Introduction to Least Squares

Least Squares methods are crucial in chemical sciences because they provide a robust approach for data fitting, error minimization, and parameter estimation in experimental and computational studies. In fields like spectroscopy, thermodynamics, and reaction kinetics, researchers often collect noisy data, and Least Squares helps identify the best-fit model by minimizing the sum of the squared differences between observed and predicted values. This technique enables accurate characterization of molecular structures, reaction pathways, and physical properties, making it essential for reliable analysis and predictions in chemistry.

1 Linear models

Let's start by defining a class of regression models known that are linear with respect to their parameters,

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^\top \mathbf{x} = \begin{bmatrix} w_0, & w_1, & \cdots, & w_d \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} = w_0 + \sum_{i=1}^{d} w_i x_i. \tag{1}$$

From these equation we can observe that any argument that depends on $\mathbf{x}$ is only multiply by a single w_i, therefore the name **linear models**. Polynomials expansions are also linear models,

$$f(x, \mathbf{w}) = \mathbf{w}^\top \Phi(x) = \begin{bmatrix} w_0, & w_1, & \cdots, & w_p \end{bmatrix} \begin{bmatrix} x^0 \\ x^1 \\ \vdots \\ x^p \end{bmatrix} = \sum_{i=1}^{p} w_i x^i. \tag{2}$$

To measure the quality of a model one can define **loss** or **error** function that quantifies the difference between the prediction of the model and the target values. Commonly, we use square errors,

$$\varepsilon^2(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2, \tag{3}$$

where $\hat{\mathrm{y}}_i$ is the target value and y_i is the predicted with our model.

Let's consider a simple linear model, $f(x) = a\ x + b$, where we chose $a = 5$ and $b = 1$ as are the parameters of our model; from here on the parameters of the model will be defined as $\mathbf{w}^\top = [b, a]$. Fig. 1 show case the error for each of the 5 points ($\epsilon(x_i, y_i)$) considered in this example case. We must stress that the *square error* weights each error differently, the larger the difference with the target value the larger the value of the error, see right panel of Fig. 1.

2 Mean Square Error

From what we can observe, each point give us an error that estimates how accurate or inaccurate is our model. Therefore to measure the quality of the the model on the entire

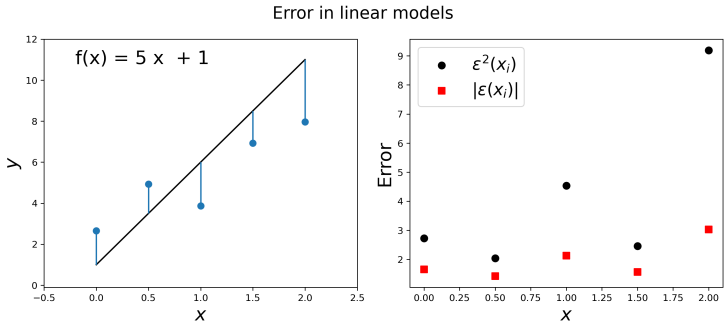

Figure 1: Linear regression model for one-dimensional data.

collection of data points, **we simply average the individual errors**. The Mean Squared Error (MSE) is defined as:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=i}^{n}\frac{1}{2}\left(\hat{y}_i - \mathbf{w}^\top x_i\right)^2 = \frac{1}{n}\sum_{i=i}^{n}\frac{1}{2}\underbrace{\left(\hat{y}_i - (a\,x_i + b)\right)^2}_{\varepsilon_i^2}. \tag{4}$$

The **goal** is simple, find the values of $\mathbf{w}$ where $\mathcal{L}(\mathbf{w})$ is **minimum**,

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}), \tag{5}$$

where $\mathbf{w}^*$ denote the optimal values where the error function has the lowest value. There are two main methods to solve for $\mathbf{w}^*$,

1. **Grid** search-based methods, where we create a grid of possible values of $\mathbf{w}$, and simply look for the lowest value of $\mathcal{L}$.
2. **Gradient**-based methods, where we use the Jacobian of $\mathcal{L}$ ($\nabla_{\mathbf{w}}\,\mathcal{L}$), to iteratively search for $\mathbf{w}^*$.

This notes will only be focused on **Gradient**-based methods for the least square problem, Eqs. 5 and 4. We must remember the following property of calculus. The first order derivative of a function, $\frac{\partial f(x)}{\partial x}$, is the slope of a function, when equal to 0 ($\frac{\partial f(x)}{\partial x} = 0$) it indicates that we are in a **maxima or minima** of a function; meaning x is maxima or minima. For any multi-variable function, $f(\mathbf{x}) = f(x_1, \cdots, x_\ell)$, the vector that composes all the first-oder partial derivatives is known as the **Jacobian**,

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_\ell} \end{bmatrix} \tag{6}$$

For the remaining of the notes, we will only consider a linear model in on-dimension,

$$f(x) = \mathbf{x}^\top \mathbf{w} = \begin{bmatrix} 1 & x \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix}, \tag{7}$$

where each point x now includes a new column with a "one"; $\mathbf{x}^\top = [1, x]$.

3 Linear regression without matrix calculus

Here, we will solve Eq. 5 without using matrix calculus. (**Step 1**) We start with computing the derivatives of $\mathcal{L}(\mathbf{w})$ with respect to each of the parameters of the linear model ($\mathbf{w}^\top = [b, a]$),

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial}{\partial b}\left(\frac{1}{n}\sum_{i=i}^{n}\frac{1}{2}\left(\hat{y}_i - (a\ x_i + b)\right)^2\right) = \frac{1}{2n}\sum_{i=i}^{n}\frac{\partial}{\partial b}\left(\hat{y}_i - (a\ x_i + b)\right)^2 \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial a} = \frac{\partial}{\partial a}\left(\frac{1}{n}\sum_{i=i}^{n}\frac{1}{2}\left(\hat{y}_i - (a\ x_i + b)\right)^2\right) = \frac{1}{2n}\sum_{i=i}^{n}\frac{\partial}{\partial a}\left(\hat{y}_i - (a\ x_i + b)\right)^2, \tag{9}$$

using the change of variable rule we can derive the following derivatives,

$$\frac{\partial}{\partial b}\left(\hat{y}_i - (a\ x_i + b)\right)^2 = 2\left(\hat{y}_i - (a\ x_i + b)\right)\left(\frac{\partial - (a\ x_i + b)}{\partial b}\right) = 2\left(\hat{y}_i - (a\ x_i + b)\right)(-1) \tag{10}$$

$$\frac{\partial}{\partial a}\left(\hat{y}_i - (a\ x_i + b)\right)^2 = 2\left(\hat{y}_i - (a\ x_i + b)\right)\left(\frac{\partial - (a\ x_i + b)}{\partial a}\right) = 2\left(\hat{y}_i - (a\ x_i + b)\right)(-x_i). \tag{11}$$

Combining Eqs. 10 and 11 with Eqs. 8 and 9 we reach the final expressions for both derivatives,

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{1}{n}\sum_{i=i}^{n}\left(\hat{y}_i - (a\ x_i + b)\right)(-1) \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{1}{n}\sum_{i=i}^{n}\left(\hat{y}_i - (a\ x_i + b)\right)(-x_i). \tag{13}$$

From these two equations, the only unknown variables are the model parameters, b and a. (**Step 2**) Set the Jacobian to zero ($\nabla_{\mathbf{w}}\,\mathcal{L}(\mathbf{w}) = 0$),

$$\nabla_{\mathbf{w}}\,\mathcal{L}(\mathbf{w}) = \begin{bmatrix}\frac{\partial \mathcal{L}}{\partial b}\\ \frac{\partial \mathcal{L}}{\partial a}\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}, \tag{14}$$

where each element of $\nabla_{\mathbf{w}}\,\mathcal{L}(\mathbf{w})$ has the following expression,

$$\frac{\partial \mathcal{L}}{\partial b} = 0 = \frac{1}{n}\sum_{i=i}^{n}\left(\hat{y}_i - (a\ x_i + b)\right)(-1) \tag{15}$$

$$\frac{\partial \mathcal{L}}{\partial a} = 0 = \frac{1}{n}\sum_{i=i}^{n}\left(\hat{y}_i - (a\ x_i + b)\right)(-x_i). \tag{16}$$

Let's define the following relations,

$$S_x = \sum_i x_i, \qquad S_y = \sum_i \hat{y}_i \qquad S_{xy} = \sum_i x_i\ \hat{y}_i \qquad S_{xx} = \sum_i x_i\ x_i, \tag{17}$$

using these, we can rewrite Eqs. 15 and 19,

$$\frac{\partial \mathcal{L}}{\partial b} = \tfrac{1}{n}\left(-S_y + a\ S_x + b(n)\right) = 0 \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial a} = \tfrac{1}{n}\left(-S_y + a\ S_{xx} + b\ S_x\right) = 0. \tag{19}$$

These two equations are a set of linear equations with two variables, b and a, which can be solved using like Gauss-Jordan elimination. Because of the simplicity of these set of equations, (2×2), we can derive the solutions for b and a. From Eq. 18, we solve for b,

$$b = \frac{S_y - a\ S_x}{n}, \tag{20}$$

then, we plug this into Eq. 19 to solve for a,

$$-S_y + a\ S_{xx} + \left(\frac{S_y - a\ S_x}{n}\right)\ S_x = 0 \tag{21}$$

$$a\ S_{xx} - \frac{a\ S_x S_x}{n} = S_y - \frac{S_y S_x}{n} \tag{22}$$

$$a\left(S_{xx} - \frac{S_x S_x}{n}\right) = S_y - \frac{S_y S_x}{n} \tag{23}$$

$$a = \frac{S_y - \frac{S_y S_x}{n}}{S_{xx} - \frac{S_x S_x}{n}} \tag{24}$$

The last step is to use Eq. 24 in Eq. 20, for the true expression of b,

$$b = \frac{S_y}{n} - \frac{nS_{xy}S_x - S_y S_x S_x}{nS_{xx} - S_x S_x}. \tag{25}$$

(**Step 3**) The final expressions for the **optimal parameters** for this linear model using mean square error are,

$$\mathbf{w} = \begin{bmatrix} b^* \\ a^* \end{bmatrix} = \begin{bmatrix} \frac{S_y}{n} - \frac{nS_{xy}S_x - S_y S_x S_x}{nS_{xx} - S_x S_x} \\ \frac{S_y - \frac{S_y S_x}{n}}{S_{xx} - \frac{S_x S_x}{n}} \end{bmatrix}. \tag{26}$$

4 Linear regression with a "little" of matrix calculus

Before we revisit the linear regression problem under the scope of matrix calculus, it is important to notice that a lot of the computations done in the prev. section can be recasted as vector-vector, matrix-vector and matrix-matrix multiplications. Many of the equations that will appear in this section will depend on some identities listed in Section 6

First let's define the following error vector $\boldsymbol{\varepsilon}$,

$$\underbrace{\boldsymbol{\varepsilon}}_{(n,1)} = \underbrace{\mathbf{Y}}_{(n,1)} - \underbrace{\mathbf{X}}_{(n,2)}\underbrace{\mathbf{w}}_{(2,1)} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} - \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \hat{y}_1 - (a\ x_1 + b) \\ \hat{y}_2 - (a\ x_2 + b) \\ \vdots \\ \hat{y}_n - (a\ x_n + b) \end{bmatrix} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \tag{27}$$

where each ε_i is the individual difference/error between the prediction of the model, $a\ x_i + b$, and the true value $\hat{y}_i$.

If we stare at Eq. 4 long enough, you would see that this Eq. resembles the dot-product of a vector, $\mathbf{x}^\top \mathbf{x} = \sum_i x_i^2$. One additional way to rewrite the mean square error is,

$$\begin{aligned} \mathcal{L}(\mathbf{w}) &= \frac{1}{2n}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = \frac{1}{2n}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{2n}\sum_{i=1}^{n}\left(\hat{y}_i - (a\ x_i + b)\right)^2 \qquad (28) \\ &= \frac{1}{2n}\begin{bmatrix} \hat{y}_1 - (a\ x_1 + b), & \hat{y}_2 - (a\ x_2 + b), & \cdots, & \hat{y}_n - (a\ x_n + b) \end{bmatrix} \begin{bmatrix} \hat{y}_1 - (a\ x_1 + b) \\ \hat{y}_2 - (a\ x_2 + b) \\ \vdots \\ \hat{y}_n - (a\ x_n + b) \end{bmatrix} \end{aligned}$$

Let's expand Eq, 28 only using the definition of $\boldsymbol{\varepsilon}$ from Eq. 27 and the identity from Eq. 49,

$$\begin{aligned} \mathcal{L}(\mathbf{w}) &= \frac{1}{2n}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = \frac{1}{2n}\left(\mathbf{Y} - \mathbf{X}\,\mathbf{w}\right)^\top \left(\mathbf{Y} - \mathbf{X}\,\mathbf{w}\right) = \frac{1}{2n}\left(\mathbf{Y}^\top - \mathbf{w}^\top \mathbf{X}^\top\right)\left(\mathbf{Y} - \mathbf{X}\,\mathbf{w}\right) \\ &= \frac{1}{2n}\left(\mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\,\mathbf{w} - \mathbf{w}^\top \mathbf{X}^\top \mathbf{Y} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\,\mathbf{w}\right). \qquad (29) \end{aligned}$$

If we check, each of the terms in Eq. 29 yield a scalar output,

- $\mathbf{Y}^\top \mathbf{Y} \rightarrow (1,n)(n,1) = (1,1)$
- $\mathbf{Y}^\top \mathbf{X}\,\mathbf{w} \rightarrow (1,n)(n,2)(2,1) = (1,1)$
- $\mathbf{w}^\top \mathbf{X}^\top \mathbf{Y} \rightarrow (1,2)(2,n)(n,1) = (1,1)$
- $\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\,\mathbf{w} \rightarrow (1,2)(2,n)(n,2)(2,1) = (1,1)$

Similar to our previous approach, let's now compute the partial derivatives of $\mathcal{L}(\mathbf{w})$ with respect to b and a. Let's start with,

$$\frac{\partial\, \mathbf{Y}^\top \mathbf{Y}}{\partial b} = \frac{\partial}{\partial b} \sum_i y_i^2 = 0 \tag{30}$$

$$\frac{\partial\, \mathbf{Y}^\top \mathbf{Y}}{\partial a} = \frac{\partial}{\partial a} \sum_i y_i^2 = 0, \tag{31}$$

both terms are equal to 0 as there isn't a single term that depends on b and a. The following term that we will consider is $\mathbf{Y}^\top \mathbf{X}\,\mathbf{w}$,

$$\frac{\partial\, \mathbf{Y}^\top \mathbf{X}\,\mathbf{w}}{\partial b} = \frac{\partial}{\partial b} \begin{bmatrix} \hat{y}_1, & \hat{y}_2, & \cdots, & \hat{y}_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \sum_i \frac{\partial}{\partial b} y_i (a\, x_i + b) = \sum_i y_i \tag{32}$$

$$\frac{\partial\, \mathbf{Y}^\top \mathbf{X}\,\mathbf{w}}{\partial a} = \frac{\partial}{\partial a} \begin{bmatrix} \hat{y}_1, & \hat{y}_2, & \cdots, & \hat{y}_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \sum_i \frac{\partial}{\partial a} y_i (a\, x_i + b) = \sum_i x_i y_i. \tag{33}$$

If you do the same procedure for the partial derivatives of $\mathbf{Y}^\top \mathbf{X}\,\mathbf{w}$, you will reach the same equations, as these two terms, $\mathbf{Y}^\top \mathbf{X}\,\mathbf{w}$ and $\mathbf{Y}^\top \mathbf{X}\,\mathbf{w}$ are equal.

The last term to consider is, $\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\,\mathbf{w}$, first let's compute this "complex" vector-matrix-matrix-vector product,

$$\begin{aligned} \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\,\mathbf{w} &= \begin{bmatrix} b, & a \end{bmatrix} \begin{bmatrix} 1, & 1, & \cdots, & 1 \\ x_1, & x_2, & \cdots, & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} \\ &= \begin{bmatrix} a\, x_1 + b, & a\, x_2 + b, & \cdots, & a\, x_n + b \end{bmatrix} \begin{bmatrix} a\, x_1 + b \\ a\, x_2 + b \\ \vdots \\ a\, x_n + b \end{bmatrix} = \sum_i (a\, x_i + b)^2 \end{aligned} \tag{34}$$

Let's use this to make the computation of the partial derivatives easier.

$$\frac{\partial\, \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\,\mathbf{w}}{\partial b} = \frac{\partial}{\partial b} \sum_i (a\, x_i + b)^2 = \sum_i 2(a\, x_i + b) = 2a \sum_i x_i + 2b(n) \tag{35}$$

$$\frac{\partial\, \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\,\mathbf{w}}{\partial a} = \frac{\partial}{\partial b} \sum_i (a\, x_i + b)^2 = \sum_i 2(a\, x_i + b)(x_i) = 2a \sum_i x_i^2 + 2b \sum_i x_i \tag{36}$$

Let's combine Eqs. 30, 31, 32, 33, 35, and 36, we get the following expression for

the Jacobian of $\mathcal{L}(\mathbf{w})$,

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{1}{2n}\left(\underbrace{0}_{\frac{\partial \mathbf{Y}^\top \mathbf{Y}}{\partial b}} \underbrace{-2\sum_i y_i}_{\frac{\partial \mathbf{Y}^\top \mathbf{X}\mathbf{w}}{\partial b}} + \underbrace{2a\sum_i x_i + 2b(n)}_{\frac{\partial \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}}{\partial b}}\right) \tag{37}$$

$$\frac{\partial \mathcal{L}}{\partial a} = \frac{1}{2n}\left(\underbrace{0}_{\frac{\partial \mathbf{Y}^\top \mathbf{Y}}{\partial a}} \underbrace{-2\sum_i x_i y_i}_{\frac{\partial \mathbf{Y}^\top \mathbf{X}\mathbf{w}}{\partial a}} + \underbrace{2a\sum_i x_i^2 + 2b\sum_i x_i}_{\frac{\partial \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}}{\partial a}}\right) \tag{38}$$

Eqs. 37 and 38, as expected, are the same equations as the ones we previously derived, Eq. 12 and 13. Furthermore, many of the terms can be simplified using the notation of Eq. 17. The only remaining steps are, i) equate the Jacobian to 0, and ii) solve the linear equations to find the values of b and a.

5 Linear regression with matrix calculus

For this final section, we will start with the definition of $\mathcal{L}(\mathbf{w})$,

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2n}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = \frac{1}{2n}\left(\mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\mathbf{w} - \mathbf{w}^\top \mathbf{X}^\top \mathbf{Y} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}\right) \tag{39}$$

then we will proof that $Y^\top \mathbf{X}\mathbf{w}$ is equal to $\mathbf{w}^\top \mathbf{X}^\top \mathbf{Y}$ using Eq. 50 and 51

$$\mathbf{w}^\top \mathbf{X}^\top \mathbf{Y} = \mathbf{Y}\left(\mathbf{X}^\top\right)^\top \left(\mathbf{w}^\top\right)^\top = \mathbf{Y}^\top \mathbf{X}\mathbf{w}\,. \tag{40}$$

this leads to,

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2n}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = \frac{1}{2n}\left(\mathbf{Y}^\top \mathbf{Y} - 2\,\mathbf{z}^\top \mathbf{w} + \mathbf{w}^\top \mathbf{A}\,\mathbf{w}\right) \tag{41}$$

where,

$$A = \mathbf{X}^\top \mathbf{X} \quad \text{and} \quad \mathbf{z}^\top = \mathbf{Y}^\top \mathbf{X}\,. \tag{42}$$

Now let's compute $\frac{\partial}{\partial \mathbf{w}}$, the Jacobian[1] of $\mathcal{L}(\mathbf{w})$, for each term in Eq. 41.

$$\begin{aligned}\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) &= \frac{1}{2n}\frac{\partial}{\partial \mathbf{w}}\left(\mathbf{Y}^\top \mathbf{Y} - 2\,\mathbf{z}^\top \mathbf{w} + \mathbf{w}^\top \mathbf{A}\,\mathbf{w}\right) \\ &= \frac{1}{2n}\left(\frac{\partial}{\partial \mathbf{w}}\mathbf{Y}^\top \mathbf{Y} - 2\frac{\partial}{\partial \mathbf{w}}\mathbf{z}^\top \mathbf{w} + \frac{\partial}{\partial \mathbf{w}}\mathbf{w}^\top \mathbf{A}\,\mathbf{w}\right)\end{aligned} \tag{43}$$

each of these terms looks like the following relations with standard derivatives,

- $\frac{\partial}{\partial \mathbf{w}}\mathbf{Y}^\top \mathbf{Y} \rightarrow \frac{\partial c}{\partial x}$, where c is a constant.
- $\frac{\partial}{\partial \mathbf{w}}\mathbf{z}^\top \mathbf{w} \rightarrow \frac{\partial ax}{\partial x}$, where a is a constant.
- $\frac{\partial}{\partial \mathbf{w}}\mathbf{w}^\top \mathbf{A}\,\mathbf{w} \rightarrow \frac{\partial ax^2}{\partial x}$, where a is a constant.

[1] We define the Jacobian in Eq. 6

Each term in Eq. 43 is,

- $\frac{\partial}{\partial \mathbf{w}} \mathbf{Y}^\top \mathbf{Y}$ is equal to zero as there is no dependence on any element of $\mathbf{w}$.
- $\frac{\partial}{\partial \mathbf{w}} \mathbf{z}^\top \mathbf{w}$ is equal to $\mathbf{z}$, using Eq. 53.
- $\frac{\partial}{\partial \mathbf{w}} \mathbf{w}^\top \mathbf{A}\, \mathbf{w}$ is equal to $2\, \mathbf{A}\, \mathbf{w}$, using Eq. 54.

Putting all these together we get,

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \quad = \quad \frac{1}{2n} \left(\mathbf{0} - 2\, \mathbf{z} + 2\, \mathbf{A}\, \mathbf{w} \right). \tag{44}$$

If we equate $\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w})$ to zero to find the optimal values of b and a, we get the following equation,

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \quad = \quad \frac{1}{n} \left(- \mathbf{z} + \mathbf{A}\, \mathbf{w} \right) = \frac{1}{n} \left(- \underbrace{\mathbf{X}^\top \mathbf{Y}}_{\mathbf{z}} + \underbrace{\mathbf{X}^\top \mathbf{X}}_{\mathbf{A}} \mathbf{w} \right) = \mathbf{0}. \tag{45}$$

This equation looks like $-c + ax = 0$, which we could simply solve by $x = a^{-1}c$. Following this we can solve for $\mathbf{w}$,

$$\begin{aligned} \mathbf{X}^\top \mathbf{X}\, \mathbf{w} &= \mathbf{X}^\top \mathbf{Y} \\ \mathbf{w} &= \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{Y} \end{aligned} \tag{46}$$

where $\mathbf{w}$ is the solution of the least square problem. For these, we need to **invert** the matrix $\mathbf{X}^\top \mathbf{X}$. Before we do it, I will like to introduce a property of $\mathbf{A}^{-1}$,

$$\mathbf{A}^{-1}\, \mathbf{A} = \mathbf{A}\ \mathbf{A}^{-1} = \mathbf{I}, \tag{47}$$

where $\mathbf{I}$ is the identity matrix. We can use this to compute the inverse of a matrix with the Gauss-Jordan elimination method,

$$\left[\begin{array}{ccc|ccc} a_{11} & a_{12} & a_{31} & 1 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & 1 & 0 \\ a_{31} & a_{32} & a_{33} & 0 & 0 & 1 \end{array}\right]$$

where the left part is the matrix $\mathbf{A}$ and the right part is $\mathbf{I}$. Let's do a simple example,

$$\left[\begin{array}{cc|cc} 5 & 5 & 1 & 0 \\ 5 & 7.5 & 0 & 1 \end{array}\right]$$

Subtract row 1 from row 2,

$$\left[\begin{array}{cc|cc} 5 & 5 & 1 & 0 \\ 0 & -2.5 & 1 & -1 \end{array}\right]$$

Subtract row 2 from row 1, and multiply row 2 by (2)

$$\left[\begin{array}{cc|cc} 5 & 0 & 3 & -2 \\ 0 & -2.5 & 1 & -1 \end{array}\right]$$

Multiply row 1 by $(\frac{1}{5})$ and row 2 by $(-\frac{1}{2.5})$

$$\left[\begin{array}{cc|cc} 1 & 0 & \frac{3}{5} & -\frac{2}{5} \\ 0 & 1 & -\frac{1}{2.5} & \frac{1}{2.5} \end{array}\right]$$

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{3}{5} & -\frac{2}{5} \\ -\frac{1}{2.5} & \frac{1}{2.5} \end{bmatrix} \tag{48}$$

6 Useful Identities

The Identities presented here are from the The Matrix Cookbook, specifically from **Chapters 1** and **2**.

$$
\begin{aligned}
(\mathbf{A}\,\mathbf{B})^{\top} &= \mathbf{B}^{\top}\,\mathbf{A}^{\top} && (49)\\
(\mathbf{A}\,\mathbf{B}\,\mathbf{C})^{\top} &= \mathbf{C}^{\top}\,\mathbf{B}^{\top}\,\mathbf{A}^{\top} && (50)\\
\left(\mathbf{A}^{\top}\right)^{\top} &= \mathbf{A} && (51)\\
& && (52)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial\,\mathbf{x}^{\top}\,\mathbf{u}}{\partial\,\mathbf{x}} &= \frac{\partial\,\mathbf{u}^{\top}\,\mathbf{x}}{\partial\,\mathbf{x}} = \mathbf{u} && (53)\\
\frac{\partial(\mathbf{x}-\mathbf{u})^{\top}\,\mathbf{A}(\mathbf{x}-\mathbf{u})}{\partial\,\mathbf{x}} &= 2\,\mathbf{A}(\mathbf{x}-\mathbf{u}) && (54)\\
& && (55)
\end{aligned}
$$

where $\mathbf{u}$ is another vector and $\mathbf{A}$ is a symmetric matrix.