Optimization Theory and Algorithm	**Lecture 4 - 05/07/2021**

Lecture 4

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

Optimization needs **Iterative Algorithms.** Why???? Let us recall the normal equation, and the solution $\mathbf{x}^* = (A^\top A)^{-1}A^\top \mathbf{b}$. Generally, the computational complexity of $(A^\top A)^{-1} \in \mathbb{R}^{n^2}$ is $O(n^3)$. why??????

The iterative algorithm usually has the following general form in Algorithm 1.

Algorithm 1 General Form of Iterative Algorithm

1: **Input:** Something you need
2: **Initialization:** a starting point $\mathbf{x}_0$, and step index $t = 0$
3: **while** a stop condition false **do**
4:
$$\mathbf{x}_{t+1} := \textbf{Iterative Algorithm}(\mathbf{x}_t),$$
and
$$t := t + 1.$$
5: **end while**
6: **Output:** The sequence $\{\mathbf{x}_t\}_{t=0}^T$.

Then we hope that $\lim_{t\to\infty} \mathbf{x}_t = \mathbf{x}^*$.

Example 1 *(Solving the Normal Equation)*

Denote that $\tilde{A} = A^\top A$ and $\tilde{\mathbf{b}} = A^\top \mathbf{b}$, then normal equation becomes that $\tilde{A}\mathbf{x} = \tilde{\mathbf{b}}$. How to compute it efficiently?

- *Jacobi Iterative Algorithm: Let $\tilde{A} = B + D$, where $D = diag(\tilde{A})$ and $B = \tilde{A} - D$. Then the normal equation is $(D + B)\mathbf{x} = \tilde{\mathbf{b}}$. Thus, $D\mathbf{x} = -B\mathbf{x} + \tilde{\mathbf{b}}$. Finally,*
$$\mathbf{x} = -D^{-1}B\mathbf{x} + D^{-1}\tilde{\mathbf{b}}. \tag{1}$$
Based on Eq.(1), Jacobi iterative algorithm is designed via
$$\mathbf{x}_{t+1} = -D^{-1}B\mathbf{x}_t + D^{-1}\tilde{\mathbf{b}}, \tag{2}$$
and the scalar form is
$$x_{t+1,i} = \frac{\tilde{b}_i - \sum_{j=1,j\neq i}^n x_{t,j}\tilde{a}_{ij}}{\tilde{a}_{ii}},$$
where we suppose that $\tilde{a}_{ii} \neq 0$ for all $i = 1, \ldots, n$.
Insights: *If $\lim_{t\to\infty} \mathbf{x}_t = \mathbf{x}^*$, then $\lim_{t\to\infty} \mathbf{x}_{t+1} = -D^{-1}B \lim_{t\to\infty} \mathbf{x}_t + D^{-1}\tilde{\mathbf{b}}$. Thus, $\mathbf{x}^* = -D^{-1}B\mathbf{x}^* + D^{-1}\tilde{\mathbf{b}}$. This indicates $\mathbf{x}^*$ satisfies the normal equation.*

- *Gauss-Seidel Algorithm: Let $\tilde{A} = L + U + D$, where $D = diag(\tilde{A})$, L is the Lower triangular matrix of $\tilde{A}$ and U is the upper triangular matrix of $\tilde{A}$. Then the normal equation is $(D + L + U)\mathbf{x} = \tilde{\mathbf{b}}$. Thus, $D\mathbf{x} = -L\mathbf{x} - U\mathbf{x} + \tilde{\mathbf{b}}$. Finally,*
$$\mathbf{x} = -D^{-1}L\mathbf{x} - D^{-1}U\mathbf{x} + D^{-1}\tilde{\mathbf{b}}. \tag{3}$$
Based on Eq.(3), Gauss-seidel iterative algorithm is designed via
$$\mathbf{x}_{t+1} = -D^{-1}L\mathbf{x}_{t+1} - D^{-1}U\mathbf{x}_t + D^{-1}\tilde{\mathbf{b}}, \tag{4}$$

and the scalar form is

$$x_{t+1,i} = \frac{\tilde{b}_i - \sum_{j=1}^{i-1} \tilde{a}_{ij} x_{t+1,j} - \sum_{j=i+1}^{n} \tilde{a}_{ij} x_{t,j}}{\tilde{a}_{ii}},$$

where we suppose that $\tilde{a}_{ii} \neq 0$ *for all* $i = 1, \ldots, n$.

Insights: *If* $\lim_{t\to\infty} \mathbf{x}_t = \mathbf{x}^*$, *then* $\lim_{t\to\infty} \mathbf{x}_{t+1} = -D^{-1}L \lim_{t\to\infty} \mathbf{x}_{t+1} - D^{-1}U \lim_{t\to\infty} \mathbf{x}_t + D^{-1}\tilde{\mathbf{b}}$. *Thus,* $\mathbf{x}^* = -D^{-1}L\mathbf{x}^* - D^{-1}U\mathbf{x}^* + D^{-1}\tilde{\mathbf{b}}$. *This indicates* $\mathbf{x}^*$ *satisfies the normal equation.*

The procedure of obtaining the iterative solution can be seen as an algorithm for solving the linear least squares problem.

Remark 1 *Algorithms in optimization can be commonly summarized as three types, but it's not limited to these.*

- *Closed Form Solution.*
- *Iterative Algorithm, see Algorithm 1.*
- *Heuristic Algorithms (e.g., genetic algorithm), which will not be covered by the course.*

1 Related Theory in Optimization

"Nothing is more practical than a good theory."– by V. Vapnik [Vapnik, 1998].

What kind of theory we have to learn in Optimization?

- Theory can support you to construct models. You have see them in many examples (e.g., MLE).
- Theory can help you develop algorithms. For example, convex analysis, KTT conditions, duality theory, optimally conditions, and among others.
- Theory can implicitly show the convergence property of the optimization algorithms. Convergence theory is to show that under what conditions the sequences $\{\mathbf{x}_t\}_{t=1}^{\infty}$ and $\{f(\mathbf{x}_t)\}_{t=1}^{\infty}$ satisfy

$$\lim_{t\to\infty} \mathbf{x}_t = \mathbf{x}^* \text{ and } \lim_{t\to\infty} f(\mathbf{x}_t) = f^* = f(\mathbf{x}^*).$$

Convergence Rate:

- linear convergence:
$$\frac{\|\mathbf{x}_{t+1} - \mathbf{x}^*\|}{\|\mathbf{x}_t - \mathbf{x}^*\|} \leq a,$$
where $a \in (0, 1)$.
- Super-linear convergence:
$$\lim_{t\to\infty} \frac{\|\mathbf{x}_{t+1} - \mathbf{x}^*\|}{\|\mathbf{x}_t - \mathbf{x}^*\|} = 0.$$
- sub-linear convergence:
$$\lim_{t\to\infty} \frac{\|\mathbf{x}_{t+1} - \mathbf{x}^*\|}{\|\mathbf{x}_t - \mathbf{x}^*\|} = 1.$$

- Others theoretical bounds:
$$\|\mathbf{x}_t - \mathbf{x}^*\| \le O(t, Q),$$
and
$$\|f(\mathbf{x}_t) - f^*\| \le O(t, Q),$$
where Q includes some constants related to the original optimization problem.

We justify the convergence theory of Jacobi and Gauss-Seidel algorithms for demonstrating an concrete example.

Theorem 1 *Suppose that we have the linear equation with form $\mathbf{x} = B\mathbf{x} + C$, then we can develop an iterative algorithm*
$$\mathbf{x}_{t+1} = B\mathbf{x}_t + C. \tag{5}$$
For any initial point $\mathbf{x}_0$, the generated sequence $\{\mathbf{x}_t\}_{t=0}^{\infty}$ converges at $\mathbf{x}^*$ if and only if $\rho(B) := \|B\|_2 = \sigma_{\max}(B) < 1$, where $\sigma_{\max}(B)$ is the biggest singular value of B and $\rho(B)$ is so-called spectral radius of B.

Proof 1 **Necessary:** *If $\lim\limits_{t\to\infty} \mathbf{x}_t = x^*$, then according to the iterative procedure (5), we have that*
$$\lim_{t\to\infty} \mathbf{x}_{t+1} = \mathbf{x}^* = B \lim_{t\to\infty} \mathbf{x}_t + C = B\mathbf{x}^* + C.$$
Thus, $\mathbf{x}^*$ is the solution of the original equation.

Sufficient: *we know that $\mathbf{x}^*$ satisfies $\mathbf{x}^* = B\mathbf{x}^* + C$, then*
$$\mathbf{x}_{t+1} - \mathbf{x}^* = B(\mathbf{x}_t - \mathbf{x}^*) = B^2(\mathbf{x}_{t-1} - \mathbf{x}^*) = \cdots = B^t(\mathbf{x}_0 - \mathbf{x}^*).$$
Thus, when $\rho(B) < 1$, then
$$\frac{\|\mathbf{x}_{t+1} - \mathbf{x}^*\|}{\|\mathbf{x}_t - \mathbf{x}^*\|} \le \rho(B) < 1,$$
this indicates the linear convergence property. In addition,
$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 = \|B^t(\mathbf{x}_0 - \mathbf{x}^*)\|_2 \le \|B\|_2^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2 = \rho^t(B)\|\mathbf{x}_0 - \mathbf{x}^*\|_2 := O(t, Q) \to 0.$$

2 Part 2: Quick Review of Linear Algebra

In this section, we will give a brief and quick review of the linear algebra that will be used in this course.

2.1 Row and Column Picture

Let us consider a set of *Simultaneous Equations*.
$$2x - y = 0,$$
$$2y - x = 3,$$
which is equivalent to
$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0.$$
The solution of these equations are the intersection point of lines $y = 2x$ and $y = \frac{1}{2}x + 3$. The lines $y = 2x$ and $y = \frac{1}{2}x + 3$ are called *row pictures* of the equations. Draw them by yourself.

These equations could be reformulated as

$$\begin{bmatrix} 2 \\ -1 \end{bmatrix} x + \begin{bmatrix} -1 \\ 2 \end{bmatrix} y = \begin{bmatrix} 0 \\ 3 \end{bmatrix}.$$

Actually, $\begin{bmatrix} 2 \\ -1 \end{bmatrix} x + \begin{bmatrix} -1 \\ 2 \end{bmatrix} y$ is the linear combination of the vectors $\begin{bmatrix} 2 \\ -1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 2 \end{bmatrix}$. Then, the *column picture* of these equations are $\mathcal{A} = \{\mathbf{z} : \mathbf{z} = \begin{bmatrix} 2 \\ -1 \end{bmatrix} x + \begin{bmatrix} -1 \\ 2 \end{bmatrix} y; x, y \in \mathbb{R}\}$.

Q: what is $\mathcal{A}$?

We similarly consider the three dimensional case and discuss the solution of the simultaneous equations.

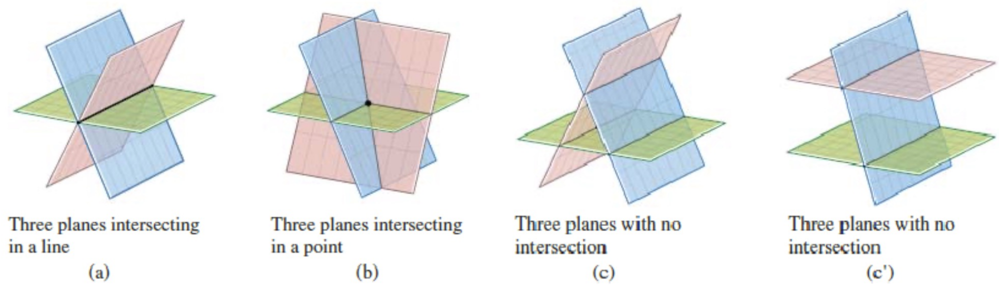

Figure 1: 3D Case

From the row pictures (see Figure 1), Figure 1(a) has infinity solutions; Figure 1(b) has an unique solution; Figure 1(c) and (c') has no solutions.

Let us consider the column picture. The equations $A\mathbf{x} = \mathbf{b}$ have solutions for any $\mathbf{b}$ if and only if the linear combination of column vectors of A can cover the 3-dimensional space $\mathbb{R}^3$.

2.2 Matrix Multiplication

Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, then $C = AB \in \mathbb{R}^{m\times p}$.

- Standard Form: $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}, i = 1, \ldots, m, j = 1, \ldots, p$.

- Column Operation: Let $A = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n), B = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p)$, where $\mathbf{a}_i \in \mathbb{R}^m$ and $\mathbf{b}_j \in \mathbb{R}^n$ are the column vector of A and B respectively. Then

  $$A\mathbf{b}_j = b_{1j} \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + b_{2j} \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + b_{nj} \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix},$$

  that is

  $$A\mathbf{b}_j = \sum_{i=1}^{n} b_{ij}\mathbf{a}_i.$$

  Thus,

  $$AB = (A\mathbf{b}_1, A\mathbf{b}_2, \ldots, A\mathbf{b}_p).$$

- Row Operation: Let $A = (\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_m)^\top = \begin{bmatrix} \tilde{a}_1^\top \\ \tilde{a}_2^\top \\ \vdots \\ \tilde{a}_m^\top \end{bmatrix}$, where $\tilde{a}_i = (a_{11}, a_{12}, \ldots, a_{1m})^\top$ is the ith row vector of A. Then
$$AB = \begin{bmatrix} \tilde{a}_1^\top \\ \tilde{a}_2^\top \\ \vdots \\ \tilde{a}_m^\top \end{bmatrix} B = \begin{bmatrix} \tilde{a}_1^\top B \\ \tilde{a}_2^\top B \\ \vdots \\ \tilde{a}_m^\top B \end{bmatrix}.$$

- Out Product:
$$AB = \sum_{i=1}^{n} \mathbf{a}_i \tilde{b}_i^\top, \tag{6}$$
where $\mathbf{a}_i$ is the ith column of A, $\tilde{b}_i$ is the ith row of B and $\mathbf{a}_i \tilde{b}_i^\top \in \mathbb{R}^{m \times p}, i = 1, \ldots, n$ are rank-1 matrices.

- Block Multiplication:
$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix},$$
where $C_{11} = A_{11}B_{11} + A_{12}B_{21},\ A_{11} \in \mathbb{R}^{m_1 \times n_1}, B_{11} \in \mathbb{R}^{n_1 \times p_1}, A_{12} \in \mathbb{R}^{m_1 \times n_2}, B_{21} \in \mathbb{R}^{n_2 \times p_1}$. Thus, $C_{11} \in \mathbb{R}^{m_1 \times p_1}, m_1 + m_2 = m, n_1 + n_2 = n, p_1 + p_2 = p$.

References

[Vapnik, 1998] Vapnik, V. (1998). *Statistical learning theory.* John Wiley, New York.