EE 126 Probability and Random Processes
April 12, 2018 MMSE

1 Lecture

SNR / LLSE

Exercise. Let $Y = \alpha X + Z$, where X and Z are independent and zero-mean. Given Y, we would like to produce an estimate $\hat{X}$ of X according to an LLSE formulation, i.e. $\hat{X} = L[X \mid Y] = bY$.

$$\begin{aligned} L[X \mid Y] &= \frac{cov(X,Y)}{var(Y)}Y \\ &= \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}Y \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}[XY] &= \mathbb{E}[\alpha X^2 + XZ] \\ &= \alpha\mathbb{E}[X^2] + \mathbb{E}[XZ] \\ &= \alpha\mathbb{E}[X^2] \\ \mathbb{E}[Y^2] &= \mathbb{E}[(\alpha X + Z)^2] \\ &= \mathbb{E}[\alpha^2 X^2 + 2\alpha XZ + Z^2] \\ &= \alpha^2\mathbb{E}[X^2] + \mathbb{E}[Z^2] \end{aligned}$$

And therefore

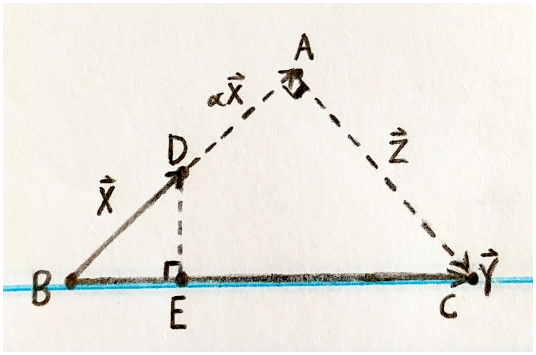
$$\begin{aligned} L[X \mid Y] &= \frac{\alpha\mathbb{E}[X^2]}{\alpha^2\mathbb{E}[X^2] + \mathbb{E}[Z^2]}Y \\ &= \frac{\alpha^{-1}}{1 + \frac{\mathbb{E}[Z^2]}{\alpha^2\mathbb{E}[X^2]}}Y \\ &= \frac{\alpha^{-1}}{1 + \frac{1}{\text{SNR}}}Y \end{aligned}$$

in which the **signal-to-noise ratio (SNR)** is defined as

$$\begin{aligned} \text{SNR} &= \frac{\alpha^2\mathbb{E}[X^2]}{\mathbb{E}[Z^2]} \\ &= \frac{\text{signal strength}}{\text{noise strength}} \end{aligned}$$

- *What if the SNR is really bad ($\ll 1$, where the signal is swamped by the noise)?*
  Then $L[X \mid Y] = 0$; we essentially ignore the signal and depend on the priors.
- *What if the SNR is really good ($\gg 1$, where the signal is strong)?*
  Then $L[X \mid Y] \propto \alpha^{-1}Y$.

Geometry of Exercise



$Y = \alpha X + Z$, so αX must extend s.t. the segment connecting it and the end of Y forms a right angle with αX.

- $\hat{X} = bY = \overline{BE}$; using similar triangles we see that

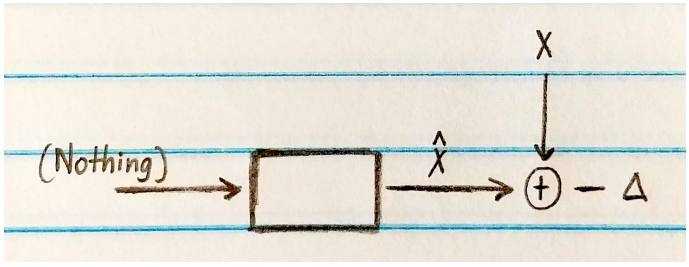
$$\frac{\overline{BE}}{\overline{BD}} = \frac{\overline{BA}}{\overline{BC}} \quad \text{because } BED \approx BAC$$
$$\overline{BE} = \frac{\overline{BA}}{\overline{BC}} \cdot \overline{BD}$$

Then

$$\begin{aligned} b &= \overline{BE}\frac{1}{\|Y\|} \\ &= \frac{\overline{BE}}{\overline{BC}} \\ &= \frac{\overline{BA} \cdot \overline{BD}}{\overline{BC}^2} \\ &= \frac{\alpha\|X\|\|X\|}{\|Y\|^2} \\ &= \frac{\alpha\|X\|^2}{\alpha^2\|X\|^2 + \|Z\|^2} \quad \text{as before.} \end{aligned}$$

Suppose We Observe Nothing

What's going on with $L[X \mid Y] = 0$? Suppose we observe nothing:



$$\text{Then } \min_{\hat{X}} \mathbb{E}[(X - \hat{X})^2] \;\rightarrow\; \hat{X} = \mathbb{E}[X]$$

i.e. we should guess the mean (0). For $Y = \alpha X + Z$, observing nothing is equivalent to just observing the noise Z (so α ends up being really small).

Non-Bayesian View of Linear Regression

In the frequentist view, we have access to the $(X_1, Y_1), ..., (X_k, Y_k)$ data points. Our goal is to construct $g(Y) = a+bY$ such that

$$\frac{1}{k}\sum_{i=1}^{k} |X_i - (a + bY_i)|^2 \quad \text{is minimized}$$

i.e. find the best fit such that the squared error is minimized.

- *How is this different from Bayesian?* This is the empirical version of the Bayesian view where each data point is equally likely (we don't know the distributions of X and Y; these are assumed to be uniform).

$$\xi(a, b) = \frac{1}{k}\sum_{i=1}^{k} |X_i - (a + bY_i)|^2$$

If we take derivatives $\frac{\partial \xi}{\partial a}$ and $\frac{\partial \xi}{\partial b}$ and optimize by setting them to 0, we find that

$$a + bY = \mathbb{E}_k[X] + \frac{cov(X, Y)}{var(Y)}(Y - \mathbb{E}_k[Y])$$

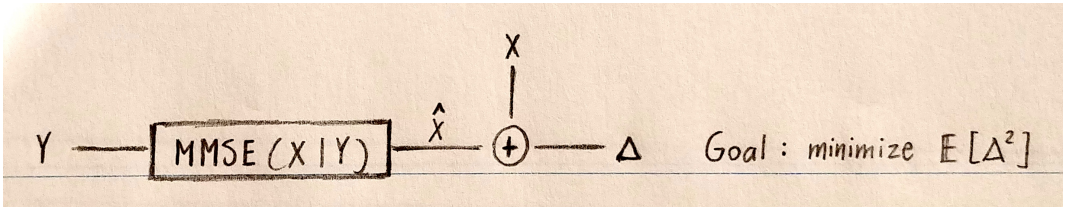
where

$$E_k[X] = \frac{1}{k}\sum_{i=1}^{k} X_i$$
$$E_k[Y] = \frac{1}{k}\sum_{i=1}^{k} Y_i$$

In other words, for this case the frequentist view of "just going by the data" is the same as the Bayesian construction.

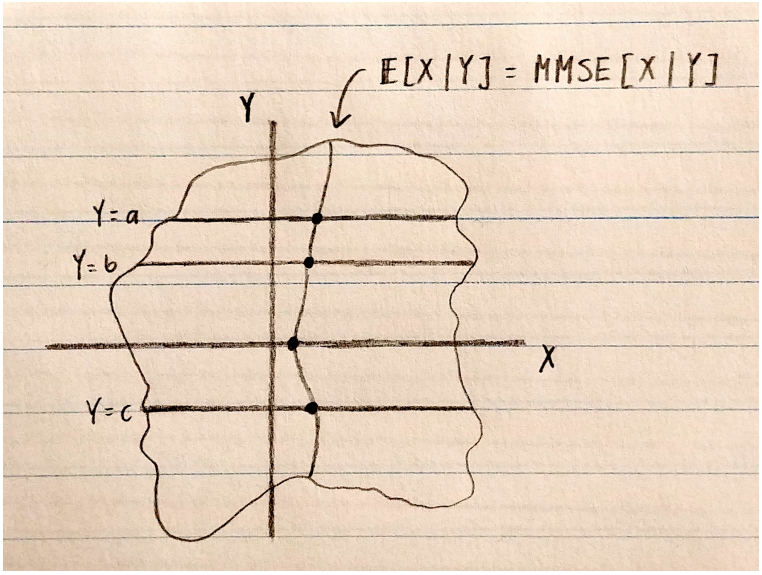
MMSE

In MMSE, we no longer have any linearity restrictions.



Theorem 7.4. *The MMSE estimate of $X \mid Y$ is $\mathbb{E}[X \mid Y]$.*

As some intuition for this, imagine that X is uniformly distributed over the cloud depicted in the following figure (as in "X is uniformly distributed along each line").

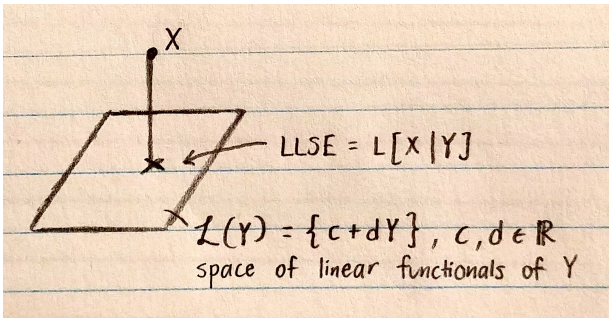


Then the mean of X given Y is the midpoint of the line at $Y = c$ for all c. Hence $\mathbb{E}[X \mid Y]$ is the curve generated by "connecting" all of the midpoints.

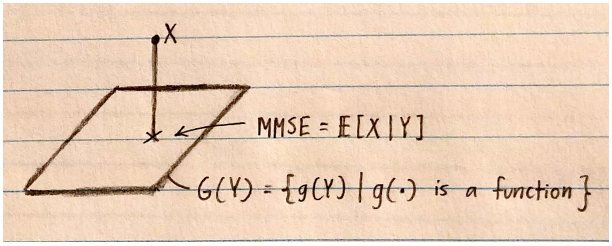
Recall that

$$\begin{aligned}\mathbb{E}[X \mid Y = y] &= \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\, dx \\ &= \int_{-\infty}^{\infty} x \frac{f_{XY}(x,y)}{f_Y(y)}\, dx\end{aligned}$$

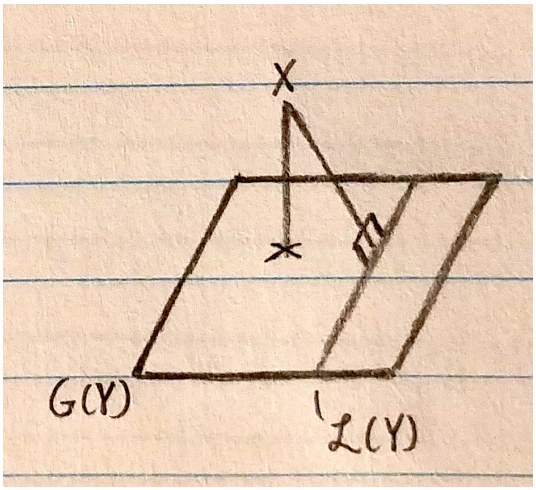
The linear estimate is



whereas the MMSE estimate is



Note that $\mathcal{L}(Y) \subset G(Y)$:



The LLSE is the projection of X onto the subspace $\mathcal{L}(Y)$. Note that this "distance" on $G(Y)$ is $\geq$ the MMSE estimate.

Lemma 7.6. *For any function $\phi(\cdot)$,*

(a) $\mathbb{E}[\underbrace{(X - \mathbb{E}[X \mid Y])}_{\Delta} \cdot \phi(Y)] = 0$

(b) If there exists a $g(Y)$ such that $\mathbb{E}[(X - g(Y)) \cdot \phi(Y)] = 0$, then $g(Y) = \mathbb{E}[X \mid Y]$.

(a) says that the error Δ is orthogonal to any function of y (as demonstrated by the previous figure).
(b) says that the orthogonal projection is unique.

Note: these statements are at the optimal point.

Proof of Theorem 7.4

Recall: theorem 7.4 states that $\text{MMSE}[X \mid Y] = g(Y) = \mathbb{E}[X \mid Y]$. As proof, we can show that

$$\mathbb{E}[(X - h(Y))^2] \geq \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2]$$

for general $h(Y)$. We begin with the left hand side, and find that

$$\begin{aligned}
\mathbb{E}[(X - h(Y))^2] &= \mathbb{E}[(X - \mathbb{E}[X \mid Y] + \mathbb{E}[X \mid Y] - h(Y))^2] \\
&= \mathbb{E}[\underbrace{(X - \mathbb{E}[X \mid Y])^2}_{\text{RHS}} + \underbrace{2\underbrace{(X - \mathbb{E}[X \mid Y])}_{\Delta}\underbrace{(E[X \mid Y] - h(Y))}_{\text{some } g(Y)}}_{= 0 \text{ by Lemma 7.6 and linearity of expectation}} + \underbrace{(\mathbb{E}[X \mid Y] - h(Y))^2}_{\geq 0}] \\
&= \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2] + \underbrace{\mathbb{E}[(\mathbb{E}[X \mid Y] - h(Y))^2]}_{\geq 0} \\
&\geq \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2]
\end{aligned}$$

as desired. Therefore $g(Y) = \mathbb{E}[X \mid Y]$ best minimizes the MSE.