1. LSE:

For convenience, consider two dimenional data (x, y).

Ideally, we would expect that $A\vec{w} = \vec{b}$:

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ & & \vdots & & \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix}}_{\substack{n \times (m+1) \\ A}} \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix}}_{\substack{(m+1) \times 1 \\ \vec{w}}} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\substack{n \times 1 \\ \vec{b}}}$$

Then $\vec{w}$ can be calculated by $\vec{w} = A^{-1}\vec{b}$.

However, A is not guaranteed to be invertible.

In fact, A doesn't even need to be a square matrix.

Therefore, we change our goal from finding $\vec{w}$ such that $A\vec{w} = \vec{b}$ to minimizing $\|A\vec{w} - \vec{b}\|^2$.

$$\begin{aligned} \text{Then } \|A\vec{w} - \vec{b}\|^2 &= (A\vec{w} - \vec{b})^T (A\vec{w} - \vec{b}) \\ &= (\vec{w}^T A^T - \vec{b}^T)(A\vec{w} - \vec{b}) \\ &= \vec{w}^T A^T A \vec{w} - \underline{\vec{w}^T A^T \vec{b}} - \underline{\vec{b}^T A \vec{w}} + \vec{b}^T \vec{b} \end{aligned}$$

$\because \vec{w}^T A^T \vec{b}$ is a scalar

$\therefore (\vec{w}^T A^T \vec{b})^T = \vec{b}^T A \vec{w} = \vec{w}^T A^T \vec{b}$

$$= \vec{w}^T A^T A \vec{w} - 2\vec{w}^T A^T \vec{b} + \vec{b}^T \vec{b}.$$

$$\frac{\partial(\vec{w}^T A^T A \vec{w})}{\partial \vec{w}} = \begin{bmatrix} \frac{\partial}{\partial w_0} \\ \frac{\partial}{\partial w_1} \\ \vdots \\ \frac{\partial}{\partial w_m} \end{bmatrix} \begin{bmatrix} w_0 & w_1 & \cdots & w_m \end{bmatrix} \begin{bmatrix} a_{0,0} & a_{0,1} & \cdots & a_{0,m} \\ a_{1,0} & a_{1,1} & \cdots & a_{1,m} \\ & \vdots & & \\ a_{m,0} & a_{m,1} & \cdots & a_{m,m} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial w_0} \\ \frac{\partial}{\partial w_1} \\ \vdots \\ \frac{\partial}{\partial w_m} \end{bmatrix} \underbrace{\begin{bmatrix} w_0(a_{0,0}w_0 + a_{0,1}w_1 + \cdots + a_{0,m}w_m) \\ + w_1(a_{1,0}w_0 + a_{1,1}w_1 + \cdots + a_{1,m}w_m) \\ + \cdots \\ + w_m(a_{m,0}w_0 + a_{m,1}w_1 + \cdots + a_{m,m}w_m) \end{bmatrix}}_{1 \times 1}$$

$$= \underbrace{\begin{bmatrix} (a_{0,0}w_0 + a_{0,1}w_1 + \cdots + a_{0,m}w_m) + (w_0 a_{0,0} + w_1 a_{1,0} + \cdots + w_m a_{m,0}) \\ (a_{1,0}w_0 + a_{1,1}w_1 + \cdots + a_{1,m}w_m) + (w_0 a_{0,1} + w_1 a_{1,1} + \cdots + w_m a_{m,1}) \\ \vdots \\ (a_{m,0}w_0 + a_{m,1}w_1 + \cdots + a_{m,m}w_m) + (w_0 a_{0,m} + w_1 a_{1,m} + \cdots + w_m a_{m,m}) \end{bmatrix}}_{(m+1) \times 1}$$

$$= \underline{(A^T A)}\vec{w} + \underline{(A^T A)^T}\vec{w}$$

$$(A^T A)^T = A^T (A^T)^T = A^T A$$

$$= 2(A^T A)\vec{w}$$

Similarly, $\frac{\partial(\vec{w}^T A^T \vec{b})}{\partial \vec{w}} = \begin{bmatrix} \frac{\partial}{\partial w_0} \\ \frac{\partial}{\partial w_1} \\ \vdots \\ \frac{\partial}{\partial w_m} \end{bmatrix} \begin{bmatrix} w_0 & w_1 & \cdots & w_m \end{bmatrix} (A^T \vec{b})$

$$= A^T \vec{b}$$

Hence, define $f(\vec{w}) = \vec{w}^T A^T A \vec{w} - 2\vec{w}^T A^T \vec{b} + \vec{b}^T \vec{b}$

$$\frac{\partial f}{\partial \vec{w}} = 2A^T A \vec{w} - 2A^T \vec{b}.$$

$$\frac{\partial f}{\partial \vec{w}} = 0 \Rightarrow \vec{w} = (A^T A)^{-1} A^T \vec{b}. \quad (*)$$

Hence $f(\vec{w})$ attains its minimum at $\vec{w} = (A^T A)^{-1} A^T \vec{b}$.

Note that we only have $\det(A^T A) \geq 0$, it's still possible that $(A^T A)$ is not invertible.

Hence we can add the regularized term λI $(\lambda > 0)$ to $(A^T A)$, and then $\det(A^T A + \lambda I) > 0$

$$\Rightarrow (A^T A + \lambda I) \text{ is invertible.}$$

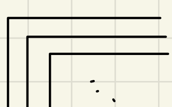
$(*)$ becomes $\vec{w} = (A^T A + \lambda I)^{-1} A^T \vec{b}$.

This is also called rLSE.

Next we'll illustrate on LU decomposition and how to apply it to finding the inverse of a given matrix.

$$\underset{n\times n}{A} = \underset{n\times n}{L}\,\underset{n\times n}{U} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{2,1} & 1 & 0 & \cdots & 0 \\ l_{3,1} & l_{3,2} & 1 & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ l_{n,1} & l_{n,2} & \cdots & l_{n,n-1} & 1 \end{bmatrix} \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ 0 & u_{2,2} & \cdots & u_{2,n} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & & \\ 0 & 0 & \cdots\ 0 & u_{n,n} \end{bmatrix}$$

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & & \ddots & \\ \vdots & & & \\ a_{n,1} & & & a_{n,n} \end{bmatrix}$$



$$\Longrightarrow\ a_{1,1} = u_{1,1},\ a_{1,2} = u_{1,2} \ \cdots\ a_{1,n} = u_{1,n}$$

$$a_{2,1} = l_{2,1}\cdot u_{1,1},\ a_{3,1} = l_{3,1}\cdot u_{1,1} \ \cdots\ a_{n,1} = l_{n,1}\cdot u_{1,1}$$

$$\Longrightarrow\ l_{2,1} = \frac{a_{2,1}}{u_{1,1}},\ l_{3,1} = \frac{a_{3,1}}{u_{1,1}},\ \cdots\ l_{n,1} = \frac{a_{n,1}}{u_{1,1}}$$

$$a_{2,2} = l_{2,1}\cdot u_{1,2} + u_{2,2},\ a_{2,3} = l_{2,1}\cdot u_{1,3} + u_{2,3} \ \cdots$$

$$\Longrightarrow\ u_{2,2} = a_{2,2} - l_{2,1}\cdot u_{1,2},\ u_{2,3} = a_{2,3} - l_{2,1}\cdot u_{1,3} \ \cdots$$

$$a_{3,2} = l_{3,1}\cdot u_{1,2} + l_{3,2}\cdot u_{2,2} \Longrightarrow l_{3,2} = \frac{a_{3,2} - l_{3,1}\cdot u_{1,2}}{u_{2,2}}$$

$$\vdots$$

Suppose that we already have $\underset{n\times n}{A} = LU$.

Define $A^{-1} = [x_1\ x_2\ \cdots\ x_n]$ where $x_1, x_2, \dots$ are column vectors.

Then $A(A^{-1}) = [Ax_1\ Ax_2\ \cdots\ Ax_n] = I = [e_1\ e_2\ \cdots\ e_n]$

$$\Rightarrow LUx_1 = e_1,\ LUx_2 = e_2\ \cdots\ LUx_n = e_n.$$

We solve for $\begin{cases} Ux_1 = y_1, & Ly_1 = e_1 \\ Ux_2 = y_2, & Ly_2 = e_2 \\ \quad\vdots & \\ Ux_n = y_n, & Ly_n = e_n \end{cases}$

Then A^{-1} can be calculated ※

2. Steepest descent method:

The formula of steepest descent, aka gradient descent method, can be written as:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

where f is the loss function.

e.g. MSE

Assume that f is Lipschitz continuous with constant $L > 0$.

Then $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|$ for any x, y.

We can perform a quadratic expansion of f around $f(x_t)$ and obtain the following inequality:

$$\begin{aligned}
f(x_{t+1}) &\le f(x_t) + \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{1}{2} \nabla^2 f(x_t) \|x_{t+1} - x_t\|^2 \\
&\le f(x_t) + \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{1}{2} L \|x_{t+1} - x_t\|^2 \\
&= f(x_t) - \eta \|\nabla f(x_t)\|^2 + \frac{1}{2} L \eta^2 \|\nabla f(x_t)\|^2 \\
&= f(x_t) - (1 - \frac{1}{2} L \eta) \eta \|\nabla f(x_t)\|^2
\end{aligned}$$

Note that $\|x_{t+1} - x_t\|$ has to be small enough, which implies that η also has to be small enough.

Choose $\eta \le \frac{1}{L}$, then $-(1-\frac{1}{2}L\eta) = \frac{1}{2}L\eta - 1$

$$\le \frac{1}{2}L\left(\frac{1}{L}\right) - 1$$

$$= -\frac{1}{2}.$$

Then $f(x_{t+1}) \le f(x_t) - \underline{\frac{1}{2}\eta\|\nabla f(x_t)\|^2}$ (**) $\rightarrow$ positive unless $\nabla f(x) = 0$

Thus the sequence $\{f(x_0), f(x_1), \dots\}$ is indeed decreasing.

Assume that f is convex and $f(x)$ attains its minimum at $x = x^*$, then we have

$$f(x^*) \ge f(x) + \nabla f(x)^T(x^* - x) \quad \text{for any } x$$

$$\Leftrightarrow f(x) \le f(x^*) + \nabla f(x)^T(x - x^*).$$

Then (**) becomes $f(x_{t+1}) \le f(x^*) + \nabla f(x_t)^T(x_t - x^*)$

$$- \frac{1}{2}\eta\|\nabla f(x_t)\|^2$$

$$\Rightarrow f(x_{t+1}) - f(x^*) \le \frac{1}{2\eta}\left(2\eta\nabla f(x_t)^T(x_t - x^*) - \eta^2\|\nabla f(x_t)\|^2\right)$$

$$= \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_t - \eta\nabla f(x_t) - x^*\|^2\right)$$

$$= \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right)$$

This implies that the sequence $\{f(x_0), f(x_1) \dots\}$ is bounded.

↰ (I skip the complicated proof of convergence :( )

In the end, consider $g(\vec{w}) = \|\vec{b} - A\vec{w}\|^2$, aka the LSE loss, in (1).

Hence $\frac{\partial g}{\partial \vec{w}} = 2A^TA\vec{w} - 2A^T\vec{b}$ ⟹ gradient

$= 2A^T(A\vec{w} - \vec{b})$

On the other hand, for the regularized term in L_1-norm, the gradient of it can be written as the sign function.

$$\text{sign}(w_i) = \begin{cases} 1, & \text{if } w_i > 0 \\ -1, & \text{if } w_i < 0 \\ 0, & \text{if } w_i = 0. \end{cases}$$

Thus the gradient in total is $2A^T(A\vec{w} - \vec{b}) + \lambda \cdot \text{sign}(\vec{w})$.

3. Newton's method:

We have the following equation (from Taylor expansion)

$$f(x) \approx f(x_0) + f'(x_0)(x-x_0) + O((x-x_0)^2)$$

If we want to find x such that $f(x)=0$, and $|x-x_0|$ is small enough, then

$$0 = f(x) \approx f(x_0) + f'(x_0)(x-x_0)$$

$$\Rightarrow x \approx x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Then we derive the formula of Newton's method:

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$$

it'll converge to the root of the original equation, and the error is up to your tolerance.

If we want to apply Newton's method to an optimization problem, we may need to solve $f'(x)=0$:

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)}$$

$$\Rightarrow x_{t+1} = x_t - (\nabla^2 f(x_t))^{-1} \nabla f(x_t) \text{ in multi-dimensional case.}$$

From (1) LSE and (2) Steepest descent method, we know that $\nabla f(\vec{w}) = 2A^T(A\vec{w} - \vec{b})$ for f: square error.

$\nabla^2 f(\vec{w}) = 2A^TA$. ※