This supplement adds information about loss functions. In particular, we first contrast the \( \ell_1 \) and \( \ell_2 \) loss functions, and show that the former is good for dealing with large outliers, while the latter yields the expected value, which is often useful when outliers are less of a concern. Next, we derive the optimal univariate linear estimator.

**Median vs. mean estimators via example:** To contrast the mean estimator, which is optimal for minimizing the \( \ell_2 \) error, with the median estimator designed for \( \ell_1 \) error, we consider a simple example. Consider the data set \( Y = \{ 2, 1, 1.5, 2, 17 \} \). Our goal is to find a model that predicts values of \( Y \). This will be a trivial model, we will compute a constant \( \beta \), which will be used as our predictor.

Let us begin by optimizing the squared error. We have

\[
\text{Error}(\beta) = \| Y - \beta \|_2^2 = \sum_{n=1}^{N} (y_n - \beta)^2.
\]

Note that \( N = 5 \), and so

\[
\text{Error}(\beta) = \sum_{n=1}^{5} (y_n - \beta)^2
= (y_1 - \beta)^2 + (y_2 - \beta)^2 + (y_3 - \beta)^2 + (y_4 - \beta)^2 + (y_5 - \beta)^2
= (2 - \beta)^2 + (1 - \beta)^2 + (1.5 - \beta)^2 + (2 - \beta)^2 + (17 - \beta)^2
= 5\beta^2 - 2\beta(2 + 1 + 1.5 + 2 + 17) + (2^2 + 1^2 + 1.5^2 + 2^2 + 17^2).
\]

We are interested in finding \( \beta^* \) that minimizes the error. To do so, we compute the derivative of \( \text{Error}(\beta) \),

\[
\frac{\partial}{\partial \beta} \text{Error}(\beta) = 5 \cdot 2\beta - 2\beta(2 + 1 + 1.5 + 2 + 17) = 10\beta - 47.
\]

To drive the derivative to zero, we need \( 10\beta^* = 47 \), and so \( \beta^* = 4.7 \). We can see that this optimal \( \beta^* \) is the expected value or average of \( Y \),

\[
\frac{1}{N} \sum_{n=1}^{N} Y_n = \frac{1}{5}(2 + 1 + 1.5 + 2 + 17) = 4.7.
\]

In this example, \( \beta^* = 4.7 \) may seem like a reasonable guess for the next value of \( Y \). But suppose instead that we had \( N = 10^4 \) data points, one of them was an outlier whose value
was quite large, while the rest of $Y$ was comprised of points in the interval $(0, 3)$. A value outside this main range might seem weird. The take home point here is that the square error is sensitive to outliers.

To combat the effect of outliers, let us consider minimizing the $\ell_1$ error function,

$$\text{Error}(\beta) = \|Y - \beta\|_1 = \sum_{n=1}^{N} |y_n - \beta|.$$  

Substituting the values of $Y$ as before (1),

$$\text{Error}(\beta) = \sum_{n=1}^{5} |y_n - \beta| = |y_1 - \beta| + \ldots + |y_5 - \beta| = |2 - \beta| + |1 - \beta| + |1.5 - \beta| + |2 - \beta| + |17 - \beta|.$$  (2)

We find the $\beta^*$ that minimizes the error function using Matlab as follows.

```matlab
beta=-10:0.001:10;
error=abs(2-beta)+abs(1-beta)+abs(1.5-beta)+abs(2-beta)+abs(17-beta);
plot(beta,error)
```

It can be seen visually that the minimum occurs for $\beta^* = 2$. This is the median value of $Y$. The reason why the error function is minimal at the median, recall that the derivative of each absolute value term of the form $|y_n - \beta|$ is -1 when $y_n < \beta$, else +1 when $y_n > \beta$. Therefore, whenever $\beta$ is below the median, at least $\frac{N}{2}$ points in $Y$ are greater than $\beta$, yielding slope more negative than $-\frac{N}{2}$; similarly, less than $\frac{N}{2}$ points are smaller than $\beta$, contributing less than $+\frac{N}{2}$ to the derivative, and the sum derivative is negative. Similarly, $\beta$ above the median yields a positive derivative. The only point where the derivative could be zero, which is the minimum, occurs at the median. While the median is less prone to outliers, there are many distributions of data for which the mean nonetheless seem better.

**Optimal univariate linear estimator:** Consider the simple univariate linear model, $Y = X\beta + \epsilon$, where $\epsilon$ is considered to be noise, and our goal is to compute the optimal $\beta$ that minimizes the squared error. That is, we will have an optimal univariate linear estimator, also known as a linear regressor.

To find the optimal regressor, we evaluate our squared error loss function,

$$\text{Error}(\beta) = \|Y - X\beta\|^2_2 = \sum_{n=1}^{N} (y_n - x_n\beta)^2 = \left[ \sum_{n=1}^{N} (y_n)^2 \right] - 2\beta \left[ \sum_{n=1}^{N} x_n y_n \right] + \beta^2 \left[ \sum_{n=1}^{N} (x_n)^2 \right] = \langle Y, Y \rangle - 2\beta \langle X, Y \rangle + \beta^2 \langle X, X \rangle.$$  (3)
To find optimal $\beta^*$, we drive the derivative to zero,

$$\frac{\partial}{\partial \beta} \text{Error}(\beta) = -2\langle X, Y \rangle + 2\beta \langle X, X \rangle = 0.$$ 

We see that

$$\beta^* = \frac{\langle X, Y \rangle}{\langle X, X \rangle},$$

which completes our derivation.