Distributed optimization in Machine learning

School of Electrical and Computer Engineering
University of Tehran
Erfan Darzi

Lecture 1 – Local Optimality, Optimality Conditions, and Convex Optimization

erfandarzi@ut.ac.ir

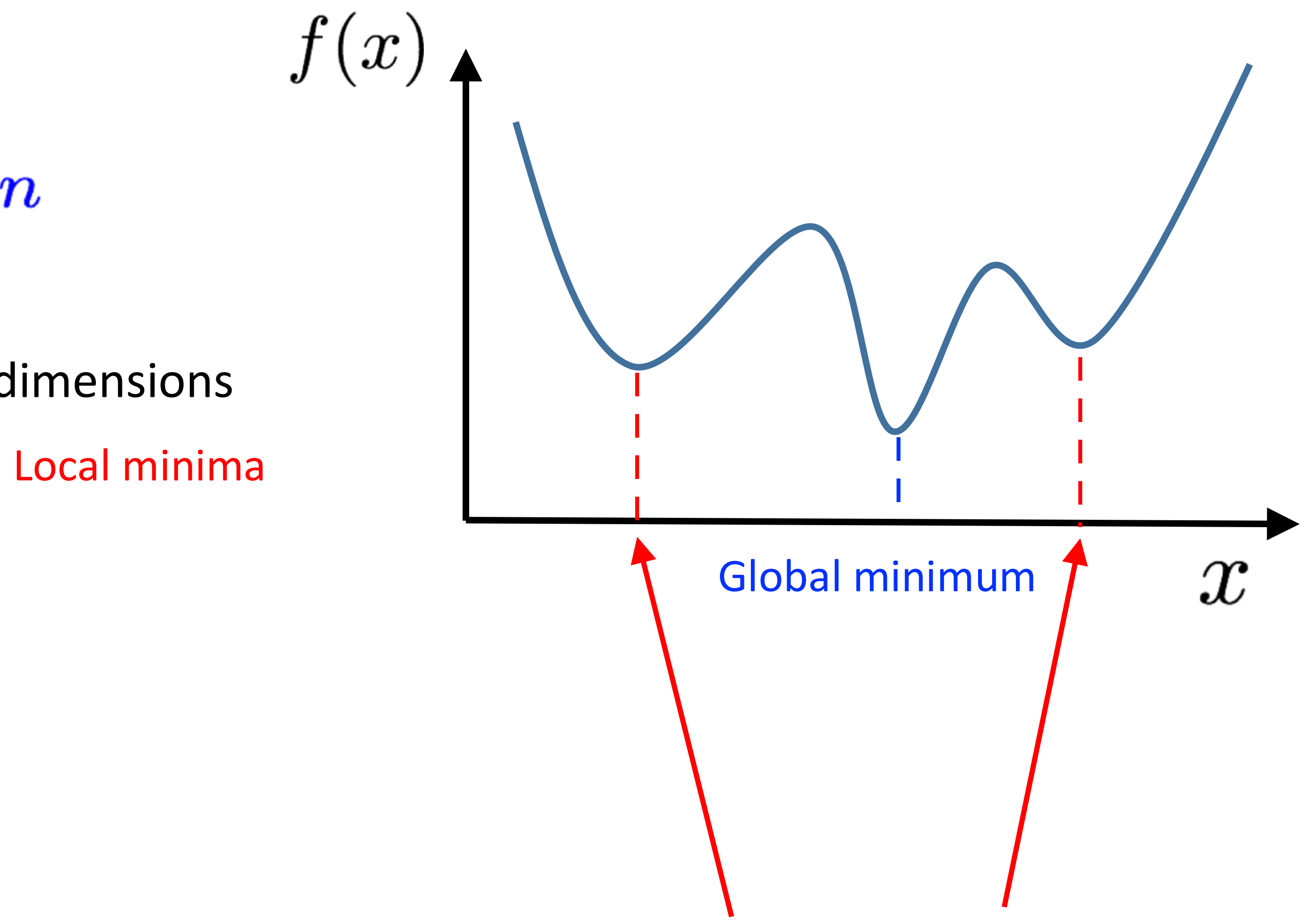
Finding a solution to an optimization problem?

- Start with unconstrained case:

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$
$$\text{s.t.} \quad \mathbf{x} \in \mathbb{R}^n$$

- Grid search?

- Requires exponential time for higher dimensions



Unconstrained Optimization

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$
$$\text{s.t.} \quad \mathbf{x} \in \mathbb{R}^n$$

- Global minimum $f(\mathbf{x}) \geq f(\mathbf{x}^*), \quad \text{for all } \mathbf{x} \in \mathbb{R}^n$
- Strict global minimum $f(\mathbf{x}) > f(\mathbf{x}^*), \quad \text{for all } \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{x}^*$
- Local minimum $\exists \epsilon > 0 \text{ s.t. } f(\mathbf{x}) \geq f(\mathbf{x}^*), \quad \text{for all } \mathbf{x} \text{ with } \|\mathbf{x} - \mathbf{x}^*\| \leq \epsilon$
- Strict local minimum
  $\exists \epsilon > 0 \text{ s.t. } f(\mathbf{x}) > f(\mathbf{x}^*), \quad \text{for all } \mathbf{x} \text{ with } \|\mathbf{x} - \mathbf{x}^*\| \leq \epsilon$

Geometrical Interpretation!

Examples: $f(x) = x^3, \ |x|^3, \ -|x|^3$

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$
$$\text{s.t.} \quad \mathbf{x} \in \mathbb{R}^n$$

Optimality Conditions

- Given a point $\mathbf{x}$, how to determine if it is a (strict) local/global minimum?
- Assume twice continuous differentiability of the objective function
- Necessary optimality condition **Is it sufficient?**

$$\nabla f(\mathbf{x}^*) = \mathbf{0}, \quad \nabla^2 f(\mathbf{x}^*) \succeq \mathbf{0}$$

- Sufficient optimality condition for local optimality **Proofs?**

$$\nabla f(\mathbf{x}^*) = \mathbf{0}, \quad \nabla^2 f(\mathbf{x}^*) \succ \mathbf{0}$$

- For above unconstrained optimization, $\bar{\mathbf{x}}$ is stationary if $\nabla f(\bar{\mathbf{x}}) = 0$. Figure

Why do we care?

- Optimality conditions are useful because

- Tractable conditions for optimality (Yes/No)

- Narrowing down the list of potential solutions

- Useful in the **design** and the **analysis** of algorithms

Existence of An Optimal Solution

- local/global minimum/maximum of $f(x) = x^2 - x^4 \quad \inf_{x \in \mathbb{R}} f(x)?$ $\quad f(x) = e^{-|x|} \quad \min_{x \in \mathbb{R}} f(x)?$

- **Bolzano-Weierstrass Theorem:** Every continuous function *f* attains its infimum over a compact set $\mathcal{X}$. In other words, $\exists\ \mathbf{x}^* \in \mathcal{X}$ s.t. $f(\mathbf{x}^*) = \inf_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$

- Consequently, if the level set $\{\mathbf{x} \in \mathbb{R}^n | f(\mathbf{x}) \leq f(\mathbf{x}^\mathbf{0})\}$ is compact, then the global min is attained

- Another sufficient condition (**coercivity**): $f(\mathbf{x}) \to \infty$ as $\|\mathbf{x}\| \to \infty$

- Coercivity + Continuity → existence of global optimal solution(s)

Unconstrained Quadratic Optimization

$$\min_{\mathbf{x}} \quad \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{b}^T\mathbf{x}$$
$$\text{s.t.} \quad \mathbf{x} \in \mathbb{R}^n$$

Q is symmeteric.

$$\mathbf{Q}\mathbf{x} + \mathbf{b} = \mathbf{0}, \quad \mathbf{Q} \succeq \mathbf{0}$$

What if not PSD?

What if not feasible?

- Necessary conditions:
- Sufficient conditions?

- **Claim**:
- The above necessary condition is also sufficient

- Any local optimum is also globally optimum (True for any convex optimization)

Convexity

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if

< : Strictly convex

- $$f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}), \quad \forall\, \alpha \in [0,1], \quad x, y$$

- A set is convex if its indicator function is convex
- Equivalent definition

Convex function

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$

- Convex optimization problem: Convex Set

$$\text{s.t.} \quad \mathbf{x} \in \mathcal{X}$$

- **Why do we care? Local optimality (or stationarity) ==> Global optimality**

Proof?

- For continuously differentiable functions Not the only class though!

$$f \;\; \text{is convex} \Leftrightarrow f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}), \;\; \forall \mathbf{x},\; \mathbf{y}$$

- For twice continuously differentiable functions

$$f \text{ is convex} \Leftrightarrow \nabla^2 f(\mathbf{x}) \succeq \mathbf{0}, \quad \forall \mathbf{x}$$

Convexity

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$
$$\text{s.t.} \quad \mathbf{x} \in \mathcal{X}$$

- **Remarks:**
- Set of optimal solutions of a convex optimization problem is convex • If the objective function is strictly convex, then the minimizer is unique

- Example:

$$\min_{x,y} \quad \frac{1}{2}(\alpha x^2 + \beta y^2) - x$$
$$\text{s.t.} \quad (x, y) \in \mathbb{R}^2$$

- Convex/Non-convex? Local/global minimum?

$\alpha, \beta > 0$ $\quad$ $\alpha = 0$ $\quad$ $\alpha > 0,\ \beta = 0$ $\quad$ $\alpha > 0,\ \beta < 0$

Linear Regression and Linear Least Squares

$$\min_{\mathbf{w}} \quad \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 \qquad \text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \qquad \text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

- Relation to quadratic optimization
- Necessary and sufficient optimality condition: $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$
- **Remarks**:
- Always has a solution
- Might have unbounded level sets

Linear Regression

Area	Crime Rate	Age	RAD	PTRATIO	Bedrooms	…	Price (K)
600	1.05	12	2.4	10.1	1	…	500
1000	2.34	10	2.5	20.1	1	…	800
1200	1.45	3	3.1	9.7	3	…	1500
1500	1.56	30	1.7	7.2	2	…	1200
…	…	…	…	…	…	…	…
2700	1.01	20	0.9	4.3	4	…	5000

$\mathbf{x}_1$ y_1

$\mathbf{x}_n$ $\mathbf{x}_i \in \mathbb{R}^d$ y_n

Model: Linear predictor
Loss: L2 difference

$$\min_{\mathbf{w}} \quad \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2$$
$$\text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

What if we want a more interpretable predictor?

Linear Regression

$$\min_{\mathbf{w}} \quad \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 \quad \text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

Imposing sparsity →

$$\min_{\mathbf{w}} \quad \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 + \lambda\|\mathbf{w}\|_0 \quad \text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

Non-convex and even not continuous

Suggestion
Tibshirani, Donoho, …

$$\min_{\mathbf{w}} \quad \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 + \lambda\|\mathbf{w}\|_1 \quad \text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

Non-smooth optimization problem

Convex and can be solved efficiently!