CSE 840: Computational Foundations of Artificial Intelligence November 29, 2023

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

Instructor: Vishnu Boddeti *Scribe: Richard Frost*

1 Product Space, Joint Distribution

Definition 1 *Consider two measureable spaces* $(\Omega_1, \mathscr{A}_1), (\Omega_2, \mathscr{A}_2)$. *The* *Product Space* *of these spaces is* $(\Omega_1 \times \Omega_2, \mathscr{A}_1 \otimes \mathscr{A}_2)$. *Where:*

$$\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) | \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$$

$$\mathscr{A}_1 \otimes \mathscr{A}_2 = \{A_1 \times A_2 | A_1 \in \mathscr{A}_1, A_2 \in \mathscr{A}_2\}$$

Consider two random variables:

$$X_1 : (\Omega, \mathscr{A}, \mathbb{P}) \to (\Omega_1, \mathscr{A}_1)$$

$$X_2 : (\Omega, \mathscr{A}, \mathbb{P}) \to (\Omega_2, \mathscr{A}_2)$$

Then,

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

$$(X_1, X_2)(\omega) = (X_1(\omega), X_2(\omega))$$

Definition 2 *For a product space* $(\Omega_1 \times \Omega_2, \mathscr{A}_1 \otimes \mathscr{A}_2)$ *with random variables* X_1 *and* X_2*, the distribution* $P_{(X_1, X_2)}$ *over that space 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.

Definition 3 *Let* $(\Omega_1, \mathscr{A}_1, P_1)$ *and* $(\Omega_2, \mathscr{A}_2, P_2)$ *be two probability spaces. The* *Product Measure* $P_1 \otimes P_2$ *on the product space* $(\Omega_1 \times \Omega_2, \mathscr{A}_1 \otimes \mathscr{A}_2)$ *is*

$$(P_1 \otimes P_2)(\mathscr{A}_1 \times A_2) := P_1(A_1) \cdot P_2(A_2)$$

Theorem 4 *Two random variables* X_1 *and* X_2 *are independent if and only if their joint distribution coincides with the product distributions:*

$$P_(X_1, X_2) = P_1 \otimes P_2$$

2 Marginal Distributions

Definition 5 *Consider the joint distribution* $P_(X_1, X_2)$ *for two random variables* $X := (X_1, X_2)$. *The* *Marginal Distribution of* X *with respect to* X_1 *is the original distribution of* X_1 *on* $(\Omega_1, \mathscr{A}_1)$*, namely* P_{X_1}. *Similarly for* X_2 *as well.*

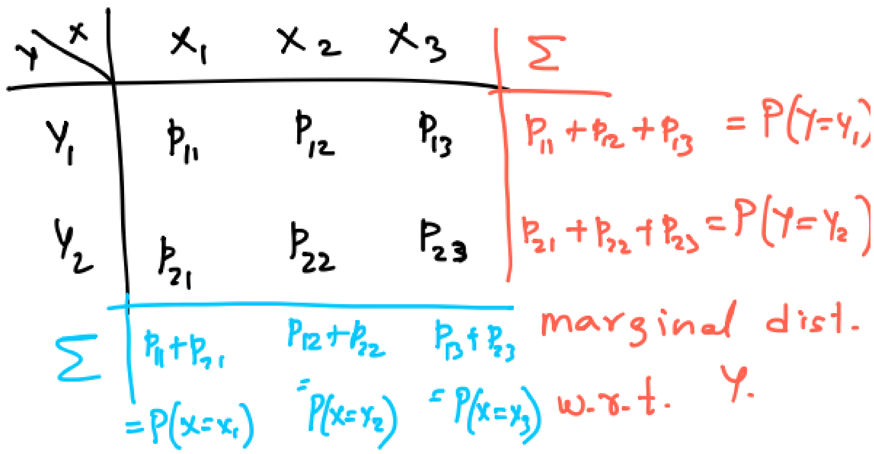

Figure 1: Example of discrete marginal distribution

2.1 Marginal Distributions in the Case of Densities

$X, Y : (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$. $z := (X, Y)$. Assume that the joint distribution of z has a density of f on $\mathbb{R}^2$. Then we have the following statements:

1. Both X and Y have densities on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ given by,

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

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

2. X and Y are independent if and only if

$$f(x, y) - f_X(x) \cdot f_Y(y) a.s.$$

2.2 Mixed Cases

There are also join distributions where the random variables are of different types. For example, consider X is a continuous random variable with density (e.g. an image (2d-continuous signal)) and Y is a discrete random variable (e.g. a classification "cat" "dog"...)

2.3 Special Case: Marginals of multivariate Normal

2.3.1 Two Dimensions

Consider a 2-dimensional normal random variable $X = \binom{X_1}{X_2}$ with mean $\mu = \binom{\mu_1}{\mu_2} \in \mathbb{R}^2$ and covariance $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{pmatrix}$ Then the marginal distribution of X with respect to X_1 is also a normal distribution with mean μ_1 and variance σ_1^2

Figure 2: Illustration of marginal distribution of X for multivariate Normal

2.3.2 n Dimensions

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n$$

Group the variables into,

$$S = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} \in \mathbb{R}^k, T = \begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^{n-k}$$

We want to look at the marginal of X with respect to S. Let the mean vector be,

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$$

Then,

$$\mu_S := \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}, \mu_T := \begin{pmatrix} \mu_{k+1} \\ \vdots \\ \mu_n \end{pmatrix}$$

We divide the covariance matrix Σ as follows:

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{S,S} & \Sigma_{S,T} \\ \Sigma_{T,S} & \Sigma_{T,T} \end{array} \right) \in \mathbb{R}^{n \times n}$$

Now the marginal of X with respect to S is a normal distribution on $\mathbb{R}^k$ with mean μ_S and covariance σ_{SS}

3 Conditional Distribution

3.1 Discrete Case

Known conditional probabilities: $P(A|B)$ defined for events $A, B \in \mathscr{A}$, and $P(B) > 0$. Let $X, Y : (\Omega, \mathscr{A}, P) \to \mathbb{R}$ be discrete random variables, $y \in \mathbb{R}$ such that $P(Y = y) > 0$. Then we can define the conditional probability measure:

$$P_{X|Y=y} : A \mapsto P(X \in A|Y = y)$$

This is a probability measure.

3.2 General Random Variables

It is very complicated and we will not cover it in this course.

3.3 Conditional distributions in the Case of Densities

Assume $Z := (X, Y)$ has a joint density $f : \mathbb{R}^2 \to \mathbb{R}$, and marginal densities $f_X, f_Y : \mathbb{R} \to \mathbb{R}$. Then the function,

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

is also a density on $\mathbb{R}$, called the conditional density of X given $Y = y$.

Example: Normal Distribution Let,

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{S,S} & \Sigma_{S,T} \\ \Sigma_{T,S} & \Sigma_{T,T} \end{pmatrix}$$

If $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$, then the conditional distributions of $X_S = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}$ conditioned on $X_T = \begin{pmatrix} x_{k+1} \\ \vdots \\ x_n \end{pmatrix}$ is given by:

$$P_{X_S|X_T} \sim \mathcal{N}(\mu_T + \Sigma_{S.T}\Sigma_{T,T}^{-1}(X_S - \mu_T), \Sigma_{T,T} - \Sigma_{S,T}^T \Sigma_{S,S}^{-1} \Sigma_{S,T})$$

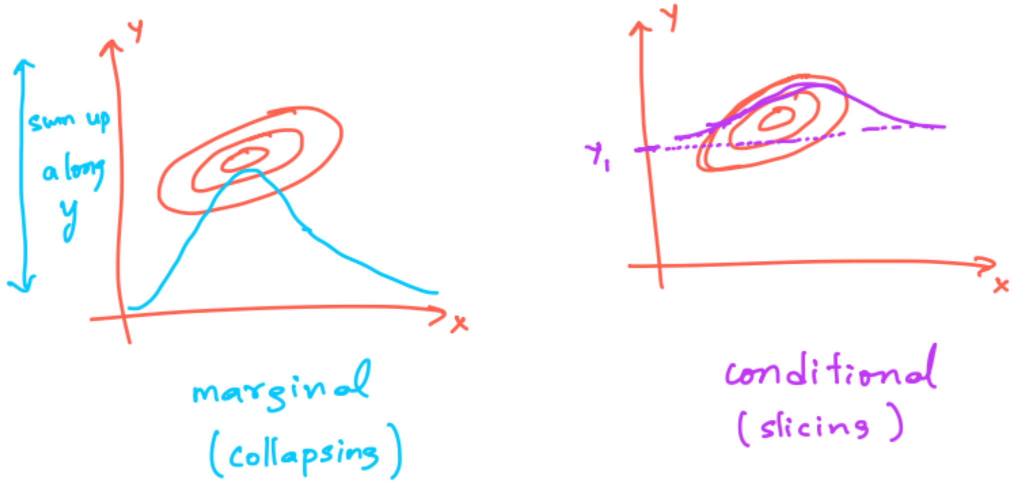

Figure 3: Visualization of conditional distributions on multivariate normal

4 Conditional Expectation

Definition 6 *Conditional Expectation in the Discrete Case Let $X, Y : (\Omega, \mathscr{A}, P) \to \mathbb{R}$. Assume X takes finitely (countably) many values $x_1, x_2, \ldots, x_n \in \mathbb{R}$ and Y takes finitely (countably) many values $y_1, y_2, \ldots, y_m in \mathbb{R}$. Always with positive probability. Then,*

$$\mathbb{E}\left(Y|X = x_i\right) := \sum_{j=1}^{m} y_j P(Y = y_j | X = x_i)$$

Example: two dice, X = value of die 1, Y = value of die 2, independent dice.

$$\mathbb{E}\left(\text{sum}|X = 1\right) = \sum_{i=1}^{12} i \cdot P(\text{sum} = i | X = 1)$$

$$= \sum_{k=1}^{6} (1 + k) \times P(Y = k | x = 1)$$

$$= \sum_{k=1}^{6} (1 + k) \times P(Y = k) = \sum_{k=1}^{6} (1 + k) \cdot \frac{1}{6} = 4.5$$

So far we defined $\mathbb{E}\left(Y|X = x_i\right)$, but often we want to consider the "function" $\mathbb{E}\left(Y|X\right)(\omega)$. This is a random variable: $\mathbb{E}\left(Y|X\right) : (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$. This leads to the following:

Definition 7 *Discrete Case X, Y as before. Then the conditional expectation is defined as follows:*

$$\mathbb{E}\left(Y|X\right) := f(x)$$

$$f(x) = \begin{cases} \mathbb{E}\left(Y|X = x\right) & \text{if} P(X = x) > 0 \\ \text{arbitrary, say 0} & \text{otherwise} \end{cases}$$

Caution: *$\mathbb{E}\left(Y|X\right)$ is only defined almost surely*

Now we want to consider the more general case. Sketch: X is a continuous random variable and Y is a discrete random variable $\sim y_1, y_2, \ldots, y_5$. We want to look at $\mathbb{E}\left(X|Y\right)$. Figure 4 gives a visualization.

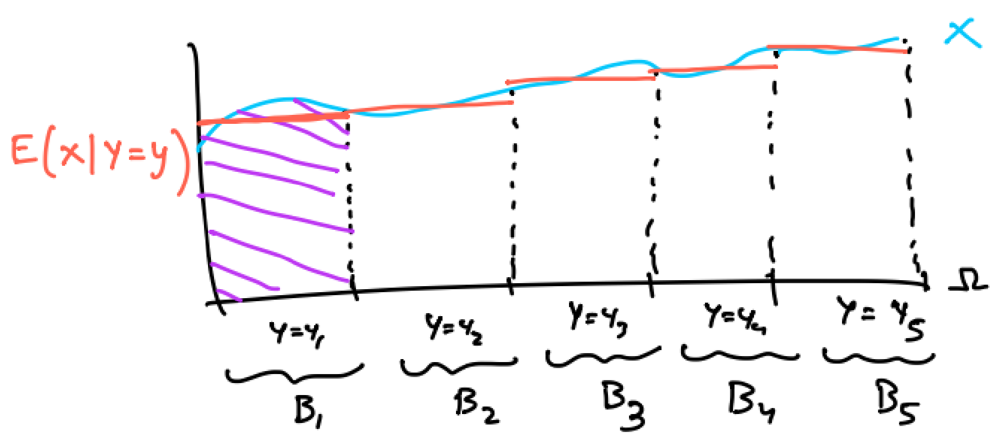

Figure 4:

We want to "define $\mathbb{E}(X|Y) := \sum_{i=1}^{5} \mathbb{E}(X|Y = y_i) \cdot \mathbb{1}_{B_i}(\omega)$, but we need to make sure that it is measurable with respect to $\sigma(Y)$ (the "bins")

Definition 8 *Consider random variables $x : (\Omega, \mathscr{A}_0, P) \to \mathbb{R}$ and $X \in L_1(\Omega, \mathscr{A}_0, P)$. Let $\mathscr{A}$ be a sub-σ-algebra of $\mathscr{A}_0$. (intuition: $\mathscr{A}$ will be the σ-algebra generated by the variable Y we want to condition on). The condition expectation of X given $\mathscr{A}$, $\mathbb{E}(X|\mathscr{A})$ is any random variable Z that satisfies:*

1. *Z is measurable with respect to $\mathscr{A}$*
2. *For all $A \in \mathscr{A}$ we have:*
$$\int_A X\, dP = \int_A Z\, dP$$

- *The existence of $\mathbb{E}(X|\mathscr{A})$ is not clear a priori; it needs to be proven.*
- $\mathbb{E}(X|Y) := \mathbb{E}(X|\sigma(Y))$

Examples (Extreme Cases):

- $X = Y$, then $\mathbb{E}(X|Y) = \mathbb{E}(X)\,(a.s)$
- $X \perp\!\!\!\perp Y$, then $\mathbb{E}(X|Y) = \mathbb{E}(X)\,(a.s)$

4.1 The Case of Join Densities

Let $X, Z : (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ have a joint density $f(x, z)$. Let $g : \mathbb{R} \to \mathbb{R}$ be a bounded function, and $Y := g(Z)$. Assume we want to compute $\mathbb{E}(Y|X) = \mathbb{E}(g(Z)|X)$.

Recall X has density $f_X(x) = \int f(x, z)\, dz$. The conditional density of Z given $X = x$ is

$$f_{X=x}(z) = \frac{f(x, z)}{f_X(x)} (\text{if} f_X(x) \neq 0)$$

Now consider,

$$h(x) := \int g(z) f_{X=x}(z)\, dz$$

and define $\mathbb{E}(Y|X) = h(x)$