Predicting the value of $Y$
- $Y$ is a RV with known pmf or pdf
- Problem: Predict (or estimate) what value of $Y$ will be observed on the next trial
- What value should we predict?
- What is a good prediction?
- We need to specify some criterion that determines what is a good/reasonable estimate
- Else any estimate is just as good as any other estimate

Minimize probability of error – I
- $\hat{e}$ denotes our estimate of the value of $Y$
- $\hat{e}$ is a number that we get to choose
- Minimum-probability-of-error criterion: choose $\alpha$ so as to minimize $P(Y \neq \hat{e})$
- If $Y$ is a discrete random variable, then $P(Y = \hat{e}) = 1 - P(Y = \hat{e}) = 1 - p_Y(\hat{e})$
- Choose $\hat{e}$ to be the location of the maximum of the pmf $p_Y(u)$
- Our estimate is wrong with probability $1 - p_Y(\hat{e})$

Minimize probability of error – II
- If $Y$ is a continuous random variable, then $P(Y = \hat{e}) = 1 - P(Y = \hat{e}) = 1$ no matter what we number we choose as $\hat{e}$
- Alternative: choose $\hat{e}$ to minimize $P(|Y - \hat{e}| > \varepsilon)$ for some suitable (small) choice of $\varepsilon$
- $P(|Y - \hat{e}| > \varepsilon) = 1 - P(|Y - \hat{e}| \leq \varepsilon)$ and so we want to maximize $P(|Y - \hat{e}| \leq \varepsilon)$
- $P(|Y - \hat{e}| \leq \varepsilon) = P(\hat{e} - \varepsilon \leq Y \leq \hat{e} + \varepsilon)$
- $= F_Y(\hat{e} + \varepsilon) - F_Y(\hat{e} - \varepsilon)$

Minimize probability of error – III
- $P(|Y - \hat{e}| \leq \varepsilon) = F_Y(\hat{e} + \varepsilon) - F_Y(\hat{e} - \varepsilon)$ has derivative $f_Y(\hat{e} + \varepsilon) - f_Y(\hat{e} - \varepsilon) = 0$ if $\hat{e}$ is chosen such that $f_Y(\hat{e} + \varepsilon) = f_Y(\hat{e} - \varepsilon)$
- Graphically, find a horizontal “chord” of length $2\varepsilon$ under the “peak” of the pdf: the midpoint of the chord is $\hat{e}$

What’s apple pie à la mode?
- In the limit as $\varepsilon \to 0$, $\hat{e}$ approaches the location of the maximum value of the pdf
- For both continuous and discrete RVs, we get the location of the maximum of the pdf or the pmf
- The location of the maximum of the pdf or pmf is called the mode of the pdf/pmf
- It is the value of $Y$ that has the “maximum probability” of occurring
- Mode = most fashionable or most frequent

Large errors are worse than small
- If our estimate is $\hat{e}$, then we make an estimation error of $Y - \hat{e}$
- Cost of making this error is $|Y - \hat{e}|$
- Large estimation errors cost us more than small estimation errors
- The absolute estimation error is $|Y - \hat{e}|$
- Average absolute estimation error (or average cost) is $E[|Y - \hat{e}|]$
- $E[|Y - \hat{e}|]$ is minimized if $\hat{e}$ is chosen to be the median value of $Y$
What’s the median of Y?

- The median of Y is usually defined to be the number m such that F_Y(m) = 1/2
- Not satisfactory: It is possible F_Y(u) ≠ 1/2 for any choice of u, e.g., if Y is a Bernoulli RV with parameter p ≠ 1/2, or it might be that F_Y(u) = 1/2 for all u in some interval
- Median = m such that P(Y ≤ m) ≥ 1/2 and P(Y ≥ m) ≥ 1/2 for discrete RV
- Median = m such that F_Y(m) = 1/2 for continuous RV (or midpoint of interval)

The mean minimizes E[(Y–ê)²]

- Let E[Y] = µ
- E[(Y–ê)²] = E[(Y–µ + µ–ê)²]
  = E[(Y–µ)² + 2(Y–µ)(µ–ê) + (µ–ê)²]
  = var(Y) + 2(µ–ê)E[Y–µ] + (µ–ê)²
  = var(Y) + (µ–ê)² > var(Y) if ë ≠ µ
- Choosing ë = µ minimizes the mean-square error of the estimate
- ë = µ is called minimum (or least) mean-square error (MMSE or LMMSE) estimate
- The minimum mean-square error is var(Y)

The median minimizes E[|Y–ê|]

- E[|Y–ê|] = ∫ (ê–v)f_Y(v)dv + 1/2 (v–ê)f_Y(v)dv
  = F_Y(ê)(1–F_Y(ê)) = 0 when F_Y(ê) = 1/2
- The derivative of E[|Y–ê|] with respect to ë is given by dE[|Y–ê|]/dë = ∫ f_Y(v)dv + 1/2 ∫ (v–ê)f_Y(v)dv
- Large errors are worser than small
  - The absolute estimation error is |Y – ê|
  - Large estimation errors cost us more if the cost function is (Y – ê)²
  - But, since x² < |x| for |x| < 1, small estimation errors are underpenalized in comparison to the cost function |Y – ê|
  - The mean-square estimation error (more simply, mean-square error) is E[(Y – ê)²]
  - E[(Y – ê)²] is minimized if ë is chosen to be the mean value of Y

Upping the ante

- Let X and Y denote random variables with known joint distribution
- When the experiment is performed, we observe random point (X, Y) in the plane
- Now suppose that the value of X becomes known to us, but not the value of Y
- What is the MMSE estimate of Y?
- Ostrich’s answer: Ignore X and, as before, estimate the value of Y as ë = Y in E[Y]
- The minimum mean-square error is var(Y)

Using all the available information

- X and Y: RVs with known joint distribution
- We know that X = α on this trial and want to find MMSE estimate of Y on this trial
- Knowing that X = α on this trial, Y has conditional pdf f_Y|X(α|v) = f_Y(α|v)/f_X(α) or conditional pmf p_{Y|X}(v|α) = p_{YX}(v|α)/p_X(α)
- MMSE estimate of Y is the mean of this conditional pdf/pmf
- ë = E[Y|X=α] = ∫ v f_Y|X(α|v) dv or ∑ v p_{Y|X}(v|α)
Using all the available information
- X and Y: RVs with known joint distribution
- Given that X = α on this trial, the MMSE estimate of Y on this trial is ë = mean of the conditional pdf/pmf of Y given X = α
- Mean-square error = conditional variance on this trial of Y given X = α
- \( \text{var}(Y|X=\alpha) = \int (v-\hat{e})^2 f_{Y|X}(v|\alpha) \, dv \)
- \( \sum (v-\hat{e})^2 p_{Y|X}(v|\alpha) \)
- Remember that the mean is \( \hat{e} = E[Y|X=\alpha] \)

An example
- The random point (X, Y) is uniformly distributed on the semicircle shown
- Joint pdf has value \( 2/\pi \) on the semicircle
- Conditional pdf of Y given that X = α is a uniform density on \( [0, (1-\alpha^2)^{1/2}] \)

Conditional mean and variance
- Conditional pdf of Y given that X = α is a uniform density on \( [0, (1-\alpha^2)^{1/2}] \)
- Hence, \( \hat{e} = E[Y|X=\alpha] = (1/2)(1-\alpha^2)^{1/2} \)
- This estimate achieves the least possible MSE of \( \text{var}(Y|X=\alpha) = (1-\alpha^2)/12 \)

Does it all make any sense?
- \( \hat{e} = E[Y|X=\alpha] = (1/2)(1-\alpha^2)^{1/2} \)
- MSE is \( \text{var}(Y|X=\alpha) = (1-\alpha^2)/12 \)
- If |α| is nearly 1, the MSE is small
- If |α| is nearly 0, the MSE is large
- Makes sense to me!

The regression curve of Y on X
- \( \hat{e} = E[Y|X=\alpha] \) as a function of α is a curve called the regression curve of Y on X
- Graph of \( (1/2)(1-\alpha^2)^{1/2} \) is a half-ellipse
- Given X value, the MMSE estimator of Y can be "read off" from the regression curve

MMSE estimate and MSE are RVs
- Given that X = α, the MMSE estimate of Y is \( \hat{e} = E[Y|X=\alpha] \) and it achieves the least possible MSE of \( \text{var}(Y|X=\alpha) \)
- If X had taken on value β, then MMSE estimate would have had a different value \( E[Y|X=\beta] \) and different MSE \( \text{var}(Y|X=\beta) \)
- The MMSE estimate \( \hat{e} \) and the MSE that it achieves both are functions of X
- The MMSE estimate and its MSE are random variables that are functions of X

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
### The RVs $E[Y|X]$ and $\text{var}(Y|X)$

- The MMSE estimate of $Y$ and the MSE achieved by this estimate are random variables that are functions of $X$.
- $E[Y|X] = g(X)$ is a random variable whose value is $E[Y|X = \alpha]$ whenever $X = \alpha$.
- $\text{var}(Y|X) = h(X)$ is a random variable with value $\text{var}(Y|X = \alpha)$ whenever $X = \alpha$.
- $E[Y|X]$ and $\text{var}(Y|X)$ look like constants but they are not: they are functions $g(X)$ and $h(X)$ of $X$. Note: They are not functions of $Y$.

---

### Professor, this is so confusing...

- You told us that expected values are constants and now you tell us that $E[Y|X]$ and $\text{var}(Y|X)$ are not constants but functions $g(X)$, $h(X)$ of $X$? And not of $Y$?
- “All the randomness due to $Y$ was integrated out” when we found the mean and variance of $Y$.
- But, we have not taken expectations with respect to $X$ as yet...
- $E[Y|X]$ and $\text{var}(Y|X)$ are random variables.

---

### Sometimes you feel like a nut ...

- Given that $X = \alpha$, the MMSE estimate of $Y$ is $\hat{\theta} = E[Y|X = \alpha]$ and it achieves the least possible MSE of $\text{var}(Y|X = \alpha)$.
- If $X$ had taken on value $\beta$, then the MMSE estimate would have a different value:
  - $E[\hat{\theta}|X = \beta]$ and different MSE $\text{var}(\hat{\theta}|X = \beta)$.
  - $P(\alpha - \varepsilon/2 \leq X \leq \alpha + \varepsilon/2) = f_X(\alpha)*\varepsilon$.
  - $P(\beta - \varepsilon/2 \leq X \leq \beta + \varepsilon/2) = f_X(\beta)*\varepsilon$.
- What is the average MMSE estimate of $Y$?

---

### … sometimes you don’t ...

- On average, the MMSE estimator has value $E[\hat{\theta}|X = \alpha] = \int g(\alpha)f_X(\alpha) \text{d}\alpha$.
- But, $g(\alpha) = E[Y|X = \alpha] = \int \nu f_{YX}(\nu|\alpha) \text{d}\nu$.
- $E[g(X)] = E[E[Y|X]] = \int g(\alpha)f_X(\alpha) \text{d}\alpha$.
- $= \int \int \nu f_{YX}(\nu|\alpha) \text{d}\nu f_X(\alpha) \text{d}\alpha$.
- $= \int \int \nu f_X(\nu, \alpha, v|\alpha) f_X(\alpha) \nu \text{d}\nu f_X(\nu) \text{d}\nu$.
- $= \int \int \nu f_{YX}(\nu, \alpha) \nu \text{d}\nu f_X(\nu) \text{d}\nu = E[Y]$!!

---

### Let’s make sure we understand...

- Given that $X = \alpha$, the MMSE estimate of $Y$ is $E[\hat{\theta}|X = \alpha] = \int \nu f_{YX}(\nu|\alpha) \text{d}\nu$.
- As we repeat the experiment over and over, we observe different values of $X$ and obtain different MMSE estimates of $Y$.
- The statement $E[g(X)] = E[E[Y|X]] = E[Y]$ is saying that the average of our MMSE estimates of $Y$ is just $E[Y]$, the MMSE estimate if we didn’t know the value of $X$.

---

### … how we are improving things...

- We can always ignore the information that $X = \alpha$, use the MMSE estimate $E[Y]$ of $Y$, and achieve MSE $\text{var}(Y)$.
- Using the estimate $E[Y|X = \alpha]$ instead of $E[Y]$ achieves MSE $\text{var}(Y|X = \alpha)$.
- Average of MMSE estimates $E[Y|X = \alpha]$ is $E[Y]$, the MMSE estimate we would have been using if we didn’t know the value of $X$.
- But, average MSE is smaller than $\text{var}(Y)$.
What's the average MSE?

- \( \text{var}(Y|X = \alpha) = E[Y^2|X = \alpha] - (E[Y|X = \alpha])^2 \)
- \( \text{var}(Y|X) = E[Y^2|X] - (E[Y|X])^2 \)
- Recall that for any random variable \( Z \), \( \text{var}(Z) = E[Z^2] - (E[Z])^2 \) and \( E[E[Z|X]] = E[Z] \)
- \( E[\text{var}(Y|X)] = E[E[Y^2|X] - E[(g(X))^2]] = E[Y^2] - E[(g(X))^2] \)
- \( \{E[Y^2] - (E[Y])^2\} - \{E[(g(X))^2] - (E[Y])^2\} \)
- \( \text{var}(Y) - \text{var}(g(X)) = \text{var}(Y) - \text{var}(E[Y|X]) \)
- because \( g(X) = E[Y|X] \) is a RV whose mean is \( E[E[Y|X]] = E[Y] \)

The average MSE is smaller

- For each observed value \( \alpha \) of \( X \), the MMSE estimator \( \hat{\alpha} = E[Y|X = \alpha] \) achieves the MSE \( \text{var}(Y|X = \alpha) \)
- On average, the MSE achieved is \( E[\text{var}(Y|X)] = \text{var}(Y) - \text{var}(E[Y|X]) \leq \text{var}(Y) \)
- The average of the MMSE estimates is \( E[Y] \), the MMSE estimate if we did not know the value of \( X \)
- The average MSE is smaller than \( \text{var}(Y) \), the MSE of the ostrich’s estimate \( E[Y] \)

Example: \( E[E[Y|X]] = E[Y] \)

- \( E[Y] = \int \int v f_{X,Y}(u, v) \, dv \, du = 4/3 \pi \) (switch to polar coordinates…)
- \( E[Y|X] = (1/2) \pi (1-X^2)^{1/2} \)
- \( f_X(u) = (2/\pi)(1-u^2)^{1/2}, \, |u| \leq 1 \)
- \( E[E[Y|X]] = (1/\pi) \int (1-u^2) \, du = 4/3 \pi \)

The conditional variance formula

- The conditional variance formula says that \( \text{var}(Y) = E[\text{var}(Y|X)] + \text{var}(E[Y|X]) \)
- We have seen this in the context of discrete random variables already, where we noted that the unconditional variance of a random variable is the mean of the conditional variance plus the variance of the conditional mean
- See Lecture 15, Slides 15-18

Saying it again and again…

- If we don’t know the value of \( X \), the MMSE estimate is \( E[Y] \) with an MSE of \( \text{var}(Y) \)
- If we know the value of \( X \), the MMSE estimate of \( Y \) is \( E[Y|X] \) and it achieves a MSE of \( \text{var}(Y|X) \)
- Average estimate = \( E[Y] \) in either case
- \( \text{var}(Y) = E[\text{var}(Y|X)] + \text{var}(E[Y|X]) \)
- By changing our estimate of \( Y \) based on knowledge of \( X \), we reduce the MSE from \( \text{var}(Y) \) to \( E[\text{var}(Y|X)] \)
Linear MMSE estimation — III
- The linear MMSE estimate $L$ is best remembered in the "symmetric" form
  $$(L - \mu_Y)/\sigma_Y = \rho\sigma_X/\sigma_Y (X - \mu_X)/\sigma_X$$
- The MSE achieved is $E[(Y - aX - b)^2] = \sigma_Y^2 + (\sigma_Y^2)(\sigma_X^2)/(\sigma_Y^2)$
- Linear MMSE estimate also reduces the MSE: from $\text{var}(Y)$ to $\text{var}(Y)(1 - \rho^2)$

When does it not work?
- If $X$ and $Y$ are uncorrelated RVs, then $\rho = 0$ and
  $$(L - \mu_Y)/\sigma_Y = \rho(X - \mu_X)/\sigma_X$$
  reduces to $L = \mu_Y = E[Y]$ just as if we did not know the value of $X$
- The MSE achieved is $(\sigma_Y^2)(1 - \rho^2) = (\sigma_Y^2)$
  just as if we did not know the value of $X$
- Reminder: Independent RVs are uncorrelated RVs

Linear MMSE estimation — II
- Suppose that we wish to estimate $Y$ as a linear function of the observation $X$
- The linear MMSE estimate of $Y$ is $aX + b$ where $a$ and $b$ are chosen to minimize the mean-square error $E[(Y - aX - b)^2]$
- Let $Z = Y - aX - b$
- $E[(Y - aX - b)^2] = E[Z^2] = \text{var}(Z) + (E[Z])^2$
  $= \text{var}(Y) + a^2\text{var}(X) - 2a\text{cov}(X, Y) + (E[Z])^2$
- What should we choose $a$ and $b$ to be?

Linear MMSE estimation — I
- Suppose $X$ and $Y$ are independent RVs
- Knowing $X$ tells us nothing about $Y$
- If $X$ and $Y$ are independent RVs, the conditional pdf/pmf is the same as the unconditional pdf/pmf
- $E[Y|X] = \text{constant} = E[Y]$ does not depend on value of $X$: $\text{var}(E[Y|X]) = 0$
- MMSE estimate of $Y$ is $E[Y]$ and has an MSE of $\text{var}(Y)$, just as if we did not know the value of $X$

When does all this not work?
- Suppose $X$ and $Y$ are independent RVs
- Knowing $X$ tells us nothing about $Y$
- If $X$ and $Y$ are independent RVs, the conditional pdf/pmf is the same as the unconditional pdf/pmf
- $E[Y|X] = \text{constant} = E[Y]$ does not depend on value of $X$: $\text{var}(E[Y|X]) = 0$
- MMSE estimate of $Y$ is $E[Y]$ and has an MSE of $\text{var}(Y)$, just as if we did not know the value of $X$
Linear MMSE versus MMSE

- In general, the linear MMSE estimate $aX + b$ has a higher MSE than the (usually nonlinear) MMSE estimate $E[Y|X]$.
- Sometimes, both estimates are the same:
  $$E[Y|X = \alpha] = \frac{(1 + \alpha)}{2} = \text{linear estimate!}$$

Gaussian MMSE = Linear MMSE

- If $X$ and $Y$ are jointly Gaussian RVs, then the conditional pdf of $Y$ given $X = \alpha$ is a Gaussian pdf with mean $\mu_Y + (\rho \sigma_Y \sigma_X)(\alpha - \mu_X)$ and variance $(\sigma_Y^2)(1 - \rho^2)$.
- Hence, $E[Y|X = \alpha] = \mu_Y + (\rho \sigma_Y \sigma_X)(\alpha - \mu_X)$ is the same as the linear MMSE and has MSE $\text{var}(Y|X = \alpha) = (\sigma_Y^2)(1 - \rho^2)$ which is the same as that of the linear MMSE.

Least-squares straight line fitting

- If $X$ and $Y$ are discrete RVs taking on $n$ values $(u_i, v_i)$ with equal probability, then the linear MMSE estimate of $Y$ given $X$ is the "least-squares straight-line fit" to the "data".
- The linear MMSE estimate of $X$ given $Y$ is also a "least-squares straight-line fit" to the "data".
- In one case, $v$ is the dependent variable; while in the other case, $u$ is …

Summary

- The MMSE estimate $E[Y|X = \alpha]$ achieves MSE of $\text{var}(Y|X = \alpha)$.
- The average MSE is $E[\text{var}(Y|X)] \leq \text{var}(Y)$ which is the MSE of the estimate $E[Y]$ that does not use knowledge of the value of $X$.
- Linear MMSE estimate $\mu_Y + (\rho \sigma_Y \sigma_X)(\alpha - \mu_X)$ has MSE $(\sigma_Y^2)(1 - \rho^2)$.
- For jointly Gaussian RVs, MMSE estimate = linear MMSE estimate.