JEAN WALRAND

# PROBABILITY IN EE & CS

## An Application-Driven Course

AMAZON

# *Preface*

This book is about applications of probability in Electrical Engineering and Computer Science. It is not a survey of all the important applications. That would be too ambitious. Rather, the course describes real, important, and representative applications that make use of a fairly wide range of probability concepts and techniques.

Probabilistic modeling and analysis are essential skills for computer scientists and electrical engineers. These skills are as important as calculus and discrete mathematics. The systems that these scientists and engineers use or design are complex and operate in an uncertain environment. Understanding and quantifying the impact of this uncertainty are critical to the design of systems.

The book was written for the upper division course EECS126 "Probability in EECS" in the Department of Electrical Engineering and Computer Sciences of the University of California, Berkeley. The students have taken an elementary course on probability. They know the concepts of event, probability, conditional probability, Bayes' rule, discrete random variables and their expectation. They also have some basic familiarity with matrix operations. The students in this class are smart, hard-working and interested in clever and sophisticated ideas. After taking this course, the students should be familiar with Markov chains, stochastic dynamic programming, detection, and estimation. They should have both an intuitive understanding and a working knowledge of these concepts and their methods.

In contrast to most introductory books on probability, the material is organized by applications. Instead of the usual sequence – probability space, random variables, expectation, detection, estimation, Markov chains – we start each topic with a concrete, real and important EECS application.

We introduce the theory as it is needed to study the applications. We believe that this approach can make the theory more relevant by demonstrating its usefulness as we introduce it. Moreover, an emphasis is on hands-on projects where the students use Matlab and Simulink to simulate and calculate. (Python versions will be available in fall 2014 from the book's website.) We designed these projects carefully to reinforce the intuitive understanding of the concepts and to prepare the students for their own investigations. The chapters, except for the last one and the appendices, are divided into two parts: A and B. Parts A contain the key ideas that should be accessible to junior level students. Parts B contain more difficult aspects of the material. It is possible to teach only the appendices and parts A. This would constitute a good junior level course. One possible approach is to teach parts A in a first course and parts B in a second course. For a more ambitious course, one may teach parts A, then parts B. It is also possible to teach the chapters in order. The last chapter is a collection of more advanced topics that the reader and instructor can choose from.

THE APPENDICES should be useful for most readers. Appendix A reviews the basic concepts of probability. Depending on the background of the students, it may be recommended to start the course with a review of that appendix. Appendix B reviews some basic Linear Algebra. Appendix C presents a number of Matlab examples. Many colleges have a Matlab site license for their students. Also, the student edition is quite affordable. Trying the algorithms and simulating the systems in Matlab is very useful to make the ideas more concrete and to develop a sense for how things work. It is also quite satisfying to see that even apparently complex looking algorithms are readily implementable with a minimum of effort.

THE THEORY STARTS WITH models of uncertain quantities. Let us denote such quantities by $\mathbf{X}$ and $\mathbf{Y}$. A model enables to calculate the expected value $E(h(\mathbf{X}))$ of a function $h(\mathbf{X})$ of $\mathbf{X}$. For instance, $\mathbf{X}$ might specify the output of a solar panel every day during one month and $h(\mathbf{X})$ the total energy that the panel produced. Then $E(h(\mathbf{X}))$ is the average energy that the panel produces per month. Other examples are the average delay of packets in a communication network or the average time a data center takes to complete one job.

$$\mathbf{X} \overset{?}{\to} E(h(\mathbf{X}))$$

Evaluation

Estimating $E(h(\mathbf{X}))$ is called performance evaluation. In many cases, the system that handles the uncertain quantities has some parameters $\theta$ that one can select to tune its operations. For instance, the orientation of the solar panels can be

adjusted. Similarly, one may be able to tune the operations of a data center. One may model the effect of the parameters by a function $h(\mathbf{X}, \theta)$ that describes the measure of performance in terms of the uncertain quantities $\mathbf{X}$ and the tuning parameters $\theta$.

$$\max_{\theta} E(h(\mathbf{X}, \theta))$$

Optimization

One important problem is then to find the values of the parameters $\theta$ that maximize $E(h(\mathbf{X}, \theta))$. This is not a simple problem if one does not have an analytical expression for this average value in terms of $\theta$. We explain such optimization problems in the book.

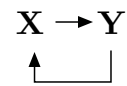
There are many situations where one observes $\mathbf{Y}$ and one is interested in guessing the value of $\mathbf{X}$, which is not observed. As an example, $\mathbf{X}$ may be the signal that a transmitter sends and $\mathbf{Y}$ the signal that the receiver gets.

$$\mathbf{Y} \overset{?}{\rightarrow} \mathbf{X}$$

Inference

The problem of guessing $\mathbf{X}$ on the basis of $\mathbf{Y}$ is an inference problem. Examples include detection problems (Is there a fire? Do you have the flu?) and estimation problems (Where is the iPhone given the GPS signal?).

Finally, there is a class of problems where one uses the observations to act upon a system that then changes. For instance, a self-driving car uses observations from laser range finders, GPS and cameras to steer the car. These are control problems.




Control

Thus, the course discusses performance evaluation, optimization, inference, and control problems. Some of these topics are called artificial intelligence in computer science and statistical signal processing in electrical engineering. Probabilists call them examples. Mathematicians may call them particular cases. The techniques used to address these topics are introduced by looking at concrete applications such as web search, multiplexing, digital communication, speech recognition, tracking, route planning and recommendation systems. Along the way we will meet some of the giants of the field.

This material is truly exciting and fun. I hope you will share my enthusiasm for the ideas.

I AM GRATEFUL to my colleagues and students who made this book possible. Special thanks to Dr. Longbo Huang and Ramtin Pedarsani for their careful reading of the manuscript, to Dr. Abhay Parekh, Professors David Aldous, Venkat Anantharam, Tom Courtade, Michael Lustig, John Musacchio, Kannan Ramchandran, Anant Sahai, David Tse, Martin Wainwright and Avideh Zakhor for their useful comments, to Stephan Adams, Vijay Kamble, Dr. Shiang Jiang, Dr. Sudeep Kamath, Jerome Thai, Dr. Baosen Zhang and Professor Antonis Dimakis for serving as TAs for the course and designing assignments, to Professors Pravin Varaiya and Eugene Wong for teaching me Probability, to Professor Tsu-Jae King Liu for her support, and to the students in EECS126 for their feed-

back.

THE WEBSITE

https://sites.google.com/site/walrandpeecs/home

provides additional resources for this book, such as an Errata, Additional Problems, Comments, Slides and more. Teachers can also use that website to request problem solutions.

A 'MUTLITOUCH' EBOOK VERSION for iPads and MACs is available from the iTunes store and can be located using the book website.

Jean Walrand
Berkeley, August 2014

# 7
# *Tracking - A*

**Background**

*A*

**Application:** Estimation, Tracking
**Topics:** LLSE, MMSE, Kalman Filter

A radar measures electromagnetic waves that an object reflects and uses the measurements to estimate the position of that object.

A GPS receiver uses the signals it gets from satellites to estimate its location. Temperature and pressure sensors provide signals that a computer uses to estimate the state of a chemical reactor.

Similarly, your car's control computer estimates the state of the car from measurements it gets from various sensors.

This chapter explains how to estimate an unobserved random variable or vector from available observations.

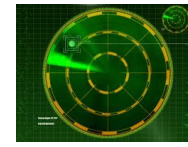


Figure 7.1: Estimating the position of an object from radar signals.

Figure 7.2: Estimating the location of a device from satellite signals.

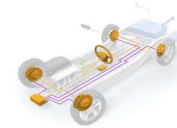


Figure 7.3: Estimating the state of a vehicle from sensor signals.

## 7.1 *Estimation Problem*

The basic estimation problem can be formulated as follows. There is a pair of continuous random variables $(X, Y)$. The problem is to estimate $X$ from the observed value of $Y$.

This problem admits a few different formulations:

- **Known Distribution**: We know the joint distribution of $(X, Y)$;
- **Off-Line**: We observe a set of sample values of $(X, Y)$;
- **On-Line**: We observe successive values of samples of $(X, Y)$;

The objective is to choose the inference function $g(\cdot)$ to minimize the expected error $C(g)$ where

$$C(g) = E(c(X, g(Y))).$$

In this expression, $c(X, \hat{X})$ is the cost of guessing $\hat{X}$ when the actual value is $X$. A standard example is

$$c(X, \hat{X}) = |X - \hat{X}|^2.$$

We will also study the case when $X \in \Re^d$ for $d > 1$. In such a situation, one uses $c(X, \hat{X}) = ||X - \hat{X}||^2$. In this case, the corresponding best guess is said to be the *Least Squares Estimate (LSE)* estimate of $X$ given $Y$. If the function $g(\cdot)$ can be arbitrary, it is the *Minimum Mean Squares Estimate (MMSE)* estimate of $X$ given $Y$. If the function $g(\cdot)$ is restricted to be linear, i.e., of the form $a + BY$, it is the *Linear Least Squares Estimate (LLSE)* estimate of $X$ given $Y$. One may also restrict $g(\cdot)$ to be a polynomial of a given degree. For instance, one may define the Quadratic Least Squares Estimate $QLSE$ of $X$ given $Y$. See Figure 7.4.

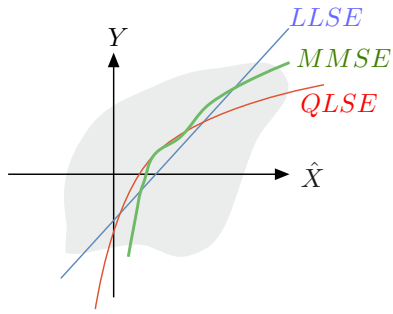



Figure 7.4: Least squares estimates of $X$ given $Y$: $LLSE$ is linear, $QLSE$ is quadratic, and $MMSE$ can be an arbitrary function.

As we will see, a general method for the off-line inference problem is to choose a parametric class of functions $\{g_w, w \in \Re^d\}$ and to then minimize the empirical error

$$\sum_{k=1}^{K} c(X_k, g_w(Y_k))$$

over the parameters $w$. Here, the $(X_k, Y_k)$ are the observed samples.

For the on-line problem, one also chooses a similar parametric family of functions and one uses a stochastic gradient algorithm of the form

$$w(k+1) = w(k) - \gamma \nabla_w c(X_{k+1}, g_w(Y_{k+1}))$$

where $\nabla$ is the gradient with respect to $w$ and $\gamma > 0$ is a small step size. The justification for this approach is that, since $\gamma$ is small, by the SLLN, the update tends to be in the direction of

$$-\sum_{i=k}^{k+K-1} \nabla_w c(X_{i+1}, g_w(Y_{i+1}) \approx -K \nabla E(c(X_k, g_w(Y_k)) = -K \nabla C(g_w),$$

which would correspond to a gradient algorithm to minimize $C(g_w)$.

## 7.2 *Linear Least Squares Estimates*

In this section, we study the linear least squares estimates. Recall the setup that we explained in the previous section. There is a pair $(X, Y)$ of random variables with some joint distribution and the problem is to find the function $g(Y) = a + bY$ that minimizes

$$C(g) = E(|X - g(Y)|^2).$$

One consider the cases where the distribution is known, or a set of samples has been observed, or one observes one sample at a time.

Assume that the joint distribution of $(X, Y)$ is known. This means that we know the *joint cumulative distribution function (j.c.d.f.)* $F_{X,Y}(x, y)$.[1]

[1] See Appendix A.

We are looking for the function $g(Y) = a + bY$ that minimizes

$$C(g) = E(|X - g(Y)|^2) = E(|X - a - bY|^2).$$

We denote this function by $L[X|Y]$. Thus, we have the following definition.

**Definition 7.1** *Linear Least Squares Estimate (LLSE) The LLSE of $X$ given $Y$, denoted by $L[X|Y]$, is the linear function $a + bY$ that minimizes*

$$E(|X - a - bY|^2).$$

◇

Note that

$$\begin{aligned} C(g) &= E(X^2 + a^2 + b^2Y^2 - 2aX - 2bXY + 2abY) \\ &= E(X^2) + a^2 + b^2E(Y^2) - 2aE(X) - 2bE(XY) + 2abE(Y). \end{aligned}$$

To find the values of $a$ and $b$ that minimize that expression, we set to zero the partial derivatives with respect to $a$ and $b$. This gives the following two equations:

$$0 = 2a - 2E(X) + 2bE(Y) \tag{7.1}$$
$$0 = 2bE(Y^2) - 2E(XY) + 2aE(Y). \tag{7.2}$$

Solving these equations for $a$ and $b$, we find that

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

where we used the identities

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

We summarize this result as a theorem.

**Theorem 7.2** *Linear Least Squares Estimate*
*One has*

$$L[X|Y] = E(X) + \frac{cov(X,Y)}{var(Y)}(Y - E(Y)). \tag{7.3}$$

□

As a first example, assume that

$$Y = \alpha X + Z \tag{7.4}$$

where $X$ and $Z$ are zero-mean and independent. In this case, we find [2]

[2] Indeed, $E(XZ) = E(X)E(Z) = 0$, by independence.

$$\begin{aligned}\text{cov}(X,Y) &= E(XY) - E(X)E(Y)\\ &= E(X(\alpha X + Z)) = \alpha E(X^2)\\ \text{var}(Y) &= \alpha^2 \text{var}(X) + \text{var}(Z) = \alpha^2 E(X^2) + E(Z^2).\end{aligned}$$

Hence,

$$L[X|Y] = \frac{\alpha E(X^2)}{\alpha^2 E(X^2) + E(Z^2)} Y = \frac{\alpha^{-1} Y}{1 + SNR^{-1}}$$

where

$$SNR := \frac{\alpha^2 E(X^2)}{\sigma^2}$$

is the *signal-to-noise ratio*, i.e., the ratio of the power $E(\alpha^2 X^2)$ of the signal in $Y$ divided by the power $E(Z^2)$ of the noise. Note that if $SNR$ is small, then $L[X|Y]$ is close to zero, which is the best guess about $X$ if one does not make any observation. Also, if $SNR$ is very large, then $L[X|Y] \approx \alpha^{-1} Y$, which is the correct guess if $Z = 0$.

As a second example, assume that

$$X = \alpha Y + \beta Y^2 \tag{7.5}$$

where[3] $Y =_D U[0,1]$. Then,

[3] Thus, $E(Y^k) = (1+k)^{-1}$.

$$\begin{aligned}E(X) &= \alpha E(Y) + \beta E(Y^2) = \alpha/2 + \beta/3;\\ \text{cov}(X,Y) &= E(XY) - E(X)E(Y)\\ &= E(\alpha Y^2 + \beta Y^3) - (\alpha/2 + \beta/3)(1/2)\\ &= \alpha/3 + \beta/4 - \alpha/4 - \beta/6\\ &= (\alpha + \beta)/12\\ \text{var}(Y) &= E(Y^2) - E(Y)^2 = 1/3 - (1/2)^2 = 1/12.\end{aligned}$$

Hence,

$$L[X|Y] = \alpha/2 + \beta/3 + (\alpha + \beta)(Y - 1/2) = -\beta/6 + (\alpha + \beta)Y.$$

This estimate is sketched in Figure 7.5. Obviously, if one observes $Y$, one can compute $X$. However, recall that $L[X|Y]$ is restricted to being a linear function of $Y$.

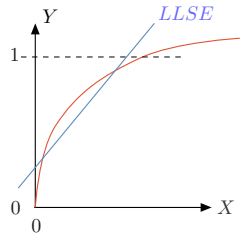



Figure 7.5: The figure shows $L[\alpha Y + \beta Y^2|Y]$ when $Y =_D U[0,1]$.

## *Projection*

There is an insightful interpretation of $L[X|Y]$ as a projection that also helps understand more complex estimates. This interpretation is that $L[X|Y]$ is the projection of $X$ onto the set $\mathcal{L}(Y)$ of linear functions of $Y$.

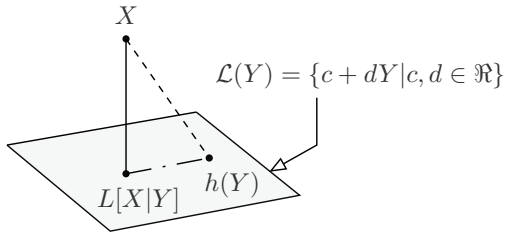



Figure 7.6: $L[X|Y]$ is the projection of $X$ onto $\mathcal{L}(Y)$.

This interpretation is sketched in Figure 7.6. In that figure, random variables are represented by points and $\mathcal{L}(Y)$ is shown as a plane since the linear combination of points in that set is again in the set. In the figure, the square of the length of a vector from a random variable $V$ to another random variable $W$ is $E(|V - W|^2)$. Also, we say that two vectors $V$ and $W$ are orthogonal if $E(VW) = 0$. Thus, $L[X|Y] = a + bY$ is the projection of $X$ onto $\mathcal{L}(Y)$ if $X - L[X|Y]$ is orthogonal to every linear function of $Y$, i.e., if

$$E((X - a - bY)(c + dY)) = 0, \forall c, d \in \Re.$$

Equivalently,

$$E(X) = a + bE(Y) \text{ and } E((X - a - bY)Y) = 0. \tag{7.6}$$

These two equations are the same as (7.1)-(7.2). We call the identities (7.6) the projection property.

FIGURE 7.7 ILLUSTRATES the projection when

$$X = \mathcal{N}(0,1) \text{ and } Y = X + Z \text{ where } Z = \mathcal{N}(0,\sigma^2).$$

In this figure, the length of $Z$ is equal to $\sqrt{E(Z^2)} = \sigma$, the length of $X$ is $\sqrt{E(X^2)} = 1$ and the vectors $X$ and $Z$ are orthogonal because $E(XZ) = 0$.

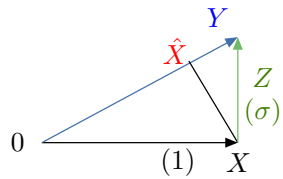



Figure 7.7: Example of projection.

We see that the triangles $0\hat{X}X$ and $0XY$ are similar. Hence,

$$\frac{||\hat{X}||}{||X||} = \frac{||X||}{||Y||},$$

so that

$$\frac{||\hat{X}||}{1} = \frac{1}{\sqrt{1+\sigma^2}} = \frac{||Y||}{1+\sigma^2},$$

since $||Y|| = \sqrt{1+\sigma^2}$. This shows that

$$\hat{X} = \frac{1}{1+\sigma^2} Y.$$

TO SEE WHY THE PROJECTION PROPERTY implies that $L[X|Y]$ is the closest point to $X$ in $\mathcal{L}(Y)$, as suggested by Figure 7.6, we verify that

$$E(|X - L[X|Y]|^2) \leq E(|X - h(Y)|^2),$$

for any given $h(Y) = c + dY$. The idea of the proof is to verify Pythagora's identity on the right triangle with vertices $X, L[X|Y]$ and $h(Y)$. We have

$$\begin{aligned} E(|X - h(Y)|^2) &= E(|X - L[X|Y] + L[X|Y] - h(Y)|^2) \\ &= E(|X - L[X|Y]|^2) + E(|L[X|Y] - h(Y)|^2) \\ &\quad + 2E((X - L[X|Y])(L[X|Y] - h(Y))). \end{aligned}$$

Now, the projection property (7.6) implies that the last term in the above expression is equal to zero. Indeed, $L[X|Y] - h(Y)$ is a linear function of $Y$. It follows that

$$\begin{aligned} E(|X - h(Y)|^2) &= E(|X - L[X|Y]|^2) + E(|L[X|Y] - h(Y)|^2) \\ &\geq E(|X - L[X|Y]|^2), \end{aligned}$$

as was to be proved.

## 7.3 *Linear Regression*

Assume now that, instead of knowing the joint distribution of $(X, Y)$, we observe $K$ i.i.d. samples $(X_1, Y_1), \ldots, (X_K, Y_K)$ of these random variables. Our goal is still to construct a function $g(Y) = a + bY$ so that

$$E(|X - a - bY|^2)$$

is minimized. We do this by choosing $a$ and $b$ to minimize the sum of the squares of the errors based on the samples. That is, we choose $a$ and $b$ to minimize

$$\sum_{k=1}^{K} |X_k - a - bY_k|^2.$$

To do this, we set to zero the derivatives of this sum with respect to $a$ and $b$. Algebra shows that the resulting values of $a$ and $b$ are such that

$$a + bY = E_K(X) + \frac{\text{cov}_K(X, Y)}{\text{var}_K(Y)}(Y - E_K(Y)) \qquad (7.7)$$

where we defined

$$\begin{aligned} E_K(X) &= \frac{1}{K}\sum_{k=1}^{K} X_k, E_K(Y) = \frac{1}{K}\sum_{k=1}^{K} Y_k, \\ \text{cov}_K(X, Y) &= \frac{1}{K}\sum_{k=1}^{K} X_k Y_k - E_K(X)E_K(Y), \\ \text{var}_K(Y) &= \frac{1}{K}\sum_{k=1}^{K} Y_k^2 - E_K(Y)^2. \end{aligned}$$

That is, the expression (7.7) is the same as (7.3), except that the expectation is replaced by the sample mean. The expression (7.7) is called the *linear regression* of $X$ over $Y$. It is shown in Figure 7.8.

One has the following result.

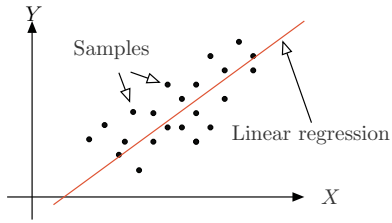



Figure 7.8: The linear regression of $X$ over $Y$.

**Theorem 7.3** *Linear Regression Converges to LLSE*

*As the number of samples increases, the linear regression approaches the LLSE.* □

**Proof:**

As $K \to \infty$, one has, by the Strong Law of Large Numbers,

$$E_K(X) \to E(X), E_K(Y) \to E(Y),$$
$$\text{cov}_K(X,Y) \to \text{cov}(X,Y), \text{var}_K(Y) \to \text{var}(Y).$$

Combined with the expressions for the linear regression and the LLSE, these properties imply the result. ♣

FORMULA (7.3) AND THE LINEAR REGRESSION PROVIDE an intuitive meaning of the covariance $\text{cov}(X,Y)$. If this covariance is zero, then $L[X|Y]$ does not depend on $Y$. If it is positive (negative), it increases (decreases, respectively) with $Y$. Thus, $\text{cov}(X,Y)$ measures a form of dependency in terms of linear regression. For instance, the random variables in Figure 7.9 are uncorrelated since $L[X|Y]$ does not depend on $X$.

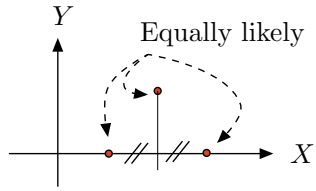



Figure 7.9: The random variables $X$ and $Y$ are uncorrelated. Note that they are not independent.

## 7.4 MMSE

In the previous section, we examined the problem of finding the linear function $a + bY$ that best approximates $X$, in the mean squared error sense. We could develop the corresponding theory for quadratic approximations $a + bY + cY^2$,

or for polynomial approximations of a given degree. The ideas would be the same and one would have a similar projection interpretation.

In principle, a higher degree polynomial approximates $X$ better than a lower degree one since there are more such polynomials. The question of fitting the parameters with a given number of observations is more complex. For instance, if one has $K$ samples, one can generically find a polynomial of degree $K$ that fits the observations exactly. However, this does not imply that this polynomial results in a smaller mean squared error than a lower-degree polynomial. This issue is called "over fitting."

For now, assume that we know the joint distribution of $(X, Y)$ and consider the problem of finding the function $g(Y)$ that minimizes

$$E(|X - g(Y)|^2),$$

per all the possible functions $g(\cdot)$. The best function is called the MMSE of $X$ given $Y$. We have the following theorem:

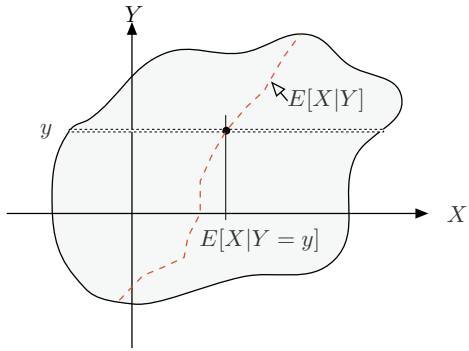



Figure 7.10: The conditional expectation $E[X|Y]$ when the pair $(X, Y)$ is picked uniformly in the shaded area.

**Theorem 7.4** *The MMSE is the Conditional Expectation*
*The MMSE of $X$ given $Y$ is given by*

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

*where $E[X|Y]$ is the* conditional expectation *of $X$ given $Y$.* □

Before proving this result, we need to define the conditional expectation.

**Definition 7.5** *Conditional Expectation*
*The conditional expectation of $X$ given $Y$ is defined by*

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}[x|y]dx$$

*where*

$$f_{X|Y}[x|y] := \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

*is the conditional density of $X$ given $Y$.* ◇

Figure 7.10 illustrates the conditional expectation. That figure assumes that the pair $(X,Y)$ is picked uniformly in the shaded area. Thus, if one observes that $Y \in (y, y+dy)$, the point $X$ is uniformly distributed along the segment that cuts the shaded area at $Y = y$. Accordingly, the average value of $X$ is the mid-point of that segment, as indicated in the figure. The dashed red line shows how that mean value depends on $Y$ and it defines $E[X|Y]$.

The following result is a direct consequence of the definition.

**Lemma 7.6** *Orthogonality Property of MMSE*
*(a) For any function $\phi(\cdot)$, one has*

$$E((X - E[X|Y])\phi(Y)) = 0. \tag{7.8}$$

*(b) Moreover, if the function $g(Y)$ is such that*

$$E((X - g(Y))\phi(Y)) = 0, \forall \phi(\cdot), \tag{7.9}$$

*then $g(Y) = E[X|Y]$.*

**Proof:**
(a) To verify (7.8) note that

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

which proves (7.8).
(b) To prove the second part of the lemma, note that

$$\begin{aligned} &E(|g(Y) - E[X|Y]|^2) \\ &\quad = E((g(Y) - E[X|Y])\{(g(Y) - X) - (E[X|Y] - X)\}) = 0, \end{aligned}$$

because of (7.8) and (7.9) with $\phi(Y) = g(Y) - E[X|Y]$.

Note that the second part of the lemma simply says that the projection property characterizes uniquely the conditional expectation. In other words, there is only one projection of $X$ onto $\mathcal{G}(Y)$. ♣

We can now prove the theorem.

**Proof of Theorem** 7.4.

The identity (7.8) is the projection property. It states that $X - E[X|Y]$ is orthogonal to the set $\mathcal{G}(Y)$ of functions of $Y$, as shown in Figure 7.11. In particular, it is orthogonal

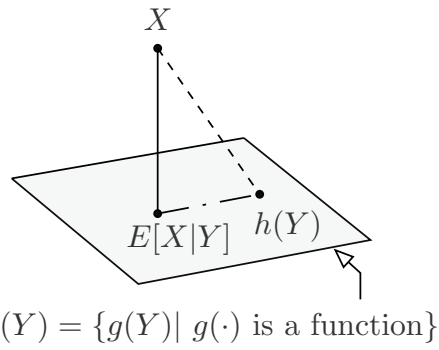



Figure 7.11: The conditional expectation $E[X|Y]$ as the projection of $X$ on the set $\mathcal{G}(Y)$ of functions of $Y$.

to $h(Y) - E[X|Y]$. As in the case of the LLSE, this projection property implies that

$$E(|X - h(Y)|^2) \geq E(|X - E[X|Y]|^2),$$

for any function $h(\cdot)$. This implies that $E[X|Y]$ is indeed the MMSE of $X$ given $Y$. ♣

From the definition, we see how to calculate $E[X|Y]$ from the conditional density of $X$ given $Y$. However, in many cases one can calculate $E[X|Y]$ more simply. One approach is to use the following properties of conditional expectation.

**Theorem 7.7** *Properties of Conditional Expectation* *(a) Linearity:*

$$E[a_1X_1 + a_2X_2|Y] = a_1E[X_1|Y] + a_2E[X_2|Y];$$

*(b) Factoring Known Values:*

$$E[h(Y)X|Y] = h(Y)E[X|Y];$$

*(c) Smoothing:*

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

*(d) Independence: If X and Y are independent, then*

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

□

**Proof:**

(a) By Lemma 7.6(b), it suffices to show that

$$a_1X_1 + a_2X_2 - (a_1E[X_1|Y] + a_2E[X_2|Y])$$

is orthogonal to $\mathcal{G}(Y)$. But this is immediate since it is the sum of two terms

$$a_i(X_i - E[X_i|Y])$$

for $i = 1, 2$ that are orthogonal to $\mathcal{G}(Y)$.

(b) By Lemma 7.6(b), it suffices to show that

$$h(Y)X - h(Y)E[X|Y]$$

is orthogonal to $\mathcal{G}(Y)$, i.e., that

$$E((h(Y)X - h(Y)E[X|Y])\phi(Y)) = 0, \forall \phi(\cdot).$$

Now,

$$E((h(Y)X - h(Y)E[X|Y])\phi(Y)) = E((X - E[X|Y])h(Y)\phi(Y)) = 0,$$

because $X - E[X|Y]$ is orthogonal to $\mathcal{G}(Y)$ and therefore to $h(Y)\phi(Y)$.

(c) Letting $\phi(Y) = 1$ in (7.8), we find

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

which is the identity we wanted to prove.

(d) By Lemma 7.6(b), it suffices to show that

$$X - E(X)$$

is orthogonal to $\mathcal{G}(Y)$. Now,

$$E((X - E(X))\phi(Y)) = E(X - E(X))E(\phi(Y)) = 0.$$

The first equality follows from the fact that $X - E(X)$ and $\phi(Y)$ are independent since they are functions of independent random variables.[4]

[4] See Appendix A.

♣

As an example, assume that $X, Y, Z$ are i.i.d. $U[0, 1]$. We want to calculate

$$E[(X + 2Y)^2|Y].$$

We find

$$\begin{aligned}
E[(X + 2Y)^2|Y] &= E[X^2 + 4Y^2 + 4XY|Y] \\
&= E[X^2|Y] + 4E[Y^2|Y] + 4E[XY|Y], \text{ by linearity} \\
&= E(X^2) + 4E[Y^2|Y] + 4E[XY|Y], \text{ by independence} \\
&= E(X^2) + 4Y^2 + 4YE[X|Y], \text{ by factoring known values} \\
&= E(X^2) + 4Y^2 + 4YE(X), \text{ by independence} \\
&= \frac{1}{3} + 4Y^2 + 2Y, \text{ since } X =_D U[0, 1].
\end{aligned}$$

Note that calculating the conditional density of $(X+2Y)^2$ given $Y$ would have been quite a bit more tedious.

In some situations, one may be able to exploit symmetry to evaluate the conditional expectation. Here is one representative example. Assume that $X, Y, Z$ are i.i.d. Then, we claim that

$$E[X|X+Y+Z] = \frac{1}{3}(X+Y+Z). \tag{7.10}$$

To see this, note that, by symmetry,

$$E[X|X+Y+Z] = E[Y|X+Y+Z] = E[Z|X+Y+Z].$$

Denote by $V$ the common value of these random variables. Note that their sum is

$$3V = E[X+Y+Z|X+Y+Z],$$

by linearity. Thus, $3V = X+Y+Z$, which proves our claim.

## *MMSE for Jointly Gaussian*

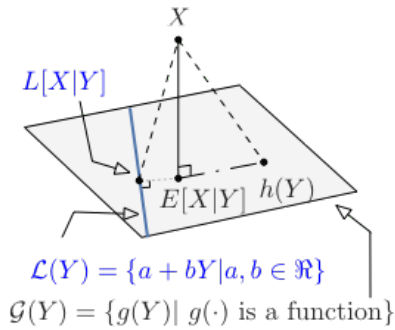



Figure 7.12: The MMSE and LLSE are generally different.

In general $L[X|Y] \neq E[X|Y]$. As a trivial example, Let $Y =_D U[-1,1]$ and $X = Y^2$. Then $E[X|Y] = Y^2$ and $L[X|Y] = E(X) = 1/3$ since $\text{cov}(X,Y) = E(XY) - E(X)E(Y) = 0$.

Figure 7.12 recalls that $E[X|Y]$ is the projection of $X$ onto $\mathcal{G}(Y)$ whereas $L[X|Y]$ is the projection of $X$ onto $\mathcal{L}(Y)$. Since $\mathcal{L}(Y)$ is a subspace of $\mathcal{G}(Y)$, one expects the two projections to be different, in general.

However, there are examples where $E[X|Y]$ happens to be linear. We saw one such example in (7.10) and it is not difficult to construct many other examples.

There is an important class of problems where this occurs. It is when $X$ and $Y$ are jointly Gaussian. We state that result as a theorem.

**Theorem 7.8** *MMSE for Jointly Gaussian RVs*
*Let $X, Y$ be jointly Gaussian random variables. Then*

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

□

**Proof:** Note that

$$X - L[X|Y] \text{ and } Y \text{ are uncorrelated.}$$

Also, $X - L[X|Y]$ and $Y$ are two linear functions of the jointly Gaussian random variables $X$ and $Y$. Consequently, they are jointly Gaussian by Theorem 6.5 and they are independent by Theorem 6.4.

Consequently,

$$X - L[X|Y] \text{ and } \phi(Y) \text{ are independent,}$$

for any $\phi(\cdot)$, because functions of independent random variables are independent by Theorem A.9 in Appendix A. Hence,

$$X - L[X|Y] \text{ and } \phi(Y) \text{ are uncorrelated,}$$

for any $\phi(\cdot)$ by Theorem A.6 of Appendix A.

This shows that

$$X - L[X|Y] \text{ is orthogonal to } \mathcal{G}(Y),$$

and, consequently, that $L[X|Y] = E[X|Y]$.

♣

## 7.5 Vector Case

So far, to keep notation at a minimum, we have considered $L[X|Y]$ and $E[X|Y]$ when $X$ and $Y$ are single random variables. In this section, we discuss the vector case, i.e., $L[\mathbf{X}|\mathbf{Y}]$ and $E[\mathbf{X}|\mathbf{Y}]$ when $\mathbf{X}$ and $\mathbf{Y}$ are random vectors. The only difficulty is one of notation. Conceptually, there is nothing new.

**Definition 7.9** *LLSE of Random Vectors*
*Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors of dimensions $m$ and $n$, respectively. Then*

$$L[\mathbf{X}|\mathbf{Y}] = A\mathbf{y} + \mathbf{b}$$

*where $A$ is the $m \times n$ matrix and $\mathbf{b}$ the vector in $\Re^m$ that minimize*

$$E(||\mathbf{X} - A\mathbf{Y} - \mathbf{b}||^2).$$

◇

Thus, as in the scalar case, the LLSE is the linear function of the observations that best approximates $\mathbf{X}$, in the mean squared error sense.

Before proceeding, review the notation of Section A.5 for $\Sigma_{\mathbf{Y}}$ and $\text{cov}(\mathbf{X}, \mathbf{Y})$.

**Theorem 7.10** *LLSE of Vectors*

*Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors such that $\Sigma_{\mathbf{Y}}$ is nonsingular.*

*(a) Then*

$$L[\mathbf{X}|\mathbf{Y}] = E(\mathbf{X}) + cov(\mathbf{X}, \mathbf{Y})\Sigma_{\mathbf{Y}}^{-1}(\mathbf{Y} - E(\mathbf{Y})). \tag{7.11}$$

*(b) Moreover,*

$$E(||\mathbf{X} - L[\mathbf{X}|\mathbf{Y}]||^2) = tr(\Sigma_{\mathbf{X}} - cov(\mathbf{X}, \mathbf{Y})\Sigma_{\mathbf{Y}}^{-1}cov(\mathbf{Y}, \mathbf{X})). \tag{7.12}$$

*In this expression, for a square matrix $M$, $tr(M) := \sum_i M_{i,i}$ is the trace of the matrix.* □

**Proof:**

(a) The proof is similar to the scalar case. Let $\mathbf{Z}$ be the right-hand side of (7.11). One shows that the error $\mathbf{X} - \mathbf{Z}$ is orthogonal to all the linear functions of $\mathbf{Y}$. One then uses that fact to show that $\mathbf{X}$ is closer to $\mathbf{Z}$ than to any other linear function $h(\mathbf{Y})$ of $\mathbf{Y}$.

First we show the orthogonality. Since $E(\mathbf{X} - \mathbf{Z}) = 0$, we have

$$E((\mathbf{X} - \mathbf{Z})(B\mathbf{Y} + \mathbf{b})') = E((\mathbf{X} - \mathbf{Z})(B\mathbf{Y})') = E((\mathbf{X} - \mathbf{Z})\mathbf{Y}')B'.$$

Next, we show that $E((\mathbf{X} - \mathbf{Z})\mathbf{Y}') = 0$. To see this, note that

$$\begin{aligned} E((\mathbf{X} - \mathbf{Z})\mathbf{Y}') &= E((\mathbf{X} - \mathbf{Z})(\mathbf{Y} - E(\mathbf{Y}))') \\ &= E((\mathbf{X} - E(\mathbf{X}))(\mathbf{Y} - E(\mathbf{Y}))') \\ &\quad - \text{cov}(\mathbf{X}, \mathbf{Y})\Sigma_{\mathbf{Y}}^{-1}E((\mathbf{Y} - E(\mathbf{Y}))(\mathbf{Y} - E(\mathbf{Y}))') \\ &= \text{cov}(\mathbf{X}, \mathbf{Y}) - \text{cov}(\mathbf{X}, \mathbf{Y})\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}} = 0. \end{aligned}$$

Second, we show that $\mathbf{Z}$ is closer to $\mathbf{X}$ than any linear $h(\mathbf{Y})$. We have

$$\begin{aligned} E(||\mathbf{X} - h(\mathbf{Y})||^2) &= E((\mathbf{X} - h(\mathbf{Y}))'(\mathbf{X} - h(\mathbf{Y}))) \\ &= E((\mathbf{X} - \mathbf{Z} + \mathbf{Z} - h(\mathbf{Y}))'(\mathbf{X} - \mathbf{Z} + \mathbf{Z} - h(\mathbf{Y}))) \\ &= E(||\mathbf{X} - \mathbf{Z}||^2) + E(||\mathbf{Z} - h(\mathbf{Y})||^2) + 2E((\mathbf{X} - \mathbf{Z})'(\mathbf{Z} - h(\mathbf{Y}))). \end{aligned}$$

We claim that the last term is equal to zero. To see this, note that

$$E((\mathbf{X} - \mathbf{Z})'(\mathbf{Z} - h(\mathbf{Y})) = \sum_{i=1}^{n} E((X_i - Z_i)(Z_i - h_i(\mathbf{Y}))).$$

Also,

$$E((X_i - Z_i)(Z_i - h_i(\mathbf{Y}))) = E((\mathbf{X} - \mathbf{Z})(\mathbf{Z} - h(\mathbf{Y}))')_{i,i}$$

and the matrix $E((\mathbf{X} - \mathbf{Z})(\mathbf{Z} - h(\mathbf{Y}))')$ is equal to zero since $\mathbf{X} - \mathbf{Y}$ is orthogonal to any linear function of $\mathbf{Y}$ and, in particular, to $\mathbf{Z} - h(\mathbf{Y})$.

(Note: an alternative way of showing that the last term is equal to zero is to write

$$E((\mathbf{X} - \mathbf{Z})'(\mathbf{Z} - h(\mathbf{Y})) = \text{tr}E((\mathbf{X} - \mathbf{Z})(\mathbf{Z} - h(\mathbf{Y}))') = 0,$$

where the first equality comes from the fact that $\text{tr}(AB) = \text{tr}(BA)$ for matrices of compatible dimensions.)

(b) Let $\tilde{\mathbf{X}} := \mathbf{X} - E[\mathbf{X}|\mathbf{Y}]$ be the estimation error. Thus,

$$\tilde{\mathbf{X}} = \mathbf{X} - E(\mathbf{X}) - \text{cov}(\mathbf{X}, \mathbf{Y})\Sigma_{\mathbf{Y}}^{-1}(\mathbf{Y} - E(\mathbf{Y})).$$

Now, if $\mathbf{V}$ and $\mathbf{W}$ are two zero-mean random vectors and $M$ a matrix,

$$\begin{aligned}\text{cov}(\mathbf{V} - M\mathbf{W}) &= E((\mathbf{V} - M\mathbf{W})(\mathbf{V} - M\mathbf{W})') \\ &= E(\mathbf{V}\mathbf{V}' - 2M\mathbf{W}\mathbf{V}' + M\mathbf{W}\mathbf{W}'M') \\ &= \text{cov}(\mathbf{V}) - 2M\text{cov}(\mathbf{W}, \mathbf{V}) + M\text{cov}(\mathbf{W})M'.\end{aligned}$$

Hence,

$$\begin{aligned}\text{cov}(\tilde{\mathbf{X}}) &= \Sigma_{\mathbf{X}} - 2\text{cov}(\mathbf{X}, \mathbf{Y})\Sigma_{\mathbf{Y}}^{-1}\text{cov}(\mathbf{Y}, \mathbf{X}) \\ &\quad + \text{cov}(\mathbf{X}, \mathbf{Y})\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}}\Sigma_{\mathbf{Y}}^{-1}\text{cov}(\mathbf{Y}, \mathbf{X}) \\ &= \Sigma_{\mathbf{X}} - \text{cov}(\mathbf{X}, \mathbf{Y})\Sigma_{\mathbf{Y}}^{-1}\text{cov}(\mathbf{Y}, \mathbf{X}).\end{aligned}$$

To conclude the proof, note that, for a zero-mean random vector $\mathbf{V}$,

$$E(||\mathbf{V}||^2) = E(\text{tr}(\mathbf{V}\mathbf{V}')) = \text{tr}(E(\mathbf{V}\mathbf{V}')) = \text{tr}(\Sigma_{\mathbf{V}}).$$

♣

## 7.6 *Kalman Filter*

C.7

The Kalman Filter is an algorithm to update the estimate of the state of a system using its output, as sketched in Figure 7.13. The system has a *state* $X(n)$ and an *output* $Y(n)$ at time $n = 0, 1, \ldots$. These variables are defined through a system of linear equations:

$$X(n+1) = AX(n) + V(n), n \geq 0; \tag{7.13}$$

$$Y(n) = CX(n) + W(n), n \geq 0. \tag{7.14}$$

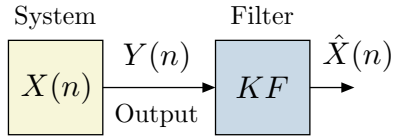



Figure 7.13: The Kalman Filter computes the LLSE of the state of a system given the past of its output.

In these equations, the random variables $\{X(0), V(n), W(n), n \geq 0\}$ are all orthogonal and zero mean. The covariance of $V(n)$ is $\Sigma_V$ and that of $W(n)$ is $\Sigma_W$. The filter is developed when the variables are random vectors and $A, C$ are matrices of compatible dimensions.

The objective is to derive recursive equations to calculate

$$\hat{X}(n) = L[X(n)|Y(0), \ldots, Y(n)], n \geq 0.$$

## *The Filter*

Figure 7.14: Rudolf Kalman, b. 1930.

Here is the result, which we prove in the next chapter. Don't panic when you see the equations!

**Theorem 7.11** *Kalman Filter*
*One has*

$$\hat{X}(n) = A\hat{X}(n-1) + K_n[Y(n) - CA\hat{X}(n-1)] \quad (7.15)$$
$$K_n = S_n C'[CS_n C' + \Sigma_W]^{-1} \quad (7.16)$$
$$S_n = A\Sigma_{n-1}A' + \Sigma_V \quad (7.17)$$
$$\Sigma_n = (I - K_n C)S_n \quad (7.18)$$

*Moreover,*

$$S_n = cov(X(n) - A\hat{X}(n-1)) \text{ and } \Sigma_n = cov(X(n) - \hat{X}(n)). \quad (7.19)$$

□

We will give a number of examples of this result. But first, let us make a few comments.

- The equations (7.15)-(7.18) are recursive: the estimate at time $n$ is a simple linear function of the estimate at time $n-1$ and of the new observation $Y(n)$.
- The matrix $K_n$ is the filter gain. It can be precomputed at time 0.
- The covariance of the error $X(n) - \hat{X}(n)$, $\Sigma_n$, can also be precomputed at time 0: it does not depend on the observations $\{Y(0), \ldots, Y(n)\}$. The estimate $\hat{X}(n)$ depends on these observations but the residual error does not.

- If $X(0)$ and the noise random variables are Gaussian, then the Kalman filter computes the MMSE.
- Finally, observe that these equations, even though they look a bit complicated, can be programmed in a few lines. This filter is elementary to implement and this explains its popularity.

## *Examples*

In this section, we examine a few examples of the Kalman filter.

**Random Walk**

The first example is a filter to track a "random walk" by making noisy observations.

Let

$$X(n+1) = X(n) + V(n) \tag{7.20}$$
$$Y(n) = X(n) + W(n) \tag{7.21}$$
$$\text{var}(V(n)) = 0.04, \text{var}(W(n)) = 0.09. \tag{7.22}$$

That is, $X(n)$ has orthogonal increments and it is observed with orthogonal noise. Figure 7.15 shows a simulation of the filter. The left-hand part of the figure shows that the estimate tracks the state with a bounded error. The middle part of the figure shows the variance of the error, which can be precomputed. The right-hand part of the figure shows the filter with the time-varying gain (in blue) and the filter with the limiting gain (in green). The filter with the constant gain performs as well as the one with the time-varying gain, in the limit, as justified by part (c) of the theorem.

**Random Walk with Unknown Drift**

In the second example, one tracks a random walk that has an unknown drift. This system is modeled by the following equations:

$$X_1(n+1) = X_1(n) + X_2(n) + V(n) \tag{7.23}$$
$$X_2(n+1) = X_2(n) \tag{7.24}$$
$$Y(n) = X_1(n) + W(n) \tag{7.25}$$
$$\text{var}(V(n)) = 1, \text{var}(W(n)) = 0.25. \tag{7.26}$$

In this model, $X_2(n)$ is the constant but unknown drift and $X_1(n)$ is the value of the "random walk." Figure 7.16 shows a simulation of the filter. It shows that the filter eventually estimates the drift and that the estimate of the position of the walk is quite accurate.

Figure 7.15: The Kalman Filter for (7.20)-(7.22).

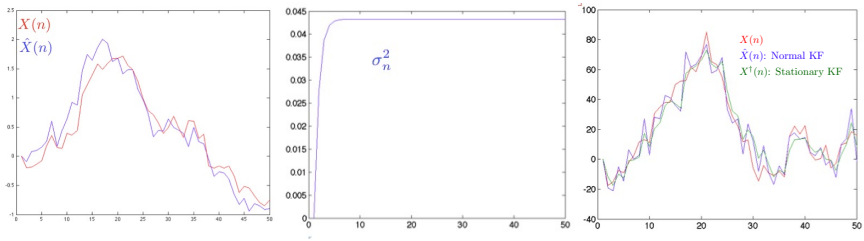
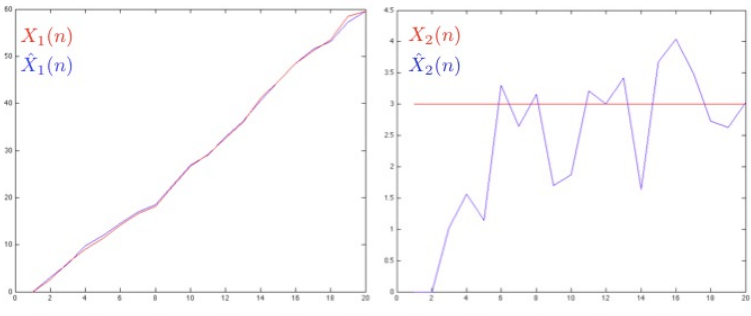



Figure 7.16: The Kalman Filter for (7.23)-(7.26).



**Random Walk with Changing Drift**

In the third example, one tracks a random walk that has changing drift. This system is modeled by the following equations:

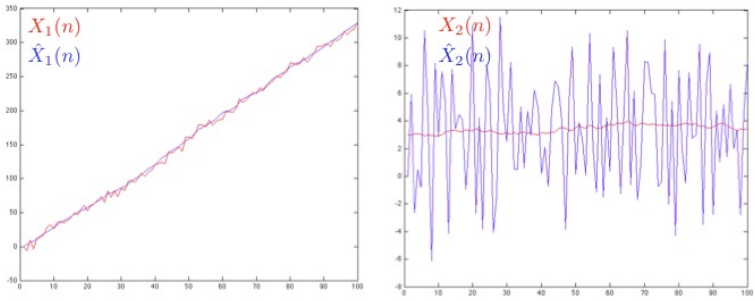



Figure 7.17: The Kalman Filter for (7.27)-(7.31).

$$X_1(n+1) = X_1(n) + X_2(n) + V_1(n) \tag{7.27}$$
$$X_2(n+1) = X_2(n) + V_2(n) \tag{7.28}$$
$$Y(n) = X_1(n) + W(n) \tag{7.29}$$
$$\text{var}(V_1(n)) = 1, \text{var}(V_2(n)) = 0.01, \tag{7.30}$$
$$\text{var}(W(n)) = 0.25. \tag{7.31}$$

In this model, $X_2(n)$ is the varying drift and $X_1(n)$ is the value of the "random walk." Figure 7.17 shows a simulation of the filter. It shows that the filter tries to track the drift and that the estimate of the position of the walk is quite accurate.

**Falling Object**

In the fourth example, one tracks a falling object. The elevation $Z(n)$ of that falling object follows the equation

$$Z(n) = Z(0) + S(0)n - gn^2/2 + V(n), n \geq 0$$

where $S(0)$ is the initial vertical velocity of the object and $g$ is the gravitational constant at the surface of the earth. In this expression, $V(n)$ is some noise that perturbs the motion. We observe $\eta(n) = Z(n) + W(n)$, where $W(n)$ is some noise.

Since the term $-gn^2/2$ is known, we consider

$$X_1(n) = Z(n) + gn^2/2 \text{ and } Y(n) = \eta(n) + gn^2/2.$$

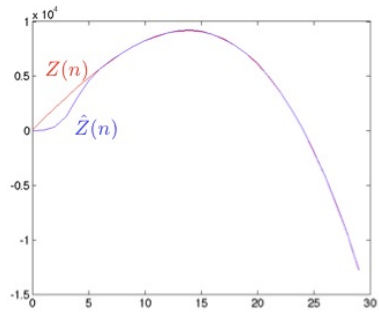



Figure 7.18: The Kalman Filter for (7.32)-(7.35).

With this change of variables, the system is described by the following equations:

$$X_1(n+1) = X_1(n) + X_2(n) + V(n) \tag{7.32}$$
$$X_2(n+1) = X_2(n) \tag{7.33}$$
$$Y(n) = X_1(n) + W(n) \tag{7.34}$$
$$\text{var}(V_1(n)) = 100 \text{ and } \text{var}(W(n)) = 1600. \tag{7.35}$$

Figure 7.18 shows a simulation of the filter that computes $\hat{X}_1(n)$ from which we subtract $gt^2/2$ to get an estimate of the actual altitude $Z(n)$ of the object.

## 7.7 *Summary*

- LLSE, linear regression and MMSE;
- Projection characterization;
- MMSE of jointly Gaussian is linear;
- Kalman Filter.

*Key equations & formulas:*

| | | |
|---|---|---|
| LLSE | $L[X\|Y] = E(X) + \text{cov}(X,Y)\text{var}(Y)^{-1}(Y - E(Y))$ | T. 7.2 |
| Orthogonality | $X - L[X\|Y] \perp a + bY$ | (7.6) |
| Linear Regression | converges to $L[X\|Y]$ | T.7.3 |
| Conditional Expectation | $E[X\|Y] = ...$ | D.7.5 |
| Orthogonality | $X - E[X\|Y] \perp g(Y)$ | L.7.6 |
| MMSE = CE | $MMSE[X\|Y] = E[X\|Y]$ | T.7.4 |
| Properties of CE | Linearity, smoothing, etc... | T.7.7 |
| CE for J.G. | If $X, Y$ J.G., then $E[X\|Y] = L[X\|Y] = \cdots$ | T.7.8 |
| LLSE vectors | $L[\mathbf{X}\|\mathbf{Y}] = E(\mathbf{X}) + \Sigma_{\mathbf{X},\mathbf{Y}}\Sigma_{\mathbf{Y}}^{-1}(\mathbf{Y} - E(\mathbf{Y}))$ | T.7.10 |
| Kalman Filter | $\hat{X}(n) = A\hat{X}(n-1) + K_n[Y(n) - CA\hat{X}(n-1)]$ | T.7.11 |

## 7.8 References

LLSE, MMSE and linear regression are covered in Chapter 4 of [Bertsekas and Tsitsiklis, 2008]. The Kalman filter was introduced in [Kalman, 1960]. The text [Brown and Hwang, 1996] is an easy introduction to Kalman filters with many examples.

## 7.9 Problems

**Problem 1** *Assume that $X_n = Y_n + 2Y_n^2 + Z_n$ where the $Y_n$ and $Z_n$ are i.i.d. $U[0,1]$. Let also $X = X_1$ and $Y = Y_1$. (a) Calculate $L[X|Y]$ and $E((X - L[X|Y])^2)$;*

*(b) Calculate $Q[X|Y]$ and $E((X - Q[X|Y])^2)$ where $Q[X|Y]$ is the quadratic least squares estimate of $X$ given $Y$.*

*(c) Design a stochastic gradient algorithm to compute $Q[X|Y]$ and implement it in Matlab.*

**Problem 2** *We want to compare the off-line and on-line methods for computing $L[X|Y]$. Use the setup of the previous problem.*

*(a) Generate $N = 1,000$ samples and compute the linear regression of $X$ given $Y$. Say that this is $X = aY + b$*

*(b) Using the same samples, compute the linear fit recursively using the stochastic gradient algorithm. Say that you obtain $X = cY + d$*

*(c ) Evaluate the quality of the two estimates your obtained by computing $E((X - aY - b)^2)$ and $E((X - cY - d)^2)$.*

**Problem 3** *The random variables $X, Y, Z$ are jointly Gaussian,*

$$(X,Y,Z)^T \sim N((0,0,0)^T, \begin{bmatrix} 2 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix})$$

*a) Find $E[X|Y,Z]$;*

*b) Find the variance of error.*

**Problem 4** *You observe three i.i.d. samples $X_1, X_2, X_3$ from the distribution $f_{X|\theta}(x) = \frac{1}{2}e^{-|x-\theta|}$, where $\theta \in \mathbb{R}$ is the parameter to estimate. Find $MLE[\theta|X_1, X_2, X_3]$.*

**Problem 5**

*(a) Given three independent $N(0,1)$ random variables $X, Y$ and $Z$, find the following minimum mean square estimator:*

$$E[X + 3Y|2Y + 5Z].$$

*(b) For the above, compute the mean squared error of the estimator.*

**Problem 6** *Given two independent $N(0,1)$ random variables $X$ and $Y$, find the following linear least square estimator:*

$$L[X|X^2 + Y].$$

*Hint: The characteristic function of a $N(0,1)$ random variable $X$ is as follows:*

$$E(e^{isX}) = e^{-\frac{1}{2}s^2}.$$

**Problem 7** *Consider a sensor network with $n$ sensors that are making observations $\mathbf{Y}^n = (Y_1, \ldots, Y_n)$ of a signal $X$ where*

$$Y_i = aX + Z_i, i = 1, \ldots, n.$$

*In this expression, $X =_D N(0,1), Z_i =_D N(0, \sigma^2)$, for $i = 1, \ldots, n$ and these random variables are mutually independent.*

*a) Compute the MMSE estimator of $X$ given $\mathbf{Y}^n$.*

*b) Compute the mean squared error $\sigma_n^2$ of the estimator.*

*c) Assume each measurement has a cost $C$ and that we want to minimize*

$$nC + \sigma_n^2.$$

*Find the best value of $n$.*

*d) Assume that we can decide at each step whether to make another measurement or to stop. Our goal is to minimize the expected value of*

$$\nu C + \sigma_\nu^2$$

*where $\nu$ is the random number of measurements. Do you think there is a decision rule that will do better than the deterministic value $n$ derived in c)? Explain.*

**Problem 8** *We want to use a Kalman filter to detect a change in the popularity of a word in twitter messages. To do this, we create a model of the number $Y_n$ of times that particular word appears in twitter messages on day n. The model is as follows:*

$$X(n+1) = X(n)$$
$$Y(n) = X(n) + W(n)$$

*where the $W(n)$ are zero-mean and uncorrelated. This model means that we are observing numbers of occurrences with an unknown mean $X(n)$ that is supposed to be constant. The idea is that if the mean actually changes, we should be able to detect it by noticing that the errors between $\hat{Y}(n)$ and $Y(n)$ are large. Propose an algorithm for detecting that change and implement it in Matlab.*

**Problem 9** *The random variable X is exponentially distributed with mean 1. Given X, the random variable Y is exponentially distributed with rate X.*

*a) Calculate $E[Y|X]$.*

*b) Calculate $E[X|Y]$.*

**Problem 10** *The random variables $X, Y, Z$ are i.i.d. $\mathcal{N}(0,1)$.*

*a) Find $L[X^2 + Y^2|X + Y]$;*

*b) Find $E[X + 2Y|X + 3Y + 4Z]$;*

*c) Find $E[(X + Y)^2|X - Y]$.*

**Problem 11** *Let $(V_n,\ n \geq 0)$ be i.i.d. $N(0, \sigma^2)$ and independent of $X_0 = N(0, u^2)$. Define*

$$X_{n+1} = aX_n + V_n,\ n \geq 0.$$

*(a) What is the distribution of $X_n$ for $n \geq 1$?*

*(b) Find $E[X_{n+m}|X_n]$ for $0 \leq n < n + m$.*

*(c) Find $u$ so that the distribution of $X_n$ is the same for all $n \geq 0$.*

**Problem 12** *Let $\theta =_D U[0,1]$, and given $\theta$, the random variable X is uniformly distributed in $[0, \theta]$. Find $E[\theta|X]$.*

**Problem 13** *Let $(X, Y)^T \sim N([0;0], [3,1;1,1])$. Find $E[X^2|Y]$.*

**Problem 14** *Let $(X, Y, Z)^T \sim N([0;0;0], [5,3,1;3,9,3;1,3,1])$. Find $E[X|Y, Z]$.*

**Problem 15** *Consider arbitrary random variables $X$ and $Y$. Prove the following property.*

$$var(Y) = E(var[Y|X]) + var(E[Y|X]).$$

**Problem 16** *Let the joint p.d.f. of two random variables $X$ and $Y$ be*

$$f_{X,Y}(x,y) = \frac{1}{4}(2x+y)1\{0 \le x \le 1\}1\{0 \le y \le 2\}.$$

*First show that this is a valid joint p.d.f. Suppose you observe $Y$ drawn from this joint density. Find $MMSE[X|Y]$.*

**Problem 17** *Given four independent $N(0,1)$ random variables $X$, $Y$, $Z$, and $V$, find the following minimum mean square estimate:*

$$E[X + 2Y + 3Z|Y + 5Z + 4V]$$

*Find the mean squared error of the estimate.*

**Problem 18** *Assume that $X, Y$ are two random variables that are such that $E[X|Y] = L[X|Y]$. Then, it must be that (choose the correct answers, if any)*

☐ *$X$ and $Y$ are jointly Gaussian;*

☐ *$X$ can be written as $X = aY + Z$ where $Z$ is a random variable that is independent of $Y$;*

☐ *$E((X - L[X|Y])Y^k) = 0$ for all $k \ge 0$;*

☐ *$E((X - L[X|Y])\sin(3Y + 5)) = 0$.*

**Problem 19** *In a linear system with independent Gaussian noise, with state $X_n$ and observation $Y_n$, the Kalman filter computes (choose the correct answers, if any)*

☐ *$MLE[Y_n|X^n]$;*

☐ *$MLE[X_n|Y^n]$;*

☐ *$MAP[Y_n|X^n]$;*

☐ *$MAP[X_n|Y^n]$;*

☐ *$E[X_n|Y^n]$;*

☐ *$E[Y_n|X^n]$;*

☐ *$E[X_n|Y_n]$;*

☐ *$E[Y_n|X_n]$.*

**Problem 20** *Let* $(X, \mathbf{Y})$ *where* $\mathbf{Y}' = [Y_1, Y_2, Y_3, Y_4]$ *be* $N(\mu, \Sigma)$ *with* $\mu' = [2, 1, 3, 4, 5]$ *and*

$$\Sigma = \begin{bmatrix} 3 & 4 & 6 & 12 & 8 \\ 4 & 6 & 9 & 18 & 12 \\ 6 & 9 & 14 & 28 & 18 \\ 12 & 18 & 28 & 56 & 36 \\ 8 & 12 & 18 & 36 & 24 \end{bmatrix}.$$

*Find* $E[X|\mathbf{Y}]$.

**Problem 21** *Let* $\mathbf{X} = A\mathbf{V}$ *and* $\mathbf{Y} = C\mathbf{V}$ *where* $\mathbf{V} = N(\mathbf{0}, \mathbf{I})$.
*Find* $E[\mathbf{X}|\mathbf{Y}]$.

**Problem 22** *Given* $\theta \in \{0, 1\}$, $\mathbf{X} = N(\mathbf{0}, \Sigma_\theta)$ *where*

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \Sigma_1 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

*where* $\rho > 0$ *is given.*
*Find* $MLE[\theta|\mathbf{X}]$.

**Problem 23** *Given two independent* $N(0, 1)$ *random variables* $X$ *and* $Y$, *find the following linear least square estimator:*

$$L[X|X^3 + Y].$$

*Hint: The characteristic function of a* $N(0, 1)$ *random variable* $X$ *is as follows:*

$$E(e^{isX}) = e^{-\frac{1}{2}s^2}.$$

**Problem 24** *Let* $X, Y, Z$ *be i.i.d.* $\mathcal{N}(0, 1)$. *Find*

$$E[X|X + Y, X + Z, Y - Z].$$

Hint: *Argue that the observation* $Y - Z$ *is redundant.*

**Problem 25** *Let* $X, Y_1, Y_S, Y_3$ *be zero-mean with covariance matrix*

$$\Sigma = \begin{bmatrix} 10 & 6 & 5 & 16 \\ 6 & 9 & 6 & 21 \\ 5 & 6 & 6 & 18 \\ 16 & 21 & 18 & 57 \end{bmatrix}$$

*Find* $L[X|Y_1, Y_2, Y_3]$. Hint: *You will observe that* $\Sigma_{\mathbf{Y}}$ *is singular. This means that at least one of the observations* $Y_1, Y_2$ *or* $Y_3$ *is redundant, i.e., is a linear combination of the others. This implies that* $L[X|Y_1, Y_2, Y_3] = L[X|Y_1, Y_2]$.