# Derivations for Linear Regression

Sayantan Auddy

This document contains derivations for linear regression with and without regularization. All the necessary linear algebra proofs have been explained in the Appendix. Refer to the links in the footnotes for more information about matrix calculus.

## 1 Linear Regression[1]

We have a *design matrix* $\mathbf{\Phi} \in \mathbb{R}^{n \times m}$ (e.g. constructed using $n$ data points and for a $m-$dimensional polynomial). We want to find a weight vector $\mathbf{w} \in \mathbb{R}^m$ such that the following is true where $\mathbf{t}$ is the vector of labels.

$$\mathbf{\Phi w} \approx \mathbf{t}$$

$$\Rightarrow \begin{bmatrix} \Phi_{11} & \Phi_{12} & \cdots & \Phi_{1m} \\ \Phi_{21} & \Phi_{22} & \cdots & \Phi_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{n1} & \Phi_{n2} & \cdots & \Phi_{nm} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} \approx \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$$

$$\Rightarrow \mathbf{\Phi w} = \begin{bmatrix} w_1\Phi_{11} + w_2\Phi_{12} + \cdots w_m\Phi_{1m} \\ w_1\Phi_{21} + w_2\Phi_{22} + \cdots w_m\Phi_{2m} \\ \vdots \\ w_1\Phi_{n1} + w_2\Phi_{n2} + \cdots w_m\Phi_{nm} \end{bmatrix} \approx \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$$

Thus, we need to minimize the differences between our predicted label vector ($\mathbf{\Phi w}$) and our target label vector $\mathbf{t}$. The vector containing the difference for each data point can be computed as

$$\mathbf{\Phi w} - \mathbf{t} = \begin{bmatrix} (\Phi w)_1 - t_1 \\ (\Phi w)_2 - t_2 \\ \vdots \\ (\Phi w)_n - t_n \end{bmatrix}$$

The squared error is a scalar number which gives the cumulative error over all the $n$ data points. This can be obtained by adding the squared errors for each prediction. In other words,

$$E(\mathbf{w}) = \frac{1}{2}\Big(\big((\Phi w)_1 - t_1\big)^2 + \big((\Phi w)_2 - t_2\big)^2 + \cdots + \big((\Phi w)_n - t_n\big)^2\Big) \ (\tfrac{1}{2} \text{ is used for mathematical convenience})$$

$$= \frac{1}{2}\begin{bmatrix} \big((\Phi w)_1 - t_1\big) & \big((\Phi w)_2 - t_2\big) & \cdots & \big((\Phi w)_n - t_n\big) \end{bmatrix} \begin{bmatrix} \big((\Phi w)_1 - t_1\big) \\ \big((\Phi w)_2 - t_2\big) \\ \vdots \\ \big((\Phi w)_n - t_n\big) \end{bmatrix}$$

$$= \frac{1}{2}\Big(\mathbf{\Phi w} - \mathbf{t}\Big)^T\Big(\mathbf{\Phi w} - \mathbf{t}\Big)$$

So, now we need to find $\mathbf{w}$ such that $E(\mathbf{w})$ is minimized. This is done by computing the gradient of $E(\mathbf{w})$ with respect to $\mathbf{w}$ and setting it to 0. Thus,

$$\mathbf{w} = \arg\min_{\mathbf{w}} E(\mathbf{w})$$

[1] This is based on `http://www.haija.org/derivation_lin_regression.pdf`, which contains a more condensed form of this proof.

$$
\begin{aligned}
\Rightarrow \nabla_{\mathbf{w}} E(\mathbf{w}) &= \nabla_{\mathbf{w}} \Big( \frac{1}{2} \big( \mathbf{\Phi w} - \mathbf{t} \big)^T \big( \mathbf{\Phi w} - \mathbf{t} \big) \Big) = 0 \\
&= \frac{1}{2} \nabla_{\mathbf{w}} \Big( \big( \mathbf{\Phi w} - \mathbf{t} \big)^T \big( \mathbf{\Phi w} - \mathbf{t} \big) \Big) = 0 \\
&= \frac{1}{2} \nabla_{\mathbf{w}} \Big( \big( \mathbf{w}^T \mathbf{\Phi}^T - \mathbf{t}^T \big) \big( \mathbf{\Phi w} - \mathbf{t} \big) \Big) = 0 \\
&\text{(since } (A-B)^T = A^T - B^T \text{ and } (AB)^T = B^T A^T \text{)} \\
&= \frac{1}{2} \nabla_{\mathbf{w}} \Big( \mathbf{w}^T \mathbf{\Phi}^T \mathbf{\Phi w} - \mathbf{t}^T \mathbf{\Phi w} - \mathbf{w}^T \mathbf{\Phi}^T \mathbf{t} + \mathbf{t}^T \mathbf{t} \Big) = 0 \\
&= \frac{1}{2} \nabla_{\mathbf{w}} \Big( \mathbf{w}^T \mathbf{\Phi}^T \mathbf{\Phi w} - \mathbf{t}^T \mathbf{\Phi w} - \mathbf{t}^T \mathbf{\Phi w} + \mathbf{t}^T \mathbf{t} \Big) = 0 \\
&\text{(since } \mathbf{t}^T \mathbf{\Phi w} = \mathbf{w}^T \mathbf{\Phi}^T \mathbf{t} = \text{same scalar number)} \\
&= \frac{1}{2} \nabla_{\mathbf{w}} \Big( \mathbf{w}^T \mathbf{\Phi}^T \mathbf{\Phi w} - 2\mathbf{t}^T \mathbf{\Phi w} + \mathbf{t}^T \mathbf{t} \Big) = 0 \\
&= \frac{1}{2} \nabla_{\mathbf{w}} \Big( \mathbf{w}^T \big( \mathbf{\Phi}^T \mathbf{\Phi} \big) \mathbf{w} \Big) - \nabla_{\mathbf{w}} \Big( \mathbf{t}^T \mathbf{\Phi w} \Big) + \frac{1}{2} \nabla_{\mathbf{w}} \Big( \mathbf{t}^T \mathbf{t} \Big) = 0 \\
&= \frac{1}{2} 2 \big( \mathbf{\Phi}^T \mathbf{\Phi} \big) \mathbf{w} - \Big( \mathbf{t}^T \mathbf{\Phi} \Big)^T + 0 = 0 \\
&\text{(for the first part note that } (\mathbf{\Phi}^T \mathbf{\Phi}) \text{ is a symmetric matrix (see section A.3), and} \\
&\text{using this fact along with the rule in section A.2 gives us } \nabla_{\mathbf{w}} \mathbf{w}^T (\mathbf{\Phi}^T \mathbf{\Phi}) \mathbf{w} = \\
&2(\mathbf{\Phi}^T \mathbf{\Phi}) \mathbf{w} \text{; for the second part see section A.1; the third part does not depend} \\
&\text{on } \mathbf{w} \text{ and so its gradient w.r.t. } \mathbf{w} \text{ is 0)} \\
&= \big( \mathbf{\Phi}^T \mathbf{\Phi} \big) \mathbf{w} - \mathbf{\Phi}^T \mathbf{t} = 0 \\
&\Rightarrow \big( \mathbf{\Phi}^T \mathbf{\Phi} \big) \mathbf{w} = \mathbf{\Phi}^T \mathbf{t} \\
&\Rightarrow \mathbf{w} = \big( \mathbf{\Phi}^T \mathbf{\Phi} \big)^{-1} \mathbf{\Phi}^T \mathbf{t} \text{ (multiplying both sides by } \big( \mathbf{\Phi}^T \mathbf{\Phi} \big)^{-1} \text{)}
\end{aligned}
$$

## 2 Linear Regression with Regularization

In linear regression without regularization, our error function was defined as

$$E(\mathbf{w}) = \frac{1}{2} \Big( \mathbf{\Phi w} - \mathbf{t} \Big)^T \Big( \mathbf{\Phi w} - \mathbf{t} \Big)$$

In order to minimize overfitting, we want to regularize the linear regression by introducing a penalty on the weights (such that low values of weights are preferred). We do this by modifying the error function as given below

$$E(\mathbf{w}) = \frac{1}{2} \Big( \mathbf{\Phi w} - \mathbf{t} \Big)^T \Big( \mathbf{\Phi w} - \mathbf{t} \Big) + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$$

(note that the sum of squared weights is given by $\mathbf{w}^T \mathbf{w} = w_1^2 + w_1^2 + ... + w_m^2$ and $\lambda$ is the regularization coefficient)

Then, similar to section 1, we compute the gradient of the error and set it to 0.

$$
\begin{aligned}
\mathbf{w} &= \arg\min_{\mathbf{w}} E(\mathbf{w}) \\
\Rightarrow \nabla_{\mathbf{w}} E(\mathbf{w}) &= \nabla_{\mathbf{w}} \Big( \frac{1}{2} \big( \mathbf{\Phi w} - \mathbf{t} \big)^T \big( \mathbf{\Phi w} - \mathbf{t} \big) + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \Big) = 0 \\
&= \nabla_{\mathbf{w}} \Big( \frac{1}{2} \big( \mathbf{\Phi w} - \mathbf{t} \big)^T \big( \mathbf{\Phi w} - \mathbf{t} \big) \Big) + \nabla_{\mathbf{w}} \Big( \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \Big) = 0 \\
&= \Big( \big( \mathbf{\Phi}^T \mathbf{\Phi} \big) \mathbf{w} - \mathbf{\Phi}^T \mathbf{t} \Big) + \frac{\lambda}{2} \nabla_{\mathbf{w}} \Big( \mathbf{w}^T \mathbf{w} \Big) = 0 \\
&\text{(the first part comes from the derivation in section 1)} \\
&= \Big( \big( \mathbf{\Phi}^T \mathbf{\Phi} \big) \mathbf{w} - \mathbf{\Phi}^T \mathbf{t} \Big) + \lambda \mathbf{w} = 0 \\
&\text{(since } \nabla_{\mathbf{w}} \mathbf{w}^T \mathbf{w} = 2\mathbf{w} \text{); see section A.4)}
\end{aligned}
$$

$$\Rightarrow \Big( \big( \mathbf{\Phi}^T \mathbf{\Phi} \big) + \lambda \mathbf{I} \Big) \mathbf{w} = \mathbf{\Phi}^T \mathbf{t}$$
$$\Rightarrow \mathbf{w} = \Big( \big( \mathbf{\Phi}^T \mathbf{\Phi} \big) + \lambda \mathbf{I} \Big)^{-1} \mathbf{\Phi}^T \mathbf{t}$$

# A Appendix: Useful Matrix Calculus rules[2]

## A.1 Rule 1

$$\nabla_{\mathbf{x}} \mathbf{b}^T \mathbf{x} = \mathbf{b} \tag{1}$$
where $\mathbf{x}$ and $\mathbf{b}$ are vectors of size $n$ ($\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^n$)

*Proof.*

$$\text{Let } f(\mathbf{x}) = \mathbf{b}^T \mathbf{x} = \sum_{i=1}^{n} b_i x_i$$
$$\text{So } \frac{\partial f(\mathbf{x})}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k$$
$$\text{Therefore } \nabla_{\mathbf{x}} \mathbf{b}^T \mathbf{x} = \begin{bmatrix} \frac{\partial}{\partial x_1} \sum_{i=1}^{n} b_i x_i = b_1 \\ \frac{\partial}{\partial x_2} \sum_{i=1}^{n} b_i x_i = b_2 \\ \vdots \\ \frac{\partial}{\partial x_n} \sum_{i=1}^{n} b_i x_i = b_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \mathbf{b}$$

■

## A.2 Rule 2

$$\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} = 2\mathbf{A}\mathbf{x} \tag{2}$$
where $\mathbf{x}$ is a vector of size $n$ ($\mathbf{x} \in \mathbb{R}^n$) and $\mathbf{A}$ is a symmetric matrix of size $n \times n$ ($\mathbf{A} \in \mathbb{R}^{n \times n}$)

*Proof.*

$$\text{Let } f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$$
$$= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$
$$= \begin{bmatrix} (x_1 A_{11} + x_2 A_{21} + \cdots + x_n A_{n1}) & \cdots & (x_1 A_{1n} + x_2 A_{2n} + \cdots + x_n A_{nn}) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$
$$= x_1 (x_1 A_{11} + x_2 A_{21} + \cdots + x_n A_{n1}) + \cdots + x_n (x_1 A_{1n} + x_2 A_{2n} + \cdots + x_n A_{nn})$$
$$= x_1 (\sum_{i=1}^{n} x_i A_{i1}) + \cdots + x_n (\sum_{i=1}^{n} x_i A_{in})$$

[2]Some of the proofs are based on `http://cs229.stanford.edu/section/cs229-linalg.pdf` but have been simplified further by showing the matrix elements where ever possible. Useful sources for matrix calculus are `https://atmos.washington.edu/~dennis/MatrixCalculus.pdf` and of course `https://en.wikipedia.org/wiki/Matrix_calculus`.

$$= \sum_{j=1}^{n}\sum_{i=1}^{n} x_j x_i A_{ij}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} x_i x_j$$

$$\frac{\partial f(\mathbf{x})}{\partial x_k} = \frac{\partial}{\partial x_k}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} x_i x_j$$

$$= \frac{\partial}{\partial x_k}\Big[\sum_{i\neq k}^{n}\sum_{j\neq k}^{n} A_{ij} x_i x_j + \sum_{i\neq k}^{n} A_{ik} x_i x_k + \sum_{j\neq k}^{n} A_{kj} x_k x_j + A_{kk} x_k x_k\Big]$$

Separating $f(\mathbf{x})$ into 4 parts: $(i \neq k, j \neq k), (i \neq k, j = k), (i = k, j \neq k), (i = k, j = k)$

$$= \sum_{i\neq k}^{n} A_{ik} x_i + \sum_{j\neq k}^{n} A_{kj} x_j + 2A_{kk} x_k$$

$$= \sum_{i\neq k}^{n} A_{ik} x_i + A_{kk} x_k + \sum_{j\neq k}^{n} A_{kj} x_j + A_{kk} x_k$$

$$= \sum_{i=1}^{n} A_{ik} x_i + \sum_{j=1}^{n} A_{kj} x_j$$

$$= 2\sum_{i=1}^{n} A_{ik} x_i \text{ (since } A \text{ is symmetric, } A_{ik} = A_{kj})$$

$$\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A}\mathbf{x} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_k} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 2\sum_{i=1}^{n} A_{i1} x_i \\ 2\sum_{i=1}^{n} A_{i2} x_i \\ \vdots \\ 2\sum_{i=1}^{n} A_{ik} x_i \\ \vdots \\ 2\sum_{i=1}^{n} A_{in} x_i \end{bmatrix} = 2\begin{bmatrix} A_{11}x_1 + A_{21}x_2 + \cdots + A_{n1}x_n \\ A_{12}x_1 + A_{22}x_2 + \cdots + A_{n2}x_n \\ \vdots \\ A_{1k}x_1 + A_{2k}x_2 + \cdots + A_{nk}x_n \\ \vdots \\ A_{1n}x_1 + A_{2n}x_2 + \cdots + A_{nn}x_n \end{bmatrix}$$

$$= 2\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= 2\begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ (since } A \text{ is symmetric, } A^T = A)$$

$$= 2A\mathbf{x}$$

■

## A.3 Rule 3

For any matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$, the matrix $\mathbf{\Phi}^T\mathbf{\Phi}$ is a symmetric matrix.

*Proof.* For any matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{B} \in \mathbb{R}^{m \times n}$, $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$

$$\begin{aligned}(\mathbf{\Phi}^T\mathbf{\Phi})^T &= \mathbf{\Phi}^T(\mathbf{\Phi}^T)^T \text{ (taking } \mathbf{A} = \mathbf{\Phi}^T \text{ and } \mathbf{B} = \mathbf{\Phi}) \\ &= \mathbf{\Phi}^T\mathbf{\Phi} \text{ (since } (\mathbf{\Phi}^T)^T = \mathbf{\Phi})\end{aligned}$$

Since the matrix $(\mathbf{\Phi}^T\mathbf{\Phi})$ is equal to its transpose, it is symmetric.

■

## A.4 Rule 4

$$\nabla_{\mathbf{w}}\left(\mathbf{w}^T\mathbf{w}\right) = 2\mathbf{w} \text{ (where } \mathbf{w} \in \mathbb{R}^m)$$

*Proof.*

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \\ \vdots \\ w_m \end{bmatrix}$$

$$\mathbf{w}^T\mathbf{w} = \begin{bmatrix} w_1 & w_2 & \cdots & w_k & \cdots & w_m \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \\ \vdots \\ w_m \end{bmatrix} = (w_1^2 + w_2^2 + \cdots + w_k^2 + \cdots + w_m^2)$$

$$\frac{\partial}{\partial w_k}\mathbf{w}^T\mathbf{w} = 2w_k$$

$$\nabla_{\mathbf{w}}\left(\mathbf{w}^T\mathbf{w}\right) = \begin{bmatrix} 2w_1 \\ 2w_2 \\ \vdots \\ 2w_k \\ \vdots \\ 2w_m \end{bmatrix} = 2\mathbf{w}$$

■