Conditional Probability Theory

Michael Betancourt

March 25, 2022

Conditional probability theory provides a rigorous way to decompose probability distributions over an ambient space X into a collection of probability distributions over subspaces of X. This can helpful for calculations – reducing a complicated calculation over all of X into a sequence of simpler calculations over the smaller subspaces – as well as the construction of useful probability distributions – building up high-dimensional probability distributions by composing together more manageable, lower-dimensional conditional probability distributions.

Throughout we will consider an ambient space X equipped with a σ-algebra $\mathcal{X}$ and a probability distribution of interest π,

$$\begin{aligned} \pi : \mathcal{X} &\to [0, 1] \subset \mathbb{R} \\ A &\mapsto \mathbb{P}_{\pi}[A] \end{aligned} \quad .$$

We begin with partitions that mathematically decompose X into subspaces and then construct decompositions of π over the subspaces defined by increasingly complicated partitions.

Because general probability distributions are difficult to visualize we will occasionally appeal the finite ambient space $X = \{x_1, \ldots, x_{16}\}$ to demonstrate some of the concepts we introduce. Because any probability distribution over a finite space can be completely specified by the probability allocated to the individual elements, $\mathbb{P}_{\pi}[x_i]$, we can visualize the entire distribution with those atomic probabilities (Figure 1).

1 Partitions

In order to rigorously define how to decompose a probability distribution we need to first consider how to rigorously decompose the ambient space X into subsets. Recall that the space of all subsets is referred to as the *power set* and denoted by 2^X.

Two sets $A \in 2^X$ and $B \in 2^X$ that do not overlap, $A \cap B = \emptyset$, are referred to as *disjoint*. A *partition* of X (Figure 2b) is a collection of sets

$$\mathcal{P} \subset 2^X$$

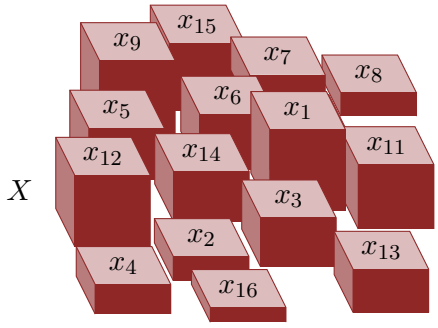


Figure 1: Any probability distribution π over the finite ambient space $X = \{x_1, \ldots, x_{16}\}$ can specified by the atomic probabilities $\mathbb{P}_\pi[x_i]$. The probability of any set is given by a finite sum of these atomic probabilities $\mathbb{P}_\pi[A] = \sum_{x_i \in A} \mathbb{P}_\pi[x_i]$.

that are mutually disjoint,

$$\begin{aligned} B \in \mathcal{P}, B' &\in \mathcal{P} \\ B \cap B' &= \emptyset, \end{aligned}$$

and whose union covers the full ambient space

$$\bigcup_{B \in \mathcal{P}} B = X.$$

A collection of sets that cover X but intersect with each other do not form a valid partition (Figure 2c), nor does a collection of disjoint sets that don't cover all of X (Figure 2d). A *measurable partition* consists of disjoint sets in the assumed σ-algebra, $\mathcal{P} \subset \mathcal{X}$.

I will refer to the individual sets that form a partition as *cells*. A partition can contain a finite number of cells, a countably infinite number of cells or even an uncountably infinite number of cells. Initially, however, we will consider partitions with at most a countably infinite number of cells. I will refer to partitions with a finite, countably infinite, and uncountable infinite number of cells as finite, countable, and uncountable partitions, respectively.

A finite partition can always be defined as an explicit list of sets, but this isn't practical for countable or uncountable partitions which would require infinitely long lists. In all of these cases, however, we can specify a partition *implicitly* through an appropriate function. To motivate the kind of function we need to implicitly specify a partition let's first consider a finite partition defined as an explicit list,

$$\mathcal{P} = \{B_1, \ldots, B_n, \ldots, B_N\}.$$

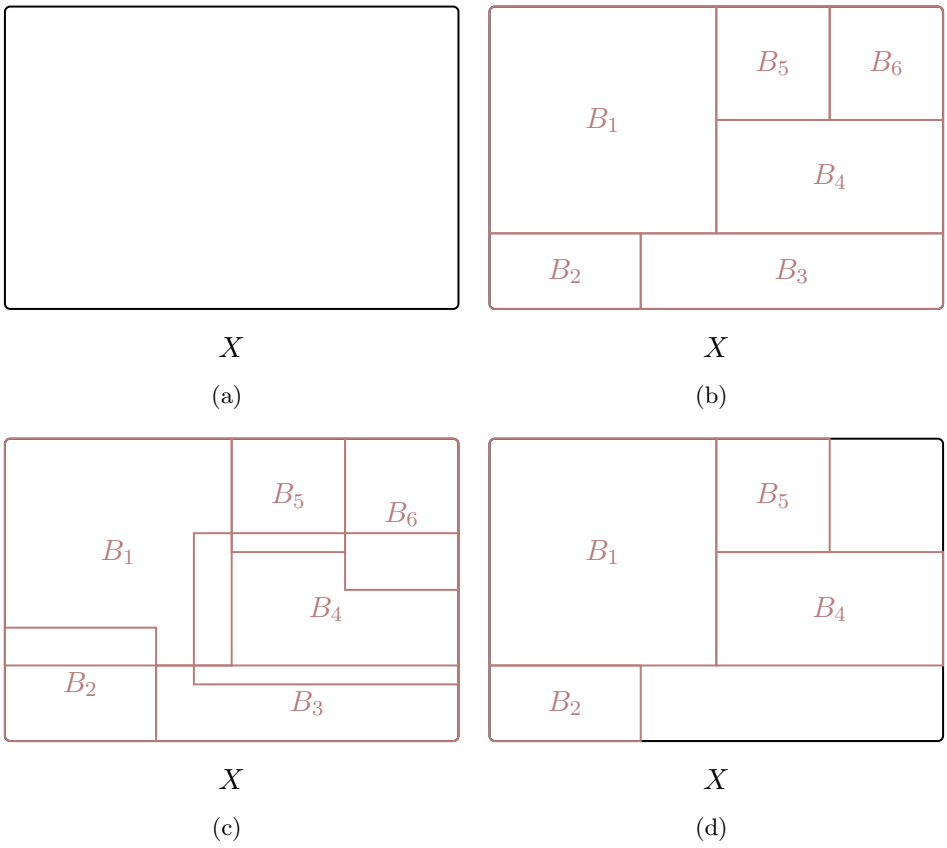


Figure 2: A partition is a decomposition of (a) an ambient space X into (b) a collection of disjoint sets. (c) Overlapping sets that cover the ambient do not form a proper partition, nor do (d) disjoint sets that do not fully cover the ambient space. A measurable partition consists of disjoint sets from the assumed σ-algebra $\mathcal{X}$.

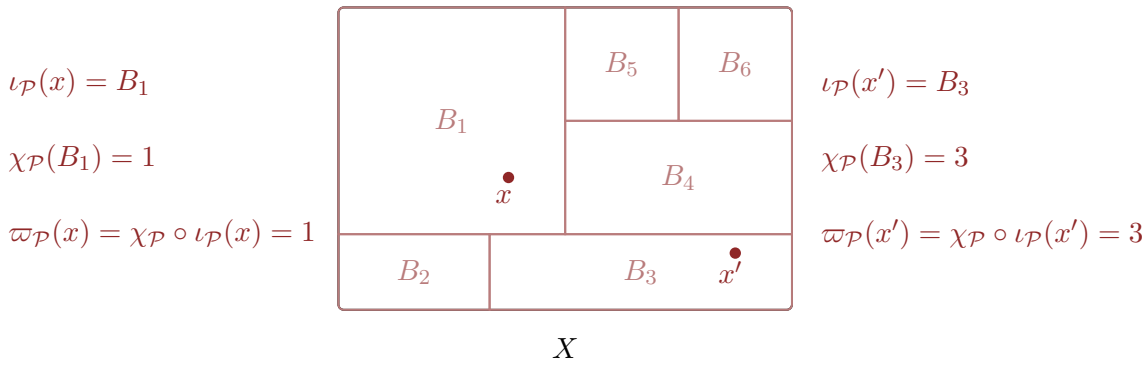


Figure 3: Any finite partition of the ambient space, for example $\mathcal{P} = \{B_1, B_2, B_3, B_4, B_5, B_6\}$, implicitly defines three functions. The function $\iota_{\mathcal{P}}$ maps each point in the ambient space to the partition cell that contains it while the function $\chi_{\mathcal{P}}$ maps each partition cell to its integer index. The composition $\varpi_{\mathcal{P}} = \chi_{\mathcal{P}} \circ \iota_{\mathcal{P}}$ maps each point directly to the corresponding index.

In this case we've numerically labeled or *indexed* the cells with the integers $\{1, \ldots, N\}$. More formally we can define a one-to-one *index function* that maps each cell to its corresponding integer index,

$$\begin{aligned} \chi_{\mathcal{P}} : \mathcal{P} &\to \{1, \ldots, N\} \\ B_n &\mapsto n \end{aligned} \quad .$$

At the same time we can also define an *inclusion* function that maps each point in the ambient space $x \in X$ into the partition cell that contains it,

$$\begin{aligned} \iota_{\mathcal{P}} : X &\to \mathcal{P} \\ x &\mapsto \{B_n \in \mathcal{P} \mid x \in B_n\}. \end{aligned}$$

Composing these two functions together then defines a map from points in the ambient space to partition cell indices (Figure 3)

$$\begin{aligned} \varpi_{\mathcal{P}} = \chi_{\mathcal{P}} \circ \iota_{\mathcal{P}} : X &\to \{1, \ldots, N\} \\ x &\mapsto \{n \in \{1, \ldots, N\} \mid x \in B_n \in \mathcal{P}\}. \end{aligned}$$

Because the partition cells are disjoint and cover the entire ambient space each point $x \in X$ falls into one and only one partition cell and hence is associated with one and only one index. Consequently $\varpi_{\mathcal{P}}$ is a *surjective* function. More importantly the preimage of this function for a given index, the set of all input points with the same function output, completely reconstructs the corresponding partition cell:

$$\varpi_{\mathcal{P}}^{-1}(n) = \{x \in X \mid \varpi_{\mathcal{P}}(x) = n\} = B_n.$$

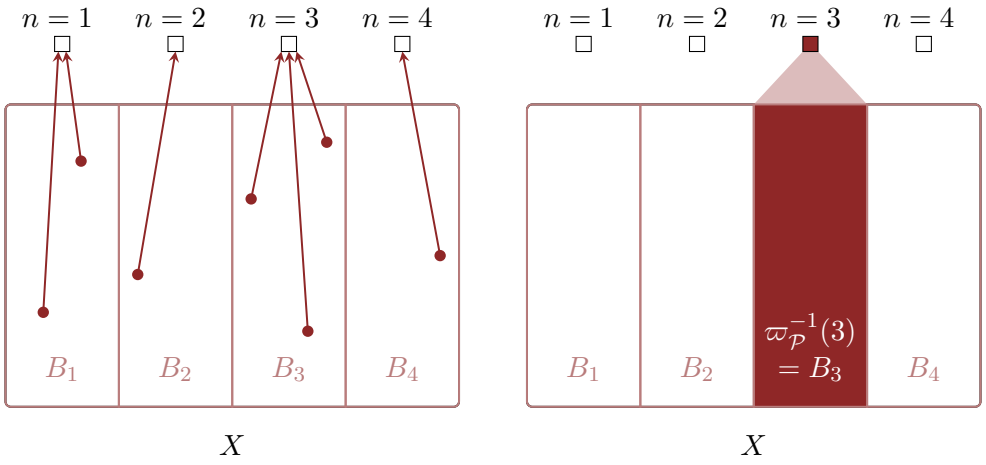


Figure 4: (a) The function $\varpi_{\mathcal{P}}$ maps points in the ambient space $x \in X$ to the indices of the partition cells that contain them. (b) The fibers $\varpi_{\mathcal{P}}^{-1}(n)$ map each index to all of the points contained in the corresponding partition cell. The set of all fibers, here $\{\varpi_{\mathcal{P}}^{-1}(1), \varpi_{\mathcal{P}}^{-1}(2), \varpi_{\mathcal{P}}^{-1}(3), \varpi_{\mathcal{P}}^{-1}(4)\}$, completely reconstructs the partition that defines $\varpi_{\mathcal{P}}$, $\mathcal{P} = \{B_1, B_2, B_3, B_4\}$.

The preimage of a surjective function is also referred to as a *level set*, i.e. the set of all input points with the same function output or "level", or a *fiber* of that function. Consequently we can associate the partition cells in $\mathcal{P}$ with the fibers of the function $\varpi_{\mathcal{P}}$ (Figure 4),

$$\mathcal{P} = \{B_1 = \varpi_{\mathcal{P}}^{-1}(1), \ldots, B_n = \varpi_{\mathcal{P}}^{-1}(n), \ldots, B_N = \varpi_{\mathcal{P}}^{-1}(N)\}!$$

In other words the finite partition $\mathcal{P}$ can be explicitly defined as a list of disjoint sets or implicitly defined by a complementary surjective function $\varpi_{\mathcal{P}}$. At the same time surjective functions can be used to implicitly define *any* partition, including countable and uncountable partitions which cannot be explicitly defined by a list of sets.

Generally any surjective function $\varpi : X \to Y$ decomposes the input space X into fibers. Each fiber $\varpi^{-1}(y)$ is mutually disjoint with all other fibers and defines a cell of the implicit partition. The union of all of these fibers for every point in the output space completely recovers the ambient space,

$$X = \bigcup_{y \in Y} \varpi^{-1}(y).$$

Consequently the collection of these fibers defines a partition $\mathcal{P}_{\varpi}$ of X.

If the output space Y contains a finite number of elements then the fibers of ϖ define a finite partition (Figure 5). Likewise if Y contains a countably infinite number of elements then the fibers define a countable partition, and if Y contains an uncountably infinite number of elements then the fibers define an uncountable partition (Figure 6). If the index space is a subset of the ambient space, $\varpi : X \to Y \subset X$ then we can also picture the indices as base points to which the fibers are attached (Figure 6c).

To demonstrate uncountable partitions let's consider a few examples over the product space $X = X_1 \times X_2$, where both X_1 and X_2 are copies of the real line $\mathbb{R}$. The surjective function

$$\begin{aligned} \varpi : X_1 \times X_2 &\to X_1 \\ (x_1, x_2) &\mapsto x_1 \end{aligned}$$

implicitly defines a partition that decomposes X into an uncountable number of copies of X_2, each of which can be visualized by a vertical line (Figure 7a). Similarly the surjective function

$$\begin{aligned} \varpi : X_1 \times X_2 &\to R \\ (x_1, x_2) &\mapsto r = \sqrt{x_1^2 + x_2^2} \end{aligned}$$

implicitly defines a partition that decomposes X into an uncountable number of concentric arcs, each with a fixed radii (Figure 7b).

When ϖ is not only surjective but also measurable then the fibers will be elements of the σ-algebra of the input space,

$$\varpi^{-1}(y) \in \mathcal{X}.$$

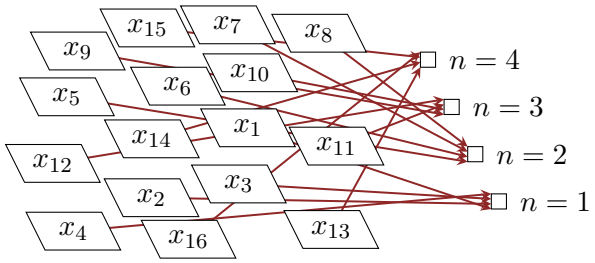
Consequently each partition cell, and hence the partition itself, will be measurable.

2 Conditioning on Explicit, Countable Partitions

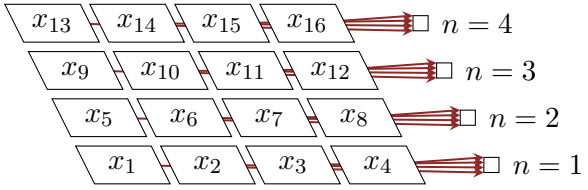
Whether defined explicitly or implicitly, a partition decomposes the ambient space X into a collection of non-overlapping subsets. Now we can consider how to decompose a probability distribution over X into a collection of probability distributions over those subsets. First let's see what we can do with a countable partition $\mathcal{P}$ explicitly defined as a list of sets.

2.1 The Law of Total Probability

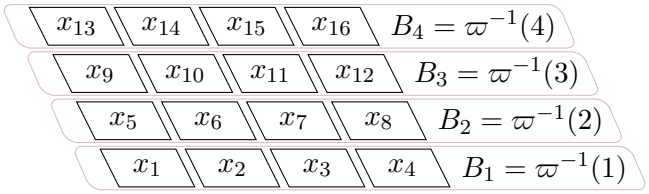
Kolmogorov's axioms define a probability distribution by its consistent allocation of allocation of probability over measurable sets $A \in \mathcal{X}$. In order to decompose a probability distribution we need to be able to decompose every measurable set, and then the probability allocated to that set.



(a)



(b)



(c)

Figure 5: On a finite ambient space (a) a surjective function ϖ (b) organizes the individual elements x_i into a table (c) where each row corresponds to a fiber $\varpi^{-1}(n)$ and hence a cell of the partition implicitly defined by ϖ.

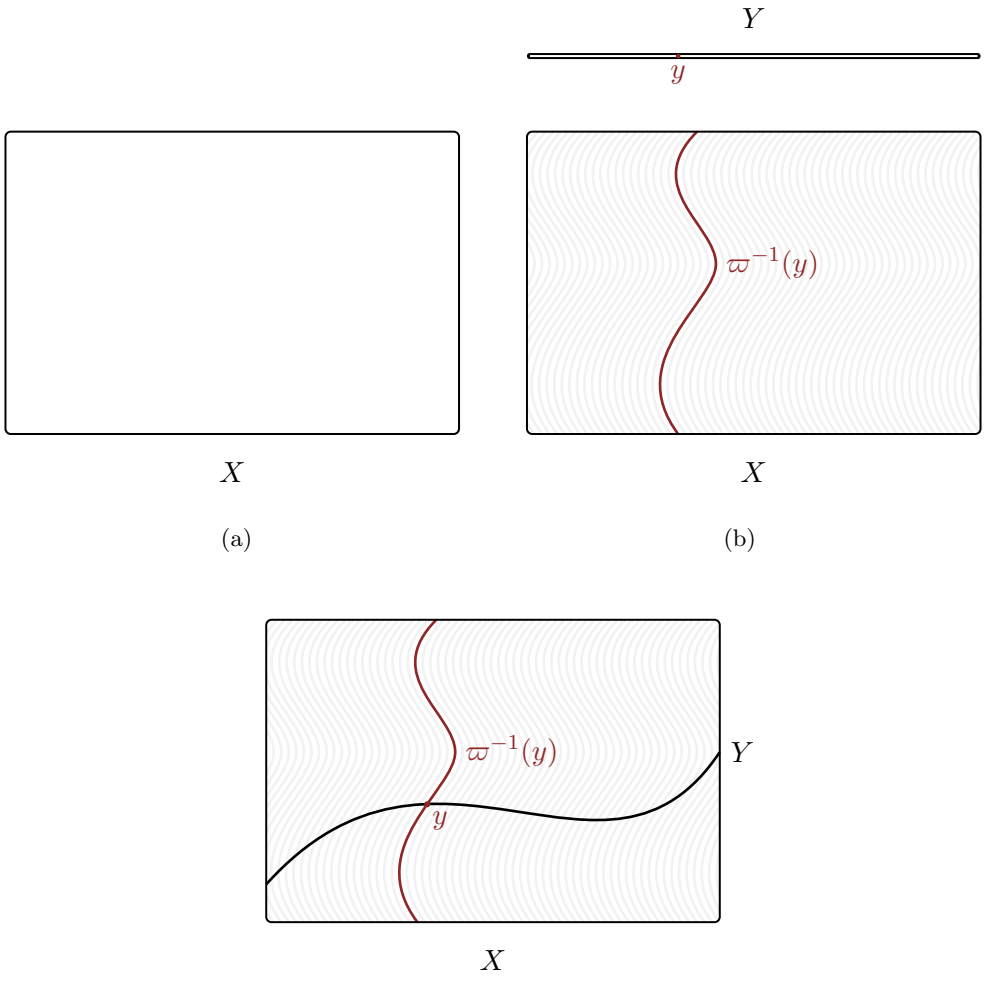


Figure 6: A surjective function $\varpi : X \to Y$ partitions the (a) input space X (b) into fibers, one for each point in the output space Y. (c) If $Y \subset X$ then ϖ can be interpreted as a *projection*, with each fiber attached to the point $y \in X$.

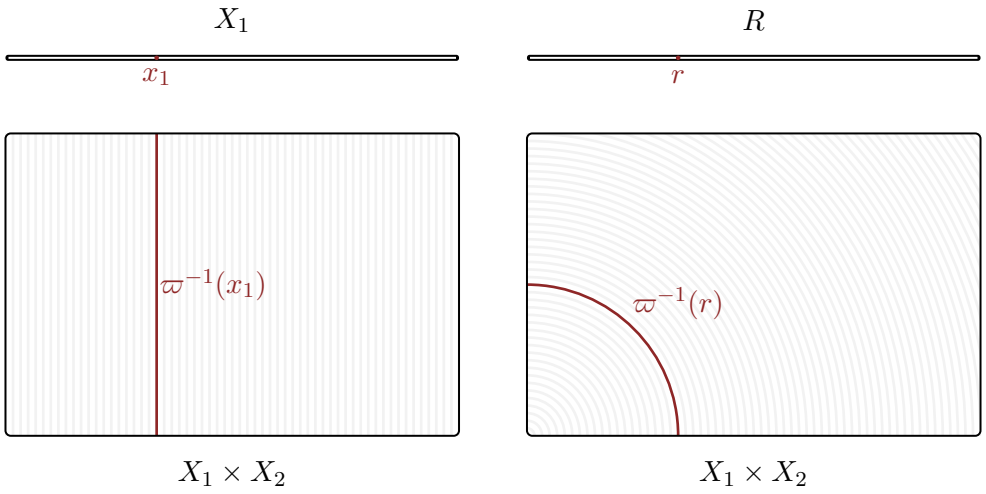


Figure 7: (a) The surjective function $\varpi : (x_1, x_2) \mapsto x_1$ decomposes the ambient space $X_1 \times X_2$ into copies of X_2, each labeled by a point $x_1 \in X_1$. (b) Likewise the surjective function $\varpi : (x_1, x_2) \mapsto \sqrt{x_1^2 + x_2^2}$ decomposes the ambient space into concentric arcs.

Any measurable set $A \in \mathcal{X}$ can be decomposed into it's intersections with the cells in a partition (Figure 8)

$$A = \bigcup_{B \in \mathcal{P}} A \cap B.$$

Because the partition cells are mutually disjoint these intersections will also be mutually disjoint: if $B, B' \in \mathcal{P}$ are two distinct partition cells then

$$\begin{aligned}(A \cap B) \cap (A \cap B') &= (A \cap B) \cap (B' \cap A) \\ &= A \cap (B \cap B') \cap A \\ &= A \cap \emptyset \cap A \\ &= \emptyset.\end{aligned}$$

If the partition $\mathcal{P}$ is countable then any measurable set $A \in \mathcal{X}$ will decompose into a countable number of intersections with the countable number of partition cells. Moreover because, by definition, σ-algebras are closed under intersections if the partition is measurable then each of these intersections will also be measurable. Consequently we can apply the countable additivity of probability distributions to the decomposition of a set induced by a measurable, countable partition.

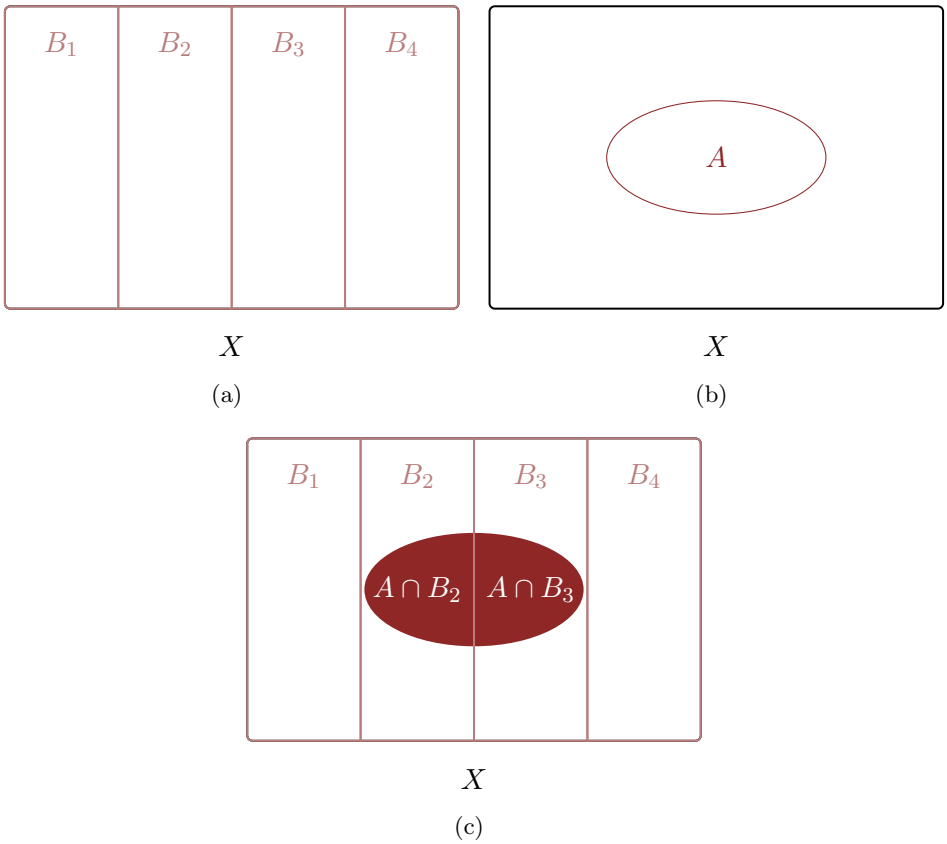


Figure 8: (a) Given the partition $\mathcal{P} = \{B_1, B_2, B_3, B_4\}$ (b) a measurable set $A \in \mathcal{X}$ (c) decomposes into the disjoint intersections, $A = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \cup (A \cap B_4)$ Here $A \cap B_1 = A \cap B_4 = \emptyset$.

In other words we can decompose the probability allocated to A into a sum of the probabilities allocated to the disjoint partition intersections,

$$\begin{aligned}\mathbb{P}_\pi[A] &= \mathbb{P}_\pi[\cup_{B \in \mathcal{P}} A \cap B] \\ &= \sum_{B \in \mathcal{P}} \mathbb{P}_\pi[A \cap B].\end{aligned}$$

This decomposition is referred to as *the law of total probability*.

2.2 Conditional Probabilities

Once we can decompose the individual probabilities allocated to measurable sets we can consider how to decompose entire probability distributions. To make our first steps towards this decomposition more manageable let's begin with a simplifying restriction on our partition.

I will refer to a partition where every cell is not only measurable but also allocated a non-zero probability by π

$$\mathbb{P}_\pi[B] > 0 \; \forall B \in \mathcal{P},$$

as a π-*non-null* partition.

When a partition $\mathcal{P}$ is countable and π-non-null we can always multiply and divide by the non-zero cell probabilities. Doing this within in each term of the law of total probability gives

$$\begin{aligned}\mathbb{P}_\pi[A] &= \sum_{B \in \mathcal{P}} \mathbb{P}_\pi[A \cap B] \\ &= \sum_{B \in \mathcal{P}} \mathbb{P}_\pi[A \cap B] \cdot \frac{\mathbb{P}_\pi[B]}{\mathbb{P}_\pi[B]} \\ &= \sum_{B \in \mathcal{P}} \frac{\mathbb{P}_\pi[A \cap B]}{\mathbb{P}_\pi[B]} \cdot \mathbb{P}_\pi[B] \\ &\equiv \sum_{B \in \mathcal{P}} \mathbb{P}_\pi[A \mid B] \cdot \mathbb{P}_\pi[B],\end{aligned}$$

where each *conditional probability*

$$\mathbb{P}_\pi[A \mid B] = \frac{\mathbb{P}_\pi[A \cap B]}{\mathbb{P}_\pi[B]}.$$

quantifies how much of the total probability allocated to the partition cell, $\mathbb{P}_\pi[B]$, is distributed to the overlap of A with that cell, $\mathbb{P}_\pi[A \cap B]$ (Figure 9).

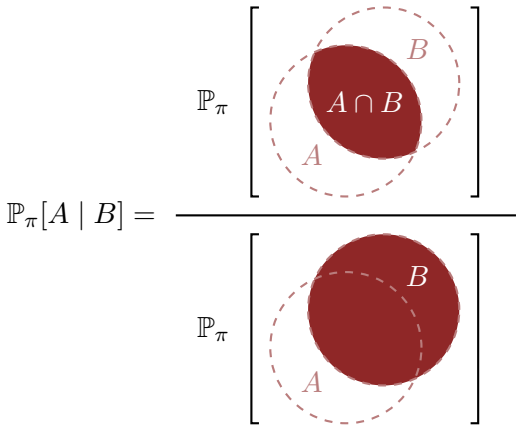


Figure 9: The conditional probability of a measurable set A given a measurable set B quantifies the proportion of probability allocated to the intersection $A \cap B$ relative to the total probability allocated to the conditioning set B.

By definition any measurable set A that doesn't overlap with the conditioning set B is allocated zero conditional probability,

$$\begin{aligned} \mathbb{P}_\pi[A \mid B] &= \frac{\mathbb{P}_\pi[A \cap B]}{\mathbb{P}_\pi[B]} \\ &= \frac{\mathbb{P}_\pi[\emptyset]}{\mathbb{P}_\pi[B]} \\ &= \frac{0}{\mathbb{P}_\pi[B]} \\ &= 0. \end{aligned}$$

At the same time any measurable set that completely overlaps with the conditioning set, $A \cap B = B$ is allocated full conditional probability,

$$\begin{aligned} &\mathbb{P}_\pi[A \mid B] \\ &\quad = \frac{\mathbb{P}_\pi[A \cap B]}{\mathbb{P}_\pi[B]} \\ &\quad = \frac{\mathbb{P}_\pi[B]}{\mathbb{P}_\pi[B]} \\ &\quad = 1. \end{aligned}$$

Conditional probabilities look suspiciously like probabilities restricted to the scope of the conditioning set. With a little more work we can show that this suspicion is correct.

2.3 Conditional Probability Distributions

Let's collect the conditional probabilities for all measurable sets $A \in \mathcal{X}$ and all conditioning sets $B \in \mathcal{P}$ into a single mathematical object. A *conditional probability distribution* with respect to a π-non-null, countable partition $\mathcal{P}$ is a binary function that maps each set A and cell $B \in \mathcal{P}$ into the corresponding conditional probabilities,

$$\begin{aligned} \mathbb{P}_{\pi|\mathcal{P}}[\cdot \mid \cdot] : \mathcal{X} \times \mathcal{P} &\to [0, 1] \subset \mathbb{R} \\ A, B \quad &\mapsto \mathbb{P}_{\pi}[A \mid B]. \end{aligned}$$

Partially evaluating this binary function on a measurable set $A \in \mathcal{X}$ in its first argument results in a measurable, unary function from each partition cell to conditional probabilities of A given that cell,

$$\begin{aligned} \mathbb{P}_{\pi|\mathcal{P}}[A \mid \cdot] : \mathcal{P} &\to [0, 1] \subset \mathbb{R} \\ B &\mapsto \mathbb{P}_{\pi}[A \mid B], \end{aligned}$$

In words this function quantifies how much the unconditional probability allocated to A contributes to the unconditional probability allocated to each partition cell.

On the other hand partially evaluating this binary function on a partition cell $B \in \mathcal{P}$ in its second argument defines what superficially looks like a probability distribution over X,

$$\begin{aligned} \mathbb{P}_{\pi|\mathcal{P}}[\cdot \mid B] : \mathcal{X} &\to [0, 1] \subset \mathbb{R} \\ A &\mapsto \mathbb{P}_{\pi}[A \mid B]. \end{aligned}$$

To confirm that this is indeed a probability distribution we have to verify all of the Kolmogorov axioms.

Immediately the inputs and output spaces of this unary function immediately satisfy the first Kolmogorov axiom: $\mathbb{P}_{\pi|\mathcal{P}}[\cdot \mid B]$ maps measurable sets to probabilities.

In order to satisfy the second Kolmogorov axiom the probability allocated to the entire ambient set must be one. Indeed

$$\begin{aligned} \mathbb{P}_{\pi|\mathcal{P}}[X \mid B] &= \frac{\mathbb{P}_{\pi}[X \cap B]}{\mathbb{P}_{\pi}[B]} \\ &= \frac{\mathbb{P}_{\pi}[B]}{\mathbb{P}_{\pi}[B]} \\ &= 1. \end{aligned}$$

Finally we need to satisfy countable additivity. For any countable collection of disjoint sets

$\{A_1, ..., A_n, \ldots\}$ with $A_n \cap A_{n'} = \emptyset$ we have

$$\begin{aligned}
\mathbb{P}_{\pi|\mathcal{P}}[\cup_n A_n \mid B] &= \frac{\mathbb{P}_\pi[(\cup_n A_n) \cap B]}{\mathbb{P}_\pi[B]} \\
&= \frac{\mathbb{P}_\pi[\cup_n (A_n \cap B)]}{\mathbb{P}_\pi[B]} \\
&= \frac{\sum_n \mathbb{P}_\pi[A_n \cap B]}{\mathbb{P}_\pi[B]} \\
&= \sum_n \frac{\mathbb{P}_\pi[A_n \cap B]}{\mathbb{P}_\pi[B]} \\
&= \sum_n \mathbb{P}_\pi[A_n \mid B]
\end{aligned}$$

as needed.

With all three Kolmogorov axioms verified we can now formally state that for each $B \in \mathcal{P}$ the partial evaluation $\mathbb{P}_{\pi|\mathcal{P}}[\cdot \mid B]$ defines a probability distribution over the ambient space X. Consequently we can interpret a conditional probability distribution as a collection of probability distributions over X, one for each cell in the partition.

Upon closer inspection, however, these probability distributions are a little bit odd Any set that doesn't intersect with B at all is allocated zero probability by $\mathbb{P}_{\pi|\mathcal{P}}[\cdot \mid B]$. Indeed all of the probability is focused onto the conditioning set B itself,

$$\begin{aligned}
\mathbb{P}_{\pi|\mathcal{P}}[B \mid B] &= \frac{\mathbb{P}_\pi[B \cap B]}{\mathbb{P}_\pi[B]} \\
&= \frac{\mathbb{P}_\pi[B]}{\mathbb{P}_\pi[B]} \\
&= 1!
\end{aligned}$$

In other words each of these probability distributions concentrates entirely within the conditioning set.

A natural question is then whether or not these probability distributions that concentrate into each conditioning set actually well define probability distributions over those sets.

To answer this question we first need to define an appropriate σ-algebra over each partition cell $B \subset X$. Fortunately there is natural way to restrict to σ-algebra over the ambient space into a σ-algebra over any measurable subset. The *subspace σ-algebra* over B is defined by the intersection of each $A \in \mathcal{X}$ with B,

$$\mathcal{X}_B = \{A \cap B \, \forall A \in \mathcal{X}\}.$$

In other words every set in the subspace σ-algebra $C \in \mathcal{X}_B$ can be written as the intersection of some set in the ambient σ-algebra $A \in \mathcal{X}$ and the defining subset B, $C = A \cap B$.

Restricting from the ambient σ-algebra $\mathcal{X}$ to the subspace σ-algebra $\mathcal{X}_B$ the partial evaluation $\mathbb{P}_{\pi|\mathcal{P}}[\cdot \mid B]$ defines a unary function

$$\begin{aligned} \mathbb{P}_{\pi|\mathcal{P}}[\cdot \mid B] : \mathcal{X}_B &\to [0, 1] \subset \mathbb{R} \\ C &\mapsto \mathbb{P}_{\pi}[A \mid B] \end{aligned}$$

with

$$\mathbb{P}_{\pi|\mathcal{P}}[B \mid B] = 1$$

and

$$\mathbb{P}_{\pi|\mathcal{P}}[\cup_n C_n \mid B] = \sum_n \mathbb{P}_{\pi|\mathcal{P}}[C_n \mid B],$$

which is exactly a probability distribution over B!

Consequently we have two valid interpretations of a conditional probability distribution. Firstly we can interpret a conditional probability distribution as a collection of probability distributions over the full ambient space X, each of which concentrates within one of the partition cells $B \in \mathcal{P}$, (Figure 10a). Alternatively we can interpret a conditional probability distribution as a collection of probability distributions over the partition cells themselves (Figure 10b). While the latter interpretation is more common, the former is more appropriate for technical results.

2.4 Marginal Probability Distributions

A critical limitation of conditional probability distributions is that they do not contain enough information to fully reconstruct a probability distribution π. In particular the law of total probability,

$$\mathbb{P}_{\pi}[A] = \sum_{B \in \mathcal{P}} \mathbb{P}_{\pi}[A \mid B] \cdot \mathbb{P}_{\pi}[B],$$

requires not just the conditional probabilities allocated by a conditional probability distribution but also the unconditional probabilities of each partition cell, $\mathbb{P}_{\pi}[B]$.

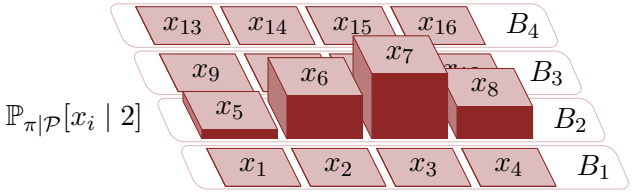
As we were able to organize the conditional probabilities into a conditional probability distribution we can also organize these unconditional cell probabilities into a probability distribution over the partition $\mathcal{P}$. When $\mathcal{P}$ is countable then we can define the *marginal probability distribution with respect to* $\mathcal{P}$ as the function

$$\mathbb{P}_{\pi_{\mathcal{P}}} : 2^{\mathcal{P}} \to [0, 1] \subset \mathbb{R}$$

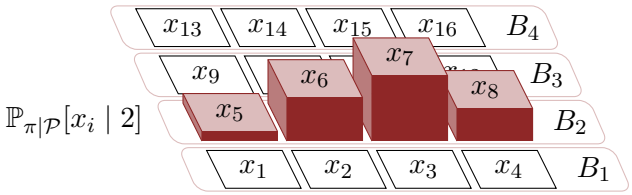
with

$$\mathbb{P}_{\pi_{\mathcal{P}}}[\mathcal{A}] = \sum_{B \in \mathcal{A}} \mathbb{P}_{\pi}[B].$$

The term "marginal" here dates back to early probability theory applications where the ambient space X consisted of a finite number of elements that could be arranged into a



(a)



(b)

Figure 10: Conditional probability distributions can be interpreted in two equally valid ways. (a) We can interpret a conditional probability distribution $\mathbb{P}_{\pi|\mathcal{P}}[x_i \mid n]$ as collection of probability distributions over the ambient space that concentrate on each partition cell. Here $\mathbb{P}_{\pi|\mathcal{P}}[x_i \mid 2]$ allocates all of its probability to the elements in $\{x_5, x_6, x_7, x_8\} = B_2$. (b) Alternatively we can interpret a $\mathbb{P}_{\pi|\mathcal{P}}[x_i \mid n]$ as a collection of probability distributions over each partition cell directly. From this perspective $\mathbb{P}_{\pi|\mathcal{P}}[x_i \mid 2]$ can allocate its total probability to only the elements $\{x_5, x_6, x_7, x_8\} = B_2$.

table with a separate row for each partition cell. Summing over each row would then give the cell probabilities which could be written in the physical margins of the table (Figure 11).

Once we've constructed a marginal probability distribution the law of total probability can be written as a marginal expectation value,

$$\begin{aligned}
\mathbb{P}_{\pi}[A] &= \sum_{B \in \mathcal{P}} \mathbb{P}_{\pi}[A \mid B] \cdot \mathbb{P}_{\pi}[B] \\
&= \sum_{B \in \mathcal{P}} \mathbb{P}_{\pi}[A \mid B] \cdot \mathbb{P}_{\pi_{\mathcal{P}}}[B] \\
&= \mathbb{E}_{\pi_{\mathcal{P}}}[p_A],
\end{aligned}$$

where p_A is the partial evaluation of the conditional probability distribution

$$p_A(B) \equiv \mathbb{P}_{\pi|\mathcal{P}}[A \mid B].$$

Consequently we can reconstruct all of the behaviors of a probability distribution π from the conditional probability distribution and marginal probability distribution induced by any π-non-null, countable partition. The marginal probability distribution quantifies how the total probability is allocated to each partition cell, and the conditional probability distribution quantifies how those allocations are further distributed within the measurable subsets of those cells (Figure 12).

In other words every π-non-null, countable partition allows us to decompose π into a conditional probability distribution and a marginal probability distribution (Figure 13). When working within the partition cells is more practical then working across the entire ambient space this decomposition can make it easier to implement the probabilistic operations defined by π.

At the same time given a π-non-null, countable partition of the ambient space X any choice of a conditional probability distribution and marginal probability distribution implicitly defines a probability distribution over X. This provides a way to build up probability distributions over complex spaces from simpler probability distributions across and within the partition cells. In this case we say that a conditional probability distribution *lifts* the marginal probability distribution across the cell partitions into a probability distribution over the full ambient space X.

2.5 Independence

Conditional probabilities also allow us to provide some intuition for the subtle concept of *independence*.

Two overlapping, measurable sets $A \in \mathcal{X}$ and $B \in \mathcal{X}$ are defined to be *independent with respect to a probability distribution* π if

$$\mathbb{P}_{\pi}[A \cap B] = \mathbb{P}_{\pi}[A] \cdot \mathbb{P}_{\pi}[B].$$

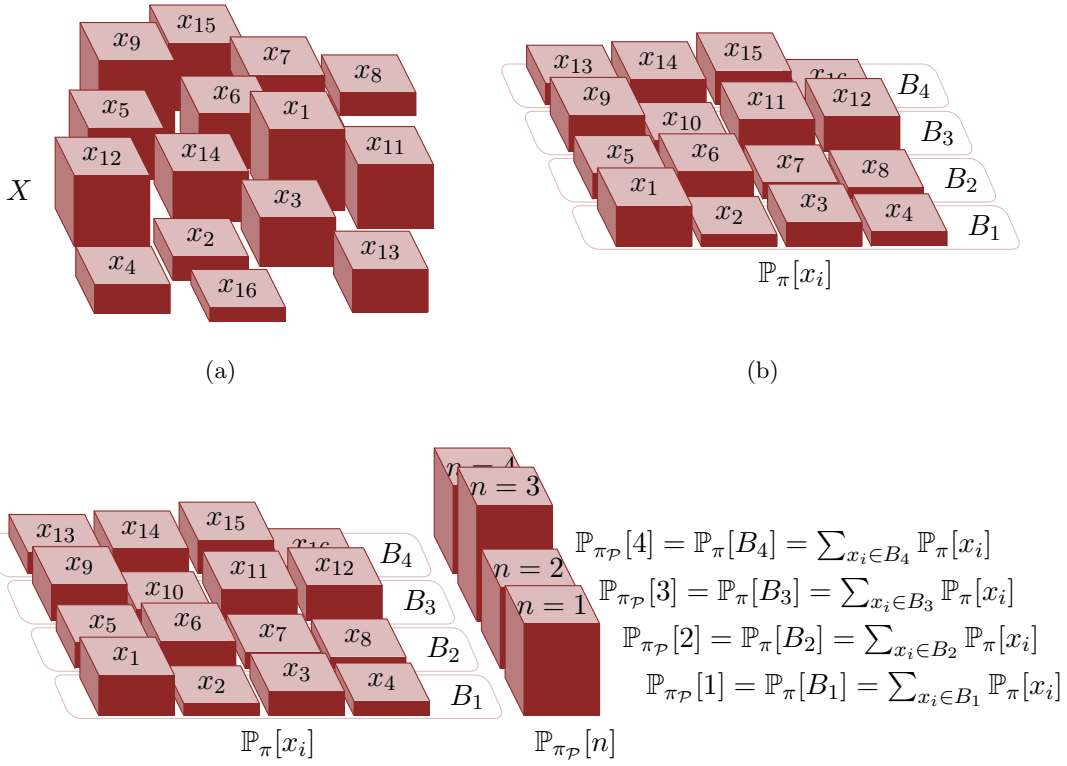


Figure 11: The term "marginal probability" for the probability of a partition cell originates in applications over ambient spaces with a finite number of elements. In this case we can arrange (a) the finite elements $X = \{x_1, \ldots, x_{16}\}$ into (b) separate rows for each partition cell $\{B_1, \ldots, B_5\}$ to form a table. (c) Summing over the probabilities allocated to the elements in each row gives the cell probabilities which can written in the "margins" of the table.

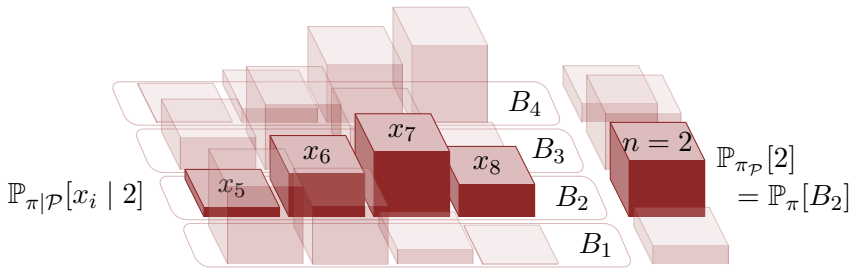


Figure 12: Marginal probability distributions quantify how much of the total probability $\mathbb{P}_\pi[X] = 1$ is allocated to each of the partition cells. Conditional probability distributions then quantify how these allocations are distributed within each partition cell. For this finite ambient space unconditional probabilities can be reconstructed as $\mathbb{P}_\pi[x_i] = \mathbb{P}_{\pi|\mathcal{P}}[x_i \mid n] \cdot \mathbb{P}_{\pi_\mathcal{P}}[n]$.

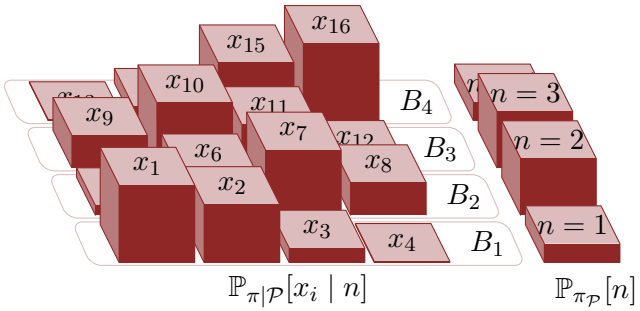


Figure 13: Given a π-non-null, countable partition, here $\mathcal{P} = \{B_1, \ldots, B_4\}$, of the ambient space, here $X = \{x_1, \ldots, x_{16}\}$, a probability distribution $\mathbb{P}_\pi[x_i]$ decomposes into a marginal probability distribution $\mathbb{P}_{\pi_\mathcal{P}}[n]$ and a conditional probability distribution $\mathbb{P}_{\pi|\mathcal{P}}[x_i \mid n]$.

The utility of this definition may not be immediately obvious but it clarifies a bit when we consider conditional probabilities.

If A and B are independent with respect to π and $\mathbb{P}_\pi[B] > 0$ then

$$\begin{aligned}
\mathbb{P}_\pi[A \mid B] &= \frac{\mathbb{P}_\pi[A \cap B]}{\mathbb{P}_\pi[B]} \\
&= \frac{\mathbb{P}_\pi[A] \cdot \mathbb{P}_\pi[B]}{\mathbb{P}_\pi[B]} \\
&= \mathbb{P}_\pi[A].
\end{aligned}$$

Similarly if $\mathbb{P}_\pi[A] > 0$ then

$$\begin{aligned}
\mathbb{P}_\pi[B \mid A] &= \frac{\mathbb{P}_\pi[B \cap A]}{\mathbb{P}_\pi[A]} \\
&= \frac{\mathbb{P}_\pi[B] \cdot \mathbb{P}_\pi[A]}{\mathbb{P}_\pi[A]} \\
&= \mathbb{P}_\pi[B].
\end{aligned}$$

In other words when two sets are independent conditioning on one the other doesn't affect how probabilities are allocated to the other.

Similarly if A is independent of all cells in partition,

$$\mathbb{P}_\pi[A \cap B] = \mathbb{P}_\pi[A] \cdot \mathbb{P}_\pi[B], \, \forall B \in \mathcal{P},$$

then

$$\mathbb{P}_\pi[B \mid A] = \mathbb{P}_\pi[B], \, \forall B \in \mathcal{P}.$$

In this case we can say that A is independent of the entire partition $\mathcal{P}$.

3 Conditioning On Implicit, Countable Partitions

The entire construction of conditional and marginal probability distributions becomes particularly elegant when we define partitions implicitly through the fibers of a surjective function.

Recall that a surjective function $\varpi : X \to Y$ implicitly defines a partition of the ambient space X where each partition cell is given by a fiber $\varpi^{-1}(y) \subset X$. When Y is countable there will be a countable number of fibers, and consequently this partition will also be countable. Similarly when ϖ is measurable the partition will also be measurable so that

$$\varpi^{-1}(y) \in \mathcal{X}, \, \forall y \in Y,$$

and we can allocate probabilities to the fibers. Moreover if every fiber is allocated a non-zero probability

$$\mathbb{P}_\pi[\varpi^{-1}(y)] > 0, \, \forall y \in Y,$$

then the partition will be π-non null.

In other words every surjective and measurable function from the ambient space to a countable output space defines a countable, measurable partition. Some of these functions will also define π-non null partitions, which then allow us to decompose π into a conditional probability distribution and a marginal probability distribution.

Given the probability distribution π the function ϖ defines a conditional probability distribution

$$\begin{aligned} \mathbb{P}_{\pi|\varpi}[\cdot \mid \cdot] : \mathcal{X} \times Y &\to [0, 1] \subset \mathbb{R} \\ A, y \quad &\mapsto \mathbb{P}_{\pi}[A \mid \varpi^{-1}(y)]. \end{aligned}$$

For each $A \in \mathcal{X}$ the partial evaluation

$$\mathbb{P}_{\pi|\varpi}[A \mid \cdot] : Y \to [0, 1]$$

defines a measurable function, and for each $y \in Y$ the partial evaluation

$$\mathbb{P}_{\pi|\varpi}[\cdot \mid y] : \mathcal{X} \to [0, 1]$$

defines a probability distribution that concentrates on the fiber,

$$\mathbb{P}_{\pi|\varpi}[\varpi^{-1}(y) \mid y] = 1.$$

Alternatively we can interpret the second partial evaluation as a probability distribution over the fiber,

$$\mathbb{P}_{\pi|\varpi}[\cdot \mid y] : \mathcal{F}_y \to [0, 1],$$

where $\mathcal{F}_y$ is the subspace σ-algebra over $\varpi^{-1}(y)$. Consequently we can think about the conditional probability distribution induced by ϖ as a collection of probability distributions over X that concentrate on each fiber or more simply as a collection of probability distributions over each fiber.

At the same time the marginal probability distribution over the countable output space Y is given by the pushforward of π along ϖ, $\varpi_* \pi$. In particular the marginal probability of $y \in Y$ is equal to the probability of the corresponding fiber,

$$\mathbb{P}_{\varpi_* \pi}[y] = \mathbb{P}_{\pi}[\varpi^{-1}(y)].$$

Consequently ϖ induces a π-non null partition if and only if the pushforward distribution $\varpi_* \pi$ allocates finite probability to every element of Y.

Together we can interpret the marginal probability distribution as quantifying how much probability π allocates to each fiber, and the conditional probability distribution as quantifying how much these allocations are further distribution within each fiber.

From this perspective the law of total probability can be formalized as an expectation with respect to the pushforward distribution,

$$\begin{aligned}
\mathbb{P}_{\pi}[A] &= \sum_{y \in Y} \mathbb{P}_{\pi|\varpi}[A \mid y] \cdot \mathbb{P}_{\pi}[\varpi^{-1}(y)] \\
&= \sum_{y \in Y} \mathbb{P}_{\pi|\varpi}[A \mid y] \cdot \mathbb{P}_{\varpi_* \pi}[y] \\
&= \mathbb{E}_{\varpi_* \pi}[p_A]
\end{aligned}$$

where

$$\begin{aligned}
p_A : y &\to \mathbb{R} \\
y &\mapsto \mathbb{P}_{\pi|\varpi}[A \mid y].
\end{aligned}$$

The law of total probability also generalizes to the *law of total expectation*: for any function $f : X \to \mathbb{R}$

$$\mathbb{E}_{\pi}[f] = \mathbb{E}_{\varpi_* \pi}[e_f]$$

where $e_f : Y \to \mathbb{R}$ is defined by the expectation values of f with respect to the probability distributions $\mathbb{P}_{\pi|\varpi}[\cdot \mid y]$. The law of total expectation is also referred to as the law of *iterated expectations*.

4 Conditioning on General Implicit Partitions

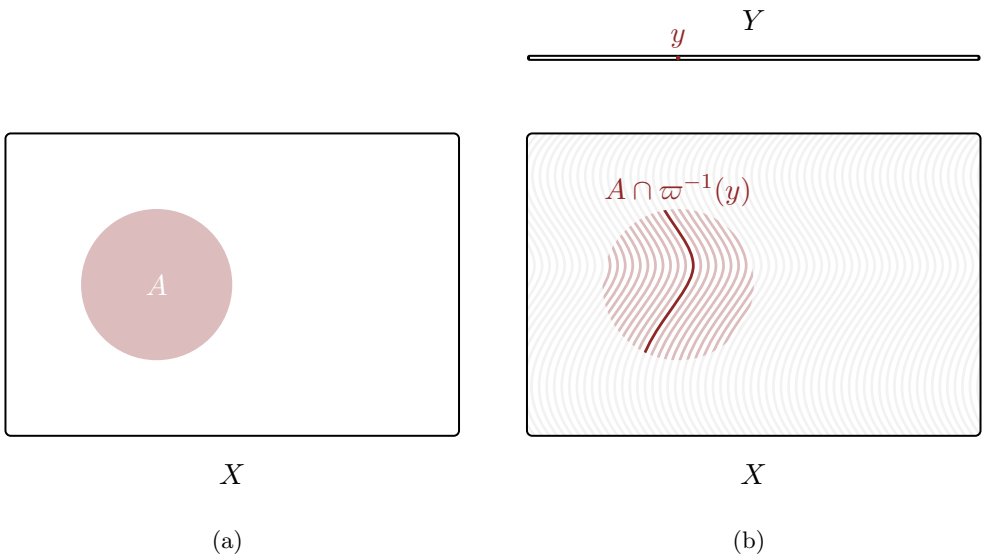
So far we have been able to define conditional probability distributions for π-non-null, countable partitions. While this construction provides useful intuition it unfortunately doesn't generalize to the continuous spaces that dominate practical applications.

Consider for example a measurable, surjective function $\varpi : X \to Y$ where both the input space X and output space Y are continuous spaces with an uncountably infinite number of elements. This function implicitly defines a partition $\mathcal{P}_{\varpi}$ of the input space X into an uncountably infinite number of fibers $\varpi^{-1}(y)$

As in the countable case we can decompose any measurable set $A \in \mathcal{X}$ into it's intersections with these fibers (Figure 14),

$$A = \bigcup_{y \in Y} A \cap \varpi^{-1}(y).$$

Because there are an uncountably infinite number of intersections, however, we cannot write $\mathbb{P}_{\pi}[A]$ as a sum over the intersection probabilities $A \cap \varpi^{-1}(y)$. Remember that probability distributions are defined to have *countable* additivity but not uncountable additivity! In other words we can't derive a law of total probability for an uncountably infinite partition.



(a) (b)

Figure 14: Given a uncountable partition (a) a set will (b) decomposed into an uncountable number of fiber intersections.

At the same time for most probability distributions the probability allocated to all of the fibers, and any subset of the fibers, will be zero,

$$\mathbb{P}_\pi[\varpi^{-1}(y)] = \mathbb{P}_{\pi_\mathcal{P}}[y] = 0.$$

Consequently we can't try to maneuver around the lack of a law of total probability and directly define conditional probabilities as the ratio

$$\mathbb{P}_{\pi|\mathcal{P}}[A \mid y] = \frac{\mathbb{P}_\pi[A \cap \varpi^{-1}(y)]}{\mathbb{P}_\pi[\varpi^{-1}(y)]}$$

because we'd be left with an indefinite $0/0$ result.

Is there hope? Can we define conditional probability distributions over continuous spaces? Fortunately the answer is yes. The key is that while we cannot sum over the fiber probabilities we can take expectations over them!

Given a measurable, surjective function $\varpi : X \to Y$ and probability distribution π one can show that in addition to the pushforward distribution

$$\mathbb{P}_{\varpi_* \pi} : \mathcal{Y} \to [0, 1],$$

there exists a binary function

$$\begin{aligned} \mathbb{P}_{\pi|\varpi}[\cdot \mid \cdot] : \mathcal{X} \times Y &\to [0, 1] \subset \mathbb{R} \\ A, y \quad &\mapsto \mathbb{P}_\pi[A \mid \varpi^{-1}(y)], \end{aligned}$$

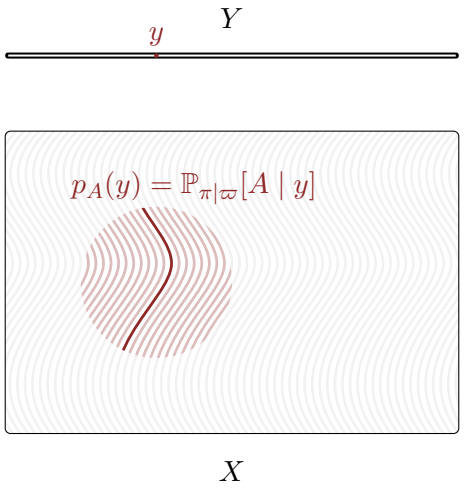


Figure 15: Although we can't sum over the vanishing probabilities allocated to the uncountable number of intersections with a measurable set $A \in \mathcal{X}$ and the fibers in an uncountable partition, we can take an expectation over the relative fiber probabilities, $\mathbb{P}_\pi[A] = \mathbb{E}_{\varpi_* \pi}[p_A]$

that not only yields a measurable function for any partial evaluation on the first argument,

$$\mathbb{P}_{\pi|\varpi}[A \mid \cdot] : Y \to [0, 1], \forall A \in \mathcal{X},$$

and a probability distribution that concentrates on the corresponding fiber for any partial evaluation on the second argument,

$$\mathbb{P}_{\pi|\varpi}[\cdot \mid y] : \mathcal{X} \to [0, 1], \forall y \in Y$$

with

$$\mathbb{P}_{\pi|\varpi}[\varpi^{-1}(y) \mid y] = 1,$$

but also satisfies (Figure 15)

$$\mathbb{P}_\pi[A] = \mathbb{E}_{\varpi_* \pi}[p_A]$$

where

$$\begin{aligned} p_A : y &\to \mathbb{R} \\ y &\mapsto \mathbb{P}_{\pi|\varpi}[A \mid y]. \end{aligned}$$

In other words the probability distribution π can be completely specified by this binary function $\mathbb{P}_{\pi|\varpi}$ and the pushforward probability distribution $\mathbb{P}_{\varpi_* \pi}$.

This latter property also implies that $\mathbb{P}_{\pi|\varpi}$ satisfies a law of total expectation: for any function $f : X \to \mathbb{R}$

$$\mathbb{E}_{\pi}[f] = \mathbb{E}_{\varpi_* \pi}[e_f]$$

where $e_f : X \to Y$ is the expectation of f with respect to the probability distribution $\mathbb{P}_{\pi|\varpi}[\cdot \mid y]$.

Any binary function satisfying these properties is denoted a *disintegration* of the probability distribution π or, less impressively, a *regular conditional probability distribution*. Unlike in the countable case, a function ϖ and probability distribution π do not uniquely define a disintegration; instead there will be an infinite number of compatible disintegrations. That said the differences between these compatible disintegrations are confined to set of zero probability and so they all define equivalent probabilities and expectation values. Consequently in practice we only ever have to consider a single disintegration.

As in the countable case we can interpret a disintegration – regular conditional probability distribution is undoubtably a more proper term but disintegration is just so fun to say – as a collection of probability distributions defined over each of the fibers $\varpi^{-1}(y)$. Once again the pushforward distribution determines how the total probability is allocated across the fibers while the disintegration determines how those allocated probabilities are further distributed along the fibers.

5 Conditional Probability Density Functions

Given a reference Lebesgue measure λ over the ambient space we can define the probability density function

$$\pi(x) = \frac{\mathrm{d}\pi}{\mathrm{d}\lambda}(x) : X \to \mathbb{R}^+,$$

from which we can evaluate expectation values as

$$\begin{aligned} \mathbb{E}_{\pi}[f(x)] &= \mathbb{E}_{\lambda}[\pi \cdot f] \\ &= \int \mathrm{d}x \, \pi(x) \, f(x). \end{aligned}$$

At the same time given the measurable function $\varpi : X \to Y$ we can define a pushforward reference measure $\varpi_* \lambda$ over Y. This allows us to define a probability density function for the pushforward of π,

$$\pi(y) = \frac{\mathrm{d}\varpi_* \pi}{\mathrm{d}\varpi_* \lambda}(y) : Y \to \mathbb{R}^+.$$

Finally each partial evaluation of a disintegration on its second argument defines the probability density function

$$\pi_y(x) = \frac{\mathrm{d}\pi_{\varpi}[\cdot \mid y]}{\mathrm{d}\lambda}(x) : X \to \mathbb{R}^+$$

that vanishes outside of the corresponding fiber,

$$\pi(x \mid y) = 0 \forall x \notin \varpi^{-1}(y).$$

The collection of these probability density functions defines a *conditional probability density function*

$$\pi(x \mid y) : X \times Y \to \mathbb{R}^{+}.$$

When Y is a finite space then these component probability density functions are given by truncating $\pi(x)$ to each fiber (Figure 16b),

$$\begin{aligned} \pi_y(x) &= \frac{\pi(x)}{\int_{\varpi^{-1}(y)} \mathrm{d}x \, \pi(x)} \\ &= \frac{\pi(x)}{\mathbb{P}_{\pi}[\varpi^{-1}(y)]}. \end{aligned}$$

The collection of these truncated probability density functions then defines a conditional probability density function (Figure 16c).

If Y is an uncountably infinite space then we can't define these component probability density functions as a truncation of $\pi(x)$ to each fiber because $\mathbb{P}_{\pi}[\varpi^{-1}(y)] = 0$. Instead the component probability density functions reduce to a ratio of probability density functions (Figure 17b),

$$\pi_y(x) = \frac{\pi(x)}{\pi(y)}.$$

The collection of these restricted probability density functions then defines a conditional probability density function (Figure 17b).

Any well-defined conditional probability density function must satisfy the law of total expectation. In particular for all measurable functions $f : X \to \mathbb{R}$ with finite expectation value we must have

$$\begin{aligned} \mathbb{E}_{\pi}[f] &= \mathbb{E}_{\varpi_* \pi}[e_f] \\ \int_X \mathrm{d}x \, \pi(x) \, f(x) &= \int_Y \mathrm{d}y \, \pi(y) \, e_f(y) \\ \int_X \mathrm{d}x \, \pi(x) \, f(x) &= \int_Y \mathrm{d}y \, \pi(y) \int_{A \cap \varpi^{-1}(y)} \mathrm{d}x \, \pi(x \mid y) \, f(x) \\ \int_X \mathrm{d}x \, \pi(x) \, f(x) &= \int_Y \mathrm{d}y \int_{A \cap \varpi^{-1}(y)} \mathrm{d}x \, \pi(y) \, \pi(x \mid y) \, f(x) \end{aligned}$$

This looks a bit ungainly, but it simplifies if we focus on the fibers. Any point $x \in X$ can be decomposed into a point $y \in Y$ and a point along the corresponding fiber $z_y \in F_y = \varpi^{-1}(y)$,

$$x = (y, z_y).$$

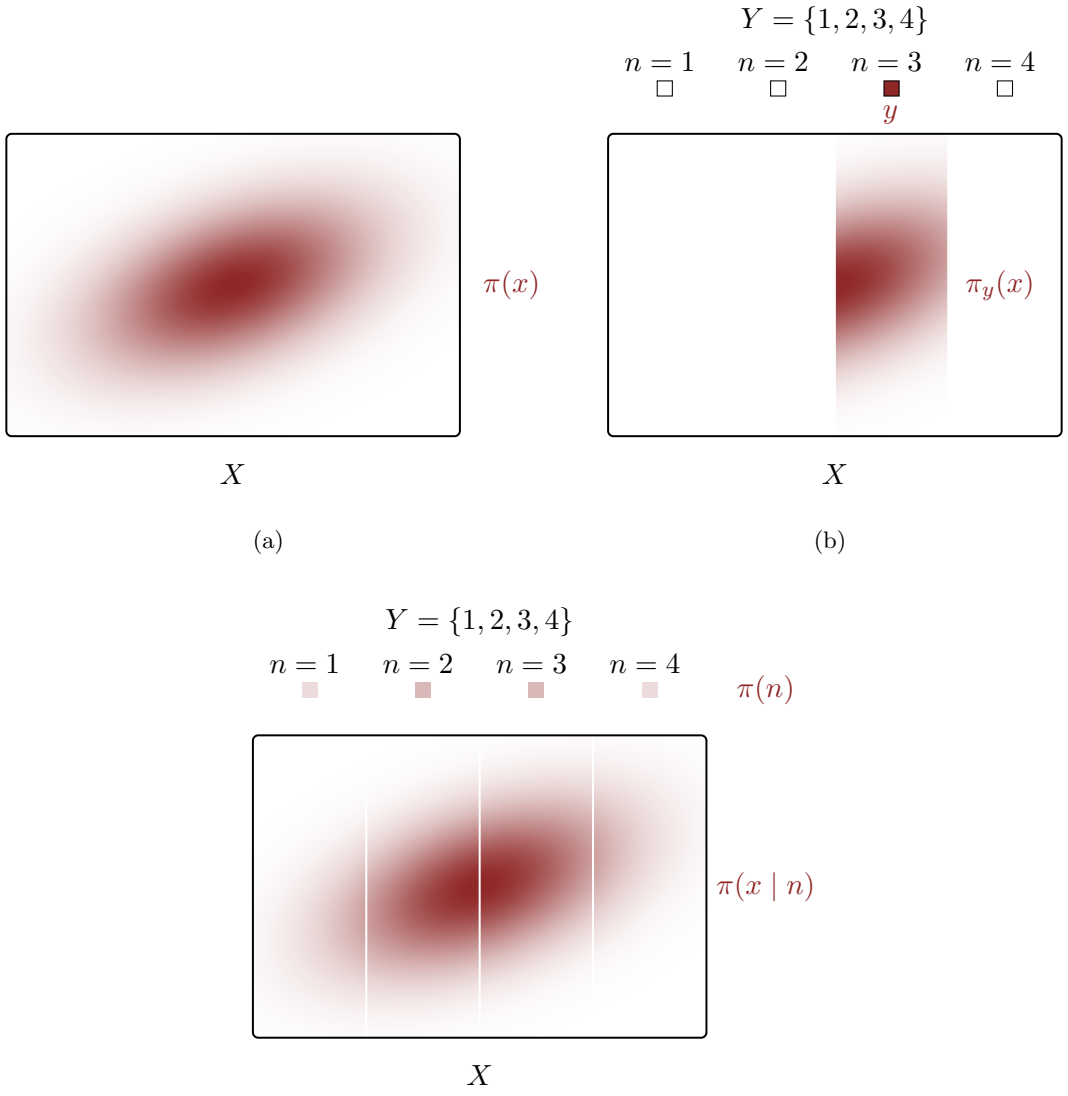


Figure 16: (a) Given a reference measure a probability distribution π over X defines a probability density function that can be used to evaluate expectation values. (b) The truncation of $\pi(x)$ to the cell of a finite partition defines a probability density function over that cell. (c) The collection of these truncated probability density functions defines a conditional probability density funtion.

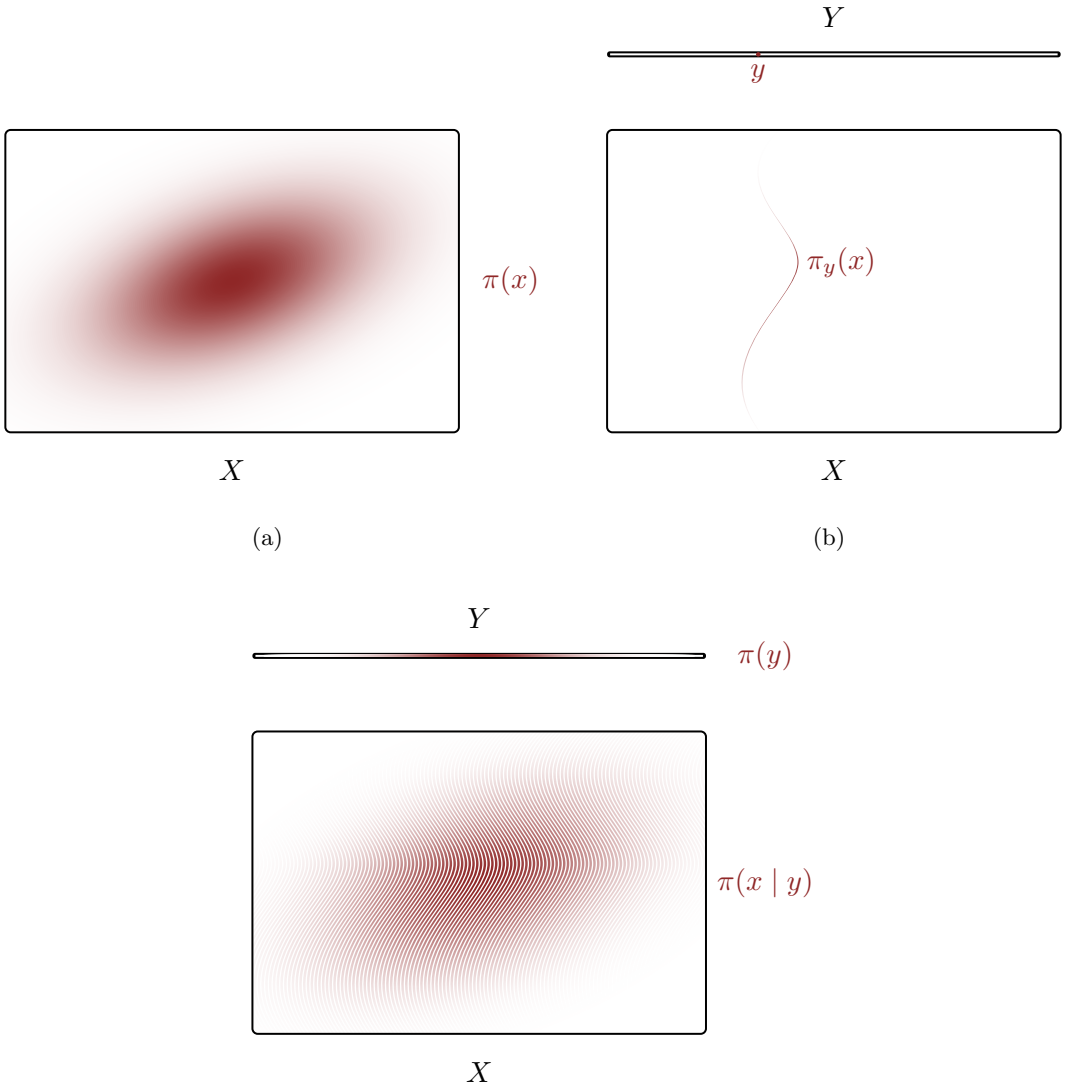


Figure 17: (a) Given a reference measure a probability distribution π over X defines a probability density function that can be used to evaluate expectation values. (b) An uncountable partition disintegrates $\pi(x)$ into a collection of probability density functions that concentrate on each of the uncountably infinite number of fibers. (c) This union of this collection defines a conditional probability density function which when coupled with the corresponding marginal probability density function completely recovers $\pi(x)$.

In general this isn't an ordered pair like we would encounter in a product space because the space in which the second component takes values depends on the the choice of the value of the first component. Mathematically this is referred to a *semi-direct product*.

From this perspective the law of total expectation requires

$$\begin{aligned}
\int_X \mathrm{d}x\, \pi(x)\, f(x) &= \int_Y \mathrm{d}y \int_{A \cap \varpi^{-1}(y)} \mathrm{d}x\, \pi(y)\, \pi(x \mid y)\, f(x) \\
\int_X \mathrm{d}x\, \pi(x)\, f(x) &= \int_Y \mathrm{d}y \int_{F_y} \mathrm{d}z_y\, \pi(y)\, \pi(z_y \mid y)\, f(y, z_y) \\
\int_Y \mathrm{d}y \int_{F_y} \mathrm{d}z_y\, \pi(y, z_y)\, f(y, z_y) &= \int_Y \mathrm{d}y \int_{F_y} \mathrm{d}z_y\, \pi(y)\, \pi(z_y \mid y)\, f(x),
\end{aligned}$$

or

$$\pi(y, z_y) = \pi(y)\, \pi(z_y \mid y)$$

for short.