Bayes Theorem: A Geometric Interpretation

Adam Cuculich

November 24, 2023

Introduction:

This document aims to provide a geometric, or visual, interpretation of Bayes' Theorem. It endeavors to guide readers from the fundamental principles to an intuitive understanding of the theorem. Starting with known probabilities, the paper derives Bayes' Rule using conditional probabilities and advances to a pragmatic application of Bayes' Theorem, addressing probability distributions represented as random variables. In doing so, I strive to bridge analytical descriptions with geometric interpretations, thus offering a comprehensive insight into the theorem's intrinsic dynamics. Ultimately, the goal is to lay down a theoretical yet intuitive foundation in Bayesian statistics.

Bayes Rule

Shown below, Bayes' Rule can be derived using known conditional probabilities. A geometric interpretation of the underlying mechanics is also provided in this section.

Conditional Probabilities:

Let A and B be distinct sets of events in set S, where S is the sample space of all possible events. Set theoretically, let $A \subseteq S$ and $B \subseteq S$.

The conditional probability of event A, given event B is defined as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

This relationship can be expressed visually as follows:

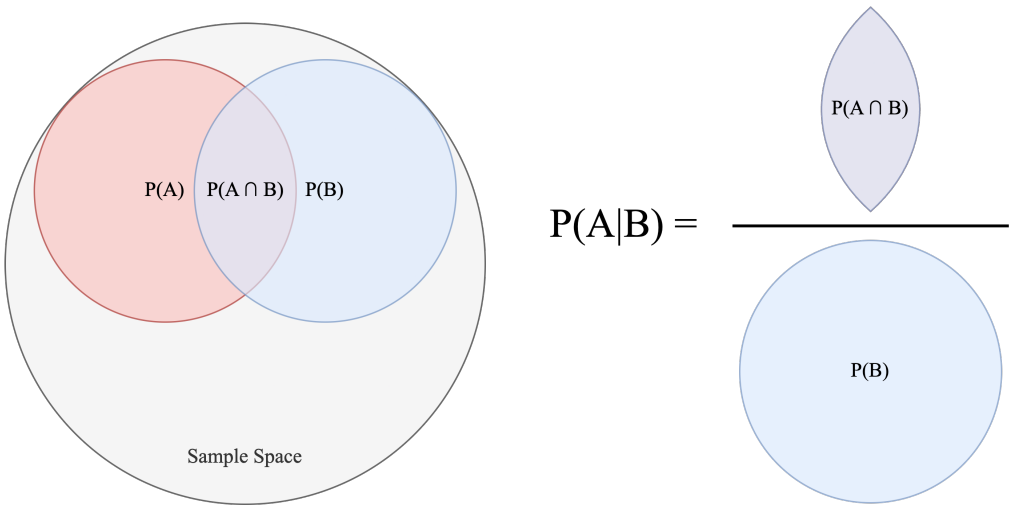

Figure 1: A visual representation of conditional probability

Shown above, some of the shapes are subsets of other shapes. Each shape represents an event, or set of outcomes, that an upcoming outcome can be categorized as. Each shape's area is used to represent the probability of an

outcome being categorized into the event the shape represents. The sample space, S, represents all possible events relevant to a specific, given context. Because all possible, mutually exclusive and collectively exhaustive events are encompassed within S, you can be 100% certain that an "upcoming outcome" will be categorized within the set of outcomes that is S. Therefore, $P(S) = 1$. As a useful representation, the probability of an upcoming outcome being categorized as a particular event is a number that describes the certainty of that outcome taking on the specific properties that define that event. When it comes to conditional probabilities, the following question provides insight into their nature: "given you know B happened, what's the probability A also happened?" The added knowledge of B's occurrence confines the set of outcomes to those in B. Succinctly, $P(A \mid B)$ is the proportion of occurrences of B where A also occurs. This proportion can be 0, 1, or anywhere in between. For example, if events A and B are mutually exclusive, then the probability that A occurs given a known B occurrence would be 0. If instead, A always occurs when B occurs, it would be 1.

Deriving Bayes' Rule:

Given the definition of conditional probability,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{2}$$

Similarly,

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} \tag{3}$$

By equivalence,

$$P(A \cap B) = P(B \cap A) \tag{4}$$

$$P(A \cap B) = P(A \mid B)P(B) \tag{5}$$

$$P(B \cap A) = P(B \mid A)P(A) \tag{6}$$

$$P(A \mid B)P(B) = P(B \mid A)P(A) \tag{7}$$

Hence, Bayes' Rule is given by:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{8}$$

With Bayes' Rule established, we transition from the abstract events, A and B, to the concepts of experimentally observed events (Evidence, E) and proposed hypotheses (H). This transition is underpinned by the principle that, akin to A and B, both observed data (evidence) and hypotheses are events within a shared sample space. Hypotheses are characterized as events that are evaluated in light of the observed data.

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \tag{9}$$

Total Probability:

Above, $P(E)$ is the total probability of the evidence occurring, taking into account all possible, mutually exclusive hypotheses. Now, suppose you constructed an array with all possible hypotheses, each with its own probability. To ensure that all cases and their compliments are accounted for, the sum of all of the hypotheses' probabilities should be 1. For each hypothesis, there will be some probability of the evidence in question occurring, given that hypothesis (a conditional probability). For clarity, here the evidence is conditioned on that hypothesis. We now have an array where each element in that array is a conditional probability: namely, probability of the evidence, E, given that element's hypothesis. This can be expressed for n hypotheses as follows:

$$P(E) = P(E \mid H_1)P(H_1) + P(E \mid H_2)P(H_2) + ... + P(E \mid H_n)P(H_n) \tag{10}$$

Or more succinctly,

$$P(E) = \sum_{i=1}^{n} P(E \mid H_i)P(H_i) \tag{11}$$

With this established, Bayes' Rule can be expressed as follows for a given hypothesis, H_j:

$$P(H_j \mid E) = \frac{P(E \mid H_j)P(H_j)}{\sum_{i=1}^{n} P(E \mid H_i)P(H_i)}, \text{ where } j \in \{1, 2, \ldots, n\} \tag{12}$$

Bayes' Rule: A Geometric Interpretation:

The foregoing states that the probability of a hypothesis, given evidence is the conditional probability of evidence given that particular hypothesis (H_j) over the total probability of the evidence, taking into account all possible hypotheses. This can be visually represented as follows:

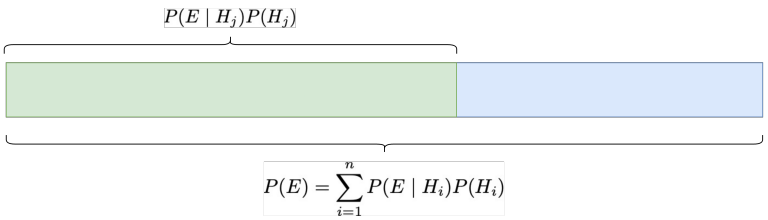

Figure 2: A visual representation of Bayes' Rule

Now, let's lay the ground work for a Geometric interpretation. Bayes' Rule can be rearranged as follows:

Given:

$$P(E \mid H) = \frac{P(E \cap H)}{P(H)}, \tag{13}$$

Substituting this for all occurrences in equation 12, it can be shown that:

$$P(H_j \mid E) = \frac{P(E \cap H_j)}{\sum_{i=1}^{n} P(E \cap H_i)}, \text{ where } j \in \{1, 2, \ldots, n\} \tag{14}$$

This states that probability of hypothesis H_j, given evidence, E, is the intersection of the evidence with H_j over the sum of all intersections between E and all hypotheses. As a visual demonstration, take the case where there are 4 hypotheses.

Suppose a hypothesis-evidence space, in which the x-axis represents all possible hypotheses and their probabilities, and the y-axis represents the probability space for evidence conditioned on a given hypothesis.

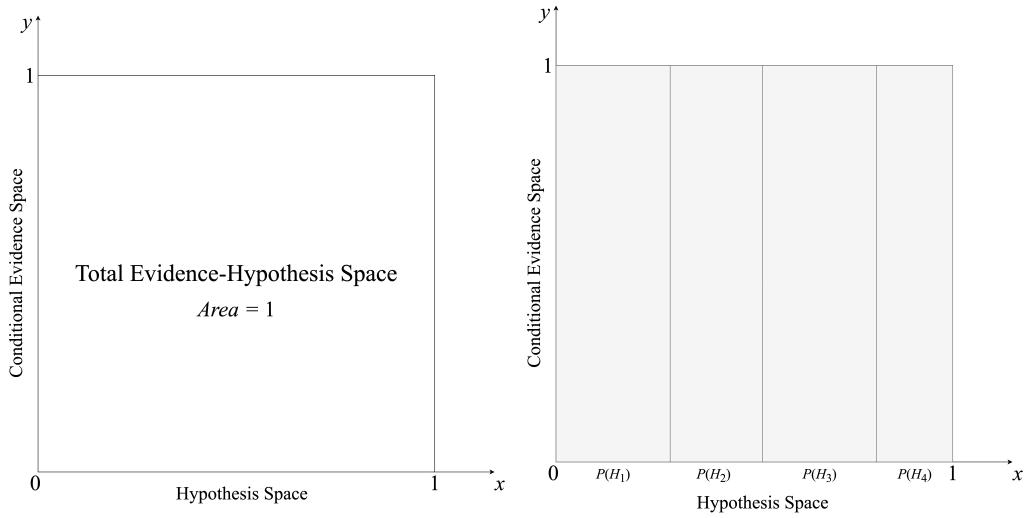

Figure 3: The hypothesis-evidence space divided by hypotheses.

Shown above, the hypothesis space can be divided into all relevant hypotheses, denoted by the subsections above. Such sections are actually one-dimensional, as they are orthogonal to the conditional evidence space.

Each hypothesis will have its own probability of evidence, previously referenced as $P(E \mid H_i)$ in the equations above. These conditional probabilities can be represented as the y-axis component of the joint probability sections below.

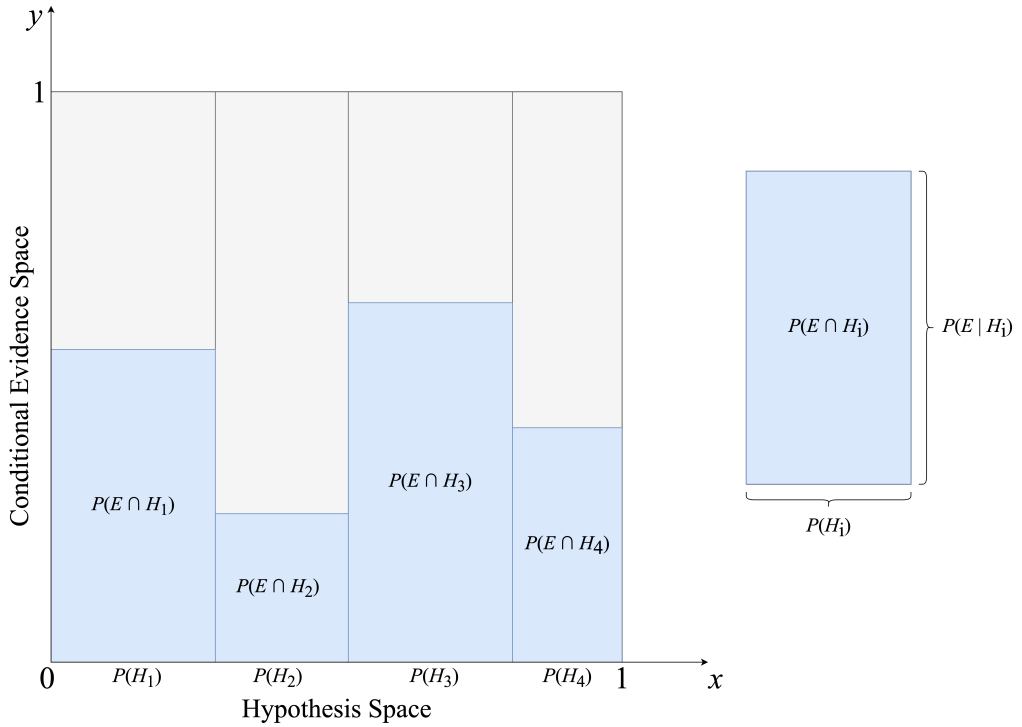

Figure 4: Left: Hypothesis-evidence intersections visualized

Examining a single, arbitrarily-chosen subsection, note that the subsection's area is equal to the probability of the hypothesis multiplied by the probability of evidence, given that specific hypothesis. This finding is consistent with equation 5 above.

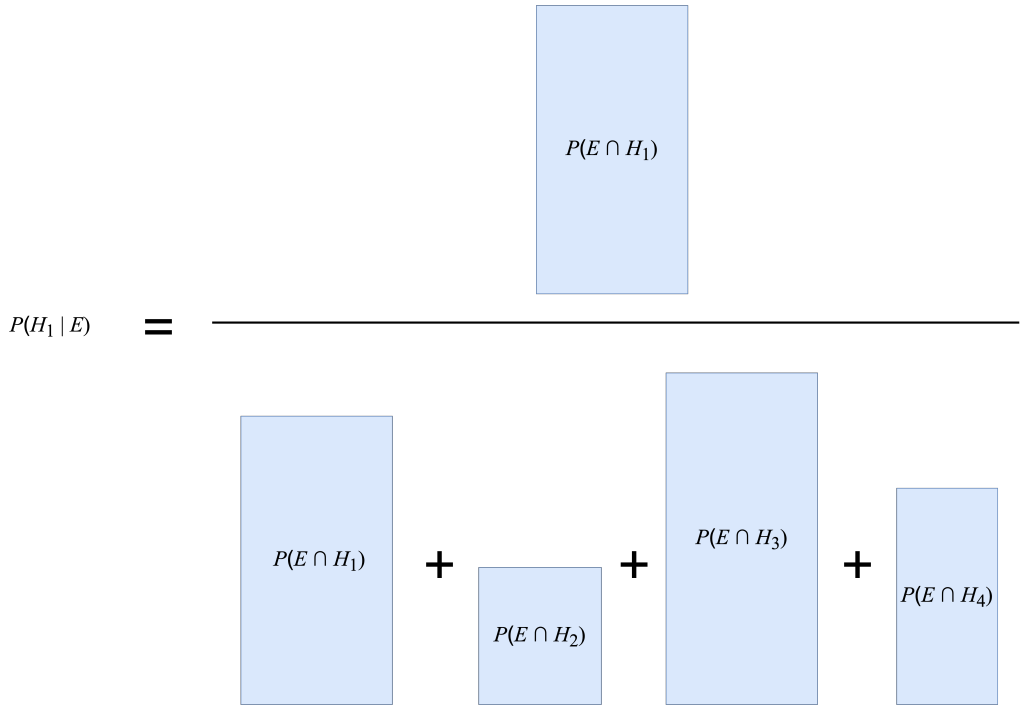

Figure 5: Probability of first hypothesis given evidence, visualized.

Bayes' Theorem with Continuous Random Variables:

The above presupposes that the probabilities involved in Bayes' Rule are known. In the real world, however, probabilities may not be known with certainty. To deal with such cases, the same mechanics can be applied with random variables. To mathematically express the uncertainty of such variables, we use probability distributions in place of known probabilities. Such probability distributions are what's meant by the term, "random variables". Using random variables, Bayes' Theorem is expressed as:

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta)\pi(\theta)d\theta} \tag{15}$$

It's worth taking some time to understand what this equation is expressing. Intuitively, we know that Bayes' Theorem incorporates information from both experiment and prior beliefs, but to what end? We've heard that a bayesian update is the process by which a prior distribution is updated to a poste-

rior distribution in light of experimental evidence, but how exactly is that achieved? In an effort to gain a deeper understanding, let's first understand the numerator of the theorem, as the denominator is simply the integral of the numerator with respect to θ.

Likelihood

The term, $f(x \mid \theta)$, is often referred to as the likelihood. Here, the likelihood function evaluates the plausibility of the data x under a statistical model parameterized by θ, and identifies the parameter values that are most probable given the observed data. This definition involves some 'statistical model', about which the data plausibility can be assessed against.

Model Selection Let's piece these concepts together with an example. Suppose that we're modeling the 'time to an event', say, the operating lifetime of a certain type of component in a machine. Also, suppose that we have reason to believe the lifetime of that component can be modeled as an exponential distribution:

$$f(x \mid \theta) = \theta e^{-\theta x}, \text{ where } \theta > 0 \text{ and } x \geq 0 \tag{16}$$

Here, $f(x \mid \theta)$ is a function of x and θ, where θ is the rate parameter. The greater the value of θ, the faster the function decays toward 0. At this point, we've proposed a statistical model as an attempt to explain the lifetime of the component type of interest. Let's call this proposal *model selection*. With a selected model, we now have something to compare experimental data against.

Experiement Suppose that an experiment is conducted to measure the lifetimes of the component type of interest. Intuitively, the empirical data outputted from such an experiment should certainly be used to inform the model we've proposed. Such data can be represented as a set of lifetimes, $\{x_1, x_2, ..., x_n\}$.

Calculating Likelihood Using the set of empirically obtained lifetimes, we can now evaluate each observed lifetime x_i against our model, which can be expressed as:

$$f(x_i \mid \theta) = \theta e^{-\theta x_i} \tag{17}$$

This expression is the likelihood function for a given x_i. Given that x_i is known (it came from the experiment), $f(x_i \mid \theta)$ condenses down to a function of θ. The statistical model we're evaluating against is indeed a probability density function over x values (component lifetimes) when θ is a fixed, constant value. That said, equation 17 above expresses the reverse: it fixes a known x_i value, and allows θ to vary, resulting in $f(x_i \mid \theta)$ as solely a function of θ. It is this act of providing known x_i values that produces a likelihood function. Likelihood functions are not probability density functions for the sole purpose that they need not integrate to 1, with respect to θ. After evaluating each x_i against our model, we have a set of likelihood functions:

$$\{f(x_1 \mid \theta), f(x_2 \mid \theta), ..., f(x_n \mid \theta)\} = \{\theta e^{-\theta x_1}, \theta e^{-\theta x_2}, ..., \theta e^{-\theta x_n}\} \tag{18}$$

Each observation x_i is entirely independent. Therefore, given θ, the joint probability density of observing all lifetimes can be represented as:

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{19}$$

Or in our example,

$$f(x_1, \ldots, x_n|\theta) = \theta^n e^{-\theta \sum_{i=1}^{n} x_i} \tag{20}$$

Now, with observed data as given and allowing theta to vary, the same expression can be interpreted through the lens of a likelihood function:

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta) = \theta^n e^{-\theta \sum_{i=1}^{n} x_i} \tag{21}$$

In summary, we've arrived an an expression, in terms of θ informed by the experimental data set of lifetimes.

The Prior

Thus far, we've selected a model, gathered observed data from an experiment, and constructed a likelihood function in terms of parameter of interest, θ. That said, prior to an experiment, we do not know the value of θ. To account for such uncertainty, we can declare it a random variable and apply a probability distribution on what we think it's value is. For example, we may represent θ as taking on a range of possible values in $\pi(\theta)$, a normal distribution with a defined mean, μ, and standard deviation, σ^2. Such a distribution is called the *prior*, as it reflects beliefs about the parameter of interest prior to gathering empirical data. It's analog in cases where probabilities are known is $P(H)$. Similar to the likelihood function in the numerator of Bayes' Theorem, the prior distribution is also a function of θ.

A Geometric Interpretation

We've demonstrated above that the numerator of Bayes' Theorem consists of two functions of θ multiplied by each other. Deviating from the specifics of the above example, the following visually demonstrates the multiplication of two arbitrary distributions:

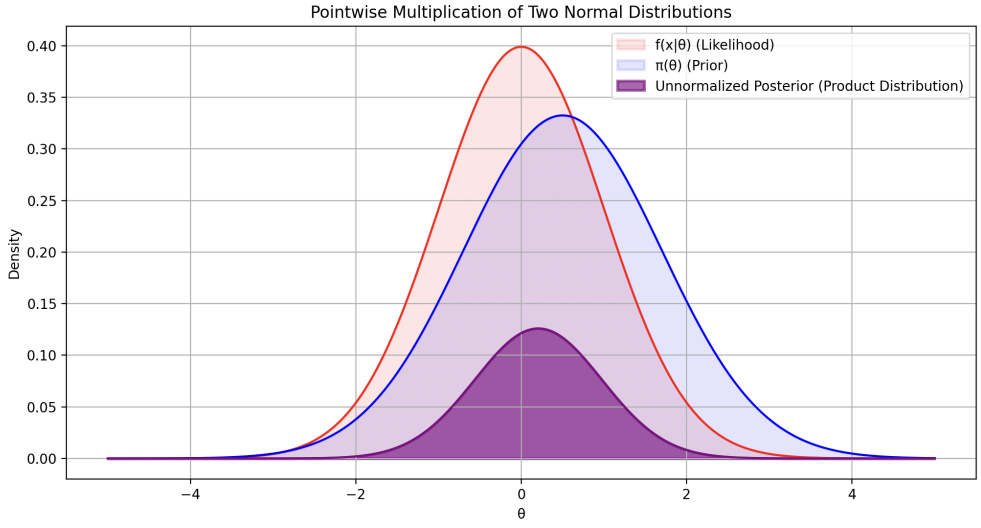

Figure 6: Two factor distributions and their product distribution

It's worth taking a moment to understand what the representation above

expresses. As shown, the product distribution appears to be centered between the hypothetical likelihood and prior distributions. This distribution positioning represents the idea that the product distribution is indeed accounting for information from both experiment (likelihood) and prior beliefs. The two distributions are multiplied together by an operation called a point-wise multiplication. A point-wise multiplication of two distributions involves multiplying the values of their representative functions at each point in their domain, resulting in a new distribution that reflects the combined characteristics of both original distributions at each point. This process is visually demonstrated for a discrete case below:

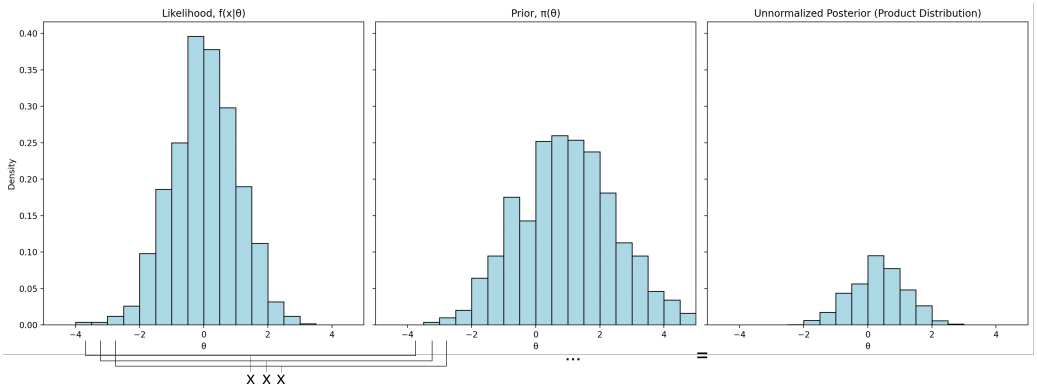

Figure 7: Point-wise multiplication visualized for a discrete case

The Posterior

While the product distribution clearly incorporates data from both prior beliefs and experiment, its evident that the product distribution's total area is significantly smaller than both of the factor distributions. This is a problem, as the distribution must integrate to 1 in order to be considered a valid probability distribution. At this moment, the product distribution is proportional to what will be the posterior distribution $\pi(\theta \mid x)$, but it does not integrate to 1. For that reason, it's called the *unnormalized posterior*. To obtain we a valid posterior probability distribution, we need a distribution that achieves the following two properties: first, it must be proportional to the unnormalized posterior distribution and second, it must integrate to 1. To achieve these properties, the unnormalized posterior can be divided by what's called,

normalizing constant, which is the scalar value, $\int_\Theta f(x \mid \theta)\pi(\theta)d\theta$. I'd like to share some intuition behind why this operation achieves the aforementioned properties we're after.

The Normalizing Constant

Suppose we're given a function, $g(x)$ that we'd like to normalize to $f(x)$, such that $f(x)$ satisfies the following:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \tag{22}$$

$$f(x) \propto g(x) \tag{23}$$

The normalizing constant mentioned above is simply the total area encompassed by the unnormalized function, which is $g(x)$ in our case. This can be represented as:

$$A_T = \int_{-\infty}^{\infty} g(x)\,dx \tag{24}$$

It's known that along the function $g(x)$, each $g(x_i)$ value has an infinitesimal dx associated with it, for which x_i is a representative value. Given that, the total area A_T can be conceptualized as the sum of individual, infinitesimal slices, each with an area, $g(x_i)dx$. With that established, it's possible to obtain each slice's percentage of the total area by dividing by A_T:

$$A_i = \frac{g(x_i)dx}{A_T} \tag{25}$$

Given that dx is of infinitesimal length, the sum of all infinitesimal slice's areas accounts for 100% of the total area, A_T. Obtaining the fractional area percentage, A_i, of each slice within the relevant x range will result in an infinite set of area percentages, the sum of which, is 100%. This achieves the property declared in equation 22 $\rightarrow$ it integrates to 1.

Furthermore, it may be useful to conceptualize $g(x)/A_T$ as a point-wise division across infinite $g(x)$ values, where any two consecutive operations are:

$g(x_i)/A_T$ and $g(x_i + dx)/A_T$. Given that A_T is a scalar value that is invariant with respect to x, the quotient $g(x)/A_T$ is indeed proportional to $g(x)$, thereby achieving the property declared in statement 23:

$$f(x) = \frac{g(x)}{A_T} \propto g(x) \tag{26}$$

Given that the resulting function $f(x)$ is proportional to $g(x)$ and integrates to 1, we can declare that $f(x)$ is a normalized version of $g(x)$. These are the same mechanics used to normalize unnormalized posterior distributions shown above. In practice, the normalizing constant may be difficult to compute, which gives rise to probabilistic sampling methods, such as the Metropolis-Hastings or Gibbs Sampling algorithms.

Conclusion

It is my hope that this document serves as a useful tool for better understanding some of the mechanics involved in Bayes' Theorem. As a fan of open-source software, the Latex for this document will be available on GitHub. If anything is incorrect or can be explained better, I strongly urge you to open a pull request! Thank you.