Lecture 5: Introduction to Probabilistic (Bayesian) Modelling and Inference

Deniz Akyildiz

MATH60047/70047 – Stochastic Simulation

October 24, 2022

Imperial College London

Announcements

Announcements

- Tomorrow, we are at HXLY 414, Maths Learning Centre. We will be coding exercises and some more things.
  - Please bring your laptops.

Announcements

- Tomorrow, we are at HXLY 414, Maths Learning Centre. We will be coding exercises and some more things.
  - Please bring your laptops.
  - Jupyter Notebooks (can you run it?)

Announcements

- Tomorrow, we are at HXLY 414, Maths Learning Centre. We will be coding exercises and some more things.
  - Please bring your laptops.
  - Jupyter Notebooks (can you run it?)
- Assignment is to be posted this Wednesday (26 October).
  - Due 9 Nov. 2022
  - 10 percent

We have seen so far

We have seen so far

- Inversion (requires F_X^{-1} to be evaluated)

We have seen so far

- Inversion (requires F_X^{-1} to be evaluated)
- Transformation (requires a good transformation)

We have seen so far

- Inversion (requires F_X^{-1} to be evaluated)
- Transformation (requires a good transformation)
- Rejection sampling (requires unnormalised/normalised density evaluation and a good proposal)

We have seen so far

- Inversion (requires F_X^{-1} to be evaluated)
- Transformation (requires a good transformation)
- Rejection sampling (requires unnormalised/normalised density evaluation and a good proposal)

While the first two conditions are far too stringent for many applications, the third one is not so bad.

We have seen so far

- Inversion (requires F_X^{-1} to be evaluated)
- Transformation (requires a good transformation)
- Rejection sampling (requires unnormalised/normalised density evaluation and a good proposal)

While the first two conditions are far too stringent for many applications, the third one is not so bad.

We will introduce today *probabilistic (Bayesian) inference* which is a very general framework for inference.

We have seen so far

- Inversion (requires F_X^{-1} to be evaluated)
- Transformation (requires a good transformation)
- Rejection sampling (requires unnormalised/normalised density evaluation and a good proposal)

While the first two conditions are far too stringent for many applications, the third one is not so bad.

We will introduce today *probabilistic (Bayesian) inference* which is a very general framework for inference.

We will see how to use rejection sampling for this purpose.

We have seen in past lectures that we could generate data from the model:

$$
\begin{aligned}
X_i &\sim p(x) \\
Y_i | X_i = x_i &\sim p(y|x).
\end{aligned}
$$

We have seen in past lectures that we could generate data from the model:

$$X_i \sim p(x)$$
$$Y_i | X_i = x_i \sim p(y|x).$$

However, in this course, our aim is not solely to simulate synthetic data.

We have seen in past lectures that we could generate data from the model:

$$X_i \sim p(x)$$
$$Y_i | X_i = x_i \sim p(y|x).$$

However, in this course, our aim is not solely to simulate synthetic data.

We want to infer hidden states or parameters *given* observed data.

The crucial tool for this purpose is the Bayes rule:

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

The crucial tool for this purpose is the Bayes rule:

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

This is the *fundamental* rule of probabilistic inference.

The crucial tool for this purpose is the Bayes rule:

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

This is the *fundamental* rule of probabilistic inference.

- x here can be a hidden variable (we can't observe) or a parameter.

The crucial tool for this purpose is the Bayes rule:

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

This is the *fundamental* rule of probabilistic inference.

- x here can be a hidden variable (we can't observe) or a parameter.
- y is the observed data.

The crucial tool for this purpose is the Bayes rule:

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

This is the *fundamental* rule of probabilistic inference.

- x here can be a hidden variable (we can't observe) or a parameter.
- y is the observed data.
- $p(x)$ is the prior distribution of x.

The crucial tool for this purpose is the Bayes rule:

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

This is the *fundamental* rule of probabilistic inference.

- x here can be a hidden variable (we can't observe) or a parameter.
- y is the observed data.
- $p(x)$ is the prior distribution of x.
- $p(y|x)$ is the likelihood of y given x.

The crucial tool for this purpose is the Bayes rule:

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

This is the *fundamental* rule of probabilistic inference.

- x here can be a hidden variable (we can't observe) or a parameter.
- y is the observed data.
- $p(x)$ is the prior distribution of x.
- $p(y|x)$ is the likelihood of y given x.
- $p(y)$ is the marginal likelihood of y.

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

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

Typically x is something unknown.

- A latent variable we cannot observe

Probabilistic Inference

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

Typically x is something unknown.

- A latent variable we cannot observe
  - The number of true cases of a disease in a population

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

Typically x is something unknown.

- A latent variable we cannot observe
  - The number of true cases of a disease in a population
  - The location of a hidden object

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

Typically x is something unknown.

- A latent variable we cannot observe
  - The number of true cases of a disease in a population
  - The location of a hidden object
  - The parameter of a distribution

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

Typically x is something unknown.

- A latent variable we cannot observe
  - The number of true cases of a disease in a population
  - The location of a hidden object
  - The parameter of a distribution
- The data y is observed (and fixed)
  - Observed number of cases in a given day

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

Typically x is something unknown.

- A latent variable we cannot observe
  - The number of true cases of a disease in a population
  - The location of a hidden object
  - The parameter of a distribution
- The data y is observed (and fixed)
  - Observed number of cases in a given day
  - Observations indicating the location of an object

Probabilistic Inference

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

Typically x is something unknown.

- A latent variable we cannot observe
  - The number of true cases of a disease in a population
  - The location of a hidden object
  - The parameter of a distribution
- The data y is observed (and fixed)
  - Observed number of cases in a given day
  - Observations indicating the location of an object
- The prior $p(x)$ is a distribution over x

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

Typically x is something unknown.

- A latent variable we cannot observe
  - The number of true cases of a disease in a population
  - The location of a hidden object
  - The parameter of a distribution
- The data y is observed (and fixed)
  - Observed number of cases in a given day
  - Observations indicating the location of an object
- The prior $p(x)$ is a distribution over x
  - This can come from prior information

Probabilistic Inference

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

Typically x is something unknown.

- A latent variable we cannot observe
  - The number of true cases of a disease in a population
  - The location of a hidden object
  - The parameter of a distribution
- The data y is observed (and fixed)
  - Observed number of cases in a given day
  - Observations indicating the location of an object
- The prior $p(x)$ is a distribution over x
  - This can come from prior information
  - This can from physical constraints or laws

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

Typically x is something unknown.

- A latent variable we cannot observe
  - The number of true cases of a disease in a population
  - The location of a hidden object
  - The parameter of a distribution
- The data y is observed (and fixed)
  - Observed number of cases in a given day
  - Observations indicating the location of an object
- The prior $p(x)$ is a distribution over x
  - This can come from prior information
  - This can from physical constraints or laws
  - Can be taken uninformative

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

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

Another important quantity is $p(y)$. In the continuous case:

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

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

Another important quantity is $p(y)$. In the continuous case:

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

In the discrete case,

$$p(y) = \sum_x p(x,y) = \sum_x p(y|x)p(x).$$

In general, the Bayes update (or rule):

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

is impossible to compute.

In general, the Bayes update (or rule):

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

is impossible to compute.

- The denominator $p(y)$ is intractable

In general, the Bayes update (or rule):

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

is impossible to compute.

- The denominator $p(y)$ is intractable

We cannot get closed form $p(x|y)$.

In general, the Bayes update (or rule):

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

is impossible to compute.

- The denominator $p(y)$ is intractable

We cannot get closed form $p(x|y)$.

But can we sample from it?

Before going to sampling, let us see examples of the Bayes rule.

Before going to sampling, let us see examples of the Bayes rule.

Consider the following discrete distribution:

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

This is the probability distribution of a roll of a pair of dice.

Before going to sampling, let us see examples of the Bayes rule.

Consider the following discrete distribution:

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

This is the probability distribution of a roll of a pair of dice.

Let us assume that we observe the sum to be 9.

Let us assume that we observe the sum to be 9.

In other words, we observe $y = 9$.

Let us assume that we observe the sum to be 9.

In other words, we observe $y = 9$.

Can we figure out the probability distribution conditioned on this observation?

What is the likelihood?

What is the likelihood?

It is noise-free (we have no error observing the sum) – could've been the case!

What is the likelihood?

It is noise-free (we have no error observing the sum) – could've been the case!

So in essence,

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

We know that $y = 9$ (observed).

Let us write the Bayes rule:

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

We know that $p(y = 9|x_1, x_2)$ is an indicator function at 9.

We know that $y = 9$ (observed).

Let us write the Bayes rule:

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

We know that $p(y = 9|x_1, x_2)$ is an indicator function at 9.

$$p(y|x_1, x_2) = \mathbf{1}(y = x_1 + x_2)$$

We know that $y = 9$ (observed).

Let us write the Bayes rule:

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

We know that $p(y = 9|x_1, x_2)$ is an indicator function at 9.

$$p(y|x_1, x_2) = \mathbf{1}(y = x_1 + x_2)$$

We can write it out explicitly for the values of x_1 and x_2.

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

What happens if we multiply this with $p(x_1)p(x_2)$?

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

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

Now we know the numerator: $p(y = 9 \mid x_1, x_2)p(x_1, x_2)$.

Recall:

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

Therefore, we need to compute $p(y = 9)$.

Recall:

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

Therefore, we need to compute $p(y = 9)$.

How is it computed?

Recall:

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

Therefore, we need to compute $p(y = 9)$.

How is it computed?

We need to sum over all possible values of x_1 and x_2:

$$p(y = 9) = \sum_{x_1, x_2} p(y = 9|x_1, x_2)p(x_1, x_2).$$

Let us write out the denominator:

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

We go back to our table of $p(y = 9|x_1, x_2)p(x_1, x_2)$ and divide by $p(y = 9)$:

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

This is our posterior distribution $p(x_1, x_2|y = 9)$!

Let us consider the following model:

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

Let us consider the following model:

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

We are given a data point y and we want to infer the value of x.

Let us consider the following model:

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

We are given a data point y and we want to infer the value of x.

What is the posterior density $p(x|y)$?

Tricks to derive posterior densities are numerous.

Tricks to derive posterior densities are numerous.

You might have heard about the list of conjugate priors.

Tricks to derive posterior densities are numerous.

You might have heard about the list of conjugate priors.

In this example, we will derive the posterior density using the Bayes rule ourselves (not looking up).

Let us write out the Bayes rule:

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

Let us write out the Bayes rule:

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

Now we write

$$p(x|y) \propto p(y|x)p(x).$$

This is the notation which means *proportional to*.

Let us write out the Bayes rule:

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

Now we write

$$p(x|y) \propto p(y|x)p(x).$$

This is the notation which means *proportional to*. Therefore, we have

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

We write

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

We write

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

We write

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

We write

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

So we've got:

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

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.

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

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

in terms of x (we can ignore the constants).

$$\frac{-y^2}{2\sigma^2} + \frac{yx}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{x^2}{2\sigma_0^2} + \frac{x\mu_0}{\sigma_0^2} = \frac{-x^2}{2\sigma_p^2} + \frac{x\mu_p}{\sigma_p^2}$$

$$\frac{-y^2}{2\sigma^2} + \frac{yx}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{x^2}{2\sigma_0^2} + \frac{x\mu_0}{\sigma_0^2} = \frac{-x^2}{2\sigma_p^2} + \frac{x\mu_p}{\sigma_p^2}$$

Match the coefficients of x^2:

$$\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} = \frac{1}{\sigma_p^2}.$$

which implies

$$\frac{-y^2}{2\sigma^2} + \frac{yx}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{x^2}{2\sigma_0^2} + \frac{x\mu_0}{\sigma_0^2} = \frac{-x^2}{2\sigma_p^2} + \frac{x\mu_p}{\sigma_p^2}$$

Match the coefficients of x^2:

$$\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} = \frac{1}{\sigma_p^2}.$$

which implies

$$\sigma_p^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + \sigma_0^2}.$$

$$\frac{-y^2}{2\sigma^2} + \frac{yx}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{x^2}{2\sigma_0^2} + \frac{x\mu_0}{\sigma_0^2} - \frac{\mu_0^2}{2\sigma_0^2} = \frac{-x^2}{2\sigma_p^2} + \frac{x\mu_p}{\sigma_p^2} - \frac{\mu_p^2}{2\sigma_p^2}$$

$$\frac{-y^2}{2\sigma^2} + \frac{yx}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{x^2}{2\sigma_0^2} + \frac{x\mu_0}{\sigma_0^2} - \frac{\mu_0^2}{2\sigma_0^2} = \frac{-x^2}{2\sigma_p^2} + \frac{x\mu_p}{\sigma_p^2} - \frac{\mu_p^2}{2\sigma_p^2}$$

Match the coefficients of x:

$$\frac{y}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} = \frac{\mu_p}{\sigma_p^2}.$$

which implies

$$\frac{-y^2}{2\sigma^2} + \frac{yx}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{x^2}{2\sigma_0^2} + \frac{x\mu_0}{\sigma_0^2} - \frac{\mu_0^2}{2\sigma_0^2} = \frac{-x^2}{2\sigma_p^2} + \frac{x\mu_p}{\sigma_p^2} - \frac{\mu_p^2}{2\sigma_p^2}$$

Match the coefficients of x:

$$\frac{y}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} = \frac{\mu_p}{\sigma_p^2}.$$

which implies

$$\mu_p = \frac{\sigma^2\mu_0 + \sigma_0^2 y}{\sigma^2 + \sigma_0^2}.$$

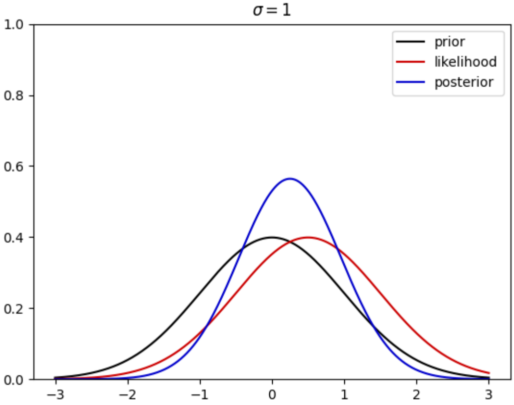
Finally, we obtain

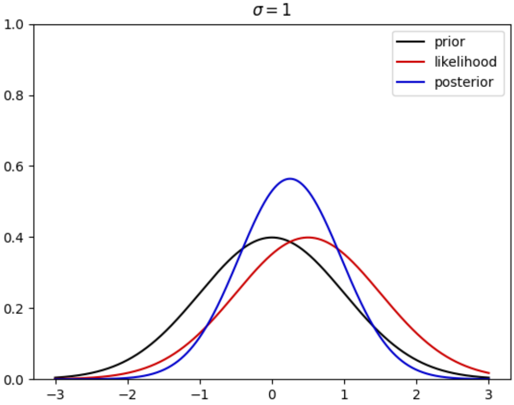
$$\mu_p = \frac{\sigma^2\mu_0 + \sigma_0^2 y}{\sigma^2 + \sigma_0^2}$$

$$\sigma_p^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + \sigma_0^2}.$$

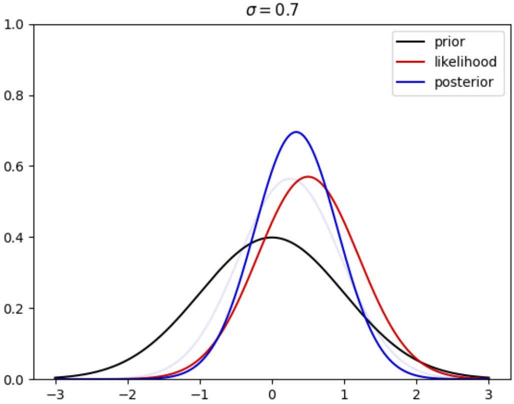
Probabilistic Inference

A continuous inference example





The likelihood is plotted for observation $y = 0.5$!

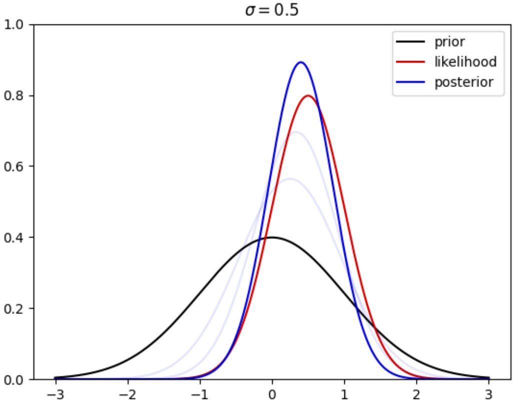


Peakier posterior for smaller likelihood variance.

Probabilistic Inference

A continuous inference example

Imperial College London



Peakier posterior for smaller likelihood variance.

We were lucky in this example that we could analytically compute the posterior. In general, we cannot do this.

We were lucky in this example that we could analytically compute the posterior. In general, we cannot do this.

Let us see how we can use rejection sampling to sample from the posterior.

We were lucky in this example that we could analytically compute the posterior. In general, we cannot do this.

Let us see how we can use rejection sampling to sample from the posterior.

We will use the same model and compare the samples to the exact posterior we computed.

Recall

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

We denote this unnormalised posterior by $\bar{p}(x|y)$.

Recall

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

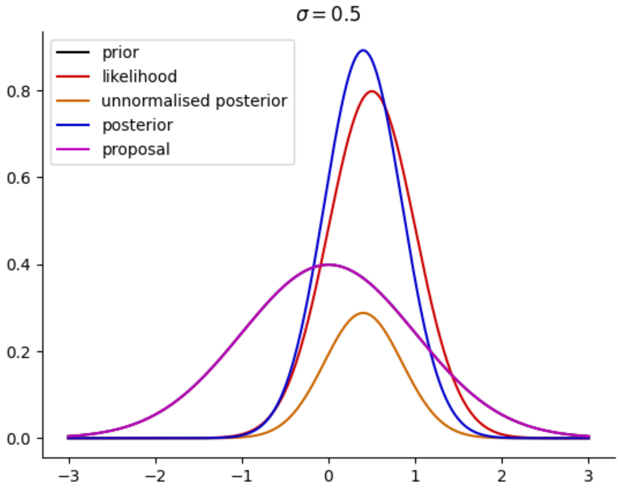
We denote this unnormalised posterior by $\bar{p}(x|y)$.

Note that this is a function of x: y is observed data and is fixed!

We need to choose a proposal. Let us choose

$$q(x) = \mathcal{N}(x; 0, 1).$$

and $M = 1$. It turns out this is enough for us.


$\sigma = 0.5$
prior
likelihood
unnormalised posterior
posterior
proposal
0.0
0.2
0.4
0.6
0.8
−3
−2
−1
0
1
2
3

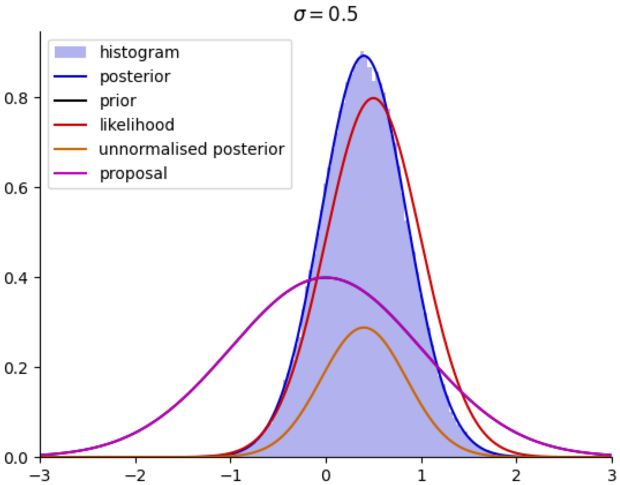
Rejection sampling:

- Sample $X' \sim q(x)$.
- Sample $U \sim \text{Unif}(0, 1)$.
- Accept if

$$U \leq \frac{p(y|X')p(X')}{Mq(X')},$$

  otherwise reject.
- Repeat.

Imperial College London



Another tractable example is the Gamma-Poisson model.

Another tractable example is the Gamma-Poisson model.

We have the prior defined as

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

Another tractable example is the Gamma-Poisson model.

We have the prior defined as

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

and the likelihood is given as

$$p(y|x) = \mathsf{Poisson}(y; x) = \frac{x^y}{y!} \exp(-x).$$

This can be seen as an observation model of a count.

Derive the posterior.

Derive the posterior.

$$p(x|y) \quad \propto \quad p(y|x)p(x)$$

Derive the posterior.

$$\begin{aligned} p(x|y) &\propto p(y|x)p(x) \\ &= \frac{x^y}{y!}\exp(-x)\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x). \end{aligned}$$

Derive the posterior.

$$\begin{aligned} p(x|y) &\propto p(y|x)p(x) \\ &= \frac{x^y}{y!}\exp(-x)\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x). \end{aligned}$$

This is proportional to

$$p(x|y) \propto x^{y+\alpha-1}\exp(-x(\beta+1)).$$

Derive the posterior.

$$\begin{aligned} p(x|y) &\propto p(y|x)p(x) \\ &= \frac{x^y}{y!}\exp(-x)\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x). \end{aligned}$$

This is proportional to

$$p(x|y) \propto x^{y+\alpha-1}\exp(-x(\beta+1)).$$

This is another Gamma density, i.e.:

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

Let us see how we can use rejection sampling to sample from the posterior.

Let us see how we can use rejection sampling to sample from the posterior.

We will use the same model and compare the samples to the exact posterior we computed.

Let us see how we can use rejection sampling to sample from the posterior.

We will use the same model and compare the samples to the exact posterior we computed.

We choose a prior

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

and a likelihood

$$p(y|x) = \mathsf{Poisson}(y; x) = \frac{x^y}{y!} \exp(-x).$$

Recall that our unnormalised posterior is written as

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

Recall that our unnormalised posterior is written as

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

Let us choose an exponential proposal:

$$q_\lambda(x) = \lambda \exp(-\lambda x).$$

Let us compute

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

Let us compute

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

The ratio is given by

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

We aim at optimising this w.r.t. x, so first compute $\log$:

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

We aim at optimising this w.r.t. x, so first compute $\log$:

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

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 $M_\lambda = \bar{p}(x^\star|y)/q_\lambda(x^\star)$::

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

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

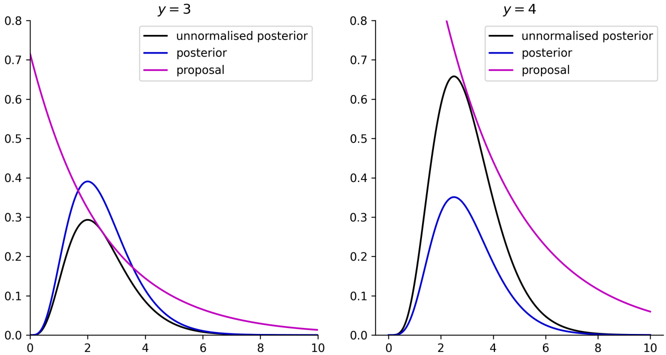
$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\log M_\lambda = -\frac{1}{\lambda} + \frac{(\alpha - 1 + y)}{2 - \lambda}$$

Setting this to zero, we obtain

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

which implies that

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


$y = 3$
unnormalised posterior
posterior
proposal
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
2
4
6
8
10
$y = 4$
unnormalised posterior
posterior
proposal

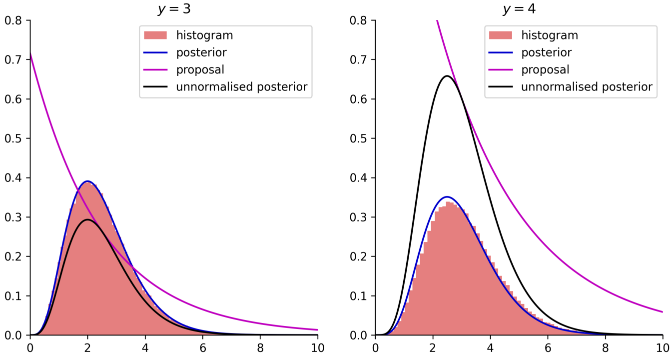


Figure: Histogram of the samples drawn using rejection sampling.

Today, we have covered:

- The Bayes rule and its applications
- Derivation of posterior distributions
- Rejection sampling for posterior distributions

We will look at more complex probabilistic models in next lectures.
We will also look at more efficient sampling methods.

See you tomorrow! (HXLY 414, Maths Learning Centre)

References I