Bias variance trade-off
Supplementary note for CSCI-UA 9473

Augustin Cosse

September 2018

1 Expected loss

Suppose our data set is given by $\mathcal{D} = \{(\boldsymbol{x}, t)\}$ where t are the labels associated to each point $\boldsymbol{x}$. We will assume that the data is distributed according to an ideal function $y(\boldsymbol{x})$ plus some perturbation ε with zero mean and variance σ^2.

$$t = y(\boldsymbol{x}) + \varepsilon \tag{1}$$

t is thus the observed label. In order to study how well a given regression model $h(\boldsymbol{x})$ is fitting the data, one can look at the expected loss

$$\mathbb{E}[L] = \int\int \left(h(\boldsymbol{x}) - t\right)^2 \; p(\boldsymbol{x}, t) \; d\boldsymbol{x} \; dt \tag{2}$$

Note that given the objective (2), the optimal model one can choose for $h(\boldsymbol{x})$ is to solve the minimization problem

$$\min_{h(\boldsymbol{x})} \int\int \left(h(\boldsymbol{x}) - t\right)^2 \; d\boldsymbol{x}dt \tag{3}$$

To do this, we set the first order derivatives to 0, and get

$$\frac{\delta \mathbb{E}[L]}{\delta h(\boldsymbol{x})} = 2\int (h(\boldsymbol{x}) - t)p(\boldsymbol{x}, t) \; d\boldsymbol{x} \; dt = 0 \tag{4}$$

from this we get

$$\int h(\boldsymbol{x})p(\boldsymbol{x}) \; d\boldsymbol{x} = \int\int tp(t, \boldsymbol{x}) \; d\boldsymbol{x} \; dt \tag{5}$$

The t disappeared in the left handside because $h(\boldsymbol{x})$ does not depend on t so we can simply integrate $p(\boldsymbol{x}, t)$ over all values of t which gives $\int p(\boldsymbol{x}, t) \; dt = p(\boldsymbol{x})$.

This last equation tells us that in general we should at every value $\boldsymbol{x}$ choose the model/prediction $h(\boldsymbol{x})$ corresponding to

$$h(\boldsymbol{x}) = \frac{\int t\, p(\boldsymbol{x}, t)\, dt}{p(\boldsymbol{x})} = \int t p(t|\boldsymbol{x})\, dt = \mathbb{E}_t \left\{ t|\boldsymbol{x} \right\}. \tag{6}$$

The last line follows from $p(\boldsymbol{x}, t) = p(t|\boldsymbol{x})p(\boldsymbol{x})$. This is however not always possible. On the other hand, we always have

$$\begin{aligned} (h(\boldsymbol{x}) - t)^2 &= (h(\boldsymbol{x}) - \mathbb{E}\left\{ t|\boldsymbol{x} \right\} + \mathbb{E}\left\{ t|\boldsymbol{x} \right\} - t)^2 & (7) \\ &= (h(\boldsymbol{x}) - \mathbb{E}\left\{ t|\boldsymbol{x} \right\})^2 + (\mathbb{E}\left\{ t|\boldsymbol{x} \right\} - t)^2 + 2\left(h(\boldsymbol{x}) - \mathbb{E}\left\{ t|\boldsymbol{x} \right\}\right)\left(\mathbb{E}\left\{ t|\boldsymbol{x} \right\} - t\right) & (8) \end{aligned}$$

If you substitute this last expression into the expected Loss (2) you can see that the cross terms will disappear as it leads to the following four terms

$$\begin{aligned} 2\int h(\boldsymbol{x})\mathbb{E}\left\{ t|\boldsymbol{x} \right\} p(\boldsymbol{x}, t)\, d\boldsymbol{x}\, dt &= 2\int h(\boldsymbol{x})\mathbb{E}\left\{ t|\boldsymbol{x} \right\} p(\boldsymbol{x})\, d\boldsymbol{x} & (9) \\ -2\int h(\boldsymbol{x}) t p(\boldsymbol{x}, t)\, dt\, d\boldsymbol{x} &= -2\int \mathbb{E}\left\{ t|\boldsymbol{x} \right\} h(\boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x} & (10) \\ -2\int \mathbb{E}\left\{ t|\boldsymbol{x} \right\}^2 p(\boldsymbol{x}, t)\ \ d\boldsymbol{x}\, dt &= -2\int \mathbb{E}\left\{ t|\boldsymbol{x} \right\}^2 p(\boldsymbol{x})\, d\boldsymbol{x} & (11) \\ 2\int \mathbb{E}\left\{ t|\boldsymbol{x} \right\} t p(\boldsymbol{x}, t)\, d\boldsymbol{x}\, dt &= 2\int \mathbb{E}\left\{ t|\boldsymbol{x} \right\}^2 p(\boldsymbol{x})\, d\boldsymbol{x} & (12) \end{aligned}$$

So that the average loss reduces to

$$\begin{aligned} \mathbb{E}[L] = \int (t - h(\boldsymbol{x}))^2\, d\boldsymbol{x}\, dt &= \int (h(\boldsymbol{x}) - \mathbb{E}\left\{ t|\boldsymbol{x} \right\})^2\, p(\boldsymbol{x})\, d\boldsymbol{x} & (13) \\ &\quad + \int (\mathbb{E}\left\{ t|\boldsymbol{x} \right\} - t)^2\, p(\boldsymbol{x}, t)\, d\boldsymbol{x}\, dt & (14) \\ &= \int (h(\boldsymbol{x}) - \mathbb{E}\left\{ t|\boldsymbol{x} \right\})^2\, p(\boldsymbol{x})\, d\boldsymbol{x} & (15) \\ &\quad + \int \sigma_x^2\, p(\boldsymbol{x})\, d\boldsymbol{x} & (16) \end{aligned}$$

Where I used

$$\begin{aligned} \int (\mathbb{E}\left\{ t|\boldsymbol{x} \right\} - t)^2 p(\boldsymbol{x}, t)\, d\boldsymbol{x}\, dt &= \int (\mathbb{E}\left\{ t|\boldsymbol{x} \right\} - t)^2 p(t|\boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x}\, dt & (17) \\ &= \int \mathrm{Var}[t|\boldsymbol{x}] p(\boldsymbol{x})\, d\boldsymbol{x} & (18) \end{aligned}$$

Once again, the t is integrated over in the first integral (13) because none of the functions $h(\boldsymbol{x})$ and $\mathbb{E}\left\{ t|\boldsymbol{x} \right\}$ depend on t anymore.

The final expression for the average loss is then

$$\mathbb{E}[L] = \int (h(\boldsymbol{x}) - \mathbb{E}\left\{ t|\boldsymbol{x} \right\})^2\, p(\boldsymbol{x})\, d\boldsymbol{x} + \int \sigma_x^2\, p(\boldsymbol{x})\, d\boldsymbol{x} \tag{19}$$

The last term is the average noise variance over all $\boldsymbol{x}$.

You see that there is a first term (MSE) that depends on the model (i.e how well the regression model captures the average data distribution $\mathbb{E}\{t|\boldsymbol{x}\} = y(x)$) and a second term that depends only on the data. Now we will show that the first term can decomposed into a *variance* and a *bias* contributions.

2 Bias variance trade-off

In the derivations above, we haven't make any assumption on the regression model learned $h(\boldsymbol{x})$. In general, when we learn a regression model, we learn it from a subset of the data. We will denote this subset $\mathcal{D}_i$ so that $\mathcal{D}_i \subset \mathcal{D}$ and I will now denote our regression model as $h(\boldsymbol{x}|\mathcal{D}_i)$ to emphasize the dependence of the model on the particular choice of $\mathcal{D}_i$. That being said, if we use $\bar{t}(\boldsymbol{x})$ to denote $\mathbb{E}\{t|\boldsymbol{x}\}$, we have the decomposition

$$\begin{aligned}
&\Big(h(\boldsymbol{x};\mathcal{D}_i) - \bar{t}(\boldsymbol{x})\Big)^2 && (20)\\
&= \Big(h(\boldsymbol{x};\mathcal{D}_i) - \mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x};\mathcal{D}_i)\} + \mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x};\mathcal{D}_i)\} - \bar{t}(\boldsymbol{x})\Big)^2 && (21)\\
&= \Big(h(\boldsymbol{x};\mathcal{D}_i) - \mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x};\mathcal{D}_i)\}\Big)^2 + \Big(\mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x};\mathcal{D}_i)\} - \bar{t}(\boldsymbol{x})\Big)^2 && (22)\\
&+ 2\Big(h(\boldsymbol{x};\mathcal{D}_i) - \mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x};\mathcal{D}_i)\}\Big)\Big(\mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x};\mathcal{D}_i)\} - \bar{t}(\boldsymbol{x})\Big) && (23)
\end{aligned}$$

If we take the expectation over all possible datasets $\mathcal{D}_i$ used to learn the model, we will get a measure on how well the family of models $h(\boldsymbol{x},\mathcal{D}_i)$ performs on average for the different datasets. Below I use $\mathbb{E}_{\mathcal{D}_i}$ to denote the average when sampling random datasets $\mathcal{D}_i$ from the full set of samples $\mathcal{D}$.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_i}\big\{h(\boldsymbol{x};\mathcal{D}_i) - \bar{t}(\boldsymbol{x})\big\}^2 && (24)\\
&= \mathbb{E}_{\mathcal{D}_i}\Big(h(\boldsymbol{x};\mathcal{D}_i) - \mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x};\mathcal{D}_i)\}\Big)^2 + \mathbb{E}_{\mathcal{D}_i}\Big(\mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x};\mathcal{D}_i)\} - \bar{t}(\boldsymbol{x})\Big)^2 && (25)\\
&+ 2\mathbb{E}_{\mathcal{D}_i}\Big\{\Big(h(\boldsymbol{x};\mathcal{D}_i) - \mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x};\mathcal{D}_i)\}\Big)\Big(\mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x};\mathcal{D}_i)\} - \bar{t}(\boldsymbol{x})\Big)\Big\} && (26)\\
&= \underbrace{\Big(\mathbb{E}_{\mathcal{D}_i}\{y(\boldsymbol{x};\mathcal{D}_i)\} - \bar{t}(\boldsymbol{x})\Big)^2}_{\text{bias}^2} + \underbrace{\mathbb{E}_{\mathcal{D}_i}\left[\left(h(\boldsymbol{x};\mathcal{D}_i) - \mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x};\mathcal{D}_i)\}\right)^2\right]}_{\text{variance}} && (27)
\end{aligned}$$

The cross term (26) is zero because the second factor does not depend on $\mathcal{D}_i$ (it can be taken out of the average, $\bar{t}(\boldsymbol{x})$ is the unknown regression function

	simpler models	complex models
squared bias	large	small
variance	small	large

Table 1: The bias variance trade-off and the model complexity. As the *model complexity increases*, the *variance* tends to *increase* and the *squared bias* tends to *decrease*. An additional illustration of this phenomenon can be found in Fig. 3.5 in [1] which is available (pdf) online.

and remains the same independently of the data we use to learn the regression model, $\bar{t}(\boldsymbol{x}) = y(\boldsymbol{x})$) and the first term has mean zero.

Complex models will usually have a large contribution to the variance term because within the class of complex models, the models will usually *vary* a lot when you change the training data. I.e if you have a very complex model with many parameters that perfectly fits the data in $\mathcal{D}_i$, then when you will take another dataset $\mathcal{D}_j$, the model that you build for $\mathcal{D}_j$ will be completely different from the model you built with $\mathcal{D}_i$ (think of a high degree polynomial). You will thus have large differences between models $h(\boldsymbol{x}; \mathcal{D}_i)$ and $h(\boldsymbol{x}; \mathcal{D}_j)$ and hence large deviations from the mean $h(\boldsymbol{x}; \mathcal{D}_i) - \mathbb{E}_{\mathcal{D}_i}\{h(\boldsymbol{x}; \mathcal{D}_i)\}$. So complex models will have larger variance.

The bias, on the other hand is a measure of how well the model "captures" the average behavior of the data $\mathbb{E}\{t|\boldsymbol{x}\}$. In general (except if the data is assumed to be very simple or linear which is the framework of the *Gauss Markov Theorem* that I discuss below), the simpler models will not be able to capture such behavior. As an example, imagine we have data of the form

$$t_\ell = y(x_\ell) + \varepsilon_\ell = \sin(x_\ell) + \varepsilon_\ell \tag{28}$$

Then without using non linear terms (and defining our new variables as $x' = \phi(x)$ where $\phi(x)$ encodes the non linearity), it is impossible to design a model of the form $h(x) = \beta_0 + \beta_1 x$ such that

$$h(x_\ell) = \sin(x_\ell) \tag{29}$$

This should illustrate the fact that simpler models have a large bias (*At least when we don't make assumptions on the data distribution*)

This relation is summarized in table 1.

3 The particular case of linearly generated data and the Gauss Markov Theorem

So far we haven't made any assumption on the data. When we don't make assumptions on the data, the relation of table 1 holds in general.

However, when we know that the data has a particular distribution, we can better describe the quality of approximation of given models. This is the framework of the *Gauss Markov Theorem* (GMT).

When the data is distributed according to a linear model, so that

$$t_\ell = y(\boldsymbol{x}_\ell) + \varepsilon = \langle \boldsymbol{\beta}, \boldsymbol{x}_\ell \rangle + \varepsilon_\ell, \tag{30}$$

with $\mathbb{E}\boldsymbol{\varepsilon} = \mathbb{E}[\varepsilon_1, \ldots, \varepsilon_L] = \mathbf{0}$ and $\mathbb{E}\left\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right\} = \sigma^2 \boldsymbol{I}$. Then the linear model has no bias (note that this might seem counterintuitive compared to table 1 but this is because here we assumed that the data is linear).

Moreover, the *Gauss Markov theorem* states that among all possible linear estimators of the $\boldsymbol{\beta}$, the estimator $\hat{\boldsymbol{\beta}}_{\text{RSS}}$ which is computed by minimizing the residual sum of squares,

$$\hat{\boldsymbol{\beta}}_{\text{RSS}} = \underset{\beta}{\operatorname{argmin}} \sum_{\ell=1}^{L} \left(t_\ell - \langle \boldsymbol{\beta}, \boldsymbol{x}_\ell \rangle\right)^2, \tag{31}$$

is the best estimator.

In the Gauss Markov framework, when we say data, we mean the t_ℓ. What we now call data are the noisy measurements $t_\ell = \langle \boldsymbol{\beta}, \boldsymbol{x}_\ell \rangle$. Because we've put ourselves in a framework where we now know how the relation between t_ℓ and $\boldsymbol{x}_\ell$, $t_\ell(\boldsymbol{x}) = \beta_0 + \langle \boldsymbol{\beta}_1, \boldsymbol{x}_\ell \rangle + \varepsilon$, we can focus on denoising the t_ℓ. In this particular framework, a linear estimator for the coefficients $\boldsymbol{\beta}$ is an estimator of the form

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{M}\boldsymbol{t} \tag{32}$$

where $\boldsymbol{t} = [t_1, t_2, \ldots, t_N]$ encode the labels. The estimator provided by the linear regression (i.e. RSS) approach (31) is a particular such estimator as it can read as (see my slides)

$$\hat{\boldsymbol{\beta}}_{\text{RSS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t} \tag{33}$$

$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \tag{34}$$

where $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_L)$ and $\boldsymbol{X}$ is the matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & \boldsymbol{x}_1^T \\ 1 & \boldsymbol{x}_2^T \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_L^T \end{bmatrix} \tag{35}$$

To prove the GMT, we consider any other linear estimator $\tilde{\boldsymbol{\beta}}$ of the coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\text{RSS}} + \Delta\boldsymbol{t}$ where Δ is a perturbation.

We first require the linear estimator $\tilde{\boldsymbol{\beta}}$ to be unbiased, which gives

$$\mathbb{E}\left\{\tilde{\boldsymbol{\beta}}\right\} = \mathbb{E}\left\{\hat{\boldsymbol{\beta}}\right\} + \mathbb{E}\left\{\Delta \boldsymbol{t}\right\} \tag{36}$$

$$= \mathbb{E}\left\{\boldsymbol{\Delta X \beta} + \boldsymbol{\Delta}\varepsilon\right\} \tag{37}$$

$$= \boldsymbol{\Delta X \beta} + 0 \tag{38}$$

$$= 0 \tag{39}$$

The only source of randomness here is the noise ε and we assumed that $\mathbb{E}\varepsilon = 0$. As we haven't made any assumption on $\boldsymbol{\beta}$, the last line implies that for the linear estimator to be unbiased, we must have

$$\Delta \boldsymbol{X} = 0 \tag{40}$$

Now, for the variance, first note that

$$\mathrm{Var}\left\{\hat{\boldsymbol{\beta}}_{\mathrm{RSS}}\right\} = \mathbb{E}\left\{(\hat{\boldsymbol{\beta}}_{\mathrm{RSS}} - \mathbb{E}\hat{\boldsymbol{\beta}}_{\mathrm{RSS}})(\hat{\boldsymbol{\beta}}_{\mathrm{RSS}} - \mathbb{E}\hat{\boldsymbol{\beta}}_{\mathrm{RSS}})^T\right\} \tag{41}$$

$$= \mathbb{E}\left\{\hat{\boldsymbol{\beta}}_{\mathrm{RSS}}\hat{\boldsymbol{\beta}}^T\right\} - \boldsymbol{\beta}\boldsymbol{\beta}^T \tag{42}$$

as the estimator is unbiased. For any other unbiased estimator $\tilde{\boldsymbol{\beta}}$ of the form $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{RSS}} + \Delta \boldsymbol{t}$ with zero bias, we have (the expectation is over the noise $\boldsymbol{\varepsilon}$)

$$\mathrm{Var}\left\{\tilde{\boldsymbol{\beta}}\right\} \tag{43}$$

$$= \mathbb{E}\left\{(\tilde{\boldsymbol{\beta}} - \mathbb{E}\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \mathbb{E}\tilde{\boldsymbol{\beta}})^T\right\} \tag{44}$$

$$= \mathbb{E}\left\{\left(\hat{\boldsymbol{\beta}}_{\mathrm{RSS}} + \boldsymbol{\Delta t} - \boldsymbol{\beta} - \mathbb{E}\left\{\Delta \boldsymbol{t}\right\}\right)\left(\hat{\boldsymbol{\beta}}_{\mathrm{RSS}} + \boldsymbol{\Delta t} - \boldsymbol{\beta} - \mathbb{E}\left\{\Delta \boldsymbol{t}\right\}\right)^T\right\} \tag{45}$$

$$= \mathbb{E}\left\{\left(\hat{\boldsymbol{\beta}}_{\mathrm{RSS}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}_{\mathrm{RSS}} - \boldsymbol{\beta}\right)^T\right\} \tag{46}$$

$$+ \mathbb{E}\left\{(\Delta \boldsymbol{t} - \mathbb{E}\left\{\Delta \boldsymbol{t}\right\})(\Delta \boldsymbol{t} - \mathbb{E}\left\{\Delta \boldsymbol{t}\right\})^T\right\} \tag{47}$$

$$- \mathbb{E}\left\{(\hat{\boldsymbol{\beta}}_{RSS} - \boldsymbol{\beta})(\Delta \boldsymbol{t} - \mathbb{E}\left\{\Delta \boldsymbol{t}\right\})^T\right\} \tag{48}$$

$$- \mathbb{E}\left\{(\Delta \boldsymbol{t} - \mathbb{E}\left\{\Delta \boldsymbol{t}\right\})(\hat{\boldsymbol{\beta}}_{RSS} - \boldsymbol{\beta})^T\right\} \tag{49}$$

The last two (cross) terms disappear simplify as

$$- \mathbb{E}\left\{(\hat{\boldsymbol{\beta}}_{RSS} - \boldsymbol{\beta})(\Delta \boldsymbol{t} - \mathbb{E}\left\{\Delta \boldsymbol{t}\right\})^T\right\} - \mathbb{E}\left\{(\Delta \boldsymbol{t} - \mathbb{E}\left\{\Delta \boldsymbol{t}\right\})(\hat{\boldsymbol{\beta}}_{RSS} - \boldsymbol{\beta})^T\right\} \tag{50}$$

$$= -\mathbb{E}\left\{(\hat{\boldsymbol{\beta}}_{RSS} - \boldsymbol{\beta})(\Delta \boldsymbol{t})^T\right\} - \mathbb{E}\left\{(\Delta \boldsymbol{t})(\hat{\boldsymbol{\beta}}_{RSS} - \boldsymbol{\beta})^T\right\} \tag{51}$$

$$= \mathbb{E}\left\{\hat{\boldsymbol{\beta}}_{RSS}(\Delta \boldsymbol{t})^T\right\} - \mathbb{E}\left\{(\Delta \boldsymbol{t})\hat{\boldsymbol{\beta}}_{RSS}^T\right\} \tag{52}$$

$$= -\mathbb{E}\left\{\hat{\boldsymbol{\beta}}_{RSS}(\Delta \boldsymbol{\varepsilon})^T\right\} - \mathbb{E}\left\{(\Delta \boldsymbol{\varepsilon})\hat{\boldsymbol{\beta}}_{RSS}^T\right\} \tag{53}$$

The second line (51) comes from $\mathbb{E}\{\boldsymbol{\Delta t}\} = \boldsymbol{\Delta X}\boldsymbol{\beta} = 0$, (52) comes from the fact that $\boldsymbol{\beta}$ does not depend on the noise ε and $\mathbb{E}\{\boldsymbol{\Delta t}\} = 0$, the third line (53) follows from developing $\boldsymbol{\Delta t} = \boldsymbol{\Delta X}\boldsymbol{\beta} + \boldsymbol{\Delta\varepsilon}$ and using $\boldsymbol{\Delta X} = 0$. For the last line, developing the expression of $\hat{\boldsymbol{\beta}}_{RSS}$, we get

$$
\begin{aligned}
&- \mathbb{E}\left\{\hat{\boldsymbol{\beta}}_{RSS}(\Delta\boldsymbol{\varepsilon})^T\right\} - \mathbb{E}\left\{(\Delta\boldsymbol{\varepsilon})\hat{\boldsymbol{\beta}}_{RSS}^T\right\} && (54)\\
&= -\mathbb{E}(\boldsymbol{X}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\varepsilon^T\boldsymbol{\Delta}^T && (55)\\
&- \mathbb{E}\boldsymbol{\Delta\varepsilon}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})^T\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X})^{-T} && (56)
\end{aligned}
$$

Then use $\mathbb{E}\boldsymbol{\varepsilon} = 0$ as well as $\mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T = \sigma^2\boldsymbol{I}$ and finally $\boldsymbol{\Delta X} = \boldsymbol{X}^T\boldsymbol{\Delta}^T = 0$. All of this gives

$$
\begin{aligned}
\mathrm{Var}\left\{\tilde{\boldsymbol{\beta}}\right\} &= \mathrm{Var}\left\{\hat{\boldsymbol{\beta}}_{RSS}\right\} + \mathbb{E}\left\{(\Delta\boldsymbol{t})(\Delta\boldsymbol{t})^T\right\} && (57)\\
&\quad - \mathbb{E}\left\{(\Delta\boldsymbol{t})\right\}\mathbb{E}\left\{(\Delta\boldsymbol{t})\right\}^T && (58)\\
&= \mathrm{Var}\left\{\hat{\boldsymbol{\beta}}_{RSS}\right\} + \Delta\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Delta}^T && (59)\\
&\quad + \boldsymbol{\Delta\varepsilon}\boldsymbol{\varepsilon}^T\boldsymbol{\Delta}^T && (60)\\
&= \mathrm{Var}\left\{\hat{\boldsymbol{\beta}}_{RSS}\right\} + \sigma^2\boldsymbol{\Delta}\boldsymbol{\Delta}^T && (61)
\end{aligned}
$$

As soon as $\Delta \neq 0$, the diagonal of $\Delta\Delta^T$ is non zero as well. This in particular means that there is one component of $\tilde{\boldsymbol{\beta}}$ for which we always have

$$
\mathbb{E}_\varepsilon\left\{(\tilde{\beta}_k - \beta_k)^2\right\} > \mathbb{E}_\varepsilon\left\{(\hat{\beta}_k - \beta_k)^2\right\} \tag{62}
$$

Both estimators have zero bias (we will overestimate and underestimate the truth equally) but the errors we make will often be larger when we don't use the RSS estimator.

In conclusion, you can thus see that any linear estimator with no bias will always increase the variance with respect to $\hat{\boldsymbol{\beta}}_{\mathrm{RSS}}$ (This is because we are in the particular framework of linearly generated data)

For this reason, the estimator $\hat{\beta}_{\mathrm{RSS}}$ is sometimes called the BLUE (Best Linear Unbiased Estimator)

3.1 Increasing the bias to reduce generalization under the assumption of linearly generated data

That being said, the fact that the BLUE is the best estimator among all unbiased estimators with respect to linearly generated data, does not mean that it is not possible to do better in terms of future predictions. And this is because a similar *bias variance trade-off* as the one discussed in section 2 applies.

Imagine that we take a new point $\boldsymbol{x}_0$ and we apply our estimator $\tilde{\boldsymbol{\beta}}$ (*now we don't assume that this estimator is unbiased anymore*, this is just a general linear estimator) to this point to get an estimated label $\langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle$ (which is a prediction for $\langle \boldsymbol{\beta}, \boldsymbol{x}_0 \rangle$). We can then write the expected mean square prediction error as before

$$\mathbb{E}_\varepsilon \left\{ \left(\langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle - \langle \boldsymbol{\beta}, \boldsymbol{x}_0 \rangle \right)^2 \right\} \tag{63}$$

$$= \mathbb{E}_\varepsilon \left\{ \left(\langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle - \mathbb{E}\left\{ \langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle \right\} + \mathbb{E}\left\{ \langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle \right\} - \langle \boldsymbol{\beta}, \boldsymbol{x}_0 \rangle \right)^2 \right\} \tag{64}$$

$$= \mathbb{E}_\varepsilon \left\{ \left(\langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle - \mathbb{E}\left\{ \langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle \right\} \right)^2 \right\} \tag{65}$$

$$+ \mathbb{E}_\varepsilon \left\{ \left(\mathbb{E}\left\{ \langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle \right\} - \langle \boldsymbol{\beta}, \boldsymbol{x}_0 \rangle \right)^2 \right\} \tag{66}$$

$$+ 2\mathbb{E}_\varepsilon \left\{ \left(\langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle - \mathbb{E}_\varepsilon \left\{ \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \right\} \right) \left(\mathbb{E}\left\{ \langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle \right\} - \langle \boldsymbol{\beta}, \boldsymbol{x}_0 \rangle \right) \right\} \tag{67}$$

The last term vanishes as the second factor is deterministic (i.e does not depend on ε) and the first factor has mean 0. We thus once again have

$$\mathbb{E}_\varepsilon \left\{ \left(\langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle - \langle \boldsymbol{\beta}, \boldsymbol{x}_0 \rangle \right)^2 \right\} = \underbrace{\mathbb{E}_\varepsilon \left\{ \left(\langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle - \mathbb{E}\left\{ \langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle \right\} \right)^2 \right\}}_{\text{variance}} \tag{68}$$

$$+ \underbrace{\mathbb{E}_\varepsilon \left\{ \left(\mathbb{E}\left\{ \langle \tilde{\boldsymbol{\beta}}, \boldsymbol{x}_0 \rangle \right\} - \langle \boldsymbol{\beta}, \boldsymbol{x}_0 \rangle \right)^2 \right\}}_{\text{bias}^2} \tag{69}$$

This shows that even in the framework of linearly generated data, despite the fact that $\hat{\boldsymbol{\beta}}$ is the BLUE estimator, it does not mean that this estimator will be the best at predicting a new value of $t = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle$ (on average over the noise). There might be an estimator $\tilde{\boldsymbol{\beta}}$ with a non zero bias (on linear data) that will have smaller variance.

If I generate many values $t_k = \langle \boldsymbol{\beta}, \boldsymbol{x}_k \rangle + \varepsilon_k$, learn my estimator $\hat{\boldsymbol{\beta}}$ from those values, then this estimator will be unbiased but when I will want to get the prediction for a new point x_0, the average prediction might be better if I lower the variance and increase the bias a little. Again this might sound counter intuitive with respect to table 1 but remember that we assumed linearly distributed data here.

3.2 To go further: Ridge regression*

Another way to understand this bias-variance trade-off in the case of linear data, is to look at the variance of the RSS estimator $\hat{\boldsymbol{\beta}}_{\text{RSS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})$ estimator. This variance (42) has the form

$$\mathbb{E}\left\{\hat{\boldsymbol{\beta}}_{\text{RSS}}\hat{\boldsymbol{\beta}}_{\text{RSS}}^T\right\} - \mathbb{E}\left\{\hat{\boldsymbol{\beta}}_{\text{RSS}}\right\}(\mathbb{E}\left\{\hat{\boldsymbol{\beta}}_{\text{RSS}}\right\})^T \tag{70}$$

$$= \mathbb{E}\left\{(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-T}\right\} - \boldsymbol{\beta}\boldsymbol{\beta}^T \tag{71}$$

When variables (entries in $\boldsymbol{x} = [x_1, \ldots, x_L]$) are correlated, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ will be badly scaled, i.e. $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ will be very large, and the prediction error for this estimator will thus be big. A biased alternative (see my slides for Lecture 3) is to add a regularization (for example ridge regression). In this case, the variance becomes

$$\mathbb{E}\left\{(\tilde{\boldsymbol{\beta}}_{RR} - \mathbb{E}\tilde{\boldsymbol{\beta}}_{RR})(\tilde{\boldsymbol{\beta}}_{RR} - \mathbb{E}\tilde{\boldsymbol{\beta}}_{RR})^T\right\} \tag{72}$$

$$= \mathbb{E}\left\{(\boldsymbol{X}\boldsymbol{X}^T + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})^T\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^T + \lambda\boldsymbol{I})^{-T}\right\} \tag{73}$$

$$- \mathbb{E}\tilde{\boldsymbol{\beta}}_{RR}(\mathbb{E}\tilde{\boldsymbol{\beta}}_{RR})^T \tag{74}$$

which will be smaller than inverting a badly conditioned matrix because of the identity which is well conditioned.

References

[1] C. M. Bishop. Pattern recognition and machine learning. *Springer*, 2006.