Basics of Numerical Optimization: Optimality Conditions

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

September 21, 2022

Supervised learning as data fitting

Step	General view	NN view
1	Gather training set $(\boldsymbol{x}_1, \boldsymbol{y}_1)$, $\ldots$, $(\boldsymbol{x}_n, \boldsymbol{y}_n)$	Gather training set $(\boldsymbol{x}_1, \boldsymbol{y}_1)$, $\ldots$, $(\boldsymbol{x}_n, \boldsymbol{y}_n)$
2	Choose a family of functions, e.g., $\mathcal{H}$, so that there is an $f \in \mathcal{H}$ to ensure $\boldsymbol{y}_i \approx f(\boldsymbol{x}_i)$, $\forall\, i$	Choose a NN with k neurons, so that there is a group of weights $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k, b_1, \ldots, b_k)$ ensuring $\boldsymbol{y}_i \approx \{\mathrm{NN}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k, b_1, \ldots, b_k)\}(\boldsymbol{x}_i)$, $\forall i$
3	Set up a loss function ℓ	Set up a loss function ℓ
4	Find an $f \in \mathcal{H}$ to minimize the average loss $$\frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{y}_i, f(\boldsymbol{x}_i))$$	Find weights $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k, b_1, \ldots, b_k)$ to minimize the average loss $$\frac{1}{n}\sum_{i=1}^{n} \ell\left[\boldsymbol{y}_i, \{\mathrm{NN}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k, b_1, \ldots, b_k)\}(\boldsymbol{x}_i)\right]$$

Three fundamental questions in DL

- **Approximation**: is it powerful, i.e., the $\mathcal{H}$ large enough for all possible weights? (last lecture)
- **Optimization**: how to solve

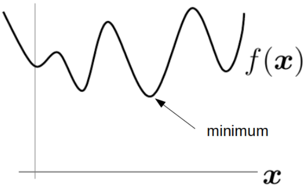
  $$\min_{\boldsymbol{w}_i' s, \boldsymbol{b}_i' s} \frac{1}{n} \sum_{i=1}^{n} \ell \left[\boldsymbol{y}_i, \left\{\mathsf{NN}\left(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k, b_1, \ldots, b_k\right)\right\}\left(\boldsymbol{x}_i\right)\right]$$

  (now)
- **Generalization**: does the learned NN work well on "similar" data? (CSCI5525, and Deep Learning Theory)

Outline

Optimality conditions of unconstrained optimization

Optimization problems



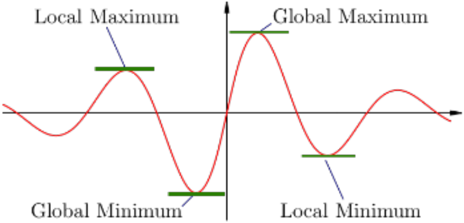
Nothing takes place in the world whose meaning is not that of some maximum or minimum. – Euler

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

- $\boldsymbol{x}$: 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

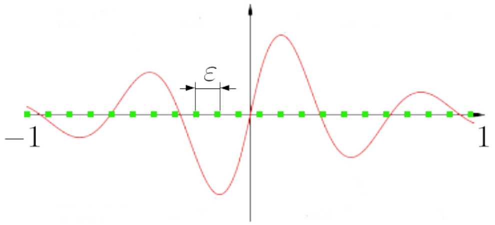
Let $f(\boldsymbol{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$,

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

- $\boldsymbol{x}_0$ 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 method for optimization

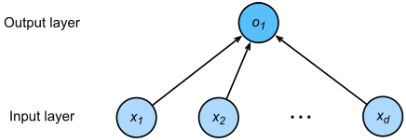
Grid search



- For 1D problem, assume we know the global min lies in $[-1, 1]$
- Take uniformly grid points in $[-1, 1]$ so that any adjacent 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.

What we do in practice



Credit: D2L

σ is the identity function

$$\min_{\boldsymbol{w}} \frac{1}{n}\sum_{i=1}^{n} \|y_i - \boldsymbol{w}^\intercal \boldsymbol{x}_i\|_2^2$$

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) \doteq \frac{1}{n}\sum_{i=1}^{n} \|y_i - \boldsymbol{w}^\intercal \boldsymbol{x}_i\|_2^2 = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \quad \text{where } \boldsymbol{X} \doteq \begin{bmatrix} \boldsymbol{x}_1^\intercal \\ \vdots \\ \boldsymbol{x}_n^\intercal \end{bmatrix}$$

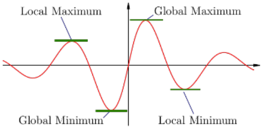
$$\implies \nabla f(\boldsymbol{w}) = \frac{2}{n}\boldsymbol{X}^\intercal(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$$

$$\nabla f(\boldsymbol{w}) = \mathbf{0} \iff \frac{2}{n}\boldsymbol{X}^\intercal(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) = \mathbf{0} \implies \boldsymbol{w} = (\boldsymbol{X}^\intercal\boldsymbol{X})^{+} + \mathrm{null}(\boldsymbol{X})$$

Optimality conditions: Reduce the search space by **characterizing** the local/global mins

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 $-\nabla f(\boldsymbol{x}_0)$

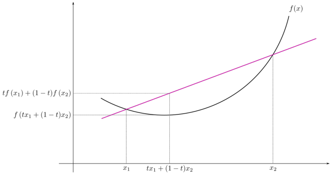
Taylor's: $f(\boldsymbol{x}_0 + \boldsymbol{\delta}) = f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle + o(\|\boldsymbol{\delta}\|_2)$. If $\boldsymbol{x}_0$ is a local min:

- For all $\boldsymbol{\delta}$ sufficiently small,
  $f(\boldsymbol{x}_0 + \boldsymbol{\delta}) - f(\boldsymbol{x}_0) = \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle + o(\|\boldsymbol{\delta}\|_2) \geq 0$
- For all $\boldsymbol{\delta}$ sufficiently small, sign of $\langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle + o(\|\boldsymbol{\delta}\|_2)$ 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 $\alpha \in [0,1]$:

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

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.

- Suppose f is twice differentiable. f is convex $\Longleftrightarrow \nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$ for all $\boldsymbol{x}$
    * Consider $f(\boldsymbol{w}) = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$ and its solutions again
    * Is it convex, $f(\boldsymbol{W}_1, \boldsymbol{W}_2) = \|\boldsymbol{y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{x}\|_2^2$?
- 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: $f(\boldsymbol{x}_0 + \boldsymbol{\delta}) = f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{\delta} \rangle + \frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}_0) \boldsymbol{\delta} \right\rangle + o\left(\|\boldsymbol{\delta}\|_2^2\right)$.

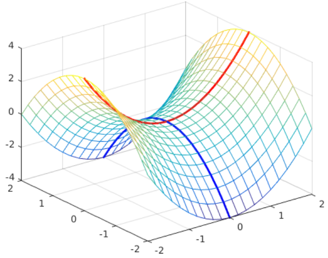
- If $\boldsymbol{x}_0$ is a local min, $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$ (1st-order condition) and $f(\boldsymbol{x}_0 + \boldsymbol{\delta}) = f(\boldsymbol{x}_0) + \frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}_0) \boldsymbol{\delta} \right\rangle + o\left(\|\boldsymbol{\delta}\|_2^2\right)$.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

What's in between?

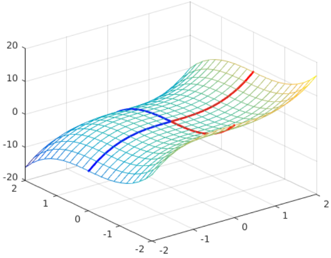
2nd order sufficient: $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}_0) \succ \mathbf{0}$

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

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



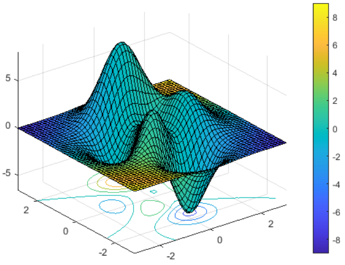
$g(x, y) = x^3 - y^3$



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

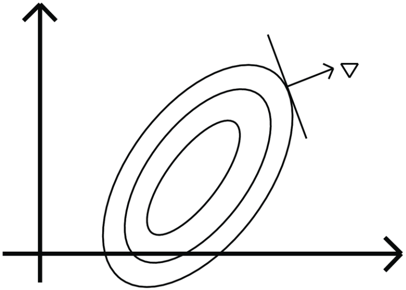
$$\nabla g = \begin{bmatrix} 3x^2 \\ -3y^2 \end{bmatrix}, \nabla^2 g = \begin{bmatrix} 6x & 0 \\ 0 & -6y \end{bmatrix}$$

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.

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