Finish Linear Regression:
Best linear function prediction of $Y$ given $X$.

MMSE: Best Function that predicts $Y$ from $S$.
Conditional Expectation.

Applications to random processes.
**Theorem**
Consider two RVs $X$, $Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

\[
L[Y|X] = \hat{Y} = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} (X - E[X]).
\]

**Proof 1:**

\[
Y - \hat{Y} = (Y - E[Y]) - \frac{\text{cov}(X,Y)}{\text{var}[X]} (X - E[X]). \quad E[Y - \hat{Y}] = 0 \text{ by linearity.}
\]

Also, $E[(Y - \hat{Y})X] = 0$, after a bit of algebra. (See next slide.)

Combine brown inequalities: $E[(Y - \hat{Y})(c + dX)] = 0$ for any $c, d$.

Since: $\hat{Y} = \alpha + \beta X$ for some $\alpha, \beta$, so $\exists c, d$ s.t. $\hat{Y} - a - bX = c + dX$.

Then, $E[(Y - \hat{Y})(\hat{Y} - a - bX)] = 0, \forall a, b$. Now,

\[
E[(Y - a - bX)^2] = E[(Y - \hat{Y} + \hat{Y} - a - bX)^2] \\
= E[(Y - \hat{Y})^2] + E[(\hat{Y} - a - bX)^2] + 0 \geq E[(Y - \hat{Y})^2].
\]

This shows that $E[(Y - \hat{Y})^2] \leq E[(Y - a - bX)^2]$, for all $(a, b)$.

Thus $\hat{Y}$ is the LLSE. \qed
A Bit of Algebra

\[ Y - \hat{Y} = (Y - E[Y]) - \frac{\text{cov}(X, Y)}{\text{var}[X]} (X - E[X]). \]

Hence, \( E[Y - \hat{Y}] = 0 \). We want to show that \( E[(Y - \hat{Y})X] = 0 \).

Note that
\[ E[(Y - \hat{Y})X] = E[(Y - \hat{Y})(X - E[X])], \]
because \( E[(Y - \hat{Y})E[X]] = 0 \).

Now,
\[
E[(Y - \hat{Y})(X - E[X])] = E[(Y - E[Y])(X - E[X])] - \frac{\text{cov}(X, Y)}{\text{var}[X]} E[(X - E[X])(X - E[X])] \\
= (*) \text{cov}(X, Y) - \frac{\text{cov}(X, Y)}{\text{var}[X]} \text{var}[X] = 0. \quad \square
\]

\( (*) \) Recall that \( \text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])] \) and \( \text{var}[X] = E[(X - E[X])^2] \).
We saw that the LLSE of $Y$ given $X$ is

$$L[Y|X] = \hat{Y} = E[Y] + \frac{\text{cov}(X, Y)}{\text{var}(X)}(X - E[X]).$$

How good is this estimator?  
Or what is the mean squared estimation error?  
We find


$$= E[(Y - E[Y])^2] - 2(\text{cov}(X, Y)/\text{var}(X))E[(Y - E[Y])(X - E[X])]$$

$$+ (\text{cov}(X, Y)/\text{var}(X))^2 E[(X - E[X])^2]$$

$$= \text{var}(Y) - \frac{\text{cov}(X, Y)^2}{\text{var}(X)}.$$

Without observations, the estimate is $E[Y]$. The error is $\text{var}(Y)$. Observing $X$ reduces the error.
Estimation Error: A Picture

We saw that
\[ L[Y|X] = \hat{Y} = E[Y] + \frac{\text{cov}(X, Y)}{\text{var}(X)}(X - E[X]) \]
and
\[ E[\| Y - L[Y|X] \|^2] = \text{var}(Y) - \frac{\text{cov}(X, Y)^2}{\text{var}(X)}. \]

Here is a picture when \( E[X] = 0, E[Y] = 0 \):
Dimensions correspond to sample points, uniform sample space.

Vector \( Y \) at dimension \( \omega \) is \( \frac{1}{\sqrt{\Omega}} Y(\omega) \)
Linear Regression Examples

Example 1:
Linear Regression Examples

Example 2:

We find:

\[ E[X] = 0; \ E[Y] = 0; \ E[X^2] = 1/2; \ E[XY] = 1/2; \]
\[ \text{var}[X] = E[X^2] - E[X]^2 = 1/2; \ 
\text{cov}(X, Y) = E[XY] - E[X]E[Y] = 1/2; \]
\[ \text{LR: } \hat{Y} = E[Y] + \frac{\text{cov}(X, Y)}{\text{var}[X]} (X - E[X]) = X. \]
Linear Regression Examples

Example 3:

We find:

\[ E[X] = 0; \ E[Y] = 0; \ E[X^2] = 1/2; \ E[XY] = -1/2; \]
\[ var[X] = E[X^2] - E[X]^2 = 1/2; \ cov(X, Y) = E[XY] - E[X]E[Y] = -1/2; \]
\[ LR: \ \hat{Y} = E[Y] + \frac{cov(X, Y)}{var[X]}(X - E[X]) = -X. \]
Linear Regression Examples

Example 4:

We find:

\[ E[X] = 3; \ E[Y] = 2.5; \ E[X^2] = (3/15)(1 + 2^2 + 3^2 + 4^2 + 5^2) = 11; \]
\[ E[XY] = (1/15)(1 \times 1 + 1 \times 2 + \cdots + 5 \times 4) = 8.4; \]
\[ \text{var}[X] = 11 - 9 = 2; \ \text{cov}(X, Y) = 8.4 - 3 \times 2.5 = 0.9; \]
\[ \text{LR}: \ \hat{Y} = 2.5 + \frac{0.9}{2} (X - 3) = 1.15 + 0.45X. \]
Note that

- the LR line goes through \((E[X], E[Y])\)
- its slope is \(\frac{\text{cov}(X,Y)}{\text{var}(X)}\).
Summary

1. Linear Regression: \( L[Y|X] = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)}(X - E[X]) \)
2. Non-Bayesian: minimize \( \sum_n (Y_n - a - bX_n)^2 \)
3. Bayesian: minimize \( E[(Y - a - bX)^2] \)
CS70: Noninear Regression.

1. Review: joint distribution, LLSE
2. Quadratic Regression
3. Definition of Conditional expectation
4. Properties of CE
5. Applications: Diluting, Mixing, Rumors
6. CE = MMSE
**Definitions** Let $X$ and $Y$ be RVs on $\Omega$.

- **Joint Distribution:** $Pr[X = x, Y = y]$
- **Marginal Distribution:** $Pr[X = x] = \sum_y Pr[X = x, Y = y]$
- **Conditional Distribution:** $Pr[Y = y | X = x] = \frac{Pr[X = x, Y = y]}{Pr[X = x]}$
- **LLSE:** $L[Y|X] = a + bX$ where $a, b$ minimize $E[(Y - a - bX)^2]$.

We saw that

$$L[Y|X] = E[Y] + \frac{cov(X, Y)}{var[X]}(X - E[X]).$$

Recall the non-Bayesian and Bayesian viewpoints.
Nonlinear Regression: Motivation

There are many situations where a good guess about $Y$ given $X$ is not linear.

E.g., (diameter of object, weight), (school years, income), (PSA level, cancer risk).

Our goal: explore estimates $\hat{Y} = g(X)$ for nonlinear functions $g(\cdot)$. 
Quadratic Regression

Let \( X, Y \) be two random variables defined on the same probability space.

**Definition:** The quadratic regression of \( Y \) over \( X \) is the random variable

\[
Q[Y|X] = a + bX + cX^2
\]

where \( a, b, c \) are chosen to minimize \( E[(Y - a - bX - cX^2)^2] \).

**Derivation:** We set to zero the derivatives w.r.t. \( a, b, c \). We get

\[
\begin{align*}
0 &= E[Y - a - bX - cX^2] \\
0 &= E[(Y - a - bX - cX^2)X] \\
0 &= E[(Y - a - bX - cX^2)X^2]
\end{align*}
\]

We solve these three equations in the three unknowns \((a, b, c)\).

**Note:** These equations imply that \( E[(Y - Q[Y|X])h(X)] = 0 \) for any \( h(X) = d + eX + fX^2 \). That is, the estimation error is orthogonal to all the quadratic functions of \( X \). Hence, \( Q[Y|X] \) is the projection of \( Y \) onto the space of quadratic functions of \( X \).
**Definition** Let $X$ and $Y$ be RVs on $\Omega$. The *conditional expectation* of $Y$ given $X$ is defined as

$$E[Y|X] = g(X)$$

where

$$g(x) := E[Y|X = x] := \sum_y y Pr[Y = y|X = x].$$

**Fact**

$$E[Y|X = x] = \sum_\omega Y(\omega) Pr[\omega|X = x].$$

**Proof:** $E[Y|X = x] = E[Y|A]$ with $A = \{\omega : X(\omega) = x\}$.  

$\square$
Have we seen this before? Yes.

Is anything new? Yes.

The idea of defining \( g(x) = E[Y|X = x] \) and then \( E[Y|X] = g(X) \).

Big deal? Quite! Simple but most convenient.

Recall that \( L[Y|X] = a + bX \) is a function of \( X \).

This is similar: \( E[Y|X] = g(X) \) for some function \( g(\cdot) \).

In general, \( g(X) \) is not linear, i.e., not \( a + bX \). It could be that \( g(X) = a + bX + cX^2 \). Or that \( g(X) = 2\sin(4X) + \exp\{-3X\} \). Or something else.
Properties of CE

\[ E[Y|X = x] = \sum_y yPr[Y = y|X = x] \]

**Theorem**
(a) \( X, Y \) independent \( \Rightarrow E[Y|X] = E[Y]; \)
(b) \( E[aY + bZ|X] = aE[Y|X] + bE[Z|X]; \)
(c) \( E[Yh(X)|X] = h(X)E[Y|X], \forall h(\cdot); \)
(d) \( E[h(X)E[Y|X]] = E[h(X)Y], \forall h(\cdot); \)
(e) \( E[E[Y|X]] = E[Y]. \)

**Proof:**
(a),(b) Obvious
(c) \[ E[Yh(X)|X = x] = \sum_\omega Y(\omega)h(X(\omega))Pr[\omega|X = x] \]
\[ = \sum_\omega Y(\omega)h(x)Pr[\omega|X = x] = h(x)E[Y|X = x] \]
Properties of CE

\[ E[Y|X = x] = \sum_y y Pr[Y = y|X = x] \]

**Theorem**

(a) \( X, Y \) independent \( \Rightarrow E[Y|X] = E[Y] \);
(b) \( E[aY + bZ|X] = aE[Y|X] + bE[Z|X] \);
(c) \( E[Yh(X)|X] = h(X)E[Y|X], \forall h(\cdot) \);
(d) \( E[h(X)E[Y|X]] = E[h(X)Y], \forall h(\cdot) \);
(e) \( E[E[Y|X]] = E[Y] \).

**Proof:** (continued)

(d) \( E[h(X)E[Y|X]] = \sum_x h(x)E[Y|X = x]Pr[X = x] \)

\[ = \sum_x h(x) \sum_y y Pr[Y = y|X = x]Pr[X = x] \]

\[ = \sum_x h(x) \sum_y y Pr[X = x, y = y] \]

\[ = \sum_{x,y} h(x)y Pr[X = x, y = y] = E[h(X)Y]. \]
Properties of CE

\[ E[Y|X = x] = \sum_y yPr[Y = y|X = x] \]

**Theorem**

(a) \( X, Y \) independent \( \Rightarrow E[Y|X] = E[Y] \);
(b) \( E[aY + bZ|X] = aE[Y|X] + bE[Z|X] \);
(c) \( E[Yh(X)|X] = h(X)E[Y|X], \forall h(\cdot) \);
(d) \( E[h(X)E[Y|X]] = E[h(X)Y], \forall h(\cdot) \);
(e) \( E[E[Y|X]] = E[Y] \).

**Proof:** (continued)

(e) Let \( h(X) = 1 \) in (d).
Properties of CE

Theorem
(a) $X, Y$ independent $\Rightarrow E[Y|X] = E[Y]$;
(b) $E[aY + bZ|X] = aE[Y|X] + bE[Z|X]$;
(c) $E[Yh(X)|X] = h(X)E[Y|X], \forall h(\cdot)$;
(d) $E[h(X)E[Y|X]] = E[h(X)Y], \forall h(\cdot)$;
(e) $E[E[Y|X]] = E[Y]$.

Note that (d) says that

$$E[(Y - E[Y|X])h(X)] = 0.$$ 

We say that the estimation error $Y - E[Y|X]$ is orthogonal to every function $h(X)$ of $X$.

We call this the projection property. More about this later.
Application: Calculating $E[Y|X]$

Let $X, Y, Z$ be i.i.d. with mean 0 and variance 1. We want to calculate

$$E[2 + 5X + 7XY + 11X^2 + 13X^3Z^2|X].$$

We find

$$E[2 + 5X + 7XY + 11X^2 + 13X^3Z^2|X]$$

$$= 2 + 5X + 7XE[Y|X] + 11X^2 + 13X^3E[Z^2|X]$$

$$= 2 + 5X + 7XE[Y] + 11X^2 + 13X^3E[Z^2]$$

$$= 2 + 5X + 11X^2 + 13X^3(var[Z] + E[Z]^2)$$

$$= 2 + 5X + 11X^2 + 13X^3.$$
Application: Diluting

Each step, pick ball from well-mixed urn. Replace with blue ball. Let $X_n$ be the number of red balls in the urn at step $n$. What is $E[X_n]$?

Given $X_n = m$, $X_{n+1} = m - 1$ w.p. $m/N$ (if you pick a red ball) and $X_{n+1} = m$ otherwise. Hence,

$$E[X_{n+1} | X_n = m] = m - (m/N) = m(N - 1)/N = X_n \rho,$$

with $\rho := (N - 1)/N$. Consequently,

$$E[X_{n+1}] = E[E[X_{n+1} | X_n]] = \rho E[X_n], n \geq 1.$$

$$\Rightarrow E[X_n] = \rho^{n-1} E[X_1] = N\left(\frac{N - 1}{N}\right)^{n-1}, n \geq 1.$$
Diluting

Here is a plot:

\[ E[X_n] \]
By analyzing $E[X_{n+1} \mid X_n]$, we found that $E[X_n] = N\left(\frac{N-1}{N}\right)^{n-1}, n \geq 1$. Here is another argument for that result.

Consider one particular red ball, say ball $k$. Each step, it remains red w.p. $(N-1)/N$, if different ball picked. $\Rightarrow$ the probability still red at step $n$ is $[(N-1)/N]^{n-1}$. Define:

$$Y_n(k) = 1\{\text{ball } k \text{ is red at step } n\}.$$ 

Then, $X_n = Y_n(1) + \cdots + Y_n(N)$. Hence,

$$E[X_n] = E[Y_n(1) + \cdots + Y_n(N)] = NE[Y_n(1)]$$

$$= NPr[Y_n(1) = 1] = N[(N-1)/N]^{n-1}.$$
Application: Mixing

Each step, pick ball from each well-mixed urn. Transfer it to other urn. Let $X_n$ be the number of red balls in the bottom urn at step $n$. What is $E[X_n]$?

Given $X_n = m$, $X_{n+1} = m + 1$ w.p. $p$ and $X_{n+1} = m - 1$ w.p. $q$

where $p = (1 - m/N)^2$ (B goes up, R down) and $q = (m/N)^2$ (R goes up, B down).

Thus,

$$E[X_{n+1}|X_n] = X_n + p - q = X_n + 1 - 2X_n/N = 1 + \rho X_n, \quad \rho := (1 - 2/N).$$
We saw that $E[X_{n+1}|X_n] = 1 + \rho X_n$, $\rho := (1 - 2/N)$. Does that make sense?

Hence,

$$E[X_{n+1}] = 1 + \rho E[X_n]$$
$$E[X_2] = 1 + \rho N; E[X_3] = 1 + \rho (1 + \rho N) = 1 + \rho + \rho^2 N$$
$$E[X_4] = 1 + \rho (1 + \rho + \rho^2 N) = 1 + \rho + \rho^2 + \rho^3 N$$
$$E[X_n] = 1 + \rho + \cdots + \rho^{n-2} + \rho^{n-1} N.$$

Hence,

$$E[X_n] = \frac{1 - \rho^{n-1}}{1 - \rho} + \rho^{n-1} N, n \geq 1.$$
Application: Mixing

Here is the plot.

\[ E[X_n] \]
Application: Going Viral

Consider a social network (e.g., Twitter).
You start a rumor (e.g., Rao is bad at making copies).
You have $d$ friends. Each of your friend retweets w.p. $p$.
Each of your friends has $d$ friends, etc.
Does the rumor spread? Does it die out (mercifully)?

In this example, $d = 4$. 
**Application: Going Viral**

**Fact:** Number of tweets $X = \sum_{n=1}^{\infty} X_n$ where $X_n$ is tweets in level $n$. Then, $E[X] < \infty$ iff $pd < 1$.

**Proof:**
Given $X_n = k$, $X_{n+1} = B(kd, p)$. Hence, $E[X_{n+1} | X_n = k] = kpd$.

Thus, $E[X_{n+1} | X_n] = pdX_n$. Consequently, $E[X_n] = (pd)^{n-1}, n \geq 1$.

If $pd < 1$, then $E[X_1 + \cdots + X_n] \leq (1 - pd)^{-1} \implies E[X] \leq (1 - pd)^{-1}$.

If $pd \geq 1$, then for all $C$ one can find $n$ s.t.

$E[X] \geq E[X_1 + \cdots + X_n] \geq C.$

In fact, one can show that $pd \geq 1 \implies Pr[X = \infty] > 0.$
An easy extension: Assume that everyone has an independent number $D_i$ of friends with $E[D_i] = d$. Then, the same fact holds.

To see this, note that given $X_n = k$, and given the numbers of friends $D_1 = d_1, \ldots, D_k = d_k$ of these $X_n$ people, one has $X_{n+1} = B(d_1 + \cdots + d_k, p)$. Hence,

$$E[X_{n+1}|X_n = k, D_1 = d_1, \ldots, D_k = d_k] = p(d_1 + \cdots + d_k).$$

Thus, $E[X_{n+1}|X_n = k, D_1, \ldots, D_k] = p(D_1 + \cdots + D_k)$.

Consequently, $E[X_{n+1}|X_n = k] = E[p(D_1 + \cdots + D_k)] = pdk$.

Finally, $E[X_{n+1}|X_n] = pdX_n$, and $E[X_{n+1}] = pdE[X_n]$.

We conclude as before.
Application: Wald’s Identity

Here is an extension of an identity we used in the last slide.

**Theorem** Wald’s Identity

Assume that $X_1, X_2, \ldots$ and $Z$ are independent, where

$Z$ takes values in \{0, 1, 2, \ldots\}

and $E[X_n] = \mu$ for all $n \geq 1$.

Then,

$$E[X_1 + \cdots + X_Z] = \mu E[Z].$$

**Proof:**

$$E[X_1 + \cdots + X_Z | Z = k] = \mu k.$$ 

Thus, $E[X_1 + \cdots + X_Z | Z] = \mu Z$.

Hence, $E[X_1 + \cdots + X_Z] = E[\mu Z] = \mu E[Z].$
Theorem
$E[Y|X]$ is the ‘best’ guess about $Y$ based on $X$.
Specifically, it is the function $g(X)$ of $X$ that
minimizes $E[(Y - g(X))^2]$. 

CE = MMSE
**Theorem** \( CE = MMSE \)

\[ g(X) := E[Y|X] \] is the function of \( X \) that minimizes \( E[(Y - g(X))^2] \).

**Proof:**

Let \( h(X) \) be any function of \( X \). Then

\[
E[(Y - h(X))^2] = E[(Y - g(X) + g(X) - h(X))^2]
\]
\[
= E[(Y - g(X))^2] + E[(g(X) - h(X))^2]
\]
\[
+ 2E[(Y - g(X))(g(X) - h(X))].
\]

But,

\[
E[(Y - g(X))(g(X) - h(X))] = 0 \text{ by the projection property.}
\]

Thus, \( E[(Y - h(X))^2] \geq E[(Y - g(X))^2] \). \( \Box \)
$E[Y|X]$ and $L[Y|X]$ as projections

$L[Y|X]$ is the projection of $Y$ on \( \{a + bX, a, b \in \mathbb{R}\} \): LLSE

$E[Y|X]$ is the projection of $Y$ on \( \{g(X), g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}\} \): MMSE.
Conditional Expectation

Definition: $E[Y|X] := \sum_y yPr[Y = y|X = x]$

Properties: Linearity, $Y - E[Y|X] \perp h(X)$; $E[E[Y|X]] = E[Y]$

Some Applications:
- Calculating $E[Y|X]$
- Diluting
- Mixing
- Rumors
- Wald

MMSE: $E[Y|X]$ minimizes $E[(Y - g(X))^2]$ over all $g(\cdot)$