Problem 1 (C, For Tutorials on 06.11 and 07.11). **Least Squares.**
Consider a simple linear regression with RSS as the error measure,

$$L(\beta_0, \beta) = \frac{1}{n} \sum_{i}^{N} \left(y_i - \beta_0 - x_i \beta\right)^2 \tag{1.1}$$

for a target y and single predictor $X \in \mathbb{R}^n$.

1. Show that the minimizers of Eq.(1.1) are given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x} \ ,$$
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

   where $\bar{x}, \bar{y}$ denote the sample means $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

   Hint: Find $\hat{\beta}_0$ first and substitute it into the expression for $\hat{\beta}$ to obtain the above.

2. Consider the example

$$\mathbf{X} = \begin{bmatrix} 2 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

   and the map

$$P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \ .$$

   (a) In a sketch, visualize the column space of $\mathbf{X}$ as a plane in $\mathbb{R}^3$.
   Reminder: the column space $S \in \mathbb{R}^3$ is spanned by the column vectors $X^{(i)}$ of $\mathbf{X}$,

$$S = \text{span}(X^{(1)}, X^{(2)}) = \{a_1 X^{(1)} + a_2 X^{(2)} \mid a_i \in \mathbb{R}\} \ .$$

   (b) Add the vector y, as well as the vector $y - Py$ to your drawing and interpret the meaning of P. What quantity is being minimized?

Solution.

1. To find β_0, we consider

$$\frac{\delta L}{\delta \beta_0} = \frac{1}{n}\sum_{i=1}^{n} 2 \cdot (y_i - x_i\beta - \beta_0)(-1)$$

and set it to zero,

$$\begin{aligned} 0 &= \frac{\delta L}{\delta \beta_0} \\ 0 &= \sum_{i=1}^{n} (y_i - x_i\beta - \beta_0) \\ \sum_{i=1}^{n} \beta_0 &= \sum_{i=1}^{n} (y_i - x_i\beta) \\ n \cdot \beta_0 &= \sum_{i=1}^{n} (y_i - x_i\beta) \\ \beta_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i\beta = \bar{y} - \beta\bar{x} \end{aligned} \tag{1.2}$$

Similarly, we consider

$$\begin{aligned} \frac{\delta L}{\delta \beta} &= \frac{\delta}{\delta \beta}\frac{1}{n}\sum_{i=1}^{n} \big(y_i - (\bar{y} - \bar{x}\beta) - x_i\beta\big)^2 \quad \text{using Eq. (1.2)} \\ &= \frac{\delta}{\delta \beta}\frac{1}{n}\sum_{i=1}^{n} \big((y_i - \bar{y}) - (x_i - \bar{x})\beta\big)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} 2\big((y_i - \bar{y}) - (x_i - \bar{x})\beta\big) \cdot (x_i - \bar{x})(-1) \end{aligned}$$

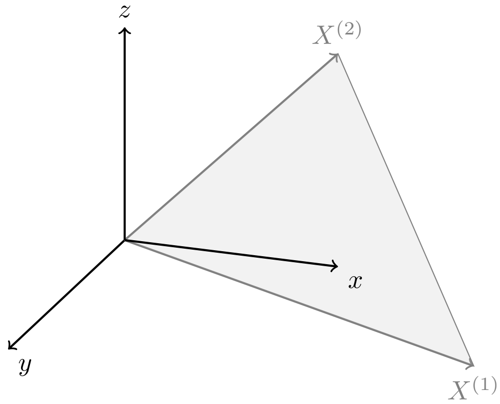
Setting this to zero, we obtain

$$\begin{aligned} 0 &= \frac{\delta L}{\delta \beta} \\ 0 &= \sum_{i=1}^{n} \big((y_i - \bar{y}) - (x_i - \bar{x})\beta\big) \cdot (x_i - \bar{x}) \\ 0 &= \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + \sum_{i=1}^{n} (x_i - \bar{x})^2\beta \end{aligned}$$

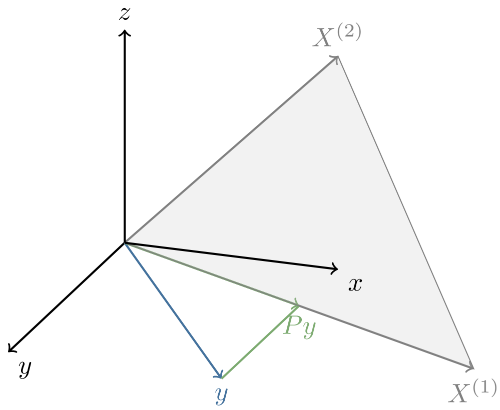
The result is

$$\beta = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \, .$$

2. (a) The vectors $X^{(1)} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}, X^{(2)} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ span the plane shown below,



(b) Py is the vector on our plane closest to y, and the connecting vector $y - Py$ is perpendicular to the plane. The OLS solution minimizes the distance $\|y - Py\|$.



Problem 2 (C, For Tutorials on 13.11 and 14.11). **Bias and Variance.**
Consider the bias and variance of a linear regression model f.

1. State the definitions of bias and variance.
2. Show that the following holds,

$$\mathbb{E}\left[(y_0 - \hat{f}(x_0))^2\right] = \mathbb{E}\left[(f(x_0) - \hat{f}(x_0))^2\right] + \mathrm{Var}(\epsilon)\ .$$

3. For k-Nearest Neighbor Regression (KNN), one can show that the following relationship holds,

$$\mathbb{E}[(y_0 - \hat{f}(x_0))^2] = \left(f(x_0) - \frac{1}{k}\sum_{i=0}^{k} f(N_i(x_0))\right)^2 + \frac{\sigma^2}{k} + \sigma^2\ .$$

   where $N_1(x_0), ..., N_k(x_0)$ are the k nearest neighbors of the sample x_0 and $\sigma^2 = \mathrm{Var}(\hat{f}(x_0))$. Conclude from this the influence of the parameter k on bias and variance.

4. Explain the difference between reducible and irreducible error.

Solution.
For brevity, we use $f = f(x_0)$, $\hat{f} = \hat{f}(x_0)$ and $y_0 = f + \epsilon$.

1. We have $\text{Var}(\hat{f}) = \mathbb{E}(\hat{f}^2) - \left[\mathbb{E}(\hat{f})\right]^2$ and $\text{Bias}(\hat{f}) = \mathbb{E}(\hat{f} - f)$.

2. We now simplify $\mathbb{E}\left[(y_0 - \hat{f})^2\right]$:

$$\begin{aligned}\mathbb{E}\left[(y_0 - \hat{f})^2\right] &= \mathbb{E}\left[(f + \epsilon - \hat{f})^2\right] \\ &= \mathbb{E}\left[\left((f - \hat{f}) + \epsilon\right)^2\right] \\ &= \mathbb{E}\left[(f - \hat{f})^2 + 2(f - \hat{f})\epsilon + \epsilon^2\right] \\ &= \mathbb{E}\left[(f - \hat{f})^2\right] + 2\mathbb{E}\left[(f - \hat{f})\epsilon\right] + \mathbb{E}\left[\epsilon^2\right]\end{aligned}$$

The last term is $\text{Var}(\epsilon)$ since $\text{Var}(\epsilon) = \mathbb{E}\left[(\epsilon - \mathbb{E}\left[\epsilon\right])^2\right] = \mathbb{E}\left[\epsilon^2\right] + [\mathbb{E}(\epsilon)]^2$ and we assume $\mathbb{E}(\epsilon) = 0$. The second term equals to zero because ϵ is independent of $(f - \hat{f})$, so $2\mathbb{E}\left[(f - \hat{f})\epsilon\right] = 2\mathbb{E}\left[(f - \hat{f})\right]\mathbb{E}\left[\epsilon\right] = 0$ with again the assumption $\mathbb{E}(\epsilon) = 0$.
Hence,

$$\mathbb{E}\left[(y_0 - \hat{f})^2\right] = \mathbb{E}\left[(f - \hat{f})^2\right] + \text{Var}(\epsilon).$$

3. Here, the first term is the bias which as we can see is monotonely increasing with the parameter k, that is, the more neighbors we allow by setting the hyperparameter k in our model, the more biased the model will be. Conversely, the the variance in the remaining term decreases as we increase k.

4. The accuracy of a prediction $\hat{Y}$ for ground truth Y depends on two quantities, the reducible and the irreducible error. The **reducible error** refers to the error resulting from the fact that a learned model is not be a perfect estimate for the true relationship. This error can be reduced by a better fit of the algorithm. The **irreducible error** refers to noise that cannot be reduced by a better fit of the algorithm (*even if it were possible to find the perfect true model*). This is, because Y is also a function of ϵ, which, by definition, cannot be predicted using X. *The irreducible* noise may come from unmeasured variables. There might be useful features in predicting Y, but if we don't measure them, the model cannot use them for its prediction. For example, the risk of an adverse reaction of a drug may depend on the patient's general feeling of well-being on the day given. Note: the irreducible error provides an upper bound on the accuracy of our prediction for Y.