

Chapter 4

Conditional Probability and the Prediction Problem

Conditional probability and conditional expectation are two fundamental concepts in stochastic modeling. They are the key causality concepts used in models of evolving dynamical system. For a randomly evolving system, if we wish to model the fact that event A causes B , this is described by requiring that $P\{B|A\}$ is increased over $P\{B\}$.

We'll see both a calculus-based description of conditional probability and a more advanced formulation. The calculus method is suited to conditioning on a finite set of r.v.s, whereas the advanced method is suited to conditioning on an infinite set of r.v.s.

We will specifically look at the *prediction problem*, as an application since conditional probability and prediction are intimately linked concepts. Two particular examples covered in great detail are: the **Best mean square prediction**, where we'll see how conditional expectation yields the best mean square predictor; and **Affine prediction**

4.1 Conditional Probability

As mentioned above, conditional probability is fundamental to modeling causality in a stochastic setting.

Figure 4.1 helps illustrate the computation of conditional probability.

Note that conditioning on B essentially restricts the sample space to B . From the picture, computing the probability $P\{A|B\}$ should be related to $P\{A \cap B\}$, and so

$$P\{\cdot|B\} \propto P\{\cdot \cap B\}. \quad (4.1)$$

Because we have $P\{B|B\} = 1$, the proportionality constant must be $P\{B\}$. Thus,

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}}. \quad (4.2)$$

If we rewrite the above equation, we also get the expression

$$P\{A \cap B\} = P\{B\}P\{A|B\}. \quad (4.3)$$

This equation is sometimes called the multiplication rule.

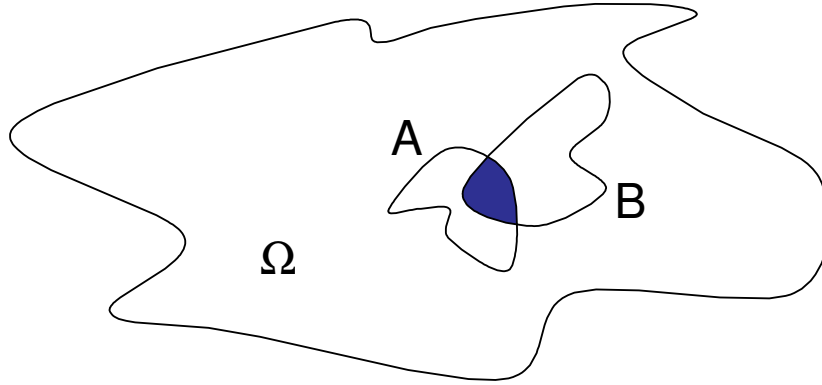


Figure 4.1: Conditional Probability

Example 4.1: In this example, we'll look at how conditional probability can be used to evaluate a medical test.

Suppose there are 60,000 people in the US population carrying a particular virus. We wish to evaluate the performance of a new diagnostic test for this virus.

Lab tests have shown the following probabilities for the diagnostic test:

	Positive Test	Negative Test
Virus present	0.99	0.01
Virus absent	0.02	0.98

For an ideal test, the above matrix would be the identity. To evaluate the test, we wish to compute

$$P \{ \text{virus} \mid \text{positive} \} .$$

This probability will tell us how likely it is that a person who tests positive under the proposed diagnostic test actually has the virus. Given the cost of treatment and additional testing for individuals who test positive, this probability is of central interest to health professionals.

To compute this probability, note that the table provides conditional probabilities in which the conditioning variable is the disease state of the patient. We wish to compute a conditional probability in which the conditioning variable is the diagnostic test outcome.

Specifically, the data in the table give us the following probabilities:

$$P \{ \text{positive} \mid \text{virus} \} = 0.99$$

$$P \{ \text{negative} \mid \text{virus} \} = 0.01$$

$$P \{ \text{positive} \mid \text{no virus} \} = 0.02$$

$$P \{ \text{negative} \mid \text{no virus} \} = 0.98$$

If we estimate the population of the US at 300,000,000, then

$$P \{ \text{virus} \} = 60,000/300,000,000 = 0.0002.$$

Now let's compute $P \{ \text{virus} \mid \text{positive} \}$:

$$\begin{aligned} P \{ \text{virus} \mid \text{positive} \} &= \frac{P \{ \text{virus, positive} \}}{P \{ \text{positive} \}} \\ &= \frac{P \{ \text{virus} \} P \{ \text{positive} \mid \text{virus} \}}{P \{ \text{positive} \}} , \end{aligned}$$

where the last step follows from the multiplication rule.

We can compute $P\{\text{positive}\}$ using the tabulated data and two applications of the multiplication rule, namely

$$\begin{aligned} P\{\text{positive}\} &= P\{\text{positive} \cap \text{virus}\} + P\{\text{positive} \cap \text{no virus}\} \\ &= P\{\text{virus}\}P\{\text{positive} \mid \text{virus}\} + P\{\text{no virus}\}P\{\text{positive} \mid \text{no virus}\}. \end{aligned}$$

Consequently,

$$\begin{aligned} P\{\text{virus} \mid \text{positive}\} &= \frac{P\{\text{virus}\}P\{\text{positive} \mid \text{virus}\}}{P\{\text{virus}\}P\{\text{positive} \mid \text{virus}\} + P\{\text{no virus}\}P\{\text{positive} \mid \text{no virus}\}} \\ &= \frac{(0.0002)(0.99)}{(0.0002)(0.99) + (0.9998)(0.02)} \\ &= 0.01. \end{aligned}$$

The specificity of this test is surprisingly low. The problem is that the likelihood of having the virus is extremely low. This makes designing an effective test difficult because false positives occur frequently.

At its core, this example was an illustration of Bayes' theorem; see the next result.

Theorem 4.1 (Bayes' theorem). *Let A_1, A_2, \dots, A_k be a collection of k mutually exclusive and exhaustive events with $P\{A_i\} > 0$ for $i = 1, \dots, k$. Then for any other event B for which $P\{B\} > 0$,*

$$P\{A_j \mid B\} = \frac{P\{A_j \cap B\}}{P\{B\}} = \frac{P\{B \mid A_j\}P\{A_j\}}{\sum_{i=1}^k P\{B \mid A_i\}P\{A_i\}}. \quad (4.4)$$

4.2 Conditional Probability for Random Variables

Discrete Random Variables We can extend our definition of conditional probability to discrete r.v.s as follows.

Let X and Y be two jointly distributed random variables. Then,

$$P\{Y = y \mid Z = z\} = \frac{P\{Y = y, Z = z\}}{P\{Z = z\}}. \quad (4.5)$$

Example 4.2: Let $X \triangleq \text{binomial}(n, p)$ and $Y \triangleq \text{binomial}(m, p)$, where X and Y are independent. We can envision these random variables in terms of coin tosses. In particular, if p is the probability of heads in the toss, then X is the number of heads in n tosses, and Y is the number of heads in m additional coin tosses. This suggests that if $Z = X + Y$, then

$$Z = X + Y = \text{binomial}(n + m, p).$$

This fact can be rigorously established by computing the convolution. We now wish to compute

$$P\{Y = y \mid Z = r\},$$

namely, the probability that y of a total of r heads came in the last m coin tosses. Note that,

$$\begin{aligned} P\{Y = y \mid Z = r\} &= \frac{P\{Y = y, Z = r\}}{P\{Z = r\}} \\ &= \frac{P\{Y = y, X = r - y\}}{P\{Z = r\}} \\ &= \frac{\binom{m}{y}p^y(1-p)^{m-y}\binom{n}{r-y}p^{r-y}(1-p)^{n-r+y}}{\binom{n+m}{r}p^r(1-p)^{n+m-r}} \\ &= \frac{\binom{m}{y}\binom{n}{r-y}}{\binom{n+m}{r}}. \end{aligned}$$

Amazingly, the final distribution does not depend on p at all! This final distribution is called the hypergeometric distribution.

If there are n ways to make a good selection and m ways to make a bad selection out of $n + m$ total possibilities, then the hypergeometric distribution can be interpreted as a probability distribution over the number of good selections in a total of r samples.

4.2.1 Conditional Probability for Continuous Random Variables

Here, we define the conditional probability for two continuous r.v.s.

Let Y and Z be jointly distributed r.v.s with joint density $f_{Y,Z}(\cdot)$ and marginal density $f_Z(\cdot)$ for the r.v. Z . Analogously to the discrete case, the conditional density of Y given $Z = z$ is given by

$$f_{Y|Z}(y|z) = \frac{f_{Y,Z}(y, z)}{f_Z(z)}. \quad (4.6)$$

Example 4.3: Bivariate Gaussian Distribution

Given parameters $\mu_Y, \mu_Z, \sigma_Y > 0, \sigma_Z > 0$, and $\rho \in (-1, 1)$, the joint distribution for two Gaussian variables is defined by

$$f_{Y,Z}(y, z) = \left(2\pi\sigma_Y\sigma_Z\sqrt{1-\rho^2}\right)^{-1} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(y-\mu_Y)^2}{\sigma_Y^2} + \frac{(z-\mu_Z)^2}{\sigma_Z^2} - 2\rho\left(\frac{y-\mu_Y}{\sigma_Y}\right)\left(\frac{z-\mu_Z}{\sigma_Z}\right)\right]\right\}.$$

It can be shown that, $Y \triangleq N(\mu_Y, \sigma_Y^2)$ and $Z \triangleq N(\mu_Z, \sigma_Z^2)$. Hence μ_Y and μ_Z can be interpreted as the means of Y and Z and σ_Y^2 and σ_Z^2 are their variances. A key feature of the Gaussian distribution is that their conditional distributions are also Gaussian. Specifically, in this bivariate setting,

$$Y|Z=z \stackrel{D}{=} N(\mu_Y + \rho\frac{\sigma_Y}{\sigma_Z}(z - \mu_Z), \sigma_Y^2(1 - \rho^2)).$$

Note that if $\rho = 0$, then Y and Z are independent.

Example 4.4: Exponential Distribution.

Let $Y \triangleq \exp(\lambda)$. This is a r.v. with distribution and density given by

$$P\{Y \leq y\} = 1 - e^{-\lambda y}, y \geq 0$$

and

$$f_Y(y) = \lambda e^{-\lambda y} I(y \geq 0).$$

The so called “tail distribution function” or “complementary distribution function” is therefore given by

$$P\{Y > y\} = e^{-\lambda y}.$$

Suppose that we interpret Y as the lifetime of the component or system. One interesting computation is

$$P\{Y > t + u | Y > t\}, u > 0.$$

In other words, given that the component has lasted t time units, what is the probability it will last at least u additional units?

Note that

$$\begin{aligned} \mathbb{P}\{Y > t + u | Y > t\} &= \frac{\mathbb{P}\{Y > t + u, Y > t\}}{\mathbb{P}\{Y > t\}} \\ &= \frac{\exp(-\lambda(t + u))}{\exp(-\lambda t)} \\ &= \exp(-\lambda u) \\ &= \mathbb{P}\{Y > u\}. \end{aligned}$$

Notice that $\mathbb{P}\{Y > t + u, Y > t\} = \mathbb{P}\{Y > t + u\}$ because if $Y > t + u$, then $Y > t$.

This example establishes the *memoryless property* of the exponential random variable. An exponential r.v. describes a component that, statistically speaking, does not age. Knowledge of how long the component has been in operation does not affect the conditional distribution of how much longer it will last.

Exercise 4.1: This problem establishes that the exponential distribution is the only memoryless distribution describing a positive r.v.

1. Suppose that F is a differentiable distribution function that satisfies $\bar{F}(t + s) = \bar{F}(t)\bar{F}(s)$ for $s, t \geq 0$, where $\bar{F}(t) \triangleq 1 - F(t)$. Prove that $\bar{F}(t) = \exp(-\lambda t)$ for some $\lambda > 0$. (Hint: Differentiate with respect to t and obtain a differential equation.)
2. **Harder!** Suppose that F is a distribution function that satisfies $\bar{F}(t + s) = \bar{F}(t)\bar{F}(s)$ for $s, t \geq 0$. Prove that $\bar{F}(t) = \exp(-\lambda t)$ for some $\lambda > 0$. (Hint: First express $\bar{F}(m/n)$ in terms of $\bar{F}(1)$ for $m, n \in \mathbb{Z}^+$. Then use right continuity of $\bar{F}(\cdot)$ to obtain $\bar{F}(\cdot)$ at the irrationals. (All distribution functions are automatically right continuous. Why?))

Example 4.5: Gaussian tails.

Let $Y \triangleq N(0, 1)$ and consider the same type of calculation as carried out for the exponential case above. Specifically, consider

$$\begin{aligned} \mathbb{P}\{Y > t + \varepsilon | Y > t\} &= \mathbb{P}\{Y > t + \varepsilon\} / \mathbb{P}\{Y > t\} \\ &= \frac{\int_{t+\varepsilon}^{\infty} e^{-y^2/2} dy}{\int_t^{\infty} e^{-y^2/2} dy}. \end{aligned}$$

If we take the limit as $t \rightarrow \infty$, we need to use L'Hopital's rule on the numerator and the denominator. This shows that

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathbb{P}\{Y > t + \varepsilon | Y > t\} &= \lim_{t \rightarrow \infty} \frac{\int_{t+\varepsilon}^{\infty} e^{-y^2/2} dy}{\int_t^{\infty} e^{-y^2/2} dy} \\ &= \lim_{t \rightarrow \infty} \frac{\frac{d}{dt} \int_{t+\varepsilon}^{\infty} e^{-y^2/2} dy}{\frac{d}{dt} \int_t^{\infty} e^{-y^2/2} dy} \\ &= \lim_{t \rightarrow \infty} \frac{e^{-(t+\varepsilon)^2/2}}{e^{-t^2/2}} \\ &= 0. \end{aligned}$$

The normal distribution is said to have *thin tails*, because its tail goes to 0 faster than any exponential. The above calculation shows that if we are told that such a thin-tailed Gaussian is greater than t (with t large), it is very likely that the Gaussian lies very close to t (specifically, in the interval $[t, t + \varepsilon]$).

Example 4.6: Pareto tails.

A r.v. Y is said to have a Pareto (c, α) distribution (with $c, \alpha > 0$), provided that

$$\mathbb{P}\{Y > y\} = (1 + cy)^{-\alpha}$$

for $y \geq 0$. Consider the same calculation as implemented above for exponential and Gaussian r.v.s:

$$\begin{aligned} \mathbb{P}\{Y > t + s | Y > t\} &= \mathbb{P}\{Y > t + s\} / \mathbb{P}\{Y > t\} \\ &= \left(\frac{1 + ct}{1 + c(t + s)} \right)^\alpha \rightarrow 1 \end{aligned}$$

as $t \rightarrow \infty$. Pareto r.v.s are said to have *fat tails*, because Pareto tails go to 0 slower than any exponential. This calculation shows that if we are told that a fat-tailed Pareto is greater than t (with t large), it is very likely that the Pareto is actually much greater than t .

Examples 3.4 through 3.6 show that exponential r.v.s lie at the critical threshold at which the conditional distribution of $Y - t$, given $Y > t$, decays neither to 0 (as for thin tails) or explodes to infinity (as for fat tails).

4.3 Reliability Modeling

We now introduce a key modeling concept that is widely used in reliability modeling and biostatistical settings (and that builds on the conditioning ideas describes above.)

Definition 4.1: Let T be a continuous positive r.v. having a density f . The *hazard rate function* (also known as the *failure rate function*) of T is the function defined by

$$r(t) = \frac{f(t)}{\bar{F}(t)}$$

for $t \geq 0$.

Note that

$$\mathbb{P}\{T \in [t, t + h] | T > t\} = r(t)h + o(h)$$

as $h \rightarrow 0$, so that $r(t)$ can be interpreted as the rate at which a component (with lifetime described by T) fails at t . Hazard rate functions are closely related to the actuarial tables used by insurance companies to price life insurance policies as a function of the policy holder's age.

Exercise 4.2: Prove that if T has hazard rate $r(\cdot)$, then

$$\bar{F}(t) = \exp\left(-\int_0^t r(u) du\right)$$

for $t \geq 0$.

Note that an exponential distribution is the unique distribution with the property that the hazard rate function is constant. This is a reflection of the memoryless character of exponential r.v.s.

Hazard rate functions are useful tools in developing an understanding of the reliability characteristics of a component (or the survival characteristics of an individual afflicted with a life-threatening disease). As an illustration, consider the “bathtub shaped” hazard function shown in Figure 4.2.

Such bathtub shaped hazard functions are typical of real-life component lifetimes. Many components experience an initial “burn-in” period (e.g. a device that was mis-assembled will not work when plugged in, so this leads to a high initial failure rate). Once the burn-in period has elapsed, the failure rate is roughly constant until such time as the component begins to “wear out,” at which time the failure rate begins increasing. These considerations lead to bathtub-shaped hazard functions.

Note that the exponential distribution may be a reasonable model for component lifetime, at least over the period in which the hazard rate is roughly constant. In any case, the shape of the hazard function for a product can have a significant impact on warranty costs (and on design of warranty products).

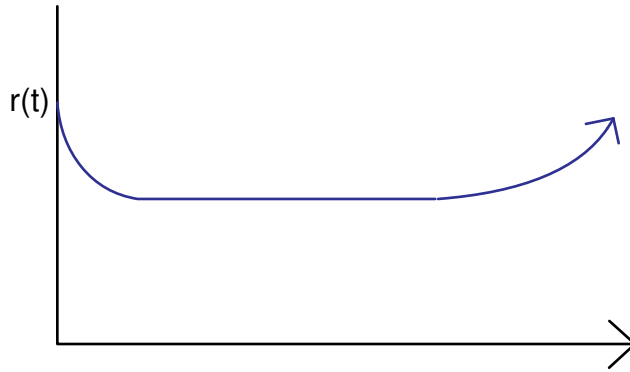


Figure 4.2: Hazard Function

4.4 The Calculus-Based View of Conditional Expectation

We now describe conditional expectation from a calculus-based viewpoint.

Let Y and Z be jointly distributed.

Definition 4.2: Discrete Conditional Expectation

$$E[Y|Z = z] = \sum_y y p_{Y|Z}(y|z).$$

Definition 4.3: Continuous Conditional Expectation

$$E[Y|Z = z] = \int_{-\infty}^{\infty} y f_{Y|Z}(y|z) dy.$$

Example 4.7: If Y and Z have a bivariate Gaussian distribution, then

$$E[Y|Z = z] = \mu_Y + \rho \frac{\sigma_Y}{\sigma_Z} (z - \mu_Z).$$

A key fact is that the conditional expectation for the Gaussian random variable is affine in the conditioning variable Z .

4.4.1 Conditional Expectation in the Presence of Multiple Conditioning Variables

The above definitions of conditional expectation generalize easily to the setting in which one conditions on a finite collection Z_1, \dots, Z_d of random variables. Specifically, suppose that

$$(Y, Z_1, \dots, Z_d) \triangleq (Y, Z)$$

is jointly distributed, where $Z = (Z_1, \dots, Z_d)^T$.

Discrete Conditional Expectation

$$E[Y|Z = z] = \sum_y y p_{Y|Z}(y|z),$$

where

$$p_{Y|Z}(y|z) = \frac{P\{Y = y, Z = z\}}{P\{Z = z\}}.$$

Continuous Conditional Expectation

$$E[Y|Z = z] = \int_{-\infty}^{\infty} y f_{Y|Z}(y|z) dy,$$

where

$$f_{Y|Z}(y|z) = \frac{f_{Y,Z}(y, z)}{f_Z(z)}.$$

Here,

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Y,Z}(y, z) dy$$

is the marginal density of Z . Note that the above definitions of conditional expectation fail to generalize to settings in which one conditions on an infinite collection of r.v.s.

4.5 Conditional Expectations and Prediction Theory

One of the major accomplishments of twentieth-century probability was a reformulation of the concept of conditional expectation that permitted one (for the first time) to rigorously define conditional expectation (and hence conditional probabilities, as a special case for indicator r.v.s) in the presence of an infinite number of conditioning variables. The extension was due to Kolmogorov. In addition to its theoretical interest, this re-formulation requires introducing a topic known as “prediction theory,” which is of significant interest in its own right.

To set the stage for prediction theory, consider a setting in which one has a great deal of past information on historical outcomes, and wishes to predict a future value. For example, one may have plenty of historical information on the number of faulty chips per wafer produced at a semiconductor manufacturing plant. Our goal is to make a prediction of tomorrow’s yield (or, equivalently, the number of faulty chips on tomorrow’s wafers).

In the presence of a great deal of historical information, we can presume (as an idealization) that the distribution of the r.v. Y is known.

Setting We are given the distribution of the r.v. Y .

Problem Based on the distribution of Y , compute a “best predictor” \hat{Y} for Y .

In the current setting, we must make the prediction based solely on historical data. There is no (correlated) observable Z available from which to improve one’s prediction (e.g. the fraction Z of complex chips that is scheduled for tomorrow’s productive run). Given the absence of such an observable Z , the prediction \hat{Y} must be deterministic. In other words, $\hat{Y} = a$ a.s. for some $a \in \mathbb{R}$. (Of course, a sensible choice for a will depend on Y ’s distribution.)

The current problem formulation involves the concept of “best.” To make this precise, let

$$\ell : \mathbb{R} \rightarrow [0, \infty)$$

be a given “loss function.” We can now re-formulate our prediction problem more precisely.

Find a predictor $\hat{Y} = a^*$ that minimizes the expected loss due to prediction error, namely,

$$E[\ell(Y - \hat{Y})]$$

over all feasible predictors $a \in \mathbb{R}$.

Clearly, the solution of this problem depends on the choice of loss function.

1. squared error ($\ell(x) = x^2$)
2. absolute error ($\ell(x) = |x|$)
3. L^p error ($\ell(x) = |x|^p$)

The optimal predictor here can be easily computed.

Example 4.8: The best mean square predictor $\hat{Y} = a^*$ (here, $\ell(x) = x^2$) is the minimizer, over a , of

$$\mathbb{E} [(Y - a)^2] = \mathbb{E} [Y^2] - 2a \mathbb{E} [Y] + a^2.$$

The minimizer of this quadratic is $a^* = \mathbb{E} [Y]$. So, $\hat{Y} = \mathbb{E} [Y]$ is the best mean square predictor.

Example 4.9: Suppose that Y is a continuous r.v. with density f . The best mean absolute predictor $\hat{Y} = a^*$ (here, $\ell(x) = |x|$) is the minimizer, over a , of

$$\mathbb{E} [|Y - a|] = \int_{-\infty}^a (a - y)f(y) dy + \int_a^{\infty} (y - a)f(y) dy.$$

The minimizer a^* must satisfy

$$\int_{-\infty}^{a^*} f(y) dy = \int_{a^*}^{\infty} f(y) dy.$$

In other words, a^* has the property that $\mathbb{P} \{Y \leq a^*\} = 1/2$ (i.e. a^* is the median of Y). So,

$$\hat{Y} = \text{median of } Y$$

is the best mean absolute predictor.

Of course, in many application settings, one has an observable Z that is correlated with the r.v. Y that can be used to enhance the prediction. For example, as suggested above, we know today the fraction Z of complex chips that will be manufactured tomorrow. As in our earlier discussion, we presume that there is sufficient historical data available that we may presume that the joint distribution of Y and Z is known.

Setting We are given the joint distribution of Y and Z .

In this context, our best predictor can depend on the observable Z . In particular, we now permit ourselves to consider predictors of the form $\hat{Y} = g(Z)$ for some (deterministic) g .

Remark For those of you who know measure theory, we are assuming here that \hat{Y} must be a measurable function of Z .

We focus our discussion here on the case in which our loss function is quadratic (i.e. squared loss). Our precise formulation of the prediction problem is:

Problem Given the joint distribution of Y and Z , find a predictor $\hat{Y} = g^*(Z)$ that minimizes

$$\mathbb{E} [(Y - \hat{Y})^2]$$

over all predictors $\hat{Y} = g(Z)$, where g is deterministic.

We can tighten our formulation somewhat. If $\mathbb{E} [Y^2] = \infty$, the minimal mean square predictor error is typically infinite, so the problem is uninteresting. So, assume $\mathbb{E} [Y^2] < \infty$. On the other hand, if $\hat{Y} = g(Z)$ has $\mathbb{E} [\hat{Y}^2] = \infty$, it is again clear that the mean square prediction error is infinite. This suggests the following re-formulation:

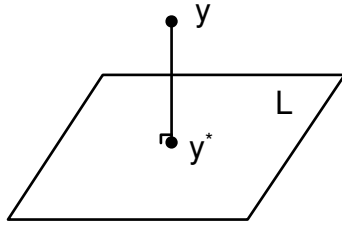


Figure 4.3: Hilber Space Projection

Suppose $E[Y^2] < \infty$. Given the joint distribution of Y and Z , find a predictor $\hat{Y} = g^*(Z)$ that minimizes

$$E[(Y - W)^2]$$

over all $W \in \mathcal{L}$, where

$$\mathcal{L} = \{W : E[W^2] < \infty, W = g(Z) \text{ for some deterministic } g\}.$$

A key observation here is that \mathcal{L} is a linear space. In fact, \mathcal{L} is a linear subspace of the vector space

$$L^2 = \{X : E[X^2] < \infty\}.$$

Note that we can measure distances between elements of L^2 via the norm

$$\|X\|_2 = \sqrt{E[X^2]};$$

the distance between X_1 and X_2 is then given by

$$\|X_1 - X_2\|_2.$$

In view of this measure of distance, our prediction problem involves finding the r.v. $\hat{Y} \in \mathcal{L}$ that minimizes the distance from Y (i.e. $\|Y - \hat{Y}\|_2$).

Of course, we know how to solve this problem when working in a finite-dimensional vector space. In \mathbb{R}^d , the closest point y^* in a linear subspace to a point y outside the subspace is obtained by “dropping the perpendicular” to that space. In other words, y^* is characterized as that point for which $y - y^*$ is orthogonal to the subspace. Figure 4.3 illustrates this concept.

We wish to generalize this idea of “dropping the perpendicular” to the setting of the infinite-dimensional space L^2 . For this purpose, we need to define an inner product on L^2 space that is consistent with the norm $\|\cdot\|_2$.

Definition 4.4: For $X_1, X_2 \in L^2$, define the *inner product* between X_1 and X_2 , written $\langle X_1, X_2 \rangle$, to be given by

$$\langle X_1, X_2 \rangle = E[X_1 X_2].$$

Observe that $\|X\|_2 = \sqrt{\langle X, X \rangle}$, so that the above inner product does indeed induce the norm $\|\cdot\|_2$.

Definition 4.5: We say that $X_1, X_2 \in L^2$ are *orthogonal* if

$$\langle X_1, X_2 \rangle = 0.$$

Remark 4.1: Recall that we can define the angle θ between any two elements x_1 and x_2 in an inner product space via

$$\cos \theta = \frac{\langle x_1, x_2 \rangle}{\|x_1\|_2 \|x_2\|_2}.$$

Hence, if $\langle x_1, x_2 \rangle = 0$, this corresponds to $\theta = \pi/2$, representing orthogonality.

With the above inner product, L^2 is a Hilbert space (i.e. a complete inner product space). The key to solving the prediction problem is the following result, known as the “projection theorem.”

Theorem 4.2 (Hilbert Space Projection Theorem). *Let \mathcal{L} be a closed linear subspace of a Hilbert space L^2 , and let $Y \in L^2$. There exists a unique $W^* \in \mathcal{L}$ that minimizes*

$$\|Y - W\| = \sqrt{\langle Y - W, Y - W \rangle}$$

over $W \in \mathcal{L}$. The unique W^* is characterized by

$$\langle Y - W^*, W \rangle = 0,$$

for $W \in \mathcal{L}$.

We now wish to apply this to the prediction problem. It can be shown that

$$\mathcal{L} = \{W \in L^2 : W = g(Z) \text{ for some deterministic } g(\cdot)\}$$

is a closed linear subspace of the space L^2 of square-integrable r.v.s. Hence, the projection theorem asserts that there exists a best mean-square predictor $W^* \in \mathcal{L}$ that is characterized via

$$\langle Y - W^*, W \rangle = 0$$

for $W \in \mathcal{L}$, i.e. if $W^* = g^*(Z)$, then

$$E[Yg(Z)] = E[g^*(Z)]g(Z)$$

for $g(Z) \in \mathcal{L}$. This gives us the basis for the following definition.

Definition 4.6: Suppose $Y \in L^2$. The *conditional expectation* of Y given Z is the r.v. $g^*(Z)$ satisfying

$$E[Yg(Z)] = E[g^*(Z)g(Z)]$$

for all r.v. $g(Z) \in \mathcal{L}$. We write

$$E[Y|Z] \triangleq g^*(Z).$$

This definition of conditional expectation makes rigorous sense regardless of whether Z is a finite-dimensional random vector or a continuum of r.v.s. Furthermore, it coincides with the calculus-based definition of conditional expectation when Z is finite-dimensional!

Exercise 4.3: Suppose Y and Z have a joint probability density $f_{Y,Z}(\cdot)$ and that $E[Y]^2 < \infty$. Put

$$g^*(z) = \frac{\int_{-\infty}^{\infty} y f_{Y,Z}(y, z) dy}{\int_{-\infty}^{\infty} f_{Y,Z}(y, z) dy}.$$

Prove that

$$E[Yg(Z)] = E[g^*(Z)g(Z)]$$

for all deterministic $g(\cdot)$ for which $E[g^2(Z)] < \infty$. (Hence, $E[Y|Z] = g^*(Z)$ a.s.)

Exercise 4.4: Suppose Y and Z are bivariate Gaussian. Prove that

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

(Hence, for Gaussian r.v.s, the conditional expectation is affine in Z .)

Exercise 4.5: Suppose that T is a component lifetime that is exponential with parameter λ . To model manufacturing variability, we assume that λ is itself random. Specifically, suppose that λ is uniform on $[0, 1]$.

1. Show that $E[T|\lambda] = 1/\lambda$. (Here, $E[T|\lambda]$ is nonlinear in λ (as opposed to Problem 3.4). This is the usual case.)
2. Compute $E[\lambda|T]$.

The following problem enumerates some of the basic properties of conditional expectation, as defined through Definition 3.4.

Exercise 4.6: Suppose that $Y_1, Y_2, \dots, Y_n \in L^2$.

1. Prove that

$$E[Y_1 + Y_2 + \dots + Y_n|Z] = \sum_{i=1}^n E[Y_i|Z] \text{ a.s.}$$

2. If $Y_1 \geq 0$ a.s., prove that

$$E[Y_1|Z] \geq 0 \text{ a.s.}$$

3. If $V = h(Z)$ for some deterministic h , prove that

$$E[Y_1|V] = E[E[Y_1|Z]|V] \text{ a.s.}$$

The above development shows that when $Y \in L^2$, the conditional expectation $E[Y|Z]$ can be viewed as the best mean square predictor based on Z . The conditional expectation can actually be uniquely extended to integrable r.v.s. The following problem indicates the steps.

Exercise 4.7: Suppose that $E[|Y|] < \infty$. To define $E[Y|Z]$, one approximates Y via square-integrable approximations. Put $Y^+ = \max(Y, 0)$.

1. Prove that $E[Y^+ \wedge n|Z]$ is a non-decreasing (in n) sequence, where $a \wedge b \triangleq \min(a, b)$.
2. Put $E[Y^+|Z] = \lim_{n \rightarrow \infty} E[Y^+ \wedge n|Z]$. Similarly, define $E[Y^-|Z]$, where $Y^- = \max(-Y, 0)$. Define $E[Y|Z] \triangleq E[Y^+|Z] - E[Y^-|Z]$.
Prove that $E[Y^+|Z]$ and $E[Y^-|Z]$ are a.s. finite valued, so $E[Y^+|Z] - E[Y^-|Z]$ is well defined.
3. Prove that $E[Yg(Z)] = E[E[Y|Z]g(Z)]$ for all deterministic bounded $g(\cdot)$.

4.6 Affine Prediction

The Hilbert space projection theorem can also be used to solve “constrained” prediction problems. For example, suppose that we have a real-time environment in which the predictions must be computed (very) rapidly. Given the integrals involved in computing $g^*(\cdot)$ and the numerical complexity of potentially evaluating a non-linear $g^*(\cdot)$, we may prefer to restrict ourselves to a smaller subclass of predictors with better computational properties.

Assume that $Z \in \mathbb{R}^d$ satisfies $E[\|Z\|_2^2] < \infty$, and put

$$\mathcal{L}_1 = \{W \in L^2 : W = a^T Z + b\}$$

where $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are deterministic. The linear subspace $\mathcal{L}_1 \subseteq L^2$ is the class of affine predictors based on Z . This subspace is closed in L^2 , so the Hilbert space projection theorem applies. In particular the best mean square predictor $\hat{Y} = a^{*T} Z + b^* \in \mathcal{L}_1$ satisfies

$$\langle Y - a^{*T} Z - b^*, W \rangle = 0$$

for $W \in \mathcal{L}_1$. In other words,

$$\mathbb{E}[YW] = a^{*T} \mathbb{E}[ZW] + b^* \mathbb{E}W$$

for each $W = a^T Z + b$ with $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Among the choices for W are $W = Z_i$ (Z_i is the i -th component of Z) for $1 \leq i \leq d$ and $W = 1$. This yields the linear system

$$\begin{pmatrix} \mathbb{E}[ZZ^T] & \mathbb{E}[Z] \\ \mathbb{E}[Z^T] & 1 \end{pmatrix} \begin{pmatrix} a^* \\ b^* \end{pmatrix} = \begin{pmatrix} \mathbb{E}[YZ] \\ \mathbb{E}[Y] \end{pmatrix}$$

If the coefficient matrix is non-singular (as will occur when the covariance matrix of Z is positive definite; see the next section on Gaussian random variables), then

$$\begin{aligned} \hat{Y} &= a^{*T} Z + b^* \\ &= (\mathbb{E}[Z^T Y] - \mathbb{E}[Z^T] \mathbb{E}[Y]) C^{-1} (Z - \mathbb{E}[Z]) \end{aligned}$$

where C is the covariance matrix of Z given by

$$C = \mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^T]$$

Note that when Y and Z are jointly Gaussian, the conditional expectation $\mathbb{E}[Y|Z]$ and the best affine predictor coincide, so there is no degradation in predictor quality by restricting our class of predictors to affine predictors. The same approach can be used when restricting to linear predictors.

Exercise 4.8: Suppose that Y is real-valued and Z is \mathbb{R}^d -valued. If $\mathbb{E}[Y^2] < \infty$ and $\mathbb{E}[Z^T Z] < \infty$, compute the best linear predictor of Y given Z (using mean square predictor error). In other words, find the best predictor in the class

$$\mathcal{L}_2 = \{W \in L^2 : W = a^T Z \text{ for some deterministic } a \in \mathbb{R}^d\}$$

