Minimizing $\|\mathbf{Ax} - \mathbf{b}\|$ by Landweber Iteration

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft : August 23, 2017
Last update : October 28, 2019

Least square

Problem setting : given $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, find $\mathbf{x} \in \mathbb{R}^n$ such that

$$f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

is minimized.

- If A is a square matrix $(m = n)$ and non-singular, the sol. is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

What about the general case $m \neq n$?

The minimizer of $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$ is $\mathbf{x}=(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b}$

$$
\begin{aligned}
f(\mathbf{x}) &= \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \\
&\overset{\text{expand}}{=} \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{b}^\top\mathbf{A}\mathbf{x} + \mathbf{b}^\top\mathbf{b} \\
\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} &= 2\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{A}^\top\mathbf{b}
\end{aligned}
$$

Set $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = 0$, we have the minimizer of $f(\mathbf{x})$ as

$$\mathbf{x} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b}.$$

i.e., $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$ can be minimized by solving a linear system of equations

$$\mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{A}^\top\mathbf{b},$$

which is called the **normal equation**.

The problems of solving $\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{b}$

Solving the normal equation becomes problematic if :

- $\mathbf{A}^\top \mathbf{A}$ is not invertible
- $\mathbf{A}^\top \mathbf{A}$ is ill-conditioned : the sol. will be numerically unstable
- n is big (so that $\mathbf{A}^\top \mathbf{A}$ is big) : the computational cost (memory and time) of inverting $A^T A$ is too high.

In these cases, we have to bypass the normal equation by using iterative approaches.

To apply MM algorithm to minimize $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, find a majorizer $G_k(\mathbf{x})$ of $f(\mathbf{x})$ that satisfies the following conditions at all iteration k :

- $G_k(\mathbf{x}_k) = f(\mathbf{x}_k)$
- $G_k(\mathbf{x}) \geq f(\mathbf{x})$ for all $\mathbf{x}$
- $G_k(\mathbf{x})$ should be “easier” to be minimized

Conceptually, $G_k(\mathbf{x})$ can be $f(\mathbf{x})$ plus a non-negative term. e.g. :

$$G_k(\mathbf{x}) = f(\mathbf{x}) + (\mathbf{x} - \mathbf{x}_k)^\top(\alpha\mathbf{I} - \mathbf{A}^\top\mathbf{A})(\mathbf{x} - \mathbf{x}_k)$$

when $\mathbf{x} = \mathbf{x}_k$, $G_k(\mathbf{x}_k) = f(\mathbf{x}_k)$.

The majorizer $G_k(\mathbf{x}) = f(\mathbf{x}) + (\mathbf{x} - \mathbf{x}_k)^\top(\alpha\mathbf{I} - \mathbf{A}^\top\mathbf{A})(\mathbf{x} - \mathbf{x}_k)$

To satisfy the condition $G_k(\mathbf{x}) \geq f(\mathbf{x})$:

$$
\begin{aligned}
& (\mathbf{x} - \mathbf{x}_k)^\top(\alpha\mathbf{I} - \mathbf{A}^\top\mathbf{A})(\mathbf{x} - \mathbf{x}_k) \text{ has to be non-negative} \\
\iff \quad & \mathbf{y}\top(\alpha\mathbf{I} - \mathbf{A}^\top\mathbf{A})\mathbf{y} \geq 0 \text{ for all } \mathbf{y} \\
\iff \quad & \alpha\mathbf{I} - \mathbf{A}^\top\mathbf{A} \text{ is positive semi-definite} \\
\iff \quad & \alpha \geq \text{ largest eigenvalue of } \mathbf{A}^\top\mathbf{A}
\end{aligned}
$$

We can set $\alpha = \lambda_{\mathsf{max}}(\mathbf{A}^\top\mathbf{A})$.

As G is quadratic, minimizer of G can be found by solving $\dfrac{\partial G}{\partial \mathbf{x}} = 0$. The solution is unique.

Minimizer of G_k and the Landweber iteration

$$
\begin{aligned}
G_k(\mathbf{x}) &= f(\mathbf{x}) + (\mathbf{x}-\mathbf{x}_k)^\top(\alpha\mathbf{I}-\mathbf{A}^\top\mathbf{A})(\mathbf{x}-\mathbf{x}_k) \\
&= \begin{array}{l} \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{b}^\top\mathbf{A}\mathbf{x} + c \\ +\mathbf{x}^\top(\alpha\mathbf{I}-\mathbf{A}^\top\mathbf{A})\mathbf{x} - \mathbf{x}^\top(\alpha\mathbf{I}-\mathbf{A}^\top\mathbf{A})\mathbf{x}_k - \mathbf{x}_k^\top(\alpha\mathbf{I}-\mathbf{A}^\top\mathbf{A})\mathbf{x} + c \end{array} \\
&= \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{b}^\top\mathbf{A}\mathbf{x} + \mathbf{x}^\top(\alpha\mathbf{I}-\mathbf{A}^\top\mathbf{A})(\mathbf{x}-2\mathbf{x}_k) + c
\end{aligned}
$$

Take derivative, set to zero

$$
2\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{A}^\top\mathbf{b} + (\alpha\mathbf{I}-\mathbf{A}^\top\mathbf{A})(2\mathbf{x}-2\mathbf{x}_k) = 0
$$

$$
\begin{aligned}
\iff \quad & -\mathbf{A}^\top\mathbf{b} + \alpha\mathbf{x} - (\alpha\mathbf{I}-\mathbf{A}^\top\mathbf{A})\mathbf{x}_k = 0 \\
\iff \quad & \mathbf{x} = \frac{1}{\alpha}\mathbf{A}^\top\mathbf{b} + \frac{1}{\alpha}(\alpha\mathbf{I}-\mathbf{A}^\top\mathbf{A})\mathbf{x}_k
\end{aligned}
$$

$$
\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{1}{\alpha}\mathbf{A}^\top(\mathbf{b}-\mathbf{A}\mathbf{x}_k)
$$

By the convergence properties of MM algorithm, the Landweber iteration guarantees the value of $f(\mathbf{x}_k)$ decreases in each iteration.
The Landweber iteration bypasses the process of inverting $\mathbf{A}^\top\mathbf{A}$. It only requires multiplying $\mathbf{A}$ by $\mathbf{A}^\top$.

Landweber iteration is a special case of Gradient Descent

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k + \frac{1}{\alpha}\mathbf{A}^\top(\mathbf{b} - \mathbf{A}\mathbf{x}_k) \\ &= \mathbf{x}_k - \frac{1}{\alpha}\mathbf{A}^\top(\mathbf{A}\mathbf{x}_k - \mathbf{b}) \\ &= \mathbf{x}_k - \frac{1}{\alpha}\nabla f(\mathbf{x}_k) \\ &= \mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\end{aligned}$$

Hence Landweber iteration = Gradient descent with constant step size

$$t = \frac{1}{\alpha} = \frac{1}{\lambda_{\text{max}}(\mathbf{A}^\top\mathbf{A})} = \frac{1}{\sigma^2_{\text{max}}(\mathbf{A})} = \frac{1}{\|\mathbf{A}\|_2^2} = \frac{1}{L}$$

where $\|\mathbf{A}\|_2^2$ is exactly the Lipschitz constant L of $\nabla f(\mathbf{x}) = \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$ (i.e. $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is L-smooth.)

As Landweber iteration is a special case of gradient descent, all the convergence properties of gradient descent apply.

Last page

- Least square $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$
- If $\mathbf{A}$ is square and non-singular : $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$
- If $\mathbf{A}$ is non-square and $\mathbf{A}^\top\mathbf{A}$ is not ill-conditioned : $\mathbf{x} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}\mathbf{b}$
- If $\mathbf{A}$ is non-square and $\mathbf{A}^\top\mathbf{A}$ is ill-conditioned or has big size : Landweber iteration $\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{1}{\lambda_{\mathsf{max}}(\mathbf{A}^\top\mathbf{A})}\mathbf{A}^\top(\mathbf{b} - \mathbf{A}\mathbf{x}_k)$
- Landweber iteration is just a special case of gradient descent

End of document