3

PROBABILISTIC MODELLING AND INFERENCE

In this chapter, we will cover probabilistic modelling in more detail and then talk about inference. We will also review probability basics and a large range of applications the Bayesian viewpoint enables.

3.1 INTRODUCTION

In the previous chapter, we have seen how to generate data from a probabilistic model. Despite we have only simulated from a linear model as an example, the idea is general. We will see more about simulating models in other parts of the course. We have seen that

$$X_i \sim p(x), \tag{3.1}$$

$$Y_i|X_i = x_i \sim p(y|x_i), \tag{3.2}$$

generates the data according to the model $p(x, y) = p(y|x)p(x)$. It is important to stress that this can describe a very general situation: x variable can be multivariate (and even be time dependent), and y can describe any other process. We will see, though, that in *Bayesian modelling* (I use it simultaneously with probabilistic modelling), x generally denotes the *latent (hidden) states* or *parameters* of a model (or both) . The variable y typically denotes the *observed data*. So seeing the model (3.1) as a generative model, simulating from it can be seen as a way of generating synthetic data[1]. Before we go into the interpretation of the variables in the model, let us first review some probability basics.

3.2 BASIC PROBABILITY THEORY

In this section, we review basic probability theory that will be required for the rest of the course. We will especially focus on Bayesian updating and conditional independence which are the key concepts for probabilistic inference methods. Let us start with a few definitions.

[1] This is a big deal in industry. Search for example for *synthetic data startups*.

3.2.1 PROBABILITY DEFINITIONS

Let X be a random variable that takes values on set $\mathcal{X}$. We do not limit ourselves to any specific set $\mathcal{X}$ for now. We call this random variable a *discrete* random variable if $\mathcal{X}$ is a discrete set, i.e., a set with a finite or countable number of elements. We call it a *continuous* random variable if $\mathcal{X}$ is a set that is not countable. We will next define associated probability distributions.

Let us start from the case where a random variable is discrete. This means the set $\mathcal{X}$ is either finite or countable. A simple example is

$$\mathsf{X} = \{1, 2, 3, 4, 5, 6\},$$

which could denote, for example, the possible outcomes of a die roll. Now we define the probability mass function.

Definition 3.1 (Probability Mass Functions)**.** *When a random variable is discrete, the probability mass function can be defined as*

$$p(x) = \mathbb{P}(X = x),$$

where $x \in \mathcal{X}$. We call $p(x)$ the probability mass function of X.

We note that in one dimensional case, the probability mass function is typically represented as a vector of probabilities when it comes to computations. Consider the following example.

Example 3.1. Assume that $\mathsf{X} = \{1, 2, 3, 4\}$ and

$$p(x) = \begin{cases} 0.1 & \text{if } x = 1, \\ 0.2 & \text{if } x = 2, \\ 0.3 & \text{if } x = 3, \\ 0.4 & \text{if } x = 4. \end{cases}$$

We can see this as a table of probabilities

X	$\mathbb{P}(X = x)$
1	0.1
2	0.2
3	0.3
4	0.4

What we did during the inversion method was to represent the probability mass function as a vector of probabilities

$$\mathbf{p} = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{bmatrix}.$$

indexed by discrete variables. Of course, one can also define a *dictionary* (Python data type) in order to have more complicated states for the random variable.

Next we define the probability density function in the case of continuous random variables.

Definition 3.2 (Measure and density)**.** *Assume $\mathcal{X} \subset \mathbb{R}$ and $X \in \mathcal{X}$ (for simplicity). Given the random variable X, we define the measure of X as*

$$\mathbb{P}(x_1 \leq X \leq x_2) = \mathbb{P}(X \in (x_1, x_2)).$$

The reason $\mathbb{P}$ called a measure is that it measures the probability of sets. We have then the probability density function which has the following relationship with the probability measure

$$\mathbb{P}(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)\,\mathrm{d}x.$$

We call $p(x)$ the probability density function of X.

3.2.2 JOINT AND CONDITIONAL PROBABILITY

We now define the joint probability distribution of two random variables X and Y. We will also focus on the discrete case first, and then move to the continuous case.

Definition 3.3 (Discrete Joint Probability Mass Function)**.** *Let X and Y be random variables and $\mathcal{X}$ and $\mathcal{Y}$ be the sets they live on. $\mathcal{X}$ and $\mathcal{Y}$ are at most countable sets. The joint probability mass function of X and Y is*

$$p(x, y) = \mathbb{P}(X = x, Y = y).$$

We call $p(x, y)$ the joint probability mass function of X and Y.

Example 3.2. Similar to the one dimensional case, we can now see the joint pmf $p(x, y)$ as a table of probabilities

	$Y=0$	$Y=1$	$Y=2$	$Y=3$	$p_X(x)$
$X=0$	1/6	1/6	0	0	2/6
$X=1$	1/6	0	1/6	0	2/6
$X=2$	0	0	1/6	0	1/6
$X=3$	0	0	0	1/6	1/6
$p_Y(y)$	2/6	1/6	2/6	1/6	1

Of course, on computer we can represent this as a matrix of probabilities

$$\mathbf{P} = \begin{bmatrix} 1/6 & 1/6 & 0 & 0 \\ 1/6 & 0 & 1/6 & 0 \\ 0 & 0 & 1/6 & 0 \\ 0 & 0 & 0 & 1/6 \end{bmatrix}.$$

This allows us to perform simple computations for marginalisation simply as sums of rows or columns. This is going to be a crucial tool when we study Markov models.

Let us finally define the probability density function $p(x, y)$ for continuous variables.

Definition 3.4 (Continuous Joint Probability Density Function)**.** *Let X and Y be random variables and $\mathcal{X}$ and $\mathcal{Y}$ be their ranges. We denote the joint probability measure as $\mathbb{P}(X \in A, Y \in B)$ and the density function $p(x, y)$ satisfies*

$$\mathbb{P}(X \in A, Y \in B) = \int_A \int_B p(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

As we have seen in previous chapter, the marginal probability densities from the joint density can be computed as

$$p(x) = \int_{\mathcal{Y}} p(x, y) \, \mathrm{d}y, \quad \text{and} \quad p(y) = \int_{\mathcal{X}} p(x, y) \, \mathrm{d}x.$$

3.2.3 CONDITIONAL PROBABILITY

We now define the conditional probability of a random variable X given another random variable Y. As usual, we will first focus on the discrete case, and then move to the continuous case.

Definition 3.5 (Discrete Conditional Probability Mass Function)**.** *Let X and Y be random variables and $\mathcal{X}$ and $\mathcal{Y}$ be their ranges. The conditional probability mass function of X given Y is*

$$p(x \mid y) = \mathbb{P}(X = x \mid Y = y).$$

We call $p(x \mid y)$ the conditional probability mass function of X given Y.

Example 3.3. We can compute the conditional probability mass function from the table of probabilities of $p(x, y)$. Consider the following joint probability mass function

	$X = 0$	$X = 1$	$X = 2$	$X = 3$	$p_Y(y)$
$Y = 0$	1/6	1/6	0	0	2/6
$Y = 1$	1/6	0	1/6	0	2/6
$Y = 2$	0	0	1/6	0	1/6
$Y = 3$	0	0	0	1/6	1/6
$p_X(x)$	2/6	1/6	2/6	1/6	1

Let us say we would like to compute $\mathbb{P}(Y = i | X = 2)$ for $i = 0, 1, 2, 3$. We can do this by simply dividing the joint probability mass function by the marginal probability mass function of X. Consider the following table

$p(x, y)$	$X = 0$	$X = 1$	$X = 2$	$X = 3$	$p_Y(y)$
$Y = 0$	1/6	1/6	**0**	0	2/6
$Y = 1$	1/6	0	**1/6**	0	2/6
$Y = 2$	0	0	**1/6**	0	1/6
$Y = 3$	0	0	**0**	1/6	1/6
$p_X(x)$	2/6	1/6	**2/6**	1/6	1

where the red entries are the joint probabilities of Y given $X = 2$. We can write the conditional probabilities as

$$\begin{aligned}
\mathbb{P}(Y = 0|X = 2) &= \frac{\mathbb{P}(Y = 0, X = 2)}{\mathbb{P}(X = 2)} = \frac{0}{2/6} = 0, \\
\mathbb{P}(Y = 1|X = 2) &= \frac{\mathbb{P}(Y = 1, X = 2)}{\mathbb{P}(X = 2)} = \frac{1/6}{2/6} = 1/2, \\
\mathbb{P}(Y = 2|X = 2) &= \frac{\mathbb{P}(Y = 2, X = 2)}{\mathbb{P}(X = 2)} = \frac{1/6}{2/6} = 1/2, \\
\mathbb{P}(Y = 3|X = 2) &= \frac{\mathbb{P}(Y = 3, X = 2)}{\mathbb{P}(X = 2)} = \frac{0}{2/6} = 0.
\end{aligned}$$

As we can see that the conditional probability can also be represented as a vector

$$\mathbf{p} = [0, 1/2, 1/2, 0].$$

for implementation purposes.

One can compute conditional probability tables from the joint probability table.

Example 3.4. We can derive the conditional probability table from the joint probability table given above. For example, the conditional probability mass function $p(y|x)$ is given below.

$p(y\|x)$	$X = 0$	$X = 1$	$X = 2$	$X = 3$
$Y = 0$	1/2	1	0	0
$Y = 1$	1/2	0	1/2	0
$Y = 2$	0	0	1/2	0
$Y = 3$	0	0	0	1

Similarly, we can compute $p(x|y)$ as

$p(x\|y)$	$X = 0$	$X = 1$	$X = 2$	$X = 3$
$Y = 0$	1/2	1/2	0	0
$Y = 1$	1/2	0	1/2	0
$Y = 2$	0	0	1	0
$Y = 3$	0	0	0	1

We next define the continuous conditional density given $p(x, y)$.

Definition 3.6 (Continuous Conditional Probability Density Function)**.** *Let X and Y be random variables and $\mathcal{X}$ and $\mathcal{Y}$ be their ranges. The conditional probability density function of X given Y is*

$$p(y|x) = \frac{p(x, y)}{p(x)},$$

where we call $p(x \mid y)$ the conditional probability density function of X given Y. Similarly, we have

$$p(x|y) = \frac{p(x,y)}{p(y)}.$$

We call $p(x \mid y)$ the conditional probability density function of X given Y.

3.3 THE BAYES RULE AND ITS USES

In this section, we will discuss the Bayes rule in depth and its uses. The Bayesian formula is at the heart of many probabilistic modelling approaches. We start with the definition of the Bayes rule.

Definition 3.7 (Bayes Theorem)**.** *Let X and Y be random variables with associated probability density functions $p(x)$ and $p(y)$, respectively. The Bayes rule is given by*

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}. \tag{3.3}$$

Note that the formula holds for continuous random variables as well as discrete random variables. Its importance comes from the fact that it provides us a natural way to incorporate or synthesise data into a probabilistic model. In this interpretation, we have three key concepts.

- **Prior:** In the formula (3.3), $p(x)$ is called the prior probability of X. Here X can be interpreted as a parameter of $p(y|x)$ or a hidden (unobserved) variable. The probability distribution $p(x)$ encodes our prior knowledge about this variable we cannot observe directly. This could be simple constraints, a distribution dictated by a real application (e.g. a physical variable can be only positive). In time series applications, $p(x)$ can be the distribution over an entire time series, it can even encode physical laws.
- **Likelihood:** $p(y|x)$ is called the likelihood of Y given X. This is the probability model of the process of *observation* – in other words, it describes how the underlying parameter or hidden variable is observed. For example, if Y is the number of observed cases of a disease in a population, then $p(y|x)$ is the probability of observing y cases given that the true number of cases is x.
- **Posterior:** $p(x|y)$ is called the posterior distribution of X given $Y = y$. This is the *updated* probability distribution after we see y observation and updated our prior knowledge $p(x)$ into $p(x|y)$.

We will see a number of examples where these quantities make sense.

Remark 3.1. Note the difference between *simulation* and *inference*. We can write down our *model* (sometimes we will call the forward model) $p(x)$ and $p(y|x)$ to describe the *data generation* process and can generate toy (synthetic) data with it as we have seen. But the essential goal of Bayes rule (also called Bayesian or probabilistic inference) is to *infer* the posterior distribution conditioned on already *observed* data. In other words, we can use a probabilistic model for two purposes:

- *Simulation:* We can generate synthetic data with a probabilistic model.
- *Inference:* We can infer the posterior distribution (implied by the model structure we impose) of a parameter or hidden variable given observed data.

Example 3.5. Let us see the Bayes' rule on a discrete example. Suppose we have two fair dice, each with six faces. Define the outcome of the first die as X_1 and the outcome of the second die as X_2. We can then describe their joint probability table as

$p(x_1, x_2)$	$X_1 = 1$	$X_1 = 2$	$X_1 = 3$	$X_1 = 4$	$X_1 = 5$	$X_1 = 6$
$X_2 = 1$	1/36	1/36	1/36	1/36	1/36	1/36
$X_2 = 2$	1/36	1/36	1/36	1/36	1/36	1/36
$X_2 = 3$	1/36	1/36	1/36	1/36	1/36	1/36
$X_2 = 4$	1/36	1/36	1/36	1/36	1/36	1/36
$X_2 = 5$	1/36	1/36	1/36	1/36	1/36	1/36
$X_2 = 6$	1/36	1/36	1/36	1/36	1/36	1/36

i.e., each combination is equally probable. Note that this is also the table of $p(x_1)p(x_2)$ due to independence. Suppose that we can only observe the sum of the two dice, $Y = X_1 + X_2$. This would result in a likelihood

$$p(y|x_1, x_2) = \begin{cases} 1 & \text{if } y = x_1 + x_2, \\ 0 & \text{otherwise.} \end{cases}$$

We can also denote this as an indicator function, i.e., let $\mathbf{1}(y = x_1 + x_2)$ be the indicator function of the event $y = x_1 + x_2$, then we have $p(y|x_1, x_2) = \mathbf{1}(y = x_1 + x_2)$. Suppose now we observe $Y = 9$ and would like to infer the posterior distribution of X_1 and X_2 given $Y = 9$. We can use the Bayes rule to write

$$\begin{aligned} p(x_1, x_2|y = 9) &= \frac{p(y = 9|x_1, x_2)p(x_1, x_2)}{p(y = 9)}, \\ &= \frac{p(y = 9|x_1, x_2)p(x_1)p(x_2)}{p(y = 9)}. \end{aligned}$$

Let us first write out $p(y = 9|x_1, x_2)$ as a table

$p(y = 9\|x_1, x_2)$	$X_1 = 1$	$X_1 = 2$	$X_1 = 3$	$X_1 = 4$	$X_1 = 5$	$X_1 = 6$
$X_2 = 1$	0	0	0	0	0	0
$X_2 = 2$	0	0	0	0	0	0
$X_2 = 3$	0	0	0	0	0	1
$X_2 = 4$	0	0	0	0	1	0
$X_2 = 5$	0	0	0	1	0	0
$X_2 = 6$	0	0	1	0	0	0

This is just the likelihood. In order to get the full joint (numerator of the Bayes theorem), we need to multiply the likelihood with the joint prior $p(x_1, x_2) = p(x_1)p(x_2)$. Multiplying this table with the joint probability table of X_1 and X_2 gives

$p(y=9\|x_1,x_2)p(x_1)p(x_2)$	$X_1=1$	$X_1=2$	$X_1=3$	$X_1=4$	$X_1=5$	$X_1=6$
$X_2=1$	0	0	0	0	0	0
$X_2=2$	0	0	0	0	0	0
$X_2=3$	0	0	0	0	0	1/36
$X_2=4$	0	0	0	0	1/36	0
$X_2=5$	0	0	0	1/36	0	0
$X_2=6$	0	0	1/36	0	0	0

This is just the numerator in the Bayes theorem, we now need to compute the probability $p(y=9)$ in order to finally arrive at the posterior distribution. We can compute this as

$$\begin{aligned}
p(y=9) &= \sum_{x_1,x_2} p(y=9|x_1,x_2)p(x_1)p(x_2) \\
&= \sum_{x_1,x_2} \mathbf{1}(y=x_1+x_2)p(x_1)p(x_2) \\
&= \sum_{x_1,x_2} \mathbf{1}(y=x_1+x_2) \times \frac{1}{6} \times \frac{1}{6} \\
&= \frac{1}{36} \sum_{x_1,x_2} \mathbf{1}(y=x_1+x_2), \\
&= \frac{1}{36} \times 4 \\
&= \frac{1}{9}.
\end{aligned}$$

Now we are ready to normalise $p(y=9|x_1,x_2)p(x_1)p(x_2)$ to obtain the posterior distribution as a table

$p(x_1,x_2\|y=9)$	$X_1=1$	$X_1=2$	$X_1=3$	$X_1=4$	$X_1=5$	$X_1=6$
$X_2=1$	0	0	0	0	0	0
$X_2=2$	0	0	0	0	0	0
$X_2=3$	0	0	0	0	0	1/4
$X_2=4$	0	0	0	0	1/4	0
$X_2=5$	0	0	0	1/4	0	0
$X_2=6$	0	0	1/4	0	0	0

Let us next see a continuous example adapted from Murphy (2007).

Example 3.6. Let

$$\begin{aligned}
p(x) &= \mathcal{N}(x;\mu_0,\sigma_0^2), \\
p(y|x) &= \mathcal{N}(y;x,\sigma^2),
\end{aligned}$$

where μ_0 and σ_0^2 are the prior mean and variance, respectively, and σ^2 is the variance of the likelihood. We have seen this example before, where we computed the marginal likelihood $p(y)$. In this example, we will instead derive the posterior distribution $p(x|y)$. Now let us

write

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

In order to derive the posterior, we first derive $p(y|x)p(x)$ as

$$\begin{aligned} p(y|x)p(x) &= \mathcal{N}(y; x, \sigma^2)\mathcal{N}(x; \mu_0, \sigma_0^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2\sigma_0^2}} \exp\left(-\frac{(y-x)^2}{2\sigma^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\right). \end{aligned}$$

We know that

$$\begin{aligned} p(x|y) &\propto p(y|x)p(x) \\ &\propto \exp\left(-\frac{(y-x)^2}{2\sigma^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\right). \end{aligned}$$

We can now use the help of the fact that the product of two Gaussians is a Gaussian. We can parameterise the posterior as

$$p(x|y) = \mathcal{N}(x; \mu_p, \sigma_p^2),$$

where μ_p and σ_p^2 are the posterior mean and variance, respectively. This means, we need to match

$$\exp\left(-\frac{(y-x)^2}{2\sigma^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\right) = \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right).$$

We can solve for μ_p and σ_p^2 as (exercise)

$$\begin{aligned} \mu_p &= \frac{\sigma^2\mu_0 + \sigma_0^2 y}{\sigma^2 + \sigma_0^2}, \\ \sigma_p^2 &= \frac{\sigma^2\sigma_0^2}{\sigma^2 + \sigma_0^2}. \end{aligned}$$

This gives us our Gaussian posterior. See Fig 3.1 for an illustration.

This is an example of a *conjugate prior*, where the posterior distribution is of the same form as the prior. In the solved examples section, we will see more examples of this. As you have seen, the derivation of the posterior took some work. As opposed to this conjugate case, in the general case, we will not be able to derive the posterior. Let us see one example now how we can avoid computing the normalised posterior but still sample from it.

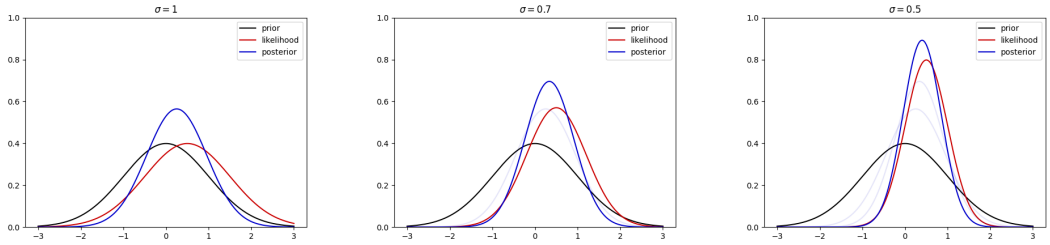

Figure 3.1: Posterior distribution of x given $\sigma = 1, \sigma = 0.7$ and $\sigma = 0.5$ respectively. One can see that as we shrink the likelihood variance, the posterior distribution becomes more peaked towards the observation $y = 0.5$. Old posteriors are also plotted in the second and third figure for comparison (in transparent blue).

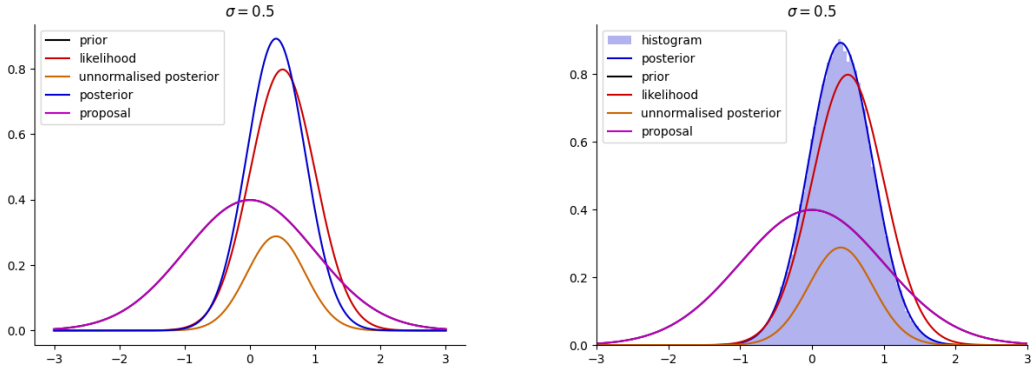

Figure 3.2: On the left, we plot all distributions of interest: prior, likelihood (with $y = 0.5$ with respect to x), the posterior, and the unnormalised posterior, and the proposal. Note that, the proposal should only cover the unnormalised posterior, even if the normalising constant is less than one. On the left, we plot the samples vs. the same quantities. One can see that we exactly sampled from the correct posterior.

Example 3.7. Let us sample from the posterior of Gaussian likelihood and prior by implementing the rejection sampling. Assume that we have a prior distribution $p(x) = \mathcal{N}(x; \mu_0, \sigma_0^2)$ and a likelihood distribution $p(y|x) = \mathcal{N}(y; x, \sigma^2)$. We want to sample the posterior distribution $p(x|y)$. We know the posterior is given by

$$p(x|y) \propto p(y|x)p(x) = \mathcal{N}(x; \mu_0, \sigma_0^2)\mathcal{N}(y; x, \sigma^2).$$

Recall that we would like to sample from the posterior $p(x|y)$ without necessarily computing the Bayes rule. We can pose this problem as a *rejection sampling* problem. We would like to sample from the posterior distribution conditioned on y. In our case, the unnormalised posterior is given by

$$\bar{p}(x|y) = p(y|x)p(x)$$

Note that we *evaluate* the likelihood at the observation y and hence it becomes a function of x. Below, for clarity, we will use the r.h.s. of above equation in acceptance rate, instead of $\bar{p}(x)$ as we usually did before. For this example, we also set $\mu_0 = 0$, $\sigma_0 = 1$, and $\sigma = 0.5$. Next, we need to design a proposal distribution $q(x)$. This could be tricky as we do not know the posterior. For now, we can choose another simple Gaussian (we could also optimise this):

$$q(x) = \mathcal{N}(x; \mu_q, \sigma_q^2).$$

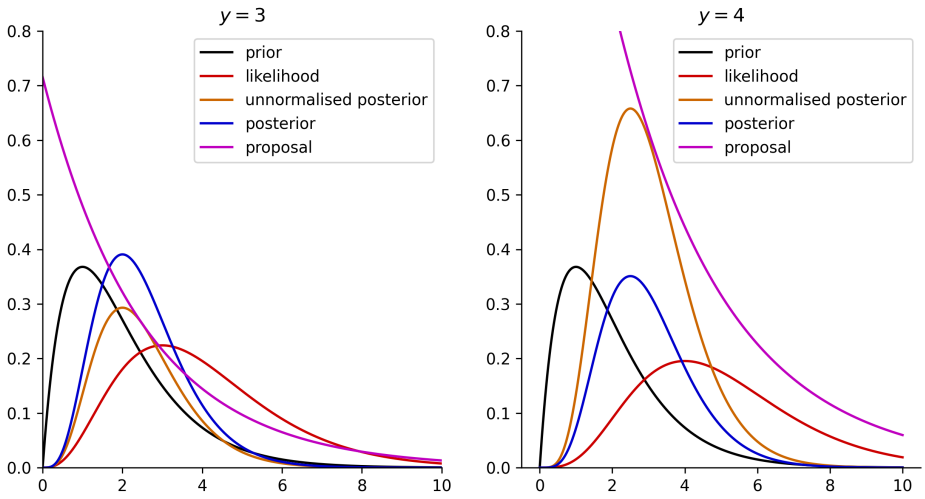

Figure 3.3: Illustration of the prior, posterior, likelihood, and the proposal distribution.

Let us choose $\mu_q = 0$ and $\sigma_q = 1$ (note again that this is the standard deviation!) and $M = 1$. An illustration of this is shown in Fig 3.2. We can now sample from the posterior

- Sample $X' \sim q(x)$
- Sample $U \sim \text{Unif}(0, 1)$
- If $U \leq \frac{p(y|X')p(X')}{Mq(X')}$, accept X'. Otherwise, reject X' and go back to step 1.

We can see the results of this procedure from Fig 3.2. As seen from the figure, we exactly sample from the posterior $p(x|y = 0.5)$ without ever computing the correct posterior. We have also plotted the correct posterior in the figure for comparison.

Let us see another example.

Example 3.8. Assume that we have a Poisson observation model:

$$p(y|x) = \text{Pois}(y; x) = \frac{x^y e^{-x}}{y!},$$

and a Gamma prior:

$$p(x) = \text{Gamma}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

We want to sample from the posterior distribution $p(x|y)$. We know the posterior is given by

$$\begin{aligned} p(x|y) &\propto p(y|x)p(x) \\ &= \text{Pois}(y; x)\text{Gamma}(x; \alpha, \beta), \\ &\propto x^{\alpha-1+y} e^{-\beta x - x}, \end{aligned}$$

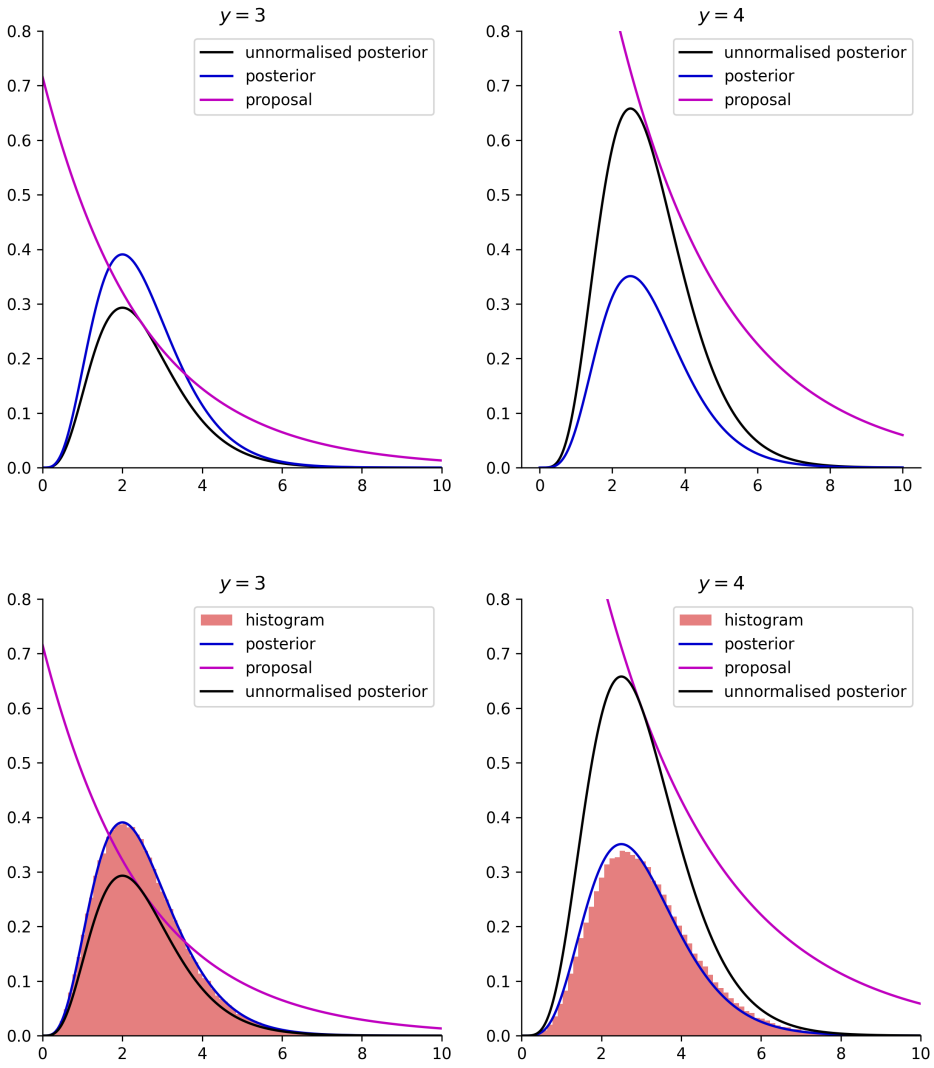

Figure 3.4: Histogram of the samples drawn using rejection sampling.

where we ignored all the normalising constants. We can see that the posterior is also a Gamma density:

$$p(x|y) = \text{Gamma}(x; \alpha + y, \beta + 1).$$

Let us sample from this posterior with rejection sampling as we did before for the Gaussian.

Example 3.9. Assume that we have a Gamma prior:

$$p(x) = \text{Gamma}(x; \alpha, 1) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x},$$

with $\alpha > 0$. Next, we define our Poisson observation model as before,

$$p(y|x) = \text{Pois}(y; x) = \frac{x^y e^{-x}}{y!}.$$

Poisson is a discrete distribution usually used to model counts with mean x. We know that the posterior is proportional to

$$\begin{aligned} p(x|y) &\propto p(y|x)p(x) = \text{Pois}(y; x)\text{Gamma}(x; \alpha, 1), \\ &\propto x^{\alpha-1+y} e^{-2x}. \end{aligned}$$

In short, we will choose this as our unnormalised posterior

$$\bar{p}(x|y) = x^{\alpha-1+y} e^{-2x}.$$

Now we will design our proposal distribution. We choose the proposal as an exponential distribution:

$$q_\lambda(x) = \text{Exp}(x; \lambda) = \lambda e^{-\lambda x}.$$

Now we derive the acceptance probability. As usual, we need to first find

$$M_\lambda = \sup_x \frac{\bar{p}(x|y)}{q_\lambda(x)}.$$

First we need to optimise the ratio:

$$\begin{aligned} \frac{\bar{p}(x|y)}{q_\lambda(x)} &= \frac{x^{\alpha-1+y} e^{-2x}}{\lambda e^{-\lambda x}} \\ &= \frac{x^{\alpha-1+y} e^{-(2-\lambda)x}}{\lambda}. \end{aligned}$$

Aiming at optimising this w.r.t. x, we first compute its log:

$$\begin{aligned} \log \frac{\bar{p}(x|y)}{q_\lambda(x)} &= \log x^{\alpha-1+y} + \log e^{-(2-\lambda)x} - \log \lambda \\ &= (\alpha - 1 + y) \log x - (2 - \lambda)x - \log \lambda. \end{aligned}$$

We now take the derivative of this w.r.t. x:

$$\frac{\mathrm{d}}{\mathrm{d}x} \left[(\alpha - 1 + y) \log x - (2 - \lambda)x - \log \lambda\right] = \frac{\alpha - 1 + y}{x} - (2 - \lambda),$$

and set it to zero:

$$\frac{\alpha - 1 + y}{x} - (2 - \lambda) = 0.$$

This gives us the maximiser

$$x^\star = \frac{\alpha - 1 + y}{2 - \lambda}.$$

We can now compute M_λ:

$$\begin{aligned} M_\lambda &= \frac{\bar{p}(x^\star|y)}{q_\lambda(x^\star)} \\ &= \frac{x^{\star\alpha-1+y}e^{-(2-\lambda)x^\star}}{\lambda} \\ &= \frac{1}{\lambda}\left(\frac{\alpha-1+y}{2-\lambda}\right)^{\alpha-1+y} e^{-(2-\lambda)\left(\frac{\alpha-1+y}{2-\lambda}\right)} \\ &= \frac{1}{\lambda}\left(\frac{\alpha-1+y}{2-\lambda}\right)^{\alpha-1+y} e^{-(\alpha-1+y)}. \end{aligned}$$

We can now optimise this further to choose our optimal proposal. We will first compute the $\log$ of M_λ:

$$\begin{aligned} \log M_\lambda &= \log\frac{1}{\lambda} + (\alpha-1+y)\log\left(\frac{\alpha-1+y}{2-\lambda}\right) - (\alpha-1+y) \\ &= -\log\lambda + (\alpha-1+y)\log\left(\frac{\alpha-1+y}{2-\lambda}\right) - (\alpha-1+y). \end{aligned}$$

Taking the derivative of this w.r.t. λ, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\log M_\lambda = -\frac{1}{\lambda} + \frac{(\alpha-1+y)}{2-\lambda}$$

Setting this to zero, we obtain

$$\frac{1}{\lambda} = \frac{(\alpha-1+y)}{2-\lambda},$$

which implies that

$$\lambda^\star = \frac{2}{\alpha+y}.$$

Therefore, we can choose our optimal proposal in terms of α and y depends on the observed sample. See Fig . 3.4 for the histogram of the samples drawn using rejection sampling.