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.

3.4 CONDITIONAL INDEPENDENCE

The step forward from the simple Bayes rule to modelling complex dependencies and interactions is to understand the notion of conditional independence. Simply put, conditional independence is a notion of independence of two random variables *conditioned* on a third random variable. Of course, this can be extended to arbitrary number of variables, defining a full probabilistic model. It is important to note that these models *everywhere* in science and engineering.

Let us first define the notion of conditional independence.

Definition 3.8. *Let X, Y and Z be random variables. We say that X and Y are conditionally*

independent given Z if

$$p(x, y|z) = p(x|z)p(y|z).$$

This definition is of course the same as plain independence, just written in terms of conditional probabilities. Note that, in general, X and Y are not independent if we do not condition on Z. We note the important corollary.

Corollary 3.1. *If X and Y are conditionally independent given Z, then*

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

and

$$p(y|x, z) = p(y|z).$$

Proof. See Exercise 4.2 solution. □

We can now describe the notion of conditional independence in terms of joint distributions.

Proposition 3.1. *Let X, Y and Z be random variables. If X and Y are conditionally independent given Z, then*

$$p(x, y, z) = p(x|z)p(y|z)p(z).$$

Proof. Recall that we have described the chain rule for conditional probabilities in Sec. 2.3.3

$$p(x_1, \ldots, x_n) = p(x_n|x_{n-1}, \ldots, x_1)p(x_{n-1}|x_{n-2}, \ldots, x_1) \cdots p(x_2|x_1)p(x_1).$$

This relationship is as important as in inference as in simulation. We can now use this to show that

$$\begin{aligned} p(x, y, z) &= p(x|y, z)p(y|z)p(z) \\ &= p(x|z)p(y|z)p(z), \end{aligned}$$

where the last line follows from Corollary 3.1. □

This idea can be extended to arbitrary number of variables. This kind of factorisations are at the core of probabilistic modelling. In other words, a probabilistic modeller (scientist) poses a set of conditional independence assumptions which then allows them to factorise the joint distribution into a product of conditional distributions. From then on, the modeller can use the conditional distributions to compute any desired marginal or conditional distributions. This is the essence of probabilistic modelling.

3.4.1 BAYES RULE FOR CONDITIONALLY INDEPENDENT OBSERVATIONS

So far, we have seen an example of prior to posterior update for a single observation in Sec. 3.3 and the definition of conditional independence. We can now combine these two ideas to obtain the Bayes update for conditionally independent observations. This is a standard use case for conditional independence: Typically, given an unobserved variable x, we can obtain multiple measurements related to a single latent variable x.

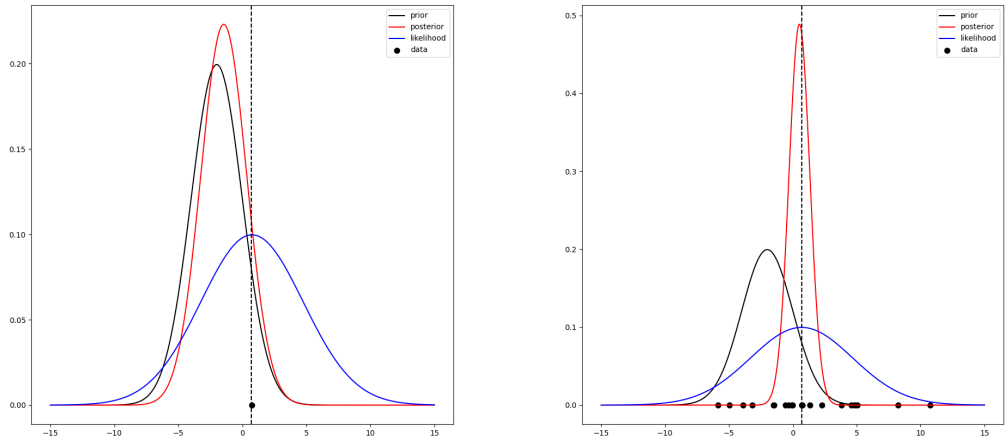

Figure 3.5: Bayes update for conditionally independent observations.

Let us define the general Bayes update for this case. Assume that we have observed $y_1, \ldots, y_n \sim p(y|x)$ (these can be thought of as conditionally i.i.d samples from the likelihood)[2]. Given a prior of x, denoted $p(x)$, we want to compute the posterior $p(x|y_1, \ldots, y_n)$. We know that the posterior is given by

$$p(x|y_{1:n}) = \frac{p(y_{1:n}|x)p(x)}{p(y_{1:n})}. \tag{3.4}$$

Under the conditional independence assumption of observations, we can just use Definition 3.8 to arrive at

$$p(y_{1:n}|x) = \prod_{i=1}^{n} p(y_i|x).$$

Plugging this in back to the Bayes update (3.4), we can see that the posterior is proportional to the product

$$\begin{aligned} p(x|y_1, \ldots, y_n) &\propto p(y_1, \ldots, y_n|x)p(x) \\ &= \prod_{i=1}^{n} p(y_i|x)p(x), \end{aligned}$$

Again, in many occasions, we will not be able to compute the normalising constant. However, we can still sample from the posterior. In this particular example, let us continue with the Gaussian prior and likelihood. In this case, we can exactly compute the posterior too.

Example 3.10 (Gaussian Bayes update for conditionally independent observations)**.** As usual, in the Gaussian case, we can compute the posterior distribution even given multiple observations. Let us assume the following probabilistic model

$$\begin{aligned} X &\sim \mathcal{N}(x; \mu_0, \sigma_0^2) \\ Y_i|X = x &\sim \mathcal{N}(y_i; x, \sigma^2), \quad i = 1, \ldots, n. \end{aligned}$$

[2]We define the following notation. Let $y_1, \ldots, y_n$ be n observations. We collectively denote these variables as $y_{1:n} := (y_1, \ldots, y_n)$. This will be also used in the following sections.

Here each observation is assumed to be conditionally independent given x. Note that this model is very different than the one where we simulated (X_i, Y_i) pairs in Example 2.20. The point in Example 2.20 was to simulate pairs exhibiting linear relationship, each (Y_i, X_i) was an independent draw from the joint distribution. Here, we assume that the observations are sampled conditioned on a *single* x – in essence, the sequence $y_1, \dots, y_n$ are dependent. They are only conditionally independent given x.

Having observed $y_1, \dots, y_n$, we would like to compute the posterior $p(x|y_1, \dots, y_n)$. Let us first compute the likelihood

$$\begin{aligned} p(y_{1:n}|x) &= \prod_{i=1}^{n} p(y_i|x) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x)^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x)^2\right). \end{aligned}$$

Using the same derivations (term matching) as in Example 3.6, we can compute the posterior

$$\begin{aligned} p(x|y_{1:n}) &= \frac{p(y_{1:n}|x)p(x)}{p(y_{1:n})} \\ &= \frac{p(y_{1:n}|x)p(x)}{\int p(y_{1:n}|x)p(x)dx} \end{aligned}$$

where $p(x|y_{1:n}) = \mathcal{N}(x; \mu_p, \sigma_p^2)$, with (Murphy, 2007)

$$\begin{aligned} \mu_p &= \frac{\sigma_0^2 \sum_{i=1}^{n} y_i + \sigma^2 \mu_0}{\sigma_0^2 n + \sigma^2} \\ \sigma_p^2 &= \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 n + \sigma^2}. \end{aligned}$$

The posterior with conditioned data can be seen from Fig. 3.5.

3.4.2 CONDITIONAL BAYES RULE

It is important to realise that the Bayes rule can be used *conditionally*. Consider three variables X, Y, Z without specifying any conditional independence assumptions. In this case, the Bayes rule for $p(x|y, z)$ can be written entirely on z (of course, this is true if we swap the variables and condition on x or y). We can write in this case the conditional Bayes rule.

Proposition 3.2. *Given X, Y, Z without any conditional independence assumptions, the conditional Bayes rule is*

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

Proof. See the solution of Exercise 4.1. □

This is of course true if we write the same rule for x or y conditioned.

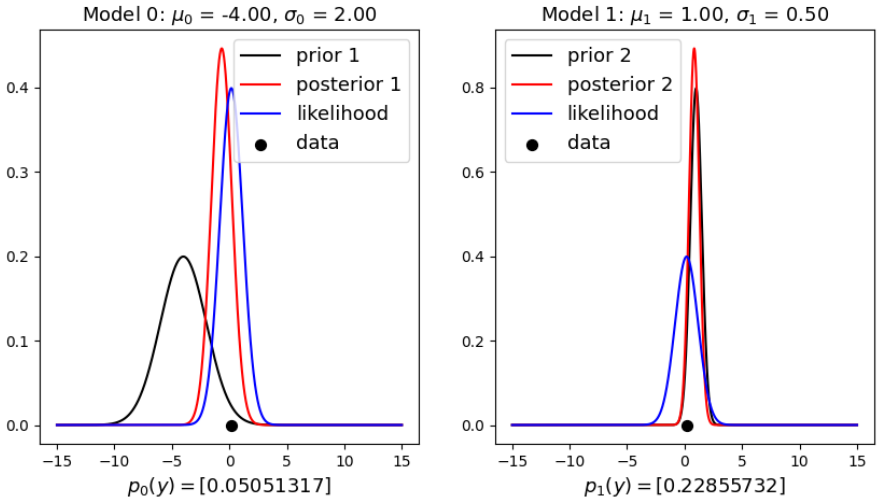

Figure 3.6: Marginal likelihood for model comparison. For observed data, we can compute the marginal likelihood for each model. The model with the highest marginal likelihood is the best model for the observed data.

3.5 MARGINAL LIKELIHOOD

The notion of marginal likelihood is left unexplored so far and we will now investigate it. We can go back to the Bayes theorem and write

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

In this formula, we have been discussing the posterior $p(x|y)$, the prior $p(x)$, and the likelihood $p(y|x)$ in past sections. However, the normalising constant, which we assumed to be intractable, is also of interest. This quantity, $p(y)$, is called the marginal likelihood and it is given by

$$p(y) = \int p(y|x)p(x)\mathrm{d}x.$$

For fixed y, the interpretation of this term is that it gives us the *probability of data* y under the model[3]. For more complicated models (where x can be multiple variables or multiple other distributions may exist), the quantity $p(y)$ becomes crucial to determine the quality of the model for the observed data. While itself does not mean much, it gives us a comparative measure to compare different models. We will discuss this with an example.

Example 3.11 (Marginal likelihood for two Gaussian models)**.** Consider two different models:

$$X \sim p_0(x) = \mathcal{N}(x; \mu_0, \sigma_0^2)$$
$$Y|X = x \sim \mathcal{N}(y; x, \sigma_y^2)$$

[3]Aside from its usual interpretation as the normalising constant.

and

$$X \sim p_1(x) = \mathcal{N}(x; \mu_1, \sigma_1^2)$$
$$Y|X = x \sim \mathcal{N}(y; x, \sigma_y^2)$$

Consider observing y (a single data point). Which model is more likely? Recall that, for these models, we have computed $p(y)$ analytically before. We can compute for both models:

$$\begin{aligned} p_0(y) &= \int p(y|x)p_0(x)\mathrm{d}x \\ &= \int \mathcal{N}(y; x, \sigma_y^2)\mathcal{N}(x; \mu_0, \sigma_0^2)\mathrm{d}x \\ &= \mathcal{N}(y; \mu_0, \sigma_0^2 + \sigma_y^2) \end{aligned}$$

and

$$\begin{aligned} p_1(y) &= \int p(y|x)p_1(x)\mathrm{d}x \\ &= \int \mathcal{N}(y; x, \sigma_y^2)\mathcal{N}(x; \mu_1, \sigma_1^2)\mathrm{d}x \\ &= \mathcal{N}(y; \mu_1, \sigma_1^2 + \sigma_y^2) \end{aligned}$$

We will say Model 1 is better than Model 0 if $p_1(y) > p_0(y)$ for fixed y. Let us choose that $\sigma = 1$, $\mu_0 = -4$, $\sigma_0 = 2$, and $\mu_1 = 1$, $\sigma_1 = 0.5$. The computed marginal likelihoods can be seen from Fig. 3.6. It can be seen that Model 1 is a much better fit to the data than Model 0.

3.6 CONCLUSION

In this section, we briefly discussed the Bayes rule and its application to probabilistic inference. This is a vast topic and we have only scratched the surface. If you are curious about the topic, Bishop (2006) is a good book to read. Some other very nices ones are Barber (2012) and Murphy (2022).

We will finish this chapter by discussing why rejection samplers as we introduced it would not be an appropriate candidate for sampling in more complicated models we discussed in this chapter.

Example 3.12 (Inadequacy of Rejection Sampling)**.** Given all these derivations, it is natural to ask whether we can use rejection samplers for Bayesian inference. Let us assume that we have $y_1, \ldots, y_n$ observed and our unnormalised posterior is given by

$$\bar{p}(x|y_{1:n}) = p(x) \prod_{i=1}^{n} p(y_i|x).$$

Let us assume that we have a proposal distribution $q(x)$ and assume that we have been lucky to identify some M such that

$$\bar{p}(x|y_{1:n}) \leq Mq(x).$$

We can now perform rejection sampling as follows:

1. Sample $X' \sim q(x)$
2. Sample $U \sim \text{Unif}(0, 1)$
3. If $U \leq \frac{\bar{p}(X'|y_{1:n})}{Mq(X')} = \frac{p(X')\prod_{i=1}^{n} p(y_i|X')}{Mq(X')}$ then accept X'
4. Otherwise reject X' and go back to step 1.

What could be an immediate problem as n grows? The multiplication $\prod_{i=1}^{n} p(y_i|X')$ would not be numerically stable. This would result in numerical underflow as the multiplication of small probabilities gets smaller and smaller. In order to mitigate this, one solution is to work with log-probabilities. This means that we can still perform rejection sampling (provided that $\bar{p}(x|y) \leq Mq(x)$) as follows:

1. Sample $X' \sim q(x)$
2. Sample $U \sim \text{Unif}(0, 1)$
3. Compute log-acceptance probability

$$\begin{aligned}\log a(X') &= \log \frac{\bar{p}(X'|y_{1:n})}{Mq(X')} = \log \frac{p(X')\prod_{i=1}^{n} p(y_i|X')}{Mq(X')}, \\ &= \log p(X') + \sum_{i=1}^{n} \log p(y_i|X') - \log M - \log q(X').\end{aligned}$$

4. If $\log U \leq \log a(X')$ then accept X'

However, this would also not often solve our issues as

- It is often impossible to find M and $q(x)$ such that $\bar{p}(x|y) \leq Mq(x)$.
- It is not easy to plot the unnormalised posterior $\bar{p}(x|y)$ (without log)
- Bounds found to log-unnormalised posterior can be very loose
  - Super low acceptance probability

This is also not the only failure mode of the rejection sampling. It is often the case that rejection sampling is very inefficient in high dimensions even if one manages to find a good proposal q. Consider the rejection sampling in 2D for sampling the circle within a square (See Lecture 1). The acceptance probability for this case:

$$a = \frac{\text{area of the circle}}{\text{area of the square}} = \frac{\pi}{4} \approx 0.78.$$

Next, consider the same sampler for the sphere and the cube (3D). The acceptance probability for this case:

$$a = \frac{\text{volume of the sphere}}{\text{volume of the cube}} = \frac{\pi}{6} \approx 0.52.$$

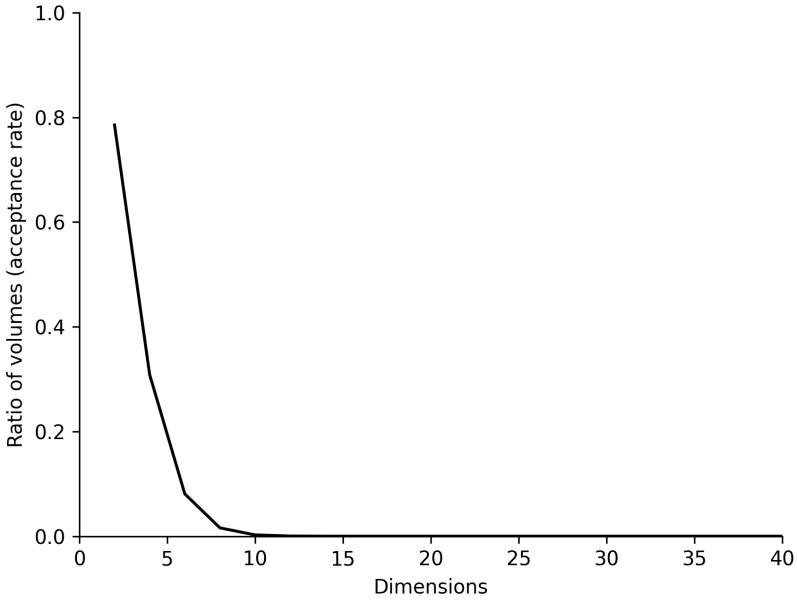

Figure 3.7: The curse of dimensionality for the sampling example for rejection sampling.

If we were doing this in d dimensions, the acceptance rate would be

$$a = \frac{\text{volume of the unit ball}}{\text{volume of the unit cube}}$$

However, this ratio goes to zero incredibly fast as d grows (see Fig. 3.7) In other words, rejection samplers have very poor acceptance rates in high dimensions. This will lead us to look at other sampling methods.