymbol{a}^\intercal \boldsymbol{c} + 2\boldsymbol{b}^\intercal \boldsymbol{c}$ to obtain the first-order approximation to $g(\boldsymbol{W} + \boldsymbol{\Delta})$. Remember that any terms lower order than $\|\boldsymbol{\Delta}\|_F$ are not interesting and we can always assume $\boldsymbol{\Delta}$ as small as needed.
- Substitute the result from **Problem 1(b)** into the 1st order Taylor expansion formula above and compare it to the result obtained here. Are they equal or not?

# Asymptotic uniqueness — why interesting?

Think of neural networks with identity activation functions

$$f(\boldsymbol{W}) = \sum_i \|\boldsymbol{y}_i - \boldsymbol{W}_k \boldsymbol{W}_{k-1} \ldots \boldsymbol{W}_2 \boldsymbol{W}_1 \boldsymbol{x}_i\|_F^2$$

How to derive the gradient?

- Scalar chain rule?
- Vector chain rule?
- First-order Taylor expansion

Why interesting? See e.g.,
[Kawaguchi, 2016, Lampinen and Ganguli, 2018]

# Directional derivatives and curvatures

Consider $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$

- **directional derivative**: $D_{\boldsymbol{v}} f(\boldsymbol{x}) \doteq \frac{d}{dt} f(\boldsymbol{x} + t\boldsymbol{v})$
- When $f$ is 1-st order differentiable at $\boldsymbol{x}$,

$$D_{\boldsymbol{v}} f(\boldsymbol{x}) = \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} \rangle .$$

- Now $D_{\boldsymbol{v}} f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$, what is $D_{\boldsymbol{u}}(D_{\boldsymbol{v}} f)(\boldsymbol{x})$?

$$D_{\boldsymbol{u}}(D_{\boldsymbol{v}} f)(\boldsymbol{x}) = \left\langle \boldsymbol{u}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{v} \right\rangle .$$

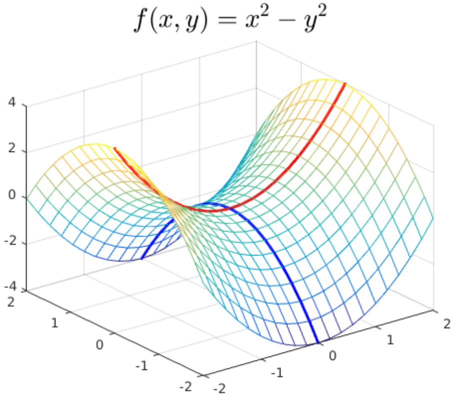
- When $\boldsymbol{u} = \boldsymbol{v}$,

$$D_{\boldsymbol{u}}(D_{\boldsymbol{u}} f)(\boldsymbol{x}) = \left\langle \boldsymbol{u}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{u} \right\rangle = \frac{d^2}{dt^2} f(\boldsymbol{x} + t\boldsymbol{u}) .$$

- $\frac{\left\langle \boldsymbol{u}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{u} \right\rangle}{\|\boldsymbol{u}\|_2^2}$ is the **directional curvature** along $\boldsymbol{u}$ independent of the norm of $\boldsymbol{u}$

# Directional curvature

$\frac{\langle \boldsymbol{u}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{u} \rangle}{\|\boldsymbol{u}\|_2^2}$ is the **directional curvature** along $\boldsymbol{u}$ independent of the norm of $\boldsymbol{u}$




Blue: negative curvature (bending down)

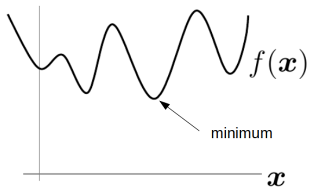
Red: positive curvature (bending up)

# Outline

Elements of multivatiate calculus

Optimality conditions of unconstrained optimization

# Optimization problems



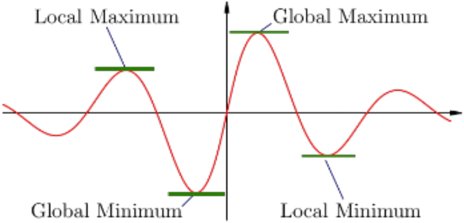

*Nothing takes place in the world whose meaning is not that of some maximum or minimum. – Euler*

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \text{ s.t. } \boldsymbol{x} \in C.$$

- $\boldsymbol{x}$: optimization variables, $f(\boldsymbol{x})$: objective function, $C$: constraint (or feasible) set
- $C$ consists of discrete values (e.g., $\{-1, +1\}^n$): discrete optimization; $C$ consists of continuous values (e.g., $\mathbb{R}^n$, $[0, 1]^n$): **continuous optimization**
- $C$ whole space $\mathbb{R}^n$: **unconstrained optimization**; $C$ a strict subset of the space: constrained optimization

We focus on continuous, unconstrained optimization here.

# Global and local mins




Credit: study.com

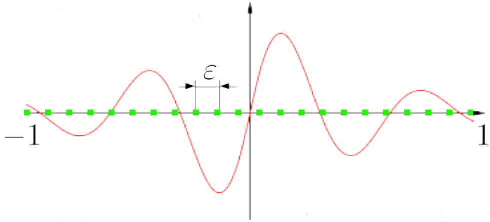
Let $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$,

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})$$

- $\boldsymbol{x}_0$ is a **local minimizer** if: $\exists \varepsilon > 0$, so that $f(\boldsymbol{x}_0) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x}$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 < \varepsilon$. The value $f(\boldsymbol{x}_0)$ is called a **local minimum**.
- $\boldsymbol{x}_0$ is a **global minimizer** if: $f(\boldsymbol{x}_0) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$. The value is $f(\boldsymbol{x}_0)$ called the **global minimum**.

# A naive solution

**Grid search**



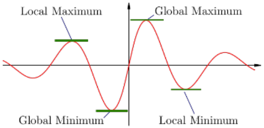

- For 1D problem, assume we know the global min lies in $[-1, 1]$
- Take uniformly grid points in $[-1, 1]$ so that any adjanent points are separated by $\varepsilon$.
- Need $O(\varepsilon^{-1})$ points to get an $\varepsilon$-close point to the global min by exhaustive search

For $N$-D problems, need $O\left(\varepsilon^{-n}\right)$ computation.

Better characterization of the local/global mins may help avoid this.

# First-order optimality condition

**Necessary condition**: Assume $f$ is 1st-order differentiable at $\boldsymbol{x}_0$. If $\boldsymbol{x}_0$ is a local minimizer, $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$.




**Intuition:** $\nabla f$ is "rate of change" of function value. If the rate is not zero at $\boldsymbol{x}_0$, possible to decrease $f$ along $-\nabla f(\boldsymbol{x}_0)$

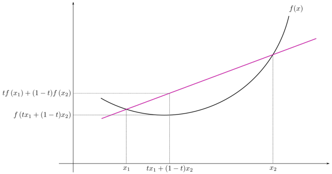
Taylor's: $f(\boldsymbol{x}_0 + \boldsymbol{\delta}) = f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle + o\left(\|\boldsymbol{\delta}\|_2\right)$. If $\boldsymbol{x}_0$ is a local min:

- For all $\boldsymbol{\delta}$ sufficiently small,
  $f(\boldsymbol{x}_0 + \boldsymbol{\delta}) - f(\boldsymbol{x}_0) = \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle + o\left(\|\boldsymbol{\delta}\|_2\right) \geq 0$
- For all $\boldsymbol{\delta}$ sufficiently small, sign of $\langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle + o\left(\|\boldsymbol{\delta}\|_2\right)$ determined by the sign of $\langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle$, i.e., $\langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle \geq 0$.
- So for all $\boldsymbol{\delta}$ sufficiently small, $\langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle \geq 0$ **and** $\langle \nabla f(\boldsymbol{x}_0), -\boldsymbol{\delta} \rangle = -\langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle \geq 0 \Longrightarrow \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle = 0$
- So $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$.

# First-order optimality condition

**Necessary condition**: Assume $f$ is 1st-order differentiable at $\boldsymbol{x}_0$. If $\boldsymbol{x}_0$ is a local minimizer, then $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$.

When sufficient? **for convex functions**




Credit: Wikipedia

- **geometric def.**: function for which any line segment connecting two points of its graph always lies above the graph
- **algebra def.**: $\forall\, \boldsymbol{x}, \boldsymbol{y}$ and $\alpha \in [0,1]$:

$$f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \le \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}).$$

Any convex function has only **one** local minimum (value!), **which is also global!**

**Proof sketch**: if $\boldsymbol{x}, \boldsymbol{z}$ are both local minimizers and $f(\boldsymbol{z}) < f(\boldsymbol{x})$,

$$f(\alpha \boldsymbol{z} + (1-\alpha)\boldsymbol{x}) \le \alpha f(\boldsymbol{z}) + (1-\alpha) f(\boldsymbol{x}) < \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{x}) = f(\boldsymbol{x}).$$

But $\alpha \boldsymbol{z} + (1-\alpha)\boldsymbol{x} \to \boldsymbol{x}$ as $\alpha \to 0$.

# First-order optimality condition

**Necessary condition**: Assume $f$ is 1st-order differentiable at $\boldsymbol{x}_0$. If $\boldsymbol{x}_0$ is a local minimizer, then $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$.

**Sufficient condition**: Assume $f$ is convex and 1st-order differentiable. If $\nabla f(\boldsymbol{x}) = \mathbf{0}$ at a point $\boldsymbol{x} = \boldsymbol{x}_0$, then $\boldsymbol{x}_0$ is a local/global minimizer.

- Convex analysis (i.e., theory) and optimization (i.e., numerical methods) are relatively mature. Recommended resources: analysis: [Hiriart-Urruty and Lemaréchal, 2001], optimization: [Boyd and Vandenberghe, 2004]
- We **don't assume** convexity unless stated, as DNN objectives are almost always nonconvex.

# Second-order optimality condition

**Necessary condition**: Assume $f(\boldsymbol{x})$ is 2-order differentiable at $\boldsymbol{x}_0$. If $\boldsymbol{x}_0$ is a local min, $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}_0) \succeq \mathbf{0}$ (i.e., positive semidefinite).

**Sufficient condition**: Assume $f(\boldsymbol{x})$ is 2-order differentiable at $\boldsymbol{x}_0$. If $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}_0) \succ \mathbf{0}$ (i.e., positive definite), $\boldsymbol{x}_0$ is a local min.

Taylor's: $f(\boldsymbol{x}_0 + \boldsymbol{\delta}) = f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle + \frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}_0) \boldsymbol{\delta} \right\rangle + o\left(\|\boldsymbol{\delta}\|_2^2\right)$.

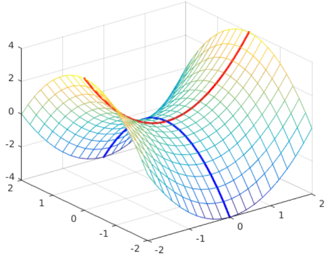
- If $\boldsymbol{x}_0$ is a local min, $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$ (1st-order condition) and $f(\boldsymbol{x}_0 + \boldsymbol{\delta}) = f(\boldsymbol{x}_0) + \frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}_0) \boldsymbol{\delta} \right\rangle + o\left(\|\boldsymbol{\delta}\|_2^2\right)$.
- So $f(\boldsymbol{x}_0 + \boldsymbol{\delta}) - f(\boldsymbol{x}_0) = \frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}_0) \boldsymbol{\delta} \right\rangle + o\left(\|\boldsymbol{\delta}\|_2^2\right) \geq 0$ for all $\boldsymbol{\delta}$ sufficiently small
- For all $\boldsymbol{\delta}$ sufficiently small, sign of $\frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}_0) \boldsymbol{\delta} \right\rangle + o\left(\|\boldsymbol{\delta}\|_2^2\right)$ determined by the sign of $\frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}_0) \boldsymbol{\delta} \right\rangle \Longrightarrow \frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}_0) \boldsymbol{\delta} \right\rangle \geq 0$
- So $\nabla^2 f(\boldsymbol{x}_0) \succeq \mathbf{0}$.

# What's in between?

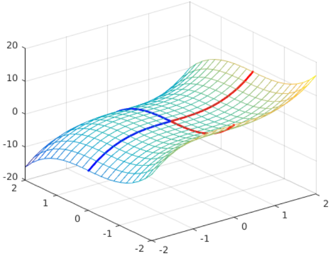
2nd order sufficient: $\nabla f\left(\boldsymbol{x}_0\right) = \mathbf{0}$ and $\nabla^2 f\left(\boldsymbol{x}_0\right) \succ \mathbf{0}$

2nd order necessary: $\nabla f\left(\boldsymbol{x}_0\right) = \mathbf{0}$ and $\nabla^2 f\left(\boldsymbol{x}_0\right) \succeq \mathbf{0}$

$f(x,y) = x^2 - y^2$




$g(x,y) = x^3 - y^3$




$$\nabla f = \begin{bmatrix} 2x \\ -2y \end{bmatrix}, \nabla^2 f = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

$$\nabla g = \begin{bmatrix} 3x^2 \\ -3y^2 \end{bmatrix}, \nabla^2 g = \begin{bmatrix} 6x & 0 \\ 0 & -6y \end{bmatrix}$$

# References i


[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). **Convex Optimization.** Cambridge University Press.

[Coleman, 2012] Coleman, R. (2012). **Calculus on Normed Vector Spaces.** Springer New York.

[Hiriart-Urruty and Lemaréchal, 2001] Hiriart-Urruty, J.-B. and Lemaréchal, C. (2001). **Fundamentals of Convex Analysis.** Springer Berlin Heidelberg.

[Kawaguchi, 2016] Kawaguchi, K. (2016). **Deep learning without poor local minima.** *arXiv:1605.07110*.

[Lampinen and Ganguli, 2018] Lampinen, A. K. and Ganguli, S. (2018). **An analytic theory of generalization dynamics and transfer learning in deep linear networks.** *arXiv:1809.10374*.

[Munkres, 1997] Munkres, J. R. (1997). **Analysis On Manifolds.** Taylor & Francis Inc.

[Zorich, 2015] Zorich, V. A. (2015). **Mathematical Analysis I.** Springer Berlin Heidelberg.