CSE 840: Computational Foundations of Artifical Intelligence April 14, 2025

Product Space and Joint Distribution,
Marginal Distribution, Conditional Distribution,
Conditional Expectation

Instructor: Vishnu Boddeti *Scribe: Muhammed Salih Kayhan, Alex Logan and Tyler Matson*

1 Product Space and Joint Distribution

1.1 Product Space

Consider two measurable spaces $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$. We can define the **product space** $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ with:

$$\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) \mid \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\} \tag{1}$$

$$\mathcal{A}_1 \otimes \mathcal{A}_2 = \{A_1 \times A_2 \mid A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\} \tag{2}$$

1.2 Random Variables on Product Spaces

Consider two random variables:

$$X_1 : (\Omega, \mathcal{A}, P) \to (\Omega_1, \mathcal{A}_1) \tag{3}$$

$$X_2 : (\Omega, \mathcal{A}, P) \to (\Omega_2, \mathcal{A}_2) \tag{4}$$

We can define a new random variable:

$$X := (X_1, X_2) : (\Omega, \mathcal{A}, P) \to (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2) \tag{5}$$

$$(X_1, X_2)(\omega) = (X_1(\omega), X_2(\omega)) \tag{6}$$

The distribution $P_{(X_1,X_2)}$ on $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ is called the **joint distribution** of X_1 and X_2.

Example from Machine Learning: (X, Y) where X is the input data and Y is the label.

1.3 Product Measure

Let $(\Omega_1, \mathcal{A}_1, P_1)$ and $(\Omega_2, \mathcal{A}_2, P_2)$ be two probability spaces. We define the **product measure** $P_1 \otimes P_2$ on the product space $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ as:

$$(P_1 \otimes P_2)(A_1 \times A_2) := P_1(A_1) \cdot P_2(A_2) \tag{7}$$

Theorem: Two RVs X_1, X_2 are independent if and only if their joint distribution coincides with the product distribution:

$$P_{(X_1,X_2)} = P_1 \otimes P_2 \tag{8}$$

2 Marginal Distributions

Consider the joint distribution $P_{(X_1,X_2)}$ of two RVs $X := (X_1, X_2)$. The **marginal distribution** of X with respect to X_1 is the original distribution of X_1 on $(\Omega_1, \mathcal{A}_1)$, namely P_{X_1}. Similarly for X_2 as well.

2.1 Example in the Discrete Case

Consider a discrete joint distribution represented as a table:

$y\backslash x$	x_1	x_2	x_3	$\sum$
y_1	p_{11}	p_{12}	p_{13}	$p_{11} + p_{12} + p_{13} = P(Y = y_1)$
y_2	p_{21}	p_{22}	p_{23}	$p_{21} + p_{22} + p_{23} = P(Y = y_2)$
$\sum$	$p_{11} + p_{21}$ $= P(X = x_1)$	$p_{12} + p_{22}$ $= P(X = x_2)$	$p_{13} + p_{23}$ $= P(X = x_3)$	

The row sums represent the marginal distribution with respect to Y, and the column sums represent the marginal distribution with respect to X.

3 Marginal Distributions in case of Densities

Let $X, Y : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $Z := (X, Y)$.

Assume that the joint distribution of Z has a density f on $\mathbb{R}^2$. Then we have the following statements:

Both X and Y have densities on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \tag{9}$$

Here, we take the joint distribution $f(x, y)$ and sum (integrate) over all values of y to obtain the marginal density of X.

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx \tag{10}$$

Similarly, we take the joint distribution $f(x, y)$ and sum (integrate) over all values of x to obtain the marginal density of Y.

X and Y are independent if and only if

$$f(x, y) = f_X(x) \cdot f_Y(y) \quad \text{almost surely} \tag{11}$$

In other words, the joint density equals the product of the marginal densities almost surely if and only if X and Y are independent.

4 Mixed Cases

For example, consider X is a continuous RV with density and Y a discrete RV.

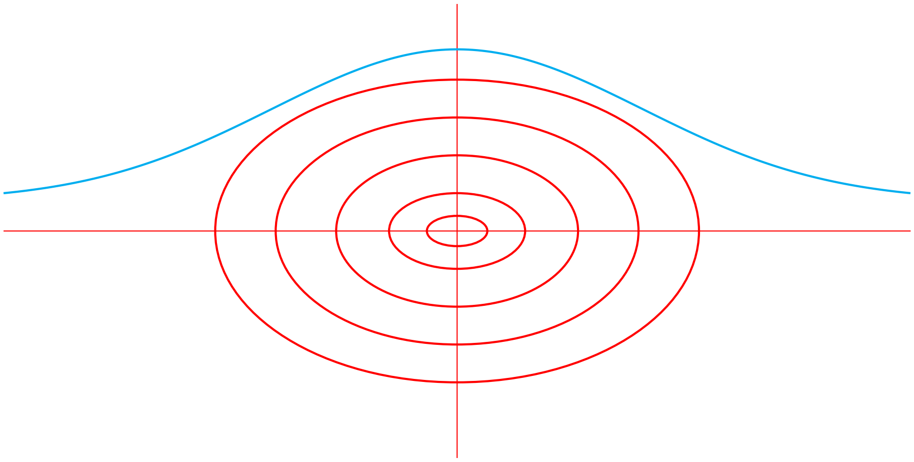
Example: X = image (2d-continuous signal), Y = "cat", "dog", ... (discrete)

4.1 Special Case: Marginals of Multivariate Normal

Consider a 2-dimensional normal random variable $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ with mean $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \in \mathbb{R}^2$ and covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$.

Then the marginal distribution of X with respect to X_1 is again a normal distribution with mean μ_1 and variance σ_1^2.

This can be visualized as "summing up" the joint distribution in the Y-direction to obtain the marginal distribution with respect to X, which results in a normal distribution.



5 Conditional Distributions

5.1 Discrete Case

We start with the discrete setting. Consider two discrete random variables X and Y defined on a probability space $(\Omega, \mathcal{A}, P)$, each taking finitely or countably many values. For $x \in \mathbb{R}$ and $y \in \mathbb{R}$ with $P(Y = y) > 0$, we define the conditional probability mass function of X given $Y = y$ as:

$$P(X = x \mid Y = y) := \frac{P(X = x, Y = y)}{P(Y = y)} \tag{12}$$

This quantity defines a new distribution over X for each fixed value of Y. More generally, this construction induces a family of conditional distributions over X indexed by values of Y.

5.2 Conditional Probability as Measure

We can reinterpret conditional distributions as conditional probability measures. For discrete random variables X and Y, and a measurable set $A \in \mathcal{A}$, the conditional distribution of X given $Y = y$ is the probability measure $P_X(\cdot \mid Y = y)$ such that:

$$P_X(A \mid Y = y) := P(X \in A \mid Y = y) \tag{13}$$

5.3 Conditional Densities

Now suppose X, Y are continuous random variables and $Z := (X, Y)$ has a joint density $f(x, y)$ on $\mathbb{R}^2$. Then the conditional density of X given $Y = y$ is defined as:

$$f_{X|Y}(x \mid y) := \frac{f(x, y)}{f_Y(y)} \quad \text{for } f_Y(y) > 0 \tag{14}$$

Here, $f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx$ is the marginal density of Y. The function $f_{X|Y}(x \mid y)$ is a valid density in x and satisfies:

$$\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)dx = 1 \tag{15}$$

5.4 Gaussian Case: Conditional of Multivariate Normal

Let $X = \begin{pmatrix} X_S \\ X_T \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$ where:

$$\mu = \begin{pmatrix} \mu_S \\ \mu_T \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{SS} & \Sigma_{ST} \\ \Sigma_{TS} & \Sigma_{TT} \end{pmatrix} \tag{16}$$

Then the conditional distribution of X_S given $X_T = x_T$ is:

$$X_S \mid X_T = x_T \sim \mathcal{N}(\mu_S + \Sigma_{ST}\Sigma_{TT}^{-1}(x_T - \mu_T),\ \Sigma_{SS} - \Sigma_{ST}\Sigma_{TT}^{-1}\Sigma_{TS}) \tag{17}$$

This is a core result in multivariate analysis. Even after conditioning, the distribution remains Gaussian, though its mean and covariance change.

5.5 Geometric Interpretation: Marginalization vs Conditioning

To understand the difference between marginalization and conditioning geometrically, imagine the joint distribution as a 3D surface over the (X, Y) plane:

- **Marginalization** integrates over one axis (e.g., summing out Y) and projects the total mass onto the X-axis. This results in the marginal distribution of X.
- **Conditioning** slices the 3D surface at a fixed value (e.g., $Y = y$), revealing the profile along the X-axis at that level. This results in a conditional distribution.

These two operations serve different inferential goals and are foundational to probabilistic modeling.

6 Conditional Expectation

6.1 Discrete Conditional Expectation

Let X and Y be discrete random variables. For a fixed value $X = x_i$ with $P(X = x_i) > 0$, the conditional expectation of Y given $X = x_i$ is:

$$\mathbb{E}[Y \mid X = x_i] = \sum_j y_j P(Y = y_j \mid X = x_i) \tag{18}$$

We can promote this to a random variable by defining:

$$\mathbb{E}[Y \mid X](\omega) := \mathbb{E}[Y \mid X = X(\omega)] \tag{19}$$

This yields a function of X and is itself a random variable measurable with respect to the σ-algebra generated by X.

6.2 Conditional Expectation: General Definition

Let X be an integrable random variable and $\mathcal{G} \subseteq \mathcal{A}$ a sub-σ-algebra. Then $\mathbb{E}[X \mid \mathcal{G}]$ is defined as the unique $\mathcal{G}$-measurable function Z such that for all $G \in \mathcal{G}$:

$$\int_G Z dP = \int_G X dP \tag{20}$$

This definition ensures that $\mathbb{E}[X \mid \mathcal{G}]$ is the best approximation of X given the information in $\mathcal{G}$.

6.3 Extreme Cases and Intuition

- If X is $\mathcal{G}$-measurable, then $\mathbb{E}[X \mid \mathcal{G}] = X$ almost surely.
- If X is independent of $\mathcal{G}$, then $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ almost surely.

These facts match our intuitive understanding: if we already know X, there's no gain from conditioning. If X and $\mathcal{G}$ are independent, then conditioning adds no value.

6.4 Conditional Expectation with Densities

Let X, Z be jointly continuous random variables with density $f(x, z)$. Let $Y = g(Z)$ for some bounded measurable function g. Then the conditional expectation:

$$\mathbb{E}[Y \mid X] = \int g(z) f_{Z|X}(z \mid x) dz = \int g(z) \frac{f(x, z)}{f_X(x)} dz \tag{21}$$

This integral defines a function of X, and captures how the expected value of Y varies as we observe different values of X.

6.5 Measurability Caveat

All conditional expectations are defined only almost surely, they may differ on sets of measure zero. This is very important in the continuous case where one must take care regarding measurability and regular conditional probabilities.

Conditional expectation is fundamental in Bayesian inference, stochastic processes, and learning theory, where it provides the basis for prediction, filtering, and updating beliefs.

7 General Conditional Expectation: Abstract Definition

We can define conditional expectation even when we do not assume the existence of densities or discrete structure. Let X be an integrable random variable defined on a probability space $(\Omega, \mathcal{A}, P)$, and let $\mathcal{G} \subseteq \mathcal{A}$ be a sub-σ-algebra. Then, the conditional expectation of X given $\mathcal{G}$ is defined as the unique $\mathcal{G}$-measurable random variable Z that satisfies:

1. Z is $\mathcal{G}$-measurable
2. For all $G \in \mathcal{G}$, we have:

$$\int_G Z \, dP = \int_G X \, dP \tag{22}$$

This formulation for the conditional expectation can be used across discrete, continuous, and mixed settings, and extends to arbitrary measurable spaces. This idea is very important for martingales, filtrations, and stochastic processes.

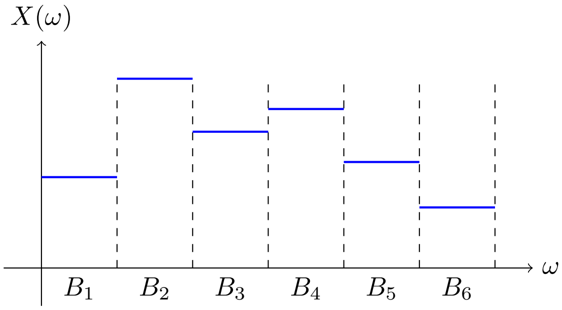
8 Special Cases and Visualization

Let us revisit two extreme special cases to ground the general theory:

- If X is $\mathcal{G}$-measurable, then $\mathbb{E}[X \mid \mathcal{G}] = X$ almost surely, nothing is learned by conditioning.
- If X is independent of $\mathcal{G}$, then $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ almost surely, knowledge of $\mathcal{G}$ does not help.

To get an intuition for conditional expectation, imagine splitting the entire sample space Ω into a bunch of non-overlapping regions, say B_1, B_2,..., where each region corresponds to a specific value $Y = y_i$. Inside each region B_i, the conditional expectation $\mathbb{E}[X \mid Y = y_i]$ tells us the average value of X when $Y = y_i$. So, we're essentially computing the average height of X over each piece of the partition.

If we imagine of X as a function over Ω, then the conditional expectation just replaces X with its average value on each chunk of the partition. The result is a new function thats flat on each region, kind of like turning X into a staircase function, where each step represents the expected value given a different value of Y.



This figure shows how conditional expectation can be viewed as "flattening" the function X within each measurable region defined by the information in $\mathcal{G}$.

9 Practical Example: Predicting Exam Scores

Suppose we model student exam performance. Let X denote a student's final exam score, and let Y be a categorical variable indicating the number of hours studied(: $Y \in \{\text{low}, \text{medium}, \text{high}\}$).

We can model the joint distribution (X, Y) using a mixture model or Gaussian assumption. Based on historical data, we may estimate:

$$\mathbb{E}[X \mid Y = \text{low}] = 60 \tag{23}$$

$$\mathbb{E}[X \mid Y = \text{medium}] = 75 \tag{24}$$

$$\mathbb{E}[X \mid Y = \text{high}] = 88 \tag{25}$$

Then, $\mathbb{E}[X \mid Y]$ is a random variable taking values 60, 75, or 88 depending on the category. We might regress X on Y using linear or nonlinear methods, and $\mathbb{E}[X \mid Y]$ becomes the predicted score for a given input.

This conditional expectation could be used in:

- **Supervised learning**: estimating target variable given features.
- **Imputation**: predicting missing values.
- **Filtering**: computing beliefs given partial observations.