Distributed optimization for Machine Learning

School of Electrical and Computer Engineering
University of Tehran

Erfan Darzi

Lecture 0 - Background

erfandarzi@ut.ac.ir

What is this course about?

- Useful optimization tools for machine learning
  -
- This is NOT a machine learning course
- Don’t expect to learn detailed ML

- This is NOT a classical optimization course
- We won’t cover many classical optimization results

- We cover some basics though
- Few weeks on optimization
- Some ML examples will be explained in details

Prerequisites

- Probability and Statistics
- Expected value, variance, statistical independence, conditional probability, maximum likelihood estimation, regression, etc.

- Linear Algebra and Mathematical Analysis
- Sets, functions, limits, liminf, limsup, derivative, gradient, subspace, linear dependence, inner product, eigenvalue, singular value, norms, etc.

- Programming skills
- Matrix/vector operations in Matlab/Python/C++
- “For, while, repeat until” loops

What is optimization?

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

Decision variable (Discrete/Continuous)

Objective/Cost function

- Existence of a solution? Feasible Region
- Checking if a candidate *x* is optimal?

Why do we care?

- Many engineering problems requires optimization

- In this course, we focus mostly on machine learning applications

Example: Regression

Samples/Data points

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

Features/independent Variables

Target/Dependent Variables/Label

Can we use this dataset to predict the price of this house?

1400	2.2	3	3.1	7.6	2	…	????

Example: 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

Training Data

Can we use this dataset to predict the price of this house?

1400	2.2	3	3.1	7.6	2	…	????

Example: 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

Training Data

Learning Algorithm

Prediction

1400	2.2	3	3.1	7.6	2	…	????

Learning Algorithms

- Various methods in ML
- Decision trees, deep learning, Bayes, empirical Bayes, linear regression, logistic regression, …

- Many methods

- Model

- Minimize the loss/Maximize the likelihood

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

Can we use this dataset to predict the price of this house?

1400	2.2	3	3.1	7.6	2	…	????

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

$\mathbf{x}_1$ $\quad y_1$

$\mathbf{x}_n$ $\quad y_n$

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

Model: Linear predictor
Loss: L2 difference

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{n} \|\mathbf{w}^T\mathbf{x}_i - y_i\|_2^2$$

$$\text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

Another Example: Classification

Radius	Texture	Area	Compactness	Symmetry	…	Rec/non-Rec
1.1	2.3	3.5	2.4	1.4	…	1
0.7	1.2	2.5	1.4	3.2	…	0
1.7	2.4	1.5	3.3	1.3	…	1
…	…	…	…	…	…	…
0.2	3.4	0.7	4.3	2.0	…	1
0.2	2.7	0.9	2.3	1.0	…	????

Logistic Regression

Radius	Texture	Area	Compactness	Symmetry	…	Rec/non-Rec
1.1	2.3	3.5	2.4	1.4	…	1
0.7	1.2	2.5	1.4	3.2	…	0
1.7	2.4	1.5	3.3	1.3	…	1
…	…	…	…	…	…	…
0.2	3.4	0.7	4.3	2.0	…	1

$\mathbf{x}_1$ $\mathbf{x}_n$ y_1 y_n

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

Model: logistic
Maximum likelihood estimator

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{n} \log\left(1+\exp\left(\mathbf{w}^T\mathbf{x}\right)\right) - \sum_{\{i:y_i=1\}} \mathbf{w}^T\mathbf{x}_i$$
$$\text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

Optimization in ML

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{n} \|\mathbf{w}^T \mathbf{x}_i - y_i\|_2^2$$
$$\text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{n} \log\left(1 + \exp\left(\mathbf{w}^T \mathbf{x}\right)\right) - \sum_{\{i:y_i=1\}} \mathbf{w}^T \mathbf{x}_i$$
$$\text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

- Many more examples (K-means, SVM, Deep learning, …)
- Efficient algorithms: CPU, Memory requirements, Parallelizable, robustness, etc.
- Other issues: Non-convexity, Sparsity, Large values of *n/d*, Online implementation, Implicit bias, Privacy concerns, Overfitting, etc.

- But first, we need to review a little bit of optimization (targeted review!)
- Even before this, let's review a bit of linear algebra and mathematical analysis

Notations

- **Sets**
  - $\mathcal{X}, \quad x \in \mathcal{X}, \quad \mathcal{X}_1 \cap \mathcal{X}_2, \quad \mathcal{X}_1 \cup \mathcal{X}_2$
- Real numbers $\mathbb{R}$, Complex numbers $\mathbb{C}$

- **Inf and Sup**
- Supremum of the set $\mathcal{X}$ is the smallest scalar y such that $y \geq x, \text{ for all } x \in \mathcal{X}$
- Infimum of the set $\mathcal{X}$ is the largest scalar y such that $y \leq x, \text{ for all } x \in \mathcal{X}$

$$\sup \mathcal{X} \in \mathcal{X} \Rightarrow \max \mathcal{X} \triangleq \sup \mathcal{X} \qquad \inf \mathcal{X} \in \mathcal{X} \Rightarrow \min \mathcal{X} \triangleq \inf \mathcal{X}$$

$\sup\{1/n : n \geq 1\} = ?$ $\quad$ $\inf\{1/n : n \geq 1\} = ?$

$\max\{1/n : n \geq 1\} = ?$ $\quad$ $\min\{1/n : n \geq 1\} = ?$

$\inf\{\sin n :$

$\sup\{\sin n : n \geq 1\} = ?$

Functions:

$$f : \mathcal{X} \mapsto \mathcal{Y}, \ \mathcal{X} \text{ is called the domain}, \ \mathcal{Y} \text{ is called the range}$$

Vectors and Subspaces

- **Linear combination:**

$$\mathbf{x}, \mathbf{y} \in \mathbb{R}^m, \quad \text{linear combination of } \mathbf{x} \text{ and } \mathbf{y} : \alpha\mathbf{x} + \beta\mathbf{y}$$

- **Subspace and linear independence**
- A set is called subspace if it is closed under linear combination
- A set of vectors is called linearly independent if no linear combination of them is equal to zero
- Inner product: $\langle \mathbf{x}, \mathbf{y} \rangle$
- Orthogonality: $\mathbf{x} \perp \mathbf{y} \text{ if } \langle \mathbf{x}, \mathbf{y} \rangle = 0$

- **Cauchy-Schwarz inequality**

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2$$

Matrices

- Matrix addition
- Matrix product
- Square matrix
- Inner product: $$\langle \mathbf{A}, \mathbf{B} \rangle \triangleq \mathrm{Tr}(\mathbf{A}\mathbf{B}^T) = \sum_{i,j} A_{ij} B_{ij}$$
- Spectral radius: $$\rho(\mathbf{A}) \triangleq \max_i \{|\lambda_i| : \lambda_i \text{ is an eigenvalue of } \mathbf{A}\}$$
- Eigenvalue decomposition of real symmetric matrices
- Positive (Semi-)definite matrices

Matrices

- Singular values: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0 : \ \sigma_i^2$ is an eigenvalue of $\mathbf{A}\mathbf{A}^T$
- Singular value decomposition of $\mathbf{A} \in \mathbb{R}^{n \times n}$:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \text{ with } \mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I} \text{ and } \boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_n)$$

$$\|\mathbf{A}\|_F = \left(\sum_{i,j} |A_{ij}|^2\right)^{1/2} = (\sum_i \sigma_i^2)^{1/2}$$

- Nuclear norm: $\|\mathbf{A}\|_* = \sum_i \sigma_i$

$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2} = \max_i \ \sigma_i$$

- Useful inequalities: $\|\mathbf{Ax}\|_2 \leq \|\mathbf{A}\|_2 \cdot \|\mathbf{x}\|_2 \qquad \|\mathbf{A}\|_* \geq \|\mathbf{A}\|_F \geq \|\mathbf{A}\|_2 \geq \rho(\mathbf{A})$
- Norms:
- Frobenius norm:

- Matrix 2-norm:

$$\langle \mathbf{A}, \mathbf{B} \rangle \leq \|\mathbf{A}\|_F \cdot \|\mathbf{B}\|_F$$

Big Oh notations

- Which one grows faster? Linear or quadratic?
- How to compare the limiting behavior of functions?

- When $x \to \infty$:

$$f(x) = O(g(x)) \quad \text{if} \quad \exists \alpha, x_0 > 0 \;\text{ s.t. }\; |f(x)| \leq \alpha |g(x)|, \;\; \forall x > x_0$$

$$f(x) = \Omega(g(x)) \quad \text{if} \quad \exists \alpha, x_0 > 0 \;\text{ s.t. }\; |f(x)| \geq \alpha |g(x)|, \;\; \forall x > x_0$$

$$f(x) = o(g(x)) \quad \text{if} \quad \forall \alpha > 0, \; \exists x_0 > 0 \;\text{ s.t. }\; |f(x)| \leq \alpha |g(x)|, \;\; \forall x > x_0$$

$$f(x) = \omega(g(x)) \quad \text{if} \quad \forall \alpha > 0, \; \exists x_0 > 0 \;\text{ s.t. }\; |f(x)| \geq \alpha |g(x)|, \;\; \forall x > x_0$$

We can also define it for $x \to a$

Examples

$$4x^4 + 3x^2 + 2 = O(x^5)???$$

$$10\sin(x) = O(1)???$$

$$4x^4 + 3x^2 + 2 = O(x^4)???$$

$$10\sin(x) = \Omega(1)???$$

$$4x^4 + 3x^2 + 2 = O(x^3)???$$

$$10\sin(x) = \Omega(x)???$$

when $x \to 0$

$$4x^4 + 3x^2 = O(x^2)???$$

$$4x^4 + 3x^2 = O(x)???$$

$$4x^4 + 3x^2 = \Omega(x^2)???$$

$$4x^4 + 3x^2 = \Omega(x)???$$

Derivatives

- Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a twice continuously differentiable function

- Derivative: $\frac{\partial f(\mathbf{x})}{\partial x_i} \triangleq \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t}$

- Gradient: $\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right)^T$

- Hessian Matrix: $\nabla^2 f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_i \partial x_j} \right]$

- Taylor Expansion:

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T\nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + o(\|\mathbf{y}-\mathbf{x}\|^2)$$

Mean Value Theorem

- There exists $\boldsymbol{\xi}, \boldsymbol{\eta}$ in the line segment connecting $\mathbf{x}$ and $\mathbf{y}$ such that

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\boldsymbol{\xi})^T(\mathbf{y} - \mathbf{x})$$

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\boldsymbol{\eta})(\mathbf{y} - \mathbf{x})$$

Chain Rule

Jacobian Matrix for $f : \mathbb{R}^n \mapsto \mathbb{R}^m$

$$\nabla f(\mathbf{x}) = [\nabla f_1(\mathbf{x}), \nabla f_2(\mathbf{x}), \ldots, \nabla f_m(\mathbf{x})]$$

Chain Rule: $f : \mathbb{R}^k \mapsto \mathbb{R}^m \qquad g : \mathbb{R}^m \mapsto \mathbb{R}^n h(\mathbf{x}) \triangleq g(f(\mathbf{x}))$

$$\nabla h(\mathbf{x}) = \nabla f(\mathbf{x}) \nabla g(f(\mathbf{x}))$$

Examples: $\nabla \left(f(\mathbf{Ax}) \right) = ? \qquad \nabla^2 \left(f(\mathbf{Ax}) \right) = ?$

Contraction Mappings

Lipschitz Continuity: $f : \mathbb{R}^n \mapsto \mathbb{R}^m$

Lipschitz constant

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y}$$

$$\gamma \leq 1 \quad \Rightarrow \quad \text{non} - \text{expansive mapping}$$

$$\gamma < 1 \quad \Rightarrow \quad \text{contraction mapping}$$

Theorem: For a contraction mapping $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ the iterated function sequence converges to a unique fixed point., i.e.,

$$\mathbf{x}, f(\mathbf{x}), f(f(\mathbf{x})), \ldots \to \mathbf{x}^* \quad \text{with} \quad \mathbf{x}^* = f(\mathbf{x}^*)$$

True for non-expansive mappings???

Probability

- Probability, Conditional probability, Random Variable, Independence
- Normal/Gaussian distribution

$$X \sim \mathcal{N}(\mu, \sigma^2), \quad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Jointly multivariate Normal distribution

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- Expected value, Variance, Covariance (Matrix)

$$X, Y \text{ independent } \Rightarrow \text{Cov}(X, Y) = 0$$

Converse?

Probability

- For $\mathbf{X}_1, \mathbf{X}_2, \ldots$ i.i.d. with $\mathbb{E}[\mathbf{X}_i] = \boldsymbol{\mu},\ \mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$ and $\mathbf{S}_n \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$

Law of large numbers $\mathbf{S}_n \to \boldsymbol{\mu}$

Central Limit Theorem $\sqrt{n}\,(\mathbf{S}_n - \boldsymbol{\mu}) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$

Markov's inequality: For RV $X \geq 0$, $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$

Chebyshev's inequality: For RV X with $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$, $\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$

Cauchy-Schwarz inequality: $|\mathbb{E}(XY)|^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$