Conditional expectation, conditional distribution and application to the Kalman filter

Conditional expectation is a fundamental concept in probability theory and statistics. It provides a rigorous way to incorporate information into computations and underpins many filtering and stochastic tools that have found widespread applications since the middle of the past century.

In this problem, we introduce conditional expectation by first considering the intuitive case of square-integrable random variables (L^2), then extending the characterization to integrable random variables (L^1) and non-negative random variables. After establishing the classical properties of conditional expectation, we turn to the concept of conditional distribution and then focus on the particularly important case of Gaussian random vectors.

We conclude with a standard yet powerful application: the Kalman filter, a tool from statistics and signal processing, that has been used in the Apollo guidance computer and is now embedded in virtually every modern smartphone for position tracking.

Conditional Expectation — The L^2 Case

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. We denote by $L^p(\Omega, \mathcal{A}, \mathbb{P})$ the space of $\mathcal{A}$-measurable real-valued random variables X such that $\mathbb{E}[|X|^p] < +\infty$.[1] All equalities and inequalities between random variables are understood almost surely (i.e., up to a set of probability zero).

Recall that $L^2(\Omega, \mathcal{A}, \mathbb{P})$, equipped with the scalar product $(X, Y) \mapsto \mathbb{E}[XY]$, is a Hilbert space.

1. Show that for any sub-σ-algebra $\mathcal{B} \subset \mathcal{A}$, the space $L^2(\Omega, \mathcal{B}, \mathbb{P})$ is a closed subspace of $L^2(\Omega, \mathcal{A}, \mathbb{P})$.
2. Let $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$. Show that the function
$$\begin{aligned} h_X : L^2(\Omega, \mathcal{B}, \mathbb{P}) &\to \mathbb{R}_+ \\ Y &\mapsto \mathbb{E}[(X-Y)^2] \end{aligned}$$
admits a unique minimizer.

   Hint: Consider a minimizing sequence and use the parallelogram identity to prove that it is a Cauchy sequence.
3. Show that the unique minimizer Y of h_X is characterized by
$$\mathbb{E}[XZ] = \mathbb{E}[YZ] \quad \text{for all } Z \in L^2(\Omega, \mathcal{B}, \mathbb{P}),$$
or equivalently,
$$\mathbb{E}[X\mathbf{1}_B] = \mathbb{E}[Y\mathbf{1}_B] \quad \text{for all } B \in \mathcal{B}.$$
From now on, we denote this minimizer by $\mathbb{E}[X \mid \mathcal{B}]$ and call it the conditional expectation of X with respect to $\mathcal{B}$.
4. Show that if $X \geq 0$, then $\mathbb{E}[X \mid \mathcal{B}] \geq 0$.
5. Show that the map $X \mapsto \mathbb{E}[X \mid \mathcal{B}]$ is linear on $L^2(\Omega, \mathcal{A}, \mathbb{P})$.
6. Let $\mathcal{B}$ and $\mathcal{C}$ be two sub-σ-algebras of $\mathcal{A}$ such that $\mathcal{C} \subset \mathcal{B}$. Prove the tower property:
$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{B}] \mid \mathcal{C}] = \mathbb{E}[X \mid \mathcal{C}].$$
7. Show that $|\mathbb{E}[X \mid \mathcal{B}]| \leq \mathbb{E}[|X| \mid \mathcal{B}]$.

   Hint: Write $X = X^+ - X^-$.

[1] We assume $p \geq 1$ throughout this problem.

Conditional Expectation — Beyond the L^2 Case

Let $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$, and let $\mathcal{B} \subset \mathcal{A}$ be a sub-σ-algebra. Consider the sequence of random variables $(Y_n)_{n\in\mathbb{N}} = (\mathbb{E}[\min(X,n) \mid \mathcal{B}])_{n\in\mathbb{N}}$.

8. Prove that $(Y_n)_n$ is a Cauchy sequence in $L^1(\Omega, \mathcal{B}, \mathbb{P})$, and deduce that $(Y_n)_n$ converges in L^1 to a random variable $Y \in L^1(\Omega, \mathcal{B}, \mathbb{P})$.

   Hint: Show that $|Y_n - Y_p| \leq \mathbb{E}[(X - \min(n,p))^+ \mid \mathcal{B}]$.

9. Prove that Y is uniquely characterized by

$$\mathbb{E}[XZ] = \mathbb{E}[YZ] \quad \text{for all bounded } \mathcal{B}\text{-measurable random variables } Z,$$

   or equivalently,

$$\mathbb{E}[X\mathbf{1}_B] = \mathbb{E}[Y\mathbf{1}_B] \quad \text{for all } B \in \mathcal{B}.$$

From now on, we denote this limit by $\mathbb{E}[X \mid \mathcal{B}]$, thereby extending the notion of conditional expectation from the L^2 setting to the L^1 setting.

10. Prove that the properties established in Questions 4–7 continue to hold for this extension.

Now let $X : \Omega \to \mathbb{R}_+ \cup \{+\infty\}$ be a nonnegative extended real-valued random variable. Define, as above, the sequence $(Y_n)_n = (\mathbb{E}[\min(X,n) \mid \mathcal{B}])_n$.

11. Prove that $(Y_n)_n$ converges almost surely to a random variable $Y : \Omega \to \mathbb{R}_+ \cup \{+\infty\}$.
12. Prove that Y is uniquely characterized by

$$\mathbb{E}[XZ] = \mathbb{E}[YZ] \quad \text{for all nonnegative } \mathcal{B}\text{-measurable random variables } Z,$$

   or equivalently,

$$\mathbb{E}[X\mathbf{1}_B] = \mathbb{E}[Y\mathbf{1}_B] \quad \text{for all } B \in \mathcal{B}.$$

From now on, we write $Y = \mathbb{E}[X \mid \mathcal{B}]$, thus extending the definition of conditional expectation to all nonnegative extended real-valued random variables.

13. Prove that the properties established in Questions 4–7 continue to hold for this extension.

Conditional Expectation Given a Random Variable

Let Y be a $\mathbb{R}^d$-values random variable and X a real-valued random variable, either nonnegative or L^1. The conditional expectation of X given Y, denoted by $\mathbb{E}[X \mid Y]$, is defined as $\mathbb{E}[X \mid \sigma(Y)]$, where $\sigma(Y)$ is the σ-algebra generated by Y.

If $X = (X_1, \ldots, X_d)'$ is a $\mathbb{R}^d$-valued random vector, then the vector $\mathbb{E}[X \mid Y]$ is defined componentwise as $(\mathbb{E}[X_1 \mid Y], \ldots, \mathbb{E}[X_d \mid Y])$.

14. Show that X is independent of Y if and only if for every bounded and nonnegative measurable function h, we have

$$\mathbb{E}[h(X) \mid Y] = \mathbb{E}[h(X)].$$

15. Let X and Y be independent random variables, and assume X has a distribution given by a probability measure ν. Let $h : \mathbb{R}^2 \to \mathbb{R}_+$ be a bounded measurable function. Show that

$$\mathbb{E}[h(X,Y) \mid Y] = \int h(x,Y)\,\nu(dx).$$

The Gaussian Case

Recall that a random vector $X = (X_1, \ldots, X_d)'$ is said to follow a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^d$ and Σ is a symmetric, positive semi-definite $d \times d$ matrix, if its characteristic function is given by

$$\mathbb{E}\left[e^{iu'X}\right] = \exp\left(iu'\mu - \tfrac{1}{2}u'\Sigma u\right), \quad \forall u \in \mathbb{R}^d.$$

Equivalently, X is a Gaussian vector if and only if every linear combination of its components is a univariate normal random variable.

We denote by $\nu(\cdot\,;\mu,\Sigma)$ the Gaussian measure with mean μ and covariance matrix Σ.

Let $(X,Y)=(X_1,\ldots,X_n,Y_1,\ldots,Y_m)'$ be a Gaussian random vector in $\mathbb{R}^{n+m}$ with covariance matrix

$$\Sigma=\begin{pmatrix}\Sigma_X & \Sigma_{XY}\\ \Sigma'_{XY} & \Sigma_Y\end{pmatrix}.$$

16. Show that X and Y are independent if and only if $\Sigma_{XY}=0$.
17. Assume that Σ_Y is invertible. Show that there exists a unique matrix B such that the random vector

$$X-\mathbb{E}[X]-B(Y-\mathbb{E}[Y])$$

is independent of $\sigma(Y)$.
18. Deduce that

$$\mathbb{E}[X\mid Y]=\mathbb{E}[X]+\Sigma_{XY}\Sigma_Y^{-1}(Y-\mathbb{E}[Y]).$$

19. Show that for every bounded measurable function $h:\mathbb{R}^n\to\mathbb{R}_+$, we have

$$\mathbb{E}[h(X)\mid Y]=\int h(x)\,\nu\left(dx\,;\,\mathbb{E}[X]+\Sigma_{XY}\Sigma_Y^{-1}(Y-\mathbb{E}[Y]),\,\Sigma_X-\Sigma_{XY}\Sigma_Y^{-1}\Sigma'_{XY}\right).$$

We say that the conditional distribution of X given Y is a multivariate Gaussian distribution with mean $\mathbb{E}[X]+\Sigma_{XY}\Sigma_Y^{-1}(Y-\mathbb{E}[Y])$ and covariance matrix $\Sigma_X-\Sigma_{XY}\Sigma_Y^{-1}\Sigma'_{XY}$.

The Kalman Filter

We consider a probability space $(\Omega,\mathcal{A},\mathbb{P})$ and a dynamical system represented by a sequence $(X_n)_{n\in\mathbb{N}}$ of $\mathbb{R}^m$-valued random variables, driven by control inputs and stochastic noise:

$$X_{n+1}=F_nX_n+U_{n+1}+W_{n+1},$$

where:

- $X_0\sim\mathcal{N}(\hat{x}_0,P_0)$, with $\hat{x}_0\in\mathbb{R}^m$ and P_0 a symmetric positive semi-definite $m\times m$ matrix;
- F_n is a given deterministic $m\times m$ matrix;
- U_{n+1} is a control input;
- $W_{n+1}\sim\mathcal{N}(0,Q_{n+1})$, with Q_{n+1} a deterministic symmetric positive semi-definite matrix.

The process $(X_n)_n$ is not directly observed. Instead, the observer/controller has access to a sequence of $\mathbb{R}^p$-valued measurements $(Z_n)_n$ defined by:

$$Z_n=H_nX_n+V_n,$$

where:

- H_n is a deterministic $p\times m$ matrix;
- $V_n\sim\mathcal{N}(0,R_n)$, with R_n symmetric positive definite;
- The random variables $(W_n)_n$, $(V_n)_n$, and X_0 are mutually independent.

The filtering problem consists in estimating X_n given the observations $(Z_1,\ldots,Z_n)$ and controls $(U_1,\ldots,U_n)$, i.e., computing the conditional distribution of X_n given the available information.

We define the filtration recursively as follows:

$$\mathcal{G}_0=\sigma(X_0),\quad \tilde{\mathcal{G}}_{n+1}=\sigma(\mathcal{G}_n\cup\sigma(U_{n+1}))=\sigma(X_0,U_1,\ldots,U_{n+1},Z_1,\ldots,Z_n),$$
$$\mathcal{G}_{n+1}=\sigma(\tilde{\mathcal{G}}_{n+1}\cup\sigma(Z_{n+1}))=\sigma(X_0,U_1,\ldots,U_{n+1},Z_1,\ldots,Z_{n+1}).$$

Assume that the conditional distribution of X_n given $\mathcal{G}_n$ is Gaussian $\mathcal{N}(\hat{x}_n,P_n)$.

20. Show that the conditional distribution of

$$\begin{pmatrix} X_{n+1} \\ Z_{n+1} \end{pmatrix} \quad \text{given } \tilde{\mathcal{G}}_{n+1}$$

is Gaussian with mean

$$\begin{pmatrix} \hat{x}_{n+1|n} \\ H_{n+1}\hat{x}_{n+1|n} \end{pmatrix}, \quad \text{where } \hat{x}_{n+1|n} = F_n \hat{x}_n + U_{n+1},$$

and covariance matrix

$$\begin{pmatrix} P_{n+1|n} & P_{n+1|n} H'_{n+1} \\ H_{n+1} P_{n+1|n} & H_{n+1} P_{n+1|n} H'_{n+1} + R_{n+1} \end{pmatrix},$$

where $P_{n+1|n} = F_n P_n F'_n + Q_{n+1}$.

Hint: Consider the linear transformation

$$\begin{pmatrix} I_m & 0 \\ -H_{n+1} & I_p \end{pmatrix} \begin{pmatrix} X_{n+1} \\ Z_{n+1} \end{pmatrix}.$$

21. Deduce that the conditional distribution of X_{n+1} given $\mathcal{G}_{n+1}$ is Gaussian with mean

$$\hat{x}_{n+1} = \hat{x}_{n+1|n} + K_{n+1} \left(Z_{n+1} - H_{n+1} \hat{x}_{n+1|n} \right),$$

and covariance matrix

$$P_{n+1} = P_{n+1|n} - K_{n+1} H_{n+1} P_{n+1|n},$$

where

$$K_{n+1} = P_{n+1|n} H'_{n+1} (H_{n+1} P_{n+1|n} H'_{n+1} + R_{n+1})^{-1}.$$

Remark: The matrix K_{n+1} is called the Kalman gain.

22. Conclude.

23. *(Bonus points)* Code in Python an implementation of the Kalman filter.