Basics of Numerical Optimization: Preliminaries

Ju Sun

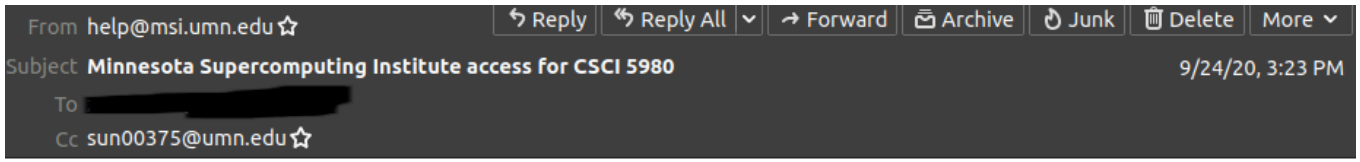
Computer Science & Engineering

University of Minnesota, Twin Cities

September 28, 2020

Logistics

- MSI account



This email is to inform you that you have received access to the Minnesota Supercomputing Institute for use in conjunction with your Fall 2020 CSCI 5980 coursework.

Before you can begin using MSI systems, you will need to review our User Agreement
by visiting the URL below and logging in with your University of Minnesota Internet ID:

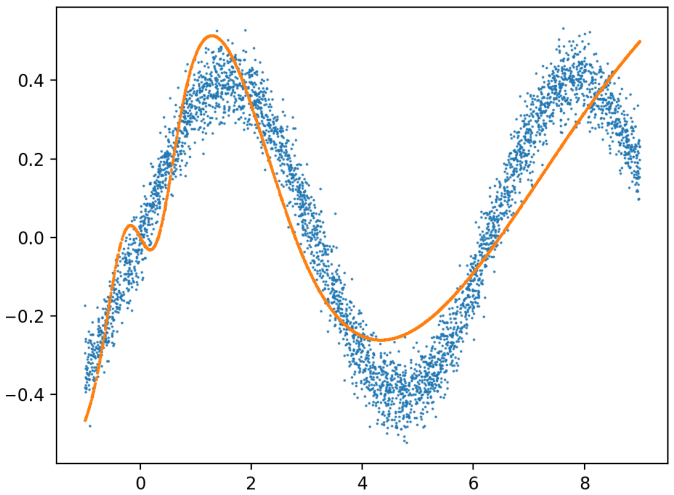
https://www.msi.umn.edu/user-agreement

The website will provide MSI login information to you after you agree to the statement.

For assistance in using your account after you have completed this process, you should work with your instructor.

- Oct 05 Zoom-only lecture: Intro to Numpy/Scipy/Colab/MSI
  + Discussion of project ideas

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

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

- **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:**

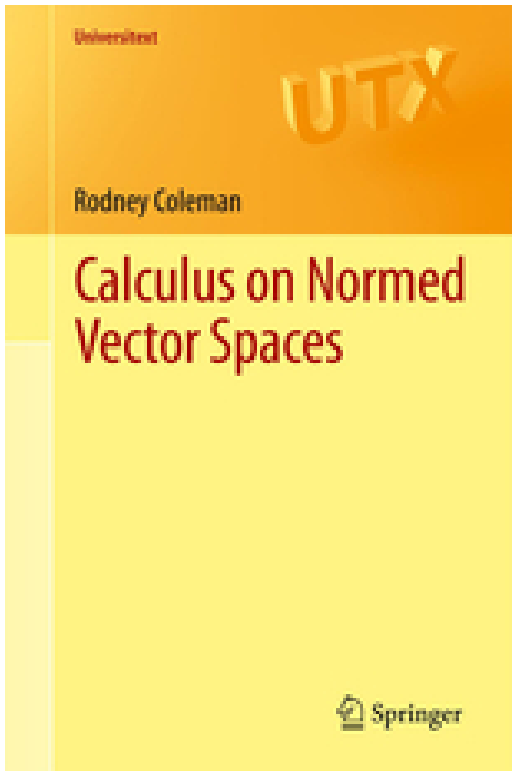
$$\min_{\boldsymbol{W}} \sum_i \ell(\boldsymbol{y}_i, \mathrm{DNN}_{\boldsymbol{W}}(\boldsymbol{x}_i)) + \Omega(\boldsymbol{W})$$

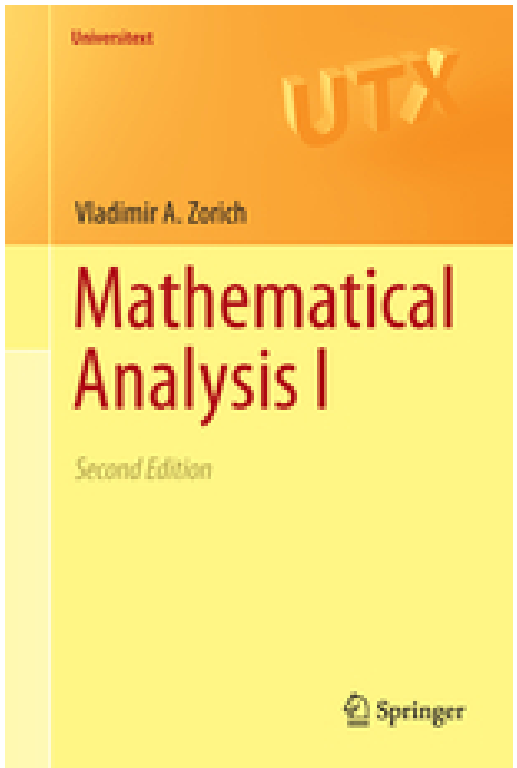
- **Generalization:** how to avoid over-complicated $\mathrm{DNN}_{\boldsymbol{W}}$ in view of UAT

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

Outline

Recommended references





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

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

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

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

- $\boldsymbol{B}$ 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 $k = 1$,

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

Put the definition in good use!

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

$$f\left(\boldsymbol{x}+\boldsymbol{\delta}\right)=f\left(\boldsymbol{x}\right)+\boldsymbol{B}\boldsymbol{\delta}+o(\|\boldsymbol{\delta}\|_2) \quad \text{as} \quad \boldsymbol{\delta} \to \mathbf{0}.$$

- prove the chain rule for $h \circ f\left(\boldsymbol{x}\right)$ (whiteboard)
- derive Jacobian (white board)
  * $f\left(\boldsymbol{x}\right)=\boldsymbol{A}\boldsymbol{x}$
  * $g\left(\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{W}_3\right)=\boldsymbol{y}-\boldsymbol{W}_1\boldsymbol{W}_2\boldsymbol{W}_3\boldsymbol{x}$

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

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

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

$$f(\boldsymbol{X}+\boldsymbol{\Delta}) = f(\boldsymbol{X}) + \langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta} \rangle + \frac{1}{2} \langle \boldsymbol{\Delta}, \nabla^2 f(\boldsymbol{X})\, \boldsymbol{\Delta} \rangle + o(\|\boldsymbol{\Delta}\|_F^2) \text{ as } \boldsymbol{\Delta} \to \mathbf{0}.$$

Put Taylor in good use!

- derive gradient and Hessian for $f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ (whiteboard)

- derive gradient (and Hessian) for

$$g(\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3) = \|\boldsymbol{y} - \boldsymbol{W}_1\boldsymbol{W}_2\boldsymbol{W}_3\boldsymbol{x}\|_F^2$$

  (whiteboard)

before: **gradient, Hessian $\Longrightarrow$ Taylor expansion**

now: **Taylor expansion $\Longrightarrow$ gradient, Hessian**

But why?

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 $P(\boldsymbol{\delta}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle + \frac{1}{2} \langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{\delta} \rangle$, i.e., the 2nd-order Taylor expansion. We can read off ∇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.

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 $D_{\boldsymbol{u}} (D_{\boldsymbol{v}} f)(\boldsymbol{x})$?

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

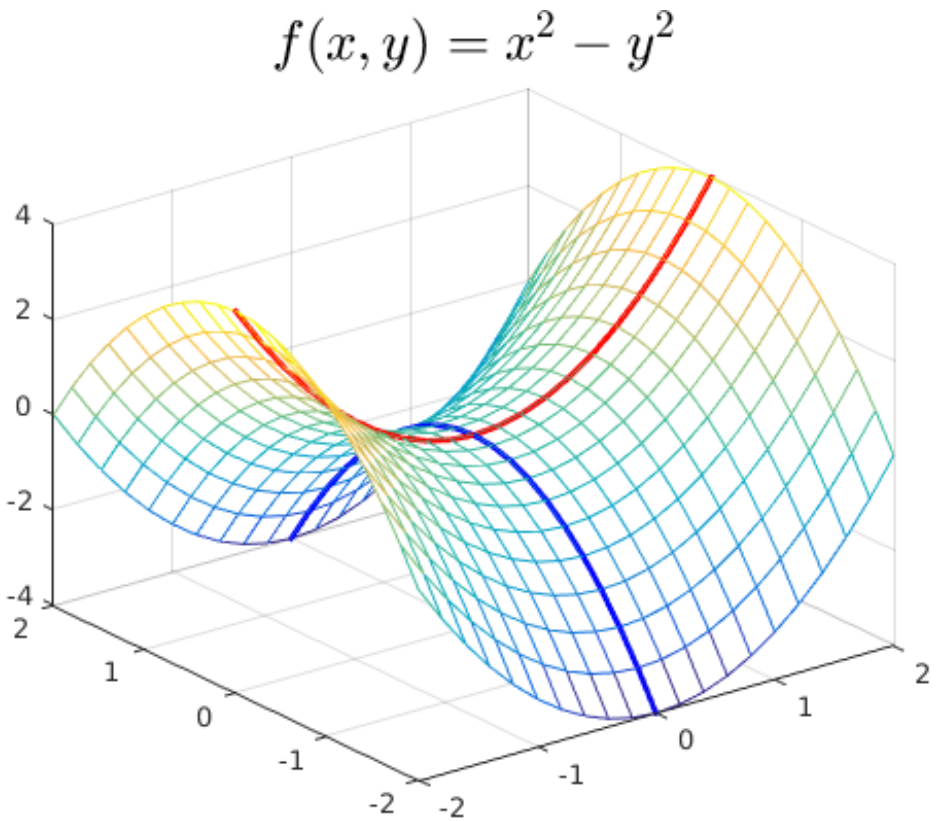
- When $\boldsymbol{u} = \boldsymbol{v}$,

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

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

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 $\boldsymbol{u}$

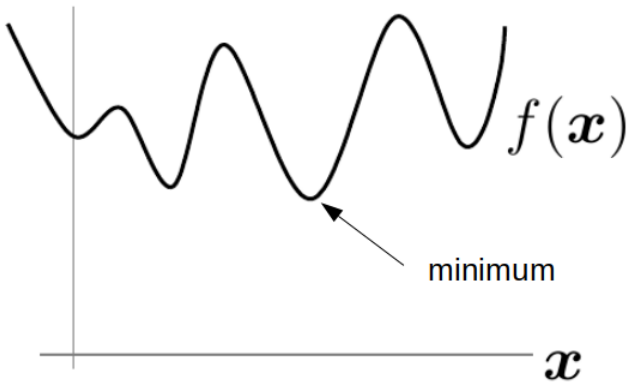


Blue: negative curvature (bending down)

Red: positive curvature (bending up)

Outline

Optimization problems



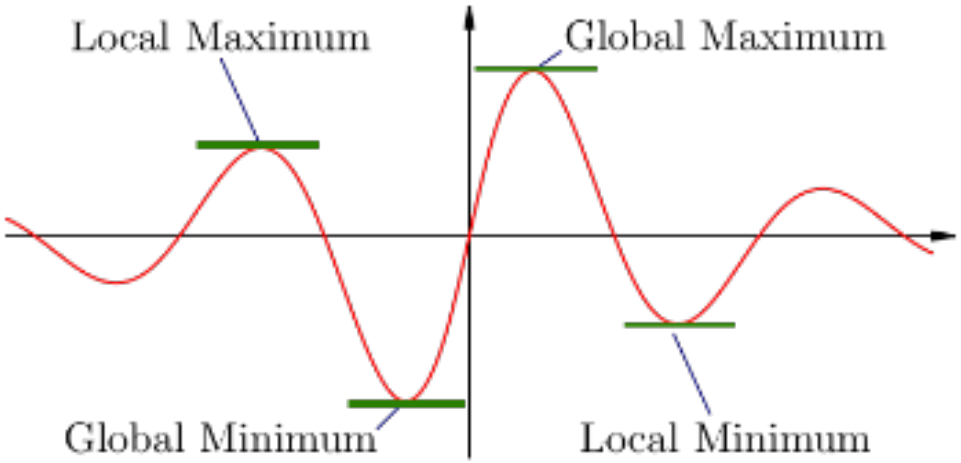
Nothing takes place in the world whose meaning is not that of some maximum or minimum. – Euler

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

- $\boldsymbol{x}$: 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

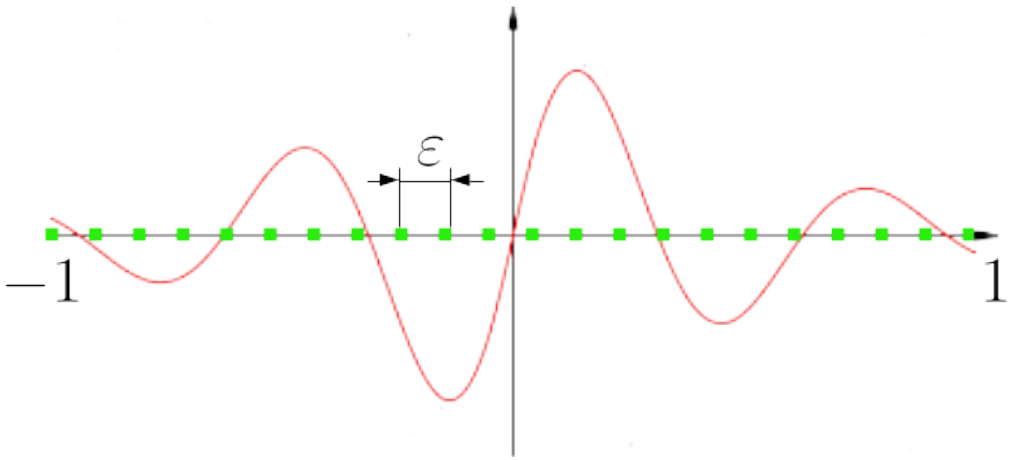
Let $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$,

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

- $\boldsymbol{x}_0$ 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



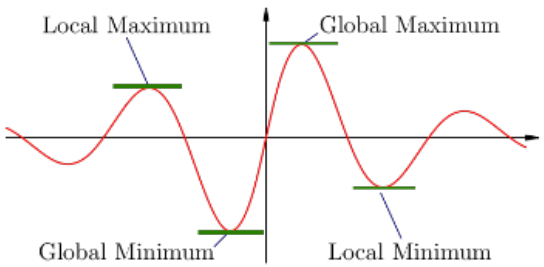
- For 1D problem, assume we know the global min lies in $[-1, 1]$
- Take uniformly grid points in $[-1, 1]$ so that any adjanent points are separated by ε.
- Need $O(\varepsilon^{-1})$ points to get an ε-close point to the global min by exhaustive search

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

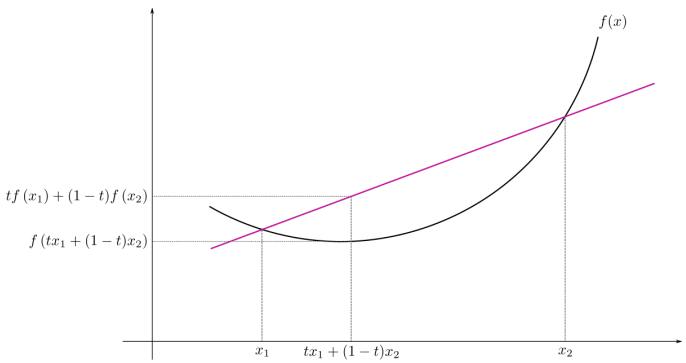
Taylor's: $f(\boldsymbol{x}_0 + \boldsymbol{\delta}) = f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle + o\left(\|\boldsymbol{\delta}\|_2\right)$. If $\boldsymbol{x}_0$ is a local min:

- 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
- **algebraic def.**: $\forall\, \boldsymbol{x}, \boldsymbol{y}$ and $\alpha \in [0,1]$:

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

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 $f(\boldsymbol{z}) < f(\boldsymbol{x})$,

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

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

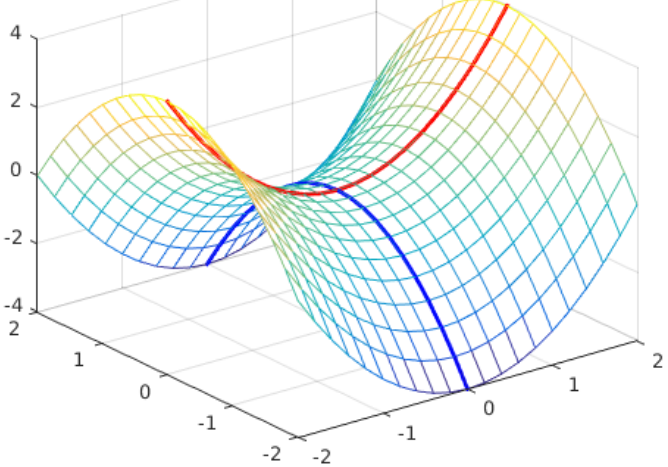
- If $\boldsymbol{x}_0$ is a local min, $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$ (1st-order condition) and $f(\boldsymbol{x}_0 + \boldsymbol{\delta}) = f(\boldsymbol{x}_0) + \frac{1}{2} \langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}_0) \boldsymbol{\delta} \rangle + o\left(\|\boldsymbol{\delta}\|_2^2\right)$.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

What's in between?

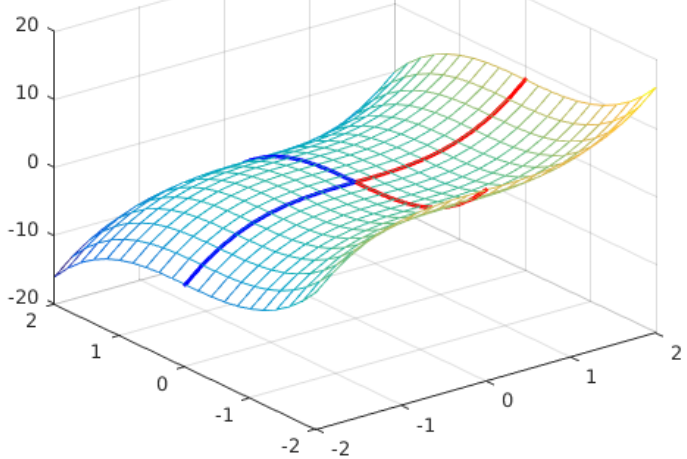
2nd order sufficient: $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}_0) \succ \mathbf{0}$

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

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



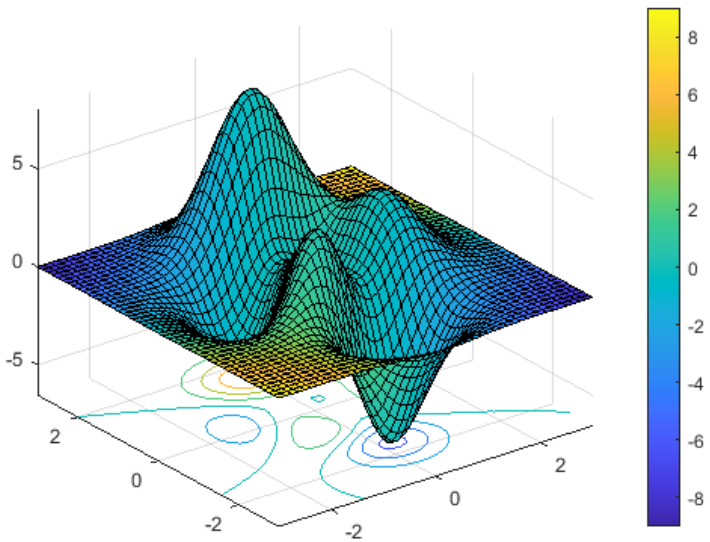
$g(x, y) = x^3 - y^3$



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

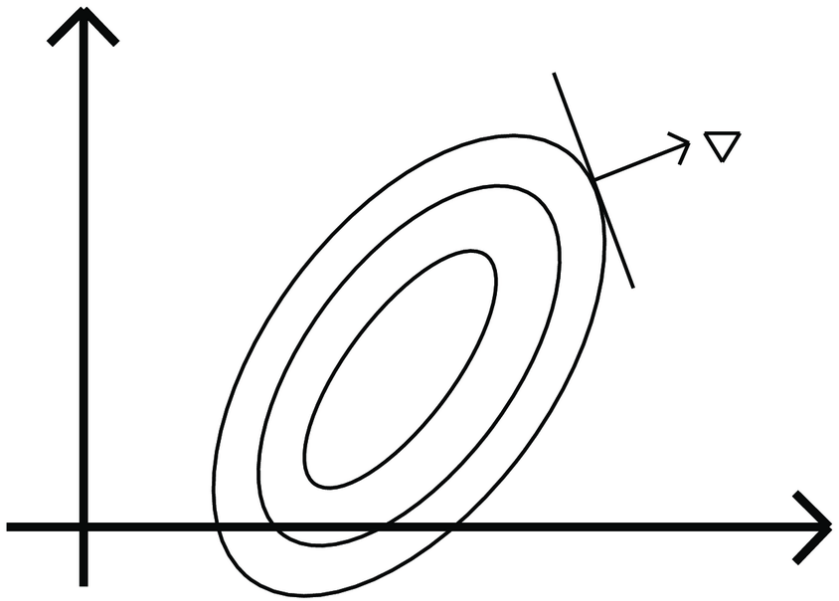
$$\nabla g = \begin{bmatrix} 3x^2 \\ -3y^2 \end{bmatrix}, \nabla^2 g = \begin{bmatrix} 6x & 0 \\ 0 & -6y \end{bmatrix}$$

Coutour plot



contour/levelset plot

(Credit: Mathworks)



gradient direction? why?

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.

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

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