Understanding Conditional Expectation via Vector Projection

Cheng-Shang Chang
Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan, R.O.C.

Jan. 14, 2008

Motivation and References

- Many students are confused with conditional expectation.
- In this talk, we explain how conditional expectation (taught in probability) is related to linear transformation and vector projection (taught in linear algebra).
- References:
  - S.J. Leon. *Linear Algebra with Applications*. New Jersey: Prentice Hall, 1998.
  - S. Ghahramani. *Fundamentals of Probability*. Pearson Prentice Hall, 2005.

Conditional Expectation

- Consider two discrete random variables X and Y.
- Let $p(x,y) = \mathsf{P}(X = x, Y = y)$ be the joint probability mass function.
- Then the marginal distribution

$$p_X(x) = \mathsf{P}(X = x) = \sum_{y \in B} p(x,y),$$

where B is the set of possible values of Y.

- Similarly,

$$p_Y(y) = \mathsf{P}(Y = y) = \sum_{x \in A} p(x,y),$$

where A is the set of possible values of X.

- Then the conditional probability mass function of X given $Y = y$ is

$$p_{X|Y}(x|y) = \mathsf{P}(X = x|Y = y) = \frac{p(x,y)}{p_Y(y)}.$$

Conditional Expectation

- The conditional expectation of X given $Y = y$ is defined as

$$\mathsf{E}[X|Y = y] = \sum_{x \in A} x p_{X|Y}(x|y). \tag{1}$$

- Consider a real-valued function h from $\mathcal{R}$ to $\mathcal{R}$.
- From the law of unconscious statistician, the conditional expectation of $h(X)$ given $Y = y$ is

$$\mathsf{E}[h(X)|Y = y] = \sum_{x \in A} h(x) p_{X|Y}(x|y).$$

- The conditional expectation of X given Y, denoted by $\mathsf{E}[X|Y]$, is the function of Y that is defined to be $\mathsf{E}[X|Y = y]$ when $Y = y$.
- Specifically, let $\delta(x)$ be the function with $\delta(0) = 1$ and $\delta(x) = 0$ for all $x \neq 0$.
- Also, let $\delta_y(Y) = \delta(Y - y)$ be the indicator random variable such that $\delta_y(Y) = 1$ if the event $\{Y = y\}$ occurs and $\delta_y(Y) = 0$ otherwise.
- Then

$$\mathsf{E}[X|Y] = \sum_{y \in B} \mathsf{E}[X|Y = y]\delta_y(Y) = \sum_{y \in B} \sum_{x \in A} x p_{X|Y}(x|y)\delta_y(Y). \tag{2}$$

Properties of Conditional Expectation

- The expectation of the conditional expectation of X given Y is the same as the expectation of X, i.e.,

$$\mathsf{E}[X] = \mathsf{E}[\mathsf{E}[X|Y]]. \tag{3}$$

- Let h be a real-valued function from $\mathcal{R}$ to $\mathcal{R}$. Then

$$\mathsf{E}[h(Y)X|Y] = h(Y)\mathsf{E}[X|Y]. \tag{4}$$

As $\mathsf{E}[X|Y]$ is a function of Y,

$$\mathsf{E}[\mathsf{E}[X|Y]|Y] = \mathsf{E}[X|Y]\mathsf{E}[1|Y] = \mathsf{E}[X|Y].$$

- This then implies

$$\mathsf{E}[X - \mathsf{E}[X|Y]|Y] = 0. \tag{5}$$

- Using (3) and (5) yields

$$\begin{aligned}\mathsf{E}[h(Y)(X - \mathsf{E}[X|Y])] &= \mathsf{E}[\mathsf{E}[h(Y)(X - \mathsf{E}[X|Y])]|Y] \\ &= \mathsf{E}[h(Y)\mathsf{E}[(X - \mathsf{E}[X|Y])]|Y] = 0.\end{aligned} \tag{6}$$

Properties of Conditional Expectation

- Let f be a real-valued function from $\mathcal{R}$ to $\mathcal{R}$.

$$\begin{aligned}
&\mathsf{E}[(X - f(Y))^2] = \mathsf{E}\left[\Big((X - \mathsf{E}[X|Y]) + (\mathsf{E}[X|Y] - f(Y))\Big)^2\right] \\
&= \mathsf{E}[(X - \mathsf{E}[X|Y])^2] + 2\mathsf{E}[(X - \mathsf{E}[X|Y])(\mathsf{E}[X|Y] - f(Y))] + \mathsf{E}[(\mathsf{E}[X|Y] - f(Y))^2] \\
&= \mathsf{E}[(X - \mathsf{E}[X|Y])^2] + \mathsf{E}[(\mathsf{E}[X|Y] - f(Y))^2],
\end{aligned}$$

  where the crossterm is 0 from (6).

- The conditional expectation of X given Y is the function of Y that minimizes $\mathsf{E}[(X - f(Y))^2]$ over the set of functions of Y, i.e.,

$$\mathsf{E}[(X - \mathsf{E}[X|Y])^2] \leq \mathsf{E}[(X - f(Y))^2], \tag{7}$$

  for any function f.

Vector Space

- Let V be the a set on which the operations of vector addition and scalar multiplication are defined.

- Axioms:

  - (Commutative law) $u + v = v + u$ for all u and v in V.
  - (Associative law(i)) $(u + v) + w = u + (v + w)$ for all u, v, w in V.
  - (Zero element) There exists an element $\mathbf{0}$ such that $u + \mathbf{0} = u$ for any $u \in V$.
  - (Inverse) For any $u \in V$, there exists an element $-u \in V$ such that $u + (-u) = \mathbf{0}$.
  - (Distributive law(i)) $\alpha(u + v) = \alpha u + \alpha v$ for any scalar α and $u, v \in V$.
  - (Distributive law(ii)) $(\alpha + \beta)u = \alpha u + \beta u$ for any scalars α and β and any $u \in V$.
  - (Associative law (ii)) $(\alpha\beta)u = \alpha(\beta u)$ for any scalars α and β and any $u \in V$.
  - (Identity) $1 \cdot u = u$ for any $u \in V$.

Vector Space

- Closure properties:
  - If $u \in V$ and α is a scalar, then $\alpha u \in V$.
  - If $u, v \in V$, then $u + v \in V$.
- Additional properties from the axioms and the closure properties:
  - $0 \cdot u = \mathbf{0}$.
  - $u + v = \mathbf{0}$ implies that $v = -u$.
  - $(-1) \cdot u = -u$.
- Example: the vector space $C[a, b]$
  - Let $C[a, b]$ be the set of real-valued functions that are defined and continuous on the closed interval $[a, b]$.
  - Vector addition: $(f + g)(x) = f(x) + g(x)$.
  - Scalar multiplication: $(\alpha f)(x) = \alpha f(x)$.

Subspace

- (Subspace) If S is a nonempty subset of a vector space V, and S satisfies the closure properties, then S is called a subspace of V.
- (Linear combination) Let $v_1, v_2, \ldots, v_n$ be vectors in a vector space V. A sum of the form $\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n$ is called a linear combination of $v_1, v_2, \ldots, v_n$.
- (Span) The set of all linear combinations of $v_1, v_2, \ldots, v_n$ is called span of $v_1, v_2, \ldots, v_n$ (denoted by Span$(v_1, v_2, \ldots, v_n)$).
- (Spanning set) The set $\{v_1, v_2, \ldots, v_n\}$ is a spanning set for V if and only if every vector in V can be written as a linear combination of $v_1, v_2, \ldots, v_n$, i.e.,

$$V \subset \text{Span}(v_1, v_2, \ldots, v_n).$$

- (Linearly independent) The vectors $v_1, v_2, \ldots, v_n$ in a vector space V are said to be linearly independent if

$$c_1 v_1 + c_2 v_2 + \ldots c_n v_n = \mathbf{0}$$

implies that all of the scalars $c_1, \ldots, c_n$ must be 0.

Basis and Dimension

- (Basis) The vectors $v_1, v_2, \ldots, v_n$ form a basis for a vector space V if and only if
  - ▶ $v_1, v_2, \ldots, v_n$ are linear independent.
  - ▶ $v_1, v_2, \ldots, v_n$ span V.
- (Dimension) If a vector space V has a basis consisting of n vectors, we say that V has dimension n.
  - ▶ finite-dimensional vector space: If there is a finite set of vectors that span the vector space.
  - ▶ infinite-dimensional vector space: for example $C[a, b]$
- Theorem: Suppose that V is a vector space of dimension $n > 0$.
  - ▶ Any set of n linear independent vectors spans V.
  - ▶ Any n vectors that span V are linear independent.
  - ▶ No set of less than n vectors can span V.

Coordinates

- Let $E = \{v_1, v_2, \ldots, v_n\}$ be an ordered basis for a vector space V.
- For any vector $v \in V$, it can be uniquely written in the form

$$v = c_1 v_1 + c_2 v_2 + \ldots c_n v_n.$$

- The vector $\mathbf{c} = (c_1, c_2, \ldots, c_n)^T$ in $\mathcal{R}^n$ is called the coordinate vector of v with respect to the ordered basis E (denoted by $[v]_E$).
- The c_i's are called the coordinates of v relative to E.
- A vector space with dimension n is isomorphic to $\mathcal{R}^n$ once a basis is found.

Random Variables on the Same Probability Space

- A probability space is a triplet $(S, \mathcal{F}, P)$, where S is the sample space, $\mathcal{F}$ is the set of (measurable) events, and P is the probability measure.
- A random variable X on a probability space $(S, \mathcal{F}, P)$ is a mapping from $X : S \mapsto \mathcal{R}$.
- The set of all random variables on the same probability space forms a vector space with each random variable being a vector.
  - ▶ Vector addition: $(X + Y)(s) = X(s) + Y(s)$ for every sample point s in the sample space S.
  - ▶ Scalar multiplication: $(\alpha X)(s) = \alpha X(s)$ for every sample point s in the sample space S.

The Set of Functions of a Discrete Random Variable

- Suppose that X is a discrete random variable with the set of possible values $A = \{x_1, x_2, \ldots, x_n\}$.
- Let $\delta_{x_i}(X) = \delta(X - x_i)$ be the indicator random variable with $\delta_{x_i}(X) = 1$ if the event $\{X = x_i\}$ occurs and 0 otherwise.
- Let $\sigma(X) = \text{Span}(\delta_{x_1}(X), \delta_{x_2}(X), \ldots, \delta_{x_n}(X))$.
  - ▶ $\delta_{x_1}(X), \delta_{x_2}(X), \ldots, \delta_{x_n}(X)$ are linearly independent. To see this, suppose s_i is a sample point such that $X(s_i) = x_i$. Then

    $$(c_1\delta_{x_1}(X) + c_2\delta_{x_2}(X) + \ldots + c_n\delta_{x_n}(X))(s_i) = \mathbf{0}(s_i) = 0$$

    implies that $c_i = 0$.
  - ▶ $\{\delta_{x_1}(X), \delta_{x_2}(X), \ldots, \delta_{x_n}(X)\}$ is a basis of $\sigma(X)$.
  - ▶ $\sigma(X)$ is a vector space with dimension n.

The Set of Functions of a Discrete Random Variable

- $\sigma(X)$ is the set of (measurable) functions of the random variable X.
  - ▶ For any real-valued function g from $\mathcal{R}$ to $\mathcal{R}$, $g(X)$ is a vector in $\sigma(X)$ as
$$g(X) = \sum_{i=1}^{n} g(x_i)\delta_{x_i}(X).$$
  - ▶ For any vector v in $\sigma(X)$, there is a real-valued function g from $\mathcal{R}$ to $\mathcal{R}$ such that $v = g(X)$. To see this, suppose that
$$v = \sum_{i=1}^{n} c_i\delta_{x_i}(X).$$
    We simply find a function g such that $g(x_i) = c_i$ for all i.
- The vector $(g(x_1), g(x_2), \ldots, g(x_n))^T \in \mathcal{R}^n$ is the coordinate vector of $g(X)$ with respect to the ordered basis $\{\delta_{x_1}(X), \delta_{x_2}(X), \ldots, \delta_{x_n}(X)\}$.
- In probability theory, $\sigma(X)$ is often called as the σ-algebra generated by the random variable X, and a random variable Y is called $\sigma(X)$-measurable if there is a (measurable) function g such that $Y = g(X)$.

Linear Transformation

- A mapping L from a vector space V into a vector space W is said to be a linear transformation if

$$L(\alpha v_1 + \beta v_2) = \alpha L(v_1) + \beta L(v_2)$$

for all $v_1, v_2 \in V$ and for all scalars α, β.

- (Matrix representation theorem) If $E = [v_1, v_2, \ldots, v_n]$ and $F = [w_1, w_2, \ldots, w_m]$ are ordered bases for vector spaces V and W, respectively, then corresponding to each linear transformation $L : V \mapsto W$ there is an $m \times n$ matrix A such that

$$[L(v)]_F = A[v]_E \quad \text{for each } v \in V.$$

- The matrix A is called the matrix representing the linear transformation L relative to the ordered bases E and F.

- The j^{th} column of the matrix A is simply of the coordinate vector of $L(v_j)$ with respect to the ordered basis F, i.e.,

$$a_j = [L(v_j)]_F.$$

Conditional Expectation As a Linear Transformation

- Suppose that X is a discrete random variable with the set of possible values $A = \{x_1, x_2, \dots, x_n\}$.
- Suppose that Y is a discrete random variable with the set of possible values $B = \{y_1, y_2, \dots, y_m\}$.
- Let $\sigma(X) = \mathsf{Span}(\delta_{x_1}(X), \delta_{x_2}(X), \dots, \delta_{x_n}(X))$ be the vector space that consists of the set of functions of the random variable X.
- Let $\sigma(Y) = \mathsf{Span}(\delta_{y_1}(Y), \delta_{y_2}(Y), \dots, \delta_{y_m}(Y))$ be the vector space that consists of the set of functions of the random variable Y.
- Consider the linear transformation $L : \sigma(X) \mapsto \sigma(Y)$ with

$$L(\delta_{x_i}(X)) = \sum_{j=1}^{m} \mathsf{P}(X = x_i | Y = y_j)\delta_{y_j}(Y), \qquad i = 1, 2, \dots, n.$$

- The linear transformation L can be represented by the $m \times n$ matrix A with

$$a_{i,j} = \mathsf{P}(X = x_i | Y = y_j).$$

Conditional Expectation As a Linear Transformation

- Since $g(X) = \sum_{i=1}^{n} g(x_i)\delta_{x_i}(X)$, we then have

$$\begin{aligned}
L(g(X)) &= L(\sum_{i=1}^{n} g(x_i)\delta_{x_i}(X)) \\
&= \sum_{i=1}^{n} g(x_i)L(\delta_{x_i}(X)) \\
&= \sum_{i=1}^{n} g(x_i)\sum_{j=1}^{m} \mathsf{P}(X = x_i|Y = y_j)\delta_{y_j}(Y) \\
&= \sum_{j=1}^{m} \Big(\sum_{i=1}^{n} g(x_i)\mathsf{P}(X = x_i|Y = y_j)\Big)\delta_{y_j}(Y) \\
&= \sum_{j=1}^{m} \mathsf{E}[g(X)|Y = y_j]\delta_{y_j}(Y) \\
&= \mathsf{E}[g(X)|Y].
\end{aligned}$$

- The linear transformation L of the random variable $g(X)$ is the condition expectation of $g(X)$ given Y.

Inner Product

- (Inner product) An inner product on a vector space V is a mapping that assigns to each pair of vectors u and v in V a real number $\langle u, v\rangle$ with the following three properties:
  - ▶ $\langle u, u\rangle \geq 0$ with equality if and only if $u = \mathbf{0}$.
  - ▶ $\langle u, v\rangle = \langle v, u\rangle$ for all u and v in V.
  - ▶ $\langle \alpha u + \beta v, w >= \alpha\langle u, w\rangle + \beta\langle v, w\rangle$ for all u, v, w in V and all scalars α and β.
- (Inner product space) A vector space with an inner product is called an inner product space.
- (Length) The length of a vector u is given by

$$||u|| = \sqrt{\langle u, u\rangle}.$$

- (Orthogonality) Two vectors u and v are orthogonal if $\langle u, v\rangle = 0$.
- (The Pythagorean law) If u and v are orthogonal vectors, then

$$||u + v||^2 = ||u||^2 + ||v||^2.$$

Inner Product on the Vector Space of Random Variables

- Consider the vector space of the set of random variables on the same probability space.
- Then
$$\langle X, Y \rangle = \mathsf{E}[XY]$$
is an inner product of that vector space.
- Note that $\mathsf{E}[X^2] = 0$ implies that $X = \mathbf{0}$ with probability 1.
- If we restrict ourselves to the set of random variables with mean 0. Then two vectors are orthogonal if and only if they are uncorrelated.
- As a direct consequence, two independent random variables with mean 0 are orthogonal.

Scalar Projection and Vector Projection

- (Scalar projection) If u and v are vectors in an inner product space V and $v \neq \mathbf{0}$, then the scalar projection of u onto v is given by

$$\alpha = \frac{\langle u, v\rangle}{||v||}.$$

- (Vector Projection) The vector projection of u onto v is given by

$$\mathbf{p} = \alpha(\frac{1}{||v||}v) = \frac{\langle u, v\rangle}{\langle v, v\rangle}v.$$

- Properties:
  - ▶ $u - \mathbf{p}$ and $\mathbf{p}$ are orthogonal.
  - ▶ $u = \mathbf{p}$ if and only if u is a scalar multiple of v.

Vector Projection on a Vector Space with an Orthogonal Basis

- An order basis $\{v_1, v_2, \dots v_n\}$ for a vector space V is said to be an orthogonal basis for V if $\langle v_i, v_j \rangle = 0$ for all $i \neq j$.
- Let S be a subspace of an inner product space V. Suppose that S has an orthogonal basis $\{v_1, v_2, \dots v_n\}$. Then the vector projection of u onto S is given by

$$\mathbf{p} = \sum_{i=1}^{n} \frac{\langle u, v_i \rangle}{\langle v_i, v_i \rangle} v_i.$$

- Properties:
  - ▶ $u - \mathbf{p}$ is orthogonal to every vector in S.
  - ▶ $u = \mathbf{p}$ if and only if $u \in S$.
- (Least square) $\mathbf{p}$ is the element of S that is closest to u, i.e.,

$$||u - v|| > ||u - \mathbf{p}||,$$

  for any $v \neq \mathbf{p}$ in S. Prove by the Pythagorean law.

$$||u - v||^2 = ||(u - \mathbf{p}) + (\mathbf{p} - v)||^2 = ||u - \mathbf{p}||^2 + ||\mathbf{p} - v||^2.$$

Conditional Expectation as a Vector Projection

- We have shown that $\mathsf{E}[g(X)|Y]$ is the linear transformation of $L(g(X))$ from $\sigma(X)$ to $\sigma(Y)$ with

$$L(\delta_{x_i}(X)) = \sum_{j=1}^{m} \mathsf{P}(X = x_i|Y = y_j)\delta_{y_j}(Y) = \mathsf{E}[\delta_{x_i}(X)|Y], \qquad i = 1, 2, \ldots, n.$$

- Note that $\delta_{y_i}(Y)\delta_{y_j}(Y) = \mathbf{0}$ for all $i \neq j$.
- Thus, $\mathsf{E}[\delta_{y_i}(Y)\delta_{y_j}(Y)] = 0$ for all $i \neq j$.
- $\{\delta_{y_1}(Y), \delta_{y_2}(Y), \ldots, \delta_{y_m}(Y)\}$ is an orthogonal basis for $\sigma(Y)$.
- The vector projection of $\delta_{x_i}(X)$ on $\sigma(Y)$ is then given by

$$\begin{aligned}
\sum_{j=1}^{m} \frac{\langle \delta_{x_i}(X), \delta_{y_j}(Y)\rangle}{\langle \delta_{y_j}(Y), \delta_{y_j}(Y)\rangle}\delta_{y_j}(Y) &= \sum_{j=1}^{m} \frac{\mathsf{E}[\delta_{x_i}(X)\delta_{y_j}(Y)]}{\mathsf{E}[\delta_{y_j}(Y)\delta_{y_j}(Y)]}\delta_{y_j}(Y) \\
= \sum_{j=1}^{m} \frac{\mathsf{E}[\delta_{x_i}(X)\delta_{y_j}(Y)]}{\mathsf{E}[\delta_{y_j}(Y)]}\delta_{y_j}(Y) &= \sum_{j=1}^{m} \frac{\mathsf{P}(X = x_i, Y = y_j)}{\mathsf{P}(Y = y_j)}\delta_{y_j}(Y) \\
= \sum_{j=1}^{m} \mathsf{P}(X = x_i|Y = y_j)\delta_{y_j}(Y) &= \mathsf{E}[\delta_{x_i}(X)|Y].
\end{aligned}$$

Conditional Expectation as a Vector Projection

- Recall that an inner product is a linear transformation for the first argument, i.e.,

$$\langle \alpha u + \beta v, w >= \alpha \langle u, w\rangle + \beta \langle v, w\rangle$$

  for all u, v, w in V and all scalars α and β.

- Since $g(X) = \sum_{i=1}^n g(x_i)\delta_{x_i}(X)$, the vector projection of $g(X)$ on $\sigma(Y)$ is then given by

$$\sum_{j=1}^m \frac{\langle g(X), \delta_{y_j}(Y)\rangle}{\langle \delta_{y_j}(Y), \delta_{y_j}(Y)\rangle}\delta_{y_j}(Y) = \sum_{j=1}^m \frac{\langle \sum_{i=1}^n g(x_i)\delta_{x_i}(X), \delta_{y_j}(Y)\rangle}{\langle \delta_{y_j}(Y), \delta_{y_j}(Y)\rangle}\delta_{y_j}(Y)$$
$$= \sum_{i=1}^n g(x_i) \sum_{j=1}^m \frac{\langle \delta_{x_i}(X), \delta_{y_j}(Y)\rangle}{\langle \delta_{y_j}(Y), \delta_{y_j}(Y)\rangle}\delta_{y_j}(Y)$$
$$= \sum_{i=1}^n g(x_i) \mathsf{E}[\delta_{x_i}(X)|Y] = \mathsf{E}[\sum_{i=1}^n g(x_i)\delta_{x_i}(X)|Y]$$
$$= \mathsf{E}[g(X)|Y].$$

- Thus, $\mathsf{E}[g(X)|Y]$ is the vector projection of $g(X)$ on $\sigma(Y)$.

Conditional Expectation as a Vector Projection

- It then follows from the properties of vector projection that
  - $g(X) - \mathsf{E}[g(X)|Y]$ is orthogonal to every random variable in $\sigma(Y)$, i.e., for any real-valued function $h : \mathcal{R} \mapsto \mathcal{R}$,

    $$\langle g(X) - \mathsf{E}[g(X)|Y], h(Y)\rangle = \mathsf{E}[(g(X) - \mathsf{E}[g(X)|Y])h(Y)] = 0.$$

  - (Least square) $\mathsf{E}[g(X)|Y]$ is the element of $\sigma(Y)$ that is closest to $g(X)$, i.e., for any real-valued function $h : \mathcal{R} \mapsto \mathcal{R}$ and $h(Y) \neq \mathsf{E}[g(X)|Y]$,

    $$\mathsf{E}[(g(X)-h(Y))^2] = ||g(X)-h(Y)|| > ||g(X)-\mathsf{E}[g(X)|Y]|| = \mathsf{E}[(g(X)-\mathsf{E}[g(X)|Y])^2].$$

Conditioning on a Set of Random Variables

- Note that Y only needs to be a random element in the previous development.
- In particular, if $Y = (Y_1, Y_2, \ldots, Y_d)$ is a d-dimensional random vector, then $\sigma(Y) = \sigma(Y_1, Y_2, \ldots, Y_d)$ is the set of functions of $Y_1, Y_2, \ldots, Y_d$.
- $\mathsf{E}[g(X)|Y] = \mathsf{E}[g(X)|Y_1, Y_2, \ldots, Y_d]$ is the vector projection of $g(X)$ on $\sigma(Y_1, Y_2, \ldots, Y_d)$.
  - ▶ $g(X) - \mathsf{E}[g(X)|Y_1, Y_2, \ldots, Y_d]$ is orthogonal to every random variable in $\sigma(Y_1, Y_2, \ldots, Y_d)$, i.e., for any function $h : \mathcal{R}^d \mapsto \mathcal{R}$,

$$\begin{aligned}&\langle g(X) - \mathsf{E}[g(X)|Y_1, Y_2, \ldots, Y_d], h(Y_1, Y_2, \ldots, Y_d)\rangle\\&= \mathsf{E}[(g(X) - \mathsf{E}[g(X)|Y_1, Y_2, \ldots, Y_d])h(Y_1, Y_2, \ldots, Y_d)] = 0.\end{aligned}$$

  - ▶ (Least square) $\mathsf{E}[g(X)|Y_1, Y_2, \ldots, Y_d]$ is the element of $\sigma(Y_1, Y_2, \ldots, Y_d)$ that is closest to $g(X)$, i.e., for any function $h : \mathcal{R}^d \mapsto \mathcal{R}$ and $h(Y_1, Y_2, \ldots, Y_d) \neq \mathsf{E}[g(X)|Y_1, Y_2, \ldots, Y_d]$,

$$\mathsf{E}[(g(X) - h(Y_1, Y_2, \ldots, Y_d))^2] > \mathsf{E}[(g(X) - \mathsf{E}[g(X)|Y_1, Y_2, \ldots, Y_d])^2].$$

General Definition of Conditional Expectation

- In some advanced probability books, conditional expectation is defined in a more general way.
- For a σ-algebra $\mathcal{G}$, $\mathsf{E}[X|\mathcal{G}]$ is defined to be the random variable that satisfies

  (i) $\mathsf{E}[X|\mathcal{G}]$ is $\mathcal{G}$-measurable, and

  (ii) $\int_A X d\mathsf{P} = \int_A \mathsf{E}[X|\mathcal{G}] d\mathsf{P}$ for all $A \in \mathcal{G}$.
- To understand this definition, consider the σ-algebra generated by the random variable Y (denoted by $\sigma(Y)$).
- The condition that $\mathsf{E}[X|Y]$ is $\sigma(Y)$-measurable is simply that $\mathsf{E}[X|Y]$ is a (measurable) function of Y, i.e., $\mathsf{E}[X|Y] = h(Y)$ for some (measurable) function.
- To understand the second condition, one may rewrite it as follows:

$$\mathsf{E}[\mathbf{1}_A X] = \mathsf{E}[\mathbf{1}_A \mathsf{E}[X|Y]], \tag{8}$$

for all event A in $\sigma(Y)$, where $\mathbf{1}_A$ is the indicator random variable with $\mathbf{1}_A = 1$ when the event A occurs.

General Definition of Conditional Expectation

- Since $\mathbf{1}_A$ is $\sigma(Y)$-measurable, it must be a function of Y. Thus, (8) is equivalent to
$$\mathsf{E}[g(Y)X] = \mathsf{E}[g(Y)\mathsf{E}[X|Y]], \tag{9}$$
for any (measurable) function g.
- Now rewriting (9) using the inner product yields
$$\langle g(Y), X - \mathsf{E}[X|Y]\rangle = 0, \tag{10}$$
for any function g.
- The condition in (10) simply says that $X - \mathsf{E}[X|Y]$ is orthogonal to every vector in $\sigma(Y)$ $\big(X - \mathsf{E}[X|Y]$ is in the orthogonal complement of $\sigma(Y)\big)$.
- To summarize, the first condition is that the vector projection should be in the projected space, and the second condition is that the difference between the vector being projected and the vector projection should be in the orthogonal complement of the projected space.
- These two conditions are exactly the same as those used to define projections in linear algebra.

Projections on the Set of Linear Functions of Y

- Recall that $\sigma(Y) = \text{Span}(\delta_{y_1}(Y), \delta_{y_2}(Y), \ldots, \delta_{y_m}(Y))$ is the set of functions of Y.
- $\sigma_L(Y) = \text{Span}(Y, 1)$ be the set of linear functions of Y, i.e., the set of functions of the form $aY + b$ for some constants a and b.
- $\sigma_L(Y)$ is a subspace of $\sigma(Y)$.
- However, Y and 1 are in general not orthogonal as $\mathsf{E}[Y \cdot 1] = \mathsf{E}[Y]$ may not be 0.
- (Gram-Schmidt orthogonalization process) $\{Y - E[Y], 1\}$ is an orthogonal basis for $\sigma_L(Y)$ as
$$E[(Y - E[Y]) \cdot 1] = E[Y] - E[Y] = 0.$$
- The projection of a random variable X on $\sigma_L(Y)$ is then given by
$$\begin{aligned} \mathbf{p_L} &= \frac{\langle X, Y - \mathsf{E}[Y] \rangle}{\langle Y - \mathsf{E}[Y], Y - \mathsf{E}[Y] \rangle}(Y - \mathsf{E}[Y]) + \frac{\langle X, 1 \rangle}{\langle 1, 1 \rangle} \cdot 1 \\ &= \frac{\mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y]}{\mathsf{E}[(Y - \mathsf{E}[Y])^2]}(Y - \mathsf{E}[Y]) + \mathsf{E}[X]. \end{aligned}$$

Projections on the Set of Linear Functions of Y

- It then follows from the properties of vector projection that
  - ▶ $X - \frac{\mathsf{E}[XY]-\mathsf{E}[X]\mathsf{E}[Y]}{\mathsf{E}[(Y-\mathsf{E}[Y])^2]}(Y - \mathsf{E}[Y]) - \mathsf{E}[X]$ is orthogonal to every random variable in $\sigma_L(Y)$, i.e., for any constants a and b,

$$\mathsf{E}\Big[\Big(X - \frac{\mathsf{E}[XY]-\mathsf{E}[X]\mathsf{E}[Y]}{\mathsf{E}[(Y-\mathsf{E}[Y])^2]}(Y - \mathsf{E}[Y]) - \mathsf{E}[X]\Big)\Big(aY+b\Big)\Big] = 0.$$

  - ▶ (Least square) $\frac{\mathsf{E}[XY]-\mathsf{E}[X]\mathsf{E}[Y]}{\mathsf{E}[(Y-\mathsf{E}[Y])^2]}(Y - \mathsf{E}[Y]) + \mathsf{E}[X]$ is the element of $\sigma_L(Y)$ that is closest to X, i.e., for any constants a and b,

$$\mathsf{E}[(X-aY-b)^2] \geq \mathsf{E}\Big[\Big(X - \frac{\mathsf{E}[XY]-\mathsf{E}[X]\mathsf{E}[Y]}{\mathsf{E}[(Y-\mathsf{E}[Y])^2]}(Y - \mathsf{E}[Y]) - \mathsf{E}[X]\Big)^2\Big].$$

- When X and Y are jointly normal, then the vector projection of X on $\sigma(Y)$ is the same as that on $\sigma_L(Y)$, i.e.,

$$\mathsf{E}[X|Y] = \frac{\mathsf{E}[XY]-\mathsf{E}[X]\mathsf{E}[Y]}{\mathsf{E}[(Y-\mathsf{E}[Y])^2]}(Y - \mathsf{E}[Y]) + \mathsf{E}[X].$$

Projections on a Subspace of $\sigma(Y)$

- Let $Y_i = \phi_i(Y)$, $i = 1, 2, \ldots, d$, where $\phi_i(\cdot)$'s are some known functions of Y.
- Let $\sigma_\phi(Y) = \text{Span}(1, Y_1, Y_2, \ldots, Y_d)$.
- $\sigma_\phi(Y)$ is a subspace of $\sigma(Y)$.
- In general, $\{1, Y_1, Y_2, \ldots, Y_d\}$ is not an orthogonal basis of $\sigma_\phi(Y)$.
- How do we find an orthogonal basis of $\sigma_\phi(Y)$?
- (Zero mean) Let $\tilde{Y}_i = Y_i - \mathsf{E}[Y_i]$. Then $\langle 1, \tilde{Y}_i \rangle = \mathsf{E}[\tilde{Y}_i] = 0$.
- (Matrix diagonalization) Let $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_d)^T$. Let $A = \mathsf{E}[\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T]$ be the $d \times d$ covariance matrix. As A is symmetric, there is an orthogonal matrix U and a diagonal matrix D such that

$$D = U^T A U.$$

  Let $\mathbf{Z} = (Z_1, Z_2 \ldots, Z_d)^T = U^T \tilde{\mathbf{Y}}$. Then

$$\mathsf{E}[\mathbf{Z}\mathbf{Z}^T] = \mathsf{E}[U^T \tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T U] = U^T \mathsf{E}[\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T] U = U^T A U = D.$$

- Thus, $\{1, Z_1, Z_2, \ldots, Z_d\}$ is an orthogonal basis of $\sigma_\phi(Y)$.

Projections on a Subspace of $\sigma(Y)$

- The projection of a random variable X on $\sigma_\phi(Y)$ is then given by

$$\begin{aligned}\mathbf{p}_\phi &= \sum_{k=1}^{d} \frac{\langle X, Z_k\rangle}{\langle Z_k, Z_k\rangle} Z_k + \frac{\langle X, 1\rangle}{\langle 1, 1\rangle}\cdot 1 \\ &= \sum_{k=1}^{d} \frac{\mathsf{E}[XZ_k]}{\mathsf{E}[Z_k^2]} Z_k + \mathsf{E}[X].\end{aligned}$$

- It then follows from the properties of vector projection that
  - ▶ $X - \mathbf{p}_\phi$ is orthogonal to every random variable in $\sigma_\phi(Y)$, i.e., for any constants a_k, $k = 1, 2, \ldots, d$, and b,

$$\mathsf{E}\Big[\Big(X - \mathbf{p}_\phi\Big)\Big(\sum_{k=1}^{d} a_k\phi_k(Y) + b\Big)\Big] = 0.$$

  - ▶ (Least square) $\mathbf{p}_\phi$ is the element of $\sigma_\phi(Y)$ that is closest to X, i.e., for any constants a_k, $k = 1, 2, \ldots, d$, and b,

$$\mathsf{E}[(X - \sum_{k=1}^{d} a_k\phi_k(Y) - b)^2] \geq \mathsf{E}\Big[\Big(X - \mathbf{p}_\phi)^2\Big].$$

Regression

- We have shown how to compute the conditional expectation (and other projections on a subspace of $\sigma(Y)$) if the point distribution of X and Y is known.
- Suppose that the point distribution of X and Y is unknown.
- Instead, a random sample of size n is given, i.e., $\{(x_k, y_k), k = 1, 2, \ldots, n\}$ is known.
- How do you find $h(Y)$ such that $\mathsf{E}[(X - h(Y))^2]$ is minimized?
- (Empirical distribution) Even though we do not know the true distribution, we still have the empirical distribution, i.e.,

$$\mathsf{P}(X = x_k, Y = y_k) = \frac{1}{n}, \qquad k = 1, 2, \ldots, n.$$

- Then one can use the empirical distribution to compute the conditional expectation (and other projections on a subspace of $\sigma(Y)$).

Linear Regression

- (Linear regression) Use the empirical distribution as the distribution of X and Y. Then

$$\mathbf{p_L} = \frac{\mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y]}{\mathsf{E}[(Y - \mathsf{E}[Y])^2]}(Y - \mathsf{E}[Y]) + \mathsf{E}[X],$$

where

$$\mathsf{E}[XY] = \frac{1}{n}\sum_{k=1}^{n} x_k y_k,$$

$$\mathsf{E}[X] = \frac{1}{n}\sum_{k=1}^{n} x_k, \qquad \mathsf{E}[Y] = \frac{1}{n}\sum_{k=1}^{n} y_k,$$

$$\mathsf{E}[Y^2] = \frac{1}{n}\sum_{k=1}^{n} y_k^2.$$

- $\mathbf{p_L}$ minimizes the empirical square error (risk)

$$\mathsf{E}[(X - aY - b)^2] = \frac{1}{n}\sum_{k=1}^{n}(x_k - ay_k - b)^2$$

for any constants a and b.