CS578 Statistical Machine Learning
02/08/2021 Lecture Notes: Linear Regression

Site Bai

1 The Linear Regression Model

Given an input $\left[\mathbf{x_1}^\top, \mathbf{x_2}^\top, ..., \mathbf{x_N}^\top\right]^\top \in \mathbb{R}^{N\times k}$, N is the number of input data points and k is the number of features we observe for each data point. We want to find a mapping between the input and the target variables $\mathbf{y} = [y_1, y_2, ..., y_N]^\top \in \mathbb{R}^N$, as shown in Figure 1. For each input data and its corresponding target variable, the linear regression model assume the target variable can be represented as a linear combination of the input with some Gaussian noise, i.e.

$$y_i = w_0 + \sum_{j=1}^{k} w_j x_{ij} + \epsilon, \tag{1}$$

in which $i = 1, 2, ..., N$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. To rewrite (1) in a matrix format,

$$y_i = \mathbf{w}^\top \begin{bmatrix} 1 \\ \mathbf{x_i} \end{bmatrix} + \epsilon. \tag{2}$$

$\mathbf{w} \in \mathbb{R}^{k+1}$, $\mathbf{x_i} \in \mathbb{R}^k$. Denote $\mathbf{x_i'} = [1, \mathbf{x_i}]^\top$, then $y_i = \mathbf{w}^\top \mathbf{x_i'} + \epsilon$, and $\mathbb{P}(y_i|\mathbf{x_i'}, \mathbf{w}) \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x_i'}, \sigma^2)$.

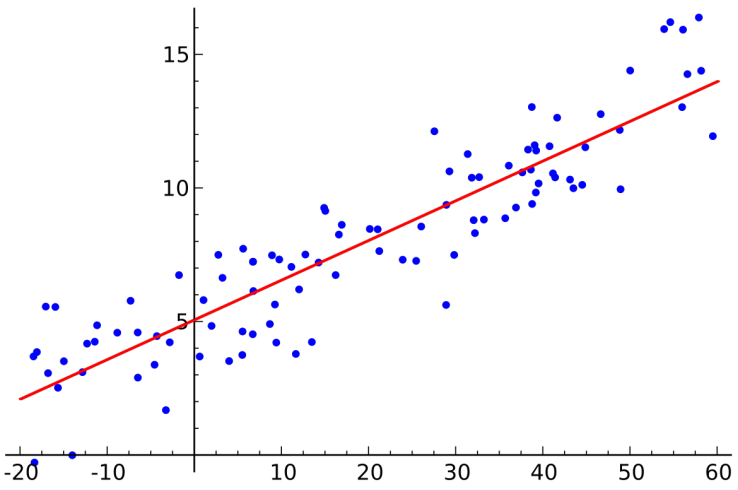

Figure 1: Linear Regression, figure from Wikipedia.

2 Maximum Likelihood Estimation

To find a proper mapping from input to target, we need to obtain the parameter $\mathbf{w}$ that fits the data, and this can be done by Maximum Likelihood Estimation (MLE). We have

$$\mathbb{P}(y_i|\mathbf{x_i'}, \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\left(y_i - \mathbf{w}^\top \mathbf{x_i'}\right)^2}{2\sigma^2}\right\} \tag{3}$$

Thus, the log likelihood function can be written as

$$\begin{aligned}
\mathcal{L}(\mathbf{w}) &= \log \prod_{i=1}^{N} \mathbb{P}(y_i|\mathbf{x_i'}, \mathbf{w}) && (4)\\
&= \sum_{i=1}^{N} \log \mathbb{P}(y_i|\mathbf{x_i'}, \mathbf{w}) && (5)\\
&= \sum_{i=1}^{N} \left(\log \frac{1}{\sqrt{2\pi}\sigma} + \log \exp\left\{-\frac{\left(y_i - \mathbf{w}^\top \mathbf{x_i'}\right)^2}{2\sigma^2}\right\}\right) && (6)\\
&= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^\top \mathbf{x_i'}\right)^2 && (7)
\end{aligned}$$

And to maximize the likelihood,

$$\begin{aligned}
\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \mathcal{L}(\mathbf{w}) && (8)\\
&= \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^\top \mathbf{x_i'}\right)^2 && (9)\\
&= \arg\min_{\mathbf{w}} ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 && (10)\\
&= \arg\min_{\mathbf{w}} \mathrm{RSS}\left(\mathbf{w}\right), && (11)
\end{aligned}$$

in which $\mathbf{X} = \left[\mathbf{x_1'}^\top, \mathbf{x_2'}^\top, ..., \mathbf{x_N'}^\top\right]^\top \in \mathbb{R}^{N\times(k+1)}$, RSS denotes the residual sum-of-squares.

3 Two Ways of Computing $\hat{\mathbf{w}}$

3.1 Derivative of RSS($\mathbf{w}$)

We can compute the derivative of RSS($\mathbf{w}$) w.r.t. $\mathbf{w}$ to obtain the optimal $\mathbf{w}$.

$$\frac{\partial \mathrm{RSS}(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^\top \left(\mathbf{y} - \mathbf{X}\mathbf{w}\right). \tag{12}$$

Then set the derivative to zero,

$$-2\mathbf{X}^\top \left(\mathbf{y} - \mathbf{X}\mathbf{w}\right) = 0. \tag{13}$$

We now obtain the solution

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{14}$$

3.2 Reformatting the Quadratic Form

$$\begin{aligned}
\mathrm{RSS}(\mathbf{w}) &= ||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2 && (15)\\
&= (\mathbf{y}-\mathbf{X}\mathbf{w})^\top (\mathbf{y}-\mathbf{X}\mathbf{w}) && (16)\\
&= \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{y}^\top\mathbf{X}\mathbf{w} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y} && (17)\\
&= \mathbf{w}^\top\left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{w} - \mathbf{y}^\top\mathbf{X}\mathbf{w} - \mathbf{w}^\top\underbrace{\left(\mathbf{X}^\top\mathbf{X}\right)\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}}_{\mathbf{I}}\mathbf{X}^\top\mathbf{y} && (18)\\
&\quad + \underbrace{\mathbf{y}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}}_{0} + \mathbf{y}^\top\mathbf{y}\\
&= \left(\mathbf{w}^\top\left(\mathbf{X}^\top\mathbf{X}\right) - \mathbf{y}^\top\mathbf{X}\right)\mathbf{w} - \left(\mathbf{w}^\top\left(\mathbf{X}^\top\mathbf{X}\right) - \mathbf{y}^\top\mathbf{X}\right)\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} && (19)\\
&\quad \underbrace{-\mathbf{y}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}}_{\text{Constant}}\\
&= \left(\mathbf{w}^\top\left(\mathbf{X}^\top\mathbf{X}\right) - \mathbf{y}^\top\mathbf{X}\right)\left(\mathbf{w} - \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\right) + Constant && (20)\\
&= \left(\mathbf{w}^\top\left(\mathbf{X}^\top\mathbf{X}\right) - \mathbf{y}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^\top\mathbf{X}\right)\right)\left(\mathbf{w} - \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\right) + Constant && (21)\\
&= \left(\mathbf{w}^\top - \mathbf{y}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)\left(\mathbf{X}^\top\mathbf{X}\right)\left(\mathbf{w} - \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\right) + Constant && (22)\\
&= \left(\underbrace{\mathbf{w} - \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}}_{\mathbf{U}}\right)^\top\left(\mathbf{X}^\top\mathbf{X}\right)\left(\underbrace{\mathbf{w} - \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}}_{\mathbf{U}}\right) + Constant && (23)\\
&= \mathbf{U}^\top\mathbf{X}^\top\mathbf{X}\mathbf{U} + Constant && (24)\\
&= (\mathbf{X}\mathbf{U})^\top(\mathbf{X}\mathbf{U}) + Constant && (25)\\
&= ||\mathbf{X}\mathbf{U}||_2^2 + Constant && (26)
\end{aligned}$$

$\mathrm{RSS}(\mathbf{w})$ is minimized when $||\mathbf{X}\mathbf{U}||_2^2 = 0$, i.e.$\mathbf{U} = 0$. Thus, we have

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}. \tag{27}$$

4 About $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$

This section poses the question that what if $\mathbf{X}^\top\mathbf{X}$ is not invertible. We know that $\mathbf{X} \in \mathbb{R}^{N\times(k+1)}$, thus $\mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{(k+1)\times(k+1)}$ is a real symmetric $(k+1)\times(k+1)$ matrix. According to the symmetric eigenvalue decomposition of $\mathbf{X}^\top\mathbf{X}$,

$$\mathbf{X}^\top\mathbf{X} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top, \tag{28}$$

in which $\mathbf{Q} \in \mathbb{R}^{(k+1)\times(k+1)}$ is orthogonal, and $\mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \lambda_2, ..., \lambda_{k+1})$.

If $\mathbf{X}^\top\mathbf{X}$ is invertible,

$$\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^\top, \tag{29}$$

in which $\mathbf{\Lambda}^{-1} = \mathbf{diag}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, ..., \frac{1}{\lambda_{k+1}}\right)$. If $\mathbf{X}^\top\mathbf{X}$ is not invertible, i.e. $\exists\lambda_i \in \{\lambda_1, \lambda_2, ..., \lambda_{k+1}\}$ that $\lambda_i = 0$, in practice, we add some small $\lambda > 0$ so that $\mathbf{\Lambda}^{-1} = \mathbf{diag}\left(\frac{1}{\lambda_1+\lambda}, \frac{1}{\lambda_2+\lambda}, ..., \frac{1}{\lambda_{k+1}+\lambda}\right)$, i.e.

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}. \tag{30}$$

5 Parametric v.s. Non-Parametric Models

As shown in previous sections, we can notice that the model of Linear Regression is based on computing a parameter $\mathbf{w}$, i.e. the mapping from the input to the target variables is determined by this parameter $\mathbf{w}$. We call this kind of Machine Learning models the Parametric Models. In comparison, those without the need to embed the mapping into parameters are called Non-Parametric Models. K-Nearest Neighbor is one example of Non-Parametric Models.

6 Regression with Non-linear Basis

The assumption that the target variables can be represented by the linear combination of the input data features with Gaussian noise may not always hold. For some obvious non-linear samples as shown in Figure 2, we can fit the data by regression with non-linear basis. In the setting of regression with non-linear basis,

$$y_i = \mathbf{w}^\top\phi\left(\mathbf{x_i}\right). \tag{31}$$

$\phi\left(\mathbf{x_i}\right)$ is some non-linear function of $\mathbf{x_i}$, polynomials to name an example. Then,

$$\mathrm{RSS}\left(\mathbf{w}\right) = \sum_{i=1}^{N}\left(y_i - \mathbf{w}^\top\phi\left(\mathbf{x_i}\right)\right)^2 \tag{32}$$

$$= ||\mathbf{y} - \mathbf{\Phi}\mathbf{w}||_2^2, \tag{33}$$

in which $\mathbf{\Phi} = \left(\phi^\top\left(\mathbf{x_1}\right), \phi^\top\left(\mathbf{x_2}\right), ..., \phi^\top\left(\mathbf{x_N}\right)\right)^\top$. Following similar derivation in Section 3, we have

$$\hat{\mathbf{w}} = \left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}. \tag{34}$$

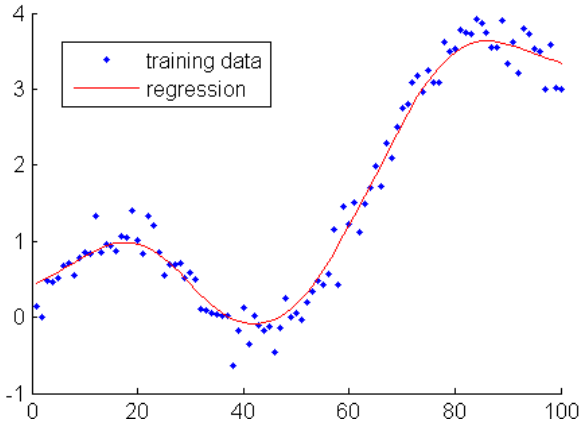

Figure 2: Non-Linear Regression, figure from Wikipedia.

7 Underfitting v.s. Overfitting

The non-linear basis of regression can be of different levels of degree or complexity. The bad choice of the complexity may lead to the problem of underfitting or overfitting. For polynomial basis, the highest order of the polynomial is the obvious representation of model complexity.As shown in Figure 3, different degrees lead to different fitting results.

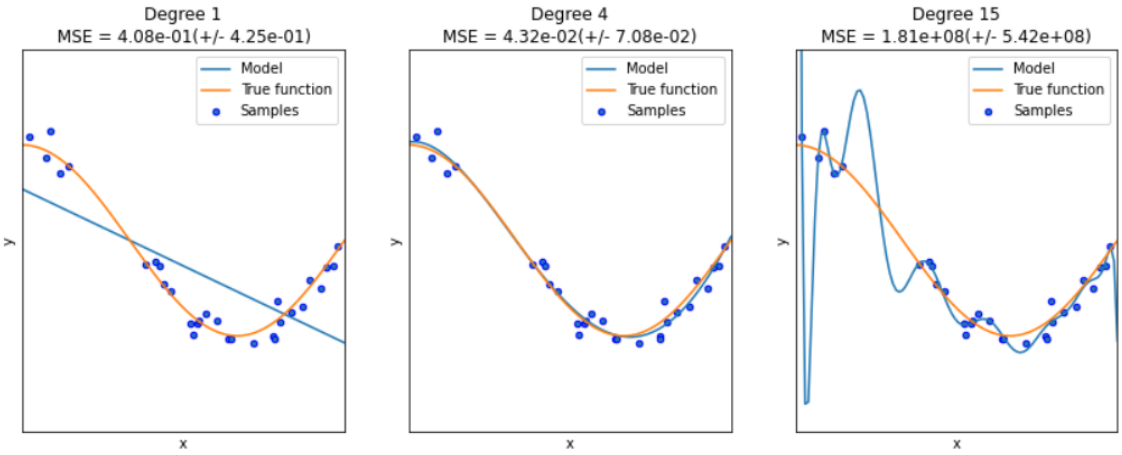

Figure 3: Underfitting v.s. Overfitting, figure from Data Science Foundation.

If the degree is too low, then the basis given is unable to fit the given sample; that is to say the bias of underfitting is high. And the result would not change much no matter how many attempts are made because it reaches the limit of the representation capability of low degree functions. Thus we can say the variance of overfitting is low.

Overfitting, on the other hand, would lure the function to pass every given data points, resulting in a possibly very low bias. How ever, if the given data changes slightly, possibly due to some random noise, the result may change dramatically, which leads to a high variance.

To find out the best complexity of the model, in practice we usually observe the error on both the training set and the evaluation set as the complexity increases. The common cases follow the trend shown in Figure 4.

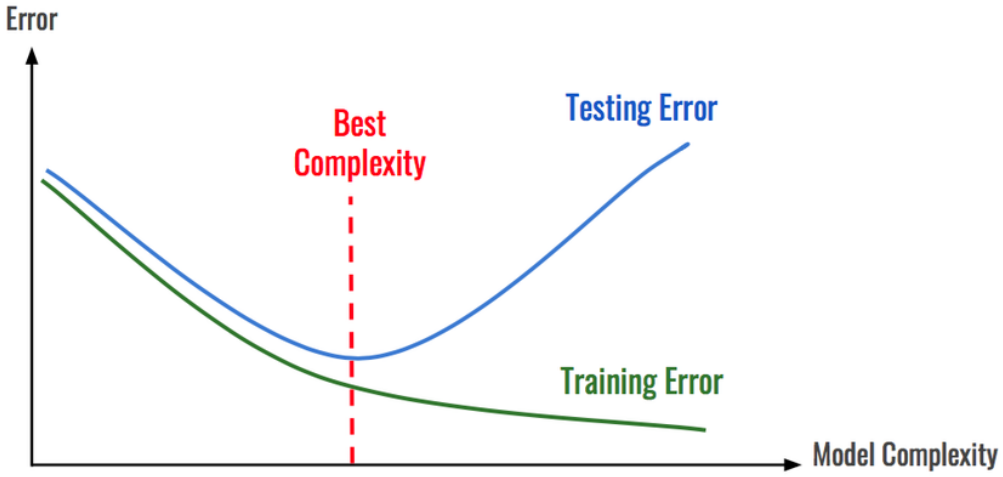

Figure 4: Best Choice of Model Complexity, figure by Robert Martin-Short.

8 Bias-Variance Trade-off

This section studies the bias-variance trade-off in regression models, which can be also regarded as a theoretical and quantitative study of the under/over-fitting problem.

We show that the error of regression is composed by a variance term and a bias term (possibly another noise). We assume that $\exists h(\mathbf{x_i})$ that is the optimal prediction function for predicting the target variable,

$$h(\mathbf{x}) = \mathop{\mathbb{E}}_{y}[\mathbb{P}(y \mid \mathbf{x})] \tag{35}$$

and

$$y_i = h(\mathbf{x_i}) + \epsilon, \tag{36}$$

$\epsilon \sim \mathcal{N}(0, \sigma^2)$. It's important to notice that for any input $\mathbf{x_i}$, $h(\mathbf{x_i})$ is deterministic. Our prediction based on some dataset $\mathcal{D}$ given the linear model in (1) is denoted by $f_{\mathcal{D}}(\mathbf{x_i})$. We can write the expectation of the loss function (Expected Square Loss) as follows:

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathbb{P}(\mathbf{x},y)}[L] = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathbb{P}(\mathbf{x},y)}\left[(y - f_{\mathcal{D}}(\mathbf{x}))^2\right]. \tag{37}$$

The expectation of the error over different datasets is

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{D},(\mathbf{x},y)}[L] &= \mathop{\mathbb{E}}_{\mathcal{D}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim p(\mathbf{x},y)}\left[(y - f_{\mathcal{D}}(\mathbf{x}))^2\right] && (38)\\
&= \mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[(h(\mathbf{x}) + \epsilon - f_{\mathcal{D}}(\mathbf{x}))^2\right] && (39)\\
&= \mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(\underbrace{h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]}_{\mathbf{A}} + \underbrace{\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f_{\mathcal{D}}(\mathbf{x})}_{\mathbf{B}} + \epsilon\right)^2\right] && (40)\\
&= \mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[\mathbf{A}^2 + \mathbf{B}^2 + \epsilon^2 + 2\epsilon\mathbf{A} + 2\epsilon\mathbf{B} + 2\mathbf{A}\mathbf{B}\right] && (41)\\
&= \mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)^2\right] + \mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f_{\mathcal{D}}(\mathbf{x})\right)^2\right] + \mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[\epsilon^2\right] && (42)\\
&\quad + 2\mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[\epsilon\left(h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)\right] + 2\mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[\epsilon\left(\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f_{\mathcal{D}}(\mathbf{x})\right)\right]\\
&\quad + 2\mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)\left(\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f_{\mathcal{D}}(\mathbf{x})\right)\right]\\
&= \left(h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)^2 + \mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f_{\mathcal{D}}(\mathbf{x})\right)^2\right] + \mathbb{E}\left[\epsilon^2\right] && (43)\\
&\quad + 2\left(h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)\mathbb{E}[\epsilon] + 2\mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f_{\mathcal{D}}(\mathbf{x})\right]\mathbb{E}[\epsilon]\\
&\quad + 2\mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)\left(\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f_{\mathcal{D}}(\mathbf{x})\right)\right].
\end{aligned}$$

Given that $\mathbb{E}[\epsilon] = 0$,

$$\mathop{\mathbb{E}}_{\mathcal{D},(\mathbf{x},y)}[L] = \left(h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)^2 + \mathop{\mathbb{E}}_{\mathbf{x}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\left(\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f_{\mathcal{D}}(\mathbf{x})\right)^2\right] + \mathrm{Var}[\epsilon] \tag{44}$$

$$+ 2\mathop{\mathbb{E}}_{\mathbf{x}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\underbrace{\left(h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)}_{Constant}\left(\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f_{\mathcal{D}}(\mathbf{x})\right)\right]$$

$$= \left(h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)^2 + \mathop{\mathbb{E}}_{\mathbf{x}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\left(\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f_{\mathcal{D}}(\mathbf{x})\right)^2\right] + \mathrm{Var}[\epsilon] \tag{45}$$

$$+ 2\left(h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)\underbrace{\mathop{\mathbb{E}}_{\mathbf{x}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\left(\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f_{\mathcal{D}}(\mathbf{x})\right)\right]}_{0}$$

$$= \underbrace{\left(h(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)^2}_{\textbf{Bias}} + \underbrace{\mathop{\mathbb{E}}_{\mathbf{x}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)^2\right]}_{\textbf{Variance}} + \underbrace{\sigma^2}_{\textbf{Noise}}. \tag{46}$$

Thus, the Mean Squared Error can be decomposed into a bias term, a variance term, and a noise term, forming a trade-off relationship, i.e. there's no way one can simultaneously minimize the two sources of error.

9 Measures to Prevent Overfitting: Regression with Regularization

Measures can be taken to prevent overfitting:

- Train the model on larger datasets;
- Control model complexity;
- Add a regularizer to the regression model.

We can add a regularization term to the regression model to control the model complexity and get

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} ||\mathbf{y} - \mathbf{\Phi}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_2^2. \tag{47}$$

To solve this model following similar ways in Section 3, we get

$$\hat{\mathbf{w}} = \left(\mathbf{\Phi}^\top\mathbf{\Phi} + \lambda\mathbf{I}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}. \tag{48}$$