Fundamental Mathematical Background for Probabilistic Circuits

Anabel Yong

8th March 2026

1 What is a Probability Distribution?

A probability distribution is a rule that assigns a chance to outcomes.

Definition 1. *Suppose a variable X, is a coin flip without outcomes where $X = \{Heads,\ Tails\}$. A distribution might be either $P(X = Heads) = 0.5$ or $P(X = Tails) = 0.5$. Now suppose there are two variables (where we have a joint distribution). X = coin flip (Heads/Tails) and Y = whether it rains tomorrow (Yes/No). A joint distribution gives probabilities for combinations like $P(X = Heads, Y = Yes)$. That is the probability of Heads and Rain.*

Definition 2. *If X is something like a real number or a continuous variable like temperature, we use a density instead of a probability at a single point. The idea is the same: it describes how likely different values are, but you need a tiny interval to talk probability. Distributions are generally a recipe for likelihood of outcomes.*

Definition 3. ***Relevant Concepts in Probability:***

- *Probability is a number from 0 to 1 describing how likely something is.*
- *Events is a set of outcomes like "roll an even number".*
- *Conditional probability, $P(A|B)$ means probability of A given B.*
- *Joint probability, $P(A, B)$ means probability that both A and B happened.*

Definition 4. *Probability Mass Function (PMF) gives probabilities of each possible value:*

$$p_X(x) = P(X = x)$$

PMF rules that all probabilities are ≥ 0 and they sum to 1. $\sum_x p_X(x) = 1$. Probability Density Function (PDF)(continuous) gives a density over values $f_X(x)$. For continuous variables, probability of a single exact point is 0.

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

PDF also rules that $f_X(x) \geq 0$ and total area equates to 1.

Definition 5. *A joint distribution tells you the probability of two variables together. For continuous variables, the joint is a joint density $f_{X,Y}(x, y)$ and probability is area/volume (integrals).*

Definition 6. *A marginal is what you get when you "ignore" the other variable. You can get $P(X)$ by summing over all values of Y.*

$$P(X = x) = \sum_y P(X = x, Y = y)$$

same with $P(Y = y) = \sum_x P(X = x, Y = y)$. The intuition here is to add the row/column to get a single-variable distribution. For continuous variables: marginalization uses integrals.

$$f_x(x) = \int f_{X,Y}(x, y)dy$$

Definition 7. *For continuous variables, the conditional probability is probability of Y given X is known. A conditional is a conditional density:*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

2 What is a graph? What is a DAG?

A graph is a collection of nodes (things) and edges (connections between things). A directed graph has a directed edge has an arrow where $A \to B$ means A points to B. A directed acyclic graph (DAG) has two properties - directed where arrows exist and acyclic where there are no loops. Therefore, in a DAG, you cannot start at a node, follow arrows, and come back to where you started. Now why this is important in circuits? Because we can compute values from bottom to top without getting stuck in a loop.

Definition 8. *Both DAG models and probabilistic circuits are really about breaking a big joint probability into smaller pieces so you can compute and learn efficiently. A DAG-based probabilistic model (often called a Bayesian network) says each variable only depends on a small set of other variables (its parents). Instead of $P(Z|X,Y)$, you assume $P(Z|Y)$, if you decide that Z depends on Y but not directly on X.* ***Structured Factorization Rule for a DAG:*** *If the DAG has nodes $X_1, \ldots, X_n$, it represents:*

$$P(X_1, \ldots, X_n) = \Pi_{i=1}^{n} P(X_i | P_a(X_i))$$

where $P_a(X_i)$ are the parents of node i in the DAG. The DAG tells you what each variable needs to "look at" to be generated.

Example 1. ***3-variable DAG:*** *Suppose DAG is:*

$$X \to Y \to Z$$

Meaning Y depends on X, Z depends on Y, Z does not directly depend on X. Then, the model factorizes as:

$$P(X,Y,Z) = P(X)P(Y|X)P(Z|Y)$$

When we compare that to chain rule:

$$P(X,Y,Z) = P(X)P(Y|X)P(Z|X,Y)$$

The DAG is effectively claiming:

$$P(Z|X,Y) = P(Z|Y)$$

Once you know Y, X adds no extra information about Z. That's the conditional independence encoded by the DAG. DAGs gives us fewer parameters (simpler conditionals), clear causal-ish interpretation and efficient inference somtimes, but not always tractable in general. DAG gives structure but inference can still be hard.

2.1 PCs: structured factorization by computation graph (sum/product)

A PC is a computation graph that builds a joint distribution by repeatedly using only **products** = factorize (independence across groups of variables) and **sums** = mixtures (latent branching/ different "modes"). So PCs implement factorizations like:

$$p(x) = \sum_{\text{some choices}} \prod_{\text{some parts}} (\text{simple leaf distributions})$$

PCs represent distributions as "mixture of factorizations". This is why they are expressive, they can say, "sometimes the world behaves like factorization A", "sometimes it behaves like factorization B", and each of those can decompose into independent parts.

2.2 Structured Factorization: what the product/sum nodes mean

A product node says:

$$p(x_{S \bigcup T}) = p(x_S)p(x_T)$$

for two disjoint sets of variables S and T. These two groups of variables don't interact (at least in this part of the model), so multiply their probabilities. A sum node says:

$$p(x) = \sum_k w_k p_k(x)$$

This is a mixture model. If you introduce a hidden variable Z saying "which child k is active", then:

$$p(x) = \sum_z p(z)p(x|z)$$

A sum node is equivalent to there is a latent choice here. Mixture is a sum of factorizations.

3 What is factorization in probability?

There is a concept in probability called independence. We start with a simple idea, if two things are independent, knowing one doesn't change the other. An example would be, X = coin flip and Y = whether a random person likes watching TV. If they have truly nothign to do with each other, then:

$$P(X,Y) = P(X) \cdot P(Y)$$

That is factorization: a joint probability becomes a product of two simpler probabilities.

3.1 Factorization (product structure)

Factorization is rewriting a joint distribution as product of simpler pieces. Most basic case is that if X and Y are independent, then $P(X,Y) = P(X)P(Y)$. More general (which works even if not independent), $P(X,Y) = P(X)P(Y|X)$. This is chain rule and is always true.

4 What is mixture model?

This is important behind sum nodes in circuits. A mixture model says "My data comes from one of several sub-populations (hidden groups) and I don't know which one each point came from.".

Example 2. *The heights of people in a room. Some are children (shorter) and some are adults (taller). If we only look at the histogram, it might look like two bumps. A mixture model says:*

- *with probability w, sample from "child-height distribution".*
- *with probability $1 - w$, sample from "adult-height distribution".*

We can model $p(x) = w \cdot p_1(x) + (1-w) \cdot p_2(x)$. That is a weighted sum of distributions.

Example 3. *We have a mixture distribution where:*

$$p(x) = 0.6p_1(x) + 0.4p_2(x)$$

Suppose $p_1(x) = 0.2$ and $p_2(x) = 0.8$, when we want to compute $p(x)$, we have $0.6(0.2) + 0.4(0.8) = 0.44$

We may also ask if mixture weights have to sum to 1. In the most simplest case, the mixture is saying "choose component 1 with probability w, and component 2 with probability $1 - w$. Those are probabilities over which component you pick, so they must sum to 1.

5 DAG Factorization vs PC Factorization

A DAG gives:

$$P(X_1, \ldots, X_n) = \prod_i P(X_i|P_a(X_i))$$

This is a product of terms, but the terms are conditionals that can still be complex. On the other hand, for PCs, a PC gives sums over latent choices, products over independent groups and leaves are simple distributions.

Definition 9. *The **product node (factorization)** indicates that if a product node combines two children that cover different variables (say, one covers X and one covers Y), it represents:*

$$p(x,y) = p(x)p(y)$$

This is a factorization (often an independence assumption.)

Definition 10. *The **Sum node (mixture:)** A sum node combines children that cover the same variables (both cover X or both cover (X,Y), and represents):*

$$p(x) = \sum_{k=1}^{K} w_k p_k(x), \text{ where } w_k \geq 0 \text{ , } \sum_k w_k = 1$$

This is a mixture model: the distribution is a weighted blend of several alternative distributions.

6 Gaussian Mixture Models Example

A 1D Gaussian normal density is:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

where $\frac{1}{\sigma\sqrt{2\pi}}$ is a normalization constant, ensures that the total area under the curve equals 1, satisfying the fundamental requirement of a probability distribution.

Example 4. *We define a 2-component mixture, where component 1:* $p_1(x) = \mathcal{N}(x|\mu_1 = 0, \sigma^2 = 1)$ *and component 2:* $p_2(x) = \mathcal{N}(x|\mu_2 = 3, \sigma^2 = 1)$*, and weights* $w_1 = 0.6$ *and* $w_2 = 0.4$*. Mixture distribution is*

$$p(x) = 0.6p_1(x) + 0.4p_2(x)$$

We introduce a hidden variable $Z \in \{1, 2\}$ *meaning "which component generated this point. Then, we have:*

$$p(x) = \sum_{z\in\{1,2\}} P(Z = z)p(x|Z = z)$$

This is exactly the sum node idea: a weighted sum over hidden choices. Now we are trying to compute a mixture density at a specific point $x = 1$*.*
Step A: compute each component density*. Component A:* $\mu_1 = 0, \sigma = 1$*:*

$$p_1(1) = 0.3989 \cdot \exp(-\frac{(1-0)^2}{2}) = 0.3989 \cdot e^{-0.5}$$

Therefore,

$$p_1(1) = 0.3989 \cdot 0.6065 \approx 0.2420$$

Component B: $\mu_2 = 3, \sigma = 1$*,*

$$p_2(1) = 0.3989 \cdot \exp(-\frac{(1-3)^2}{2}) = 0.3989 \cdot e^{-2} = 0.054$$

Step B: weighted sum (the mixture)

$$p(1) = 0.6 \cdot 0.2420 + 0.4 \cdot 0.054 = 0.1668$$

The mixture density at $x = 1$ *is:*

$$p(1) \approx 0.1668$$

6.1 Responsibility/posterior: "which component likely generated $x = 1$?"

This is the Bayes step:

$$P(Z = 1|x) = \frac{P(Z = 1)p(x|Z = 1)}{\sum_j P(Z = j)p(x|Z = j)}$$

Here, the numerator of the equation is $0.6 \cdot p_1(1) = 0.6 \cdot 0.2420 = 0.1452$. The denominator is $p(1) = 0.1668$. Therefore:

$$P(Z = 1|x = 1) = \frac{0.1452}{0.1668} \approx 0.8705$$

$$P(Z = 2|x = 1) = 1 - 0.8705 = 0.1295$$

Eventhough component 1 prior is at 0.6, at $x = 1$, it fits much better than component 2, so the posterior becomes 0.8705.

6.2 Mixture is a sum of factorizations - Gaussian Mixtures

PC: sum over components, each component is a product (factorization).

Example 5. ***Component 1 (independent Gaussians):***

$$p_1(x, y) = \mathcal{N}(x|0, 1) \cdot \mathcal{N}(y|0, 1)$$

Component 2 (independent Gaussians):

$$p_2(x, y) = \mathcal{N}(x|3, 1) \cdot \mathcal{N}(y| - 2, 1)$$

We will go onto E and M step for computing responsibilities and updating weights on a tiny dataset. For now these are the fundamentals required for learning probabilistic circuits.