Bayesian Updating with Continuous Priors
Class 13, 18.05, Spring 2014
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Understand a parameterized family of distributions as representing a continuous range of hypotheses for the observed data.
2. Be able to apply Bayes' theorem to update a prior pdf to a posterior pdf given data and a likelihood function.
3. Be able to interpret and compute probabilities using the posterior.

2 Introduction

Up to now we have only done Bayesian updating when we had a finite number of hypothesis, e.g. our dice example had five hypotheses (4, 6, 8, 12 or 20 sides). Now we will study Bayesian updating when there is a continuous range of hypothesis. The Bayesian update process will be essentially the same as in the discrete case. As usual when moving from discrete to continuous we will need to replace pmf by pdf and sums by integrals.

Here are three standard examples with continuous ranges of hypotheses.

Example 1. Suppose you have a system that can succeed or fail with probability p. Then we can hypothesize that p is anywhere in the range $[0, 1]$. That is, we have a continuous range of hypotheses. We will often model this example with a 'bent' coin with unknown probability p of heads.

Example 2. The lifetime of a certain isotope is modeled by an exponential distribution $\exp(\lambda)$. In principal, the mean lifetime $1/\lambda$ can be any real number in $(0, \infty)$.

Example 3. We are not restricted to a single parameter. In principle, the parameters μ and σ of a normal distribution can be any real numbers in $(-\infty, \infty)$ and $(0, \infty)$, respectively. If we model gestational length for single births by a normal distribution, then from millions of data points we know that μ is about 40 weeks and σ is about one week.

In all of these examples a hypothesis is a model for the random process giving rise to the data (successes and failures, atomic lifetimes, gestational lengths). If we specify a parameterized family of distributions, then a hypothesis may be regarded as a choice of parameter(s).

3 Notational conventions

3.1 Parametrized models

As in the examples above our hypotheses will often take the form 'a certain parameter has value θ'. We will often use the letter θ to stand for an arbitrary hypothesis. This will leave

symbols like p, f, and x to take there usual meanings as pmf, pdf, and data. Rather than saying 'the hypothesis that the parameter of interest has value θ' we will say simply 'the hypothesis θ'.

3.2 Big and little letters

We have two parallel notations for outcomes and probability.

1. Event A, probability function $P(A)$.
2. Value x, pmf $p(x)$ or pdf $f(x)$.

These notations are related by $P(X = x) = p(x)$, where x is a value the discrete random variable X and '$X = x$' is the corresponding event.

We carry these notations over to the conditional probabilities used in Bayesian updating.

1. Hypotheses $\mathcal{H}$ and data $\mathcal{D}$ have associated probabilities

$$P(\mathcal{H}),\ P(\mathcal{D}),\ P(\mathcal{H}|\mathcal{D}),\ P(\mathcal{D}|\mathcal{H}).$$

In the coin example we might have $\mathcal{H}$ = 'the chosen coin has probability .6 of heads', $\mathcal{D}$ = 'the flip was heads', and $P(\mathcal{D}|\mathcal{H}) = .6$

2. Hypotheses (values) θ and data values x have probabilities or probability densities:

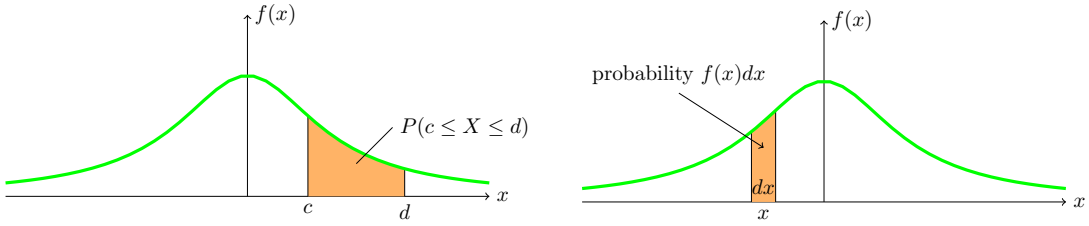
$$\begin{array}{cccc} p(\theta) & p(x) & p(\theta|x) & p(x|\theta) \\ f(\theta) & f(x) & f(\theta|x) & f(x|\theta) \end{array}$$

In the coin example we might have $\theta = .6$ and $x = 1$, so $p(x|\theta) = .6$. We might also write $p(x = 1|\theta = .6)$ to emphasize the values of x and θ, but we will never just write $p(1|.6)$ since here its unclear which value is x and which is θ.

Although we will still use both types of notation, we will mostly use the type involving pmf's and pdf's from now on. Hypotheses will usually be parameters represented by Greek letters $(\theta, \lambda, \mu, \sigma, \dots)$ while data values will usually be represented by English letters $(x, x_i, y, \dots)$.

4 Quick review of pdf and probability

Suppose X is a random variable with pdf $f(x)$. Recall $f(x)$ is a density; its units are probability/(units of x).



The probability that the value of X is in $[c, d]$ is given by

$$\int_c^d f(x)\,dx.$$

The probability that X is in an infitesimal range dx around x is $f(x)\,dx$. In fact, the integral formula is just the 'sum' of these infinitesimal probabilities. We can visualize these probabilities by viewing the integral as area under the graph of $f(x)$.

In order to manipulate probabilities instead of densities, in what follows we will make frequent use of the notion that $f(x)\,dx$ is the probability that X is in an infinitesimal range around x of width dx. Please make sure that you fully understand this notion.

5 Bayesian updating with continuous priors

The table for continuous priors is very simple. We cannot have a row for each of an infinite number of hypotheses, so instead we have just one row with variable hypothesis θ. After laying out this table, we will explain how it arises naturally by refining our table for a finite number of hypotheses as the number of hypotheses grows to infinity.

In cases with a discrete set of hypotheses we had a prior probability for each hypothesis. Now suppose our hypotheses are that the value of the parameter θ lies in the range $[a, b]$. In this case we need a prior pdf $f(\theta)$ which gives a probability density at each hypothesis θ. In order to use probabilities we will use infinitesimal ranges and state our hypothesis as:

$$\text{The parameter lies in a range } d\theta \text{ around } \theta \quad .$$

We can then write

$$\mathcal{H} : \theta \pm d\theta/2, \qquad P(\mathcal{H}) = f(\theta)\,d\theta.$$

This is a little cumbersome and it is easy to be sloppy. The advantage of this notation is two-fold. First using $d\theta$ will provide a clue as to when you need to do an integral. Second, when it comes to simulating with a computer it tells you exactly how to discretize your model.

For today we will assume that our data can only take a discrete set of values. In this case our given data x and a hypothesis θ the likelihood function is $p(x|\theta)$, i.e. the probability of x given θ. Next time we will consider continuous data distributions where our likelihood will have the form $f(x|\theta)\,dx$. Our table becomes

hypothesis	prior	likelihood	unnormalized posterior	posterior
$\theta \pm d\theta/2$	$f(\theta)\,d\theta$	$p(x\|\theta)$	$p(x\|\theta)f(\theta)\,d\theta$	$\frac{1}{T}p(x\|\theta)f(\theta)\,d\theta$
total	$\int_a^b f(\theta)\,d\theta = 1$		$T = \int_a^b p(x\|\theta)f(\theta)\,d\theta$	1

Notes:

1. The 'sum' T of the unnormalized posterior column is given by an integral. In practice, computing this integral is difficult and best left to computers to do numerically.

2. By including $d\theta$, all the entries in the table are probabilities. The posterior pdf for θ is found by removing the $d\theta$:

$$f(\theta|x) = \frac{1}{T}p(x|\theta)\,f(\theta)$$

3. If the unnormalized posterior turns out to be a multiple of a familiar type of distribution, then we can often avoid computing T. We will see several such examples in 18.05.

4. $T = p(x)$, the *prior predictive probability* of x, i.e. the *a priori* probability of observing x. By the law of total probability (see section below), $p(x)$ is a weighted average of likelihoods over all hypotheses:

$$p(x) = \int_a^b p(x|\theta) f(\theta)\, d\theta.$$

5. With $d\theta$ removed, the table organizes the continuous version of Bayes' theorem. Namely, the posterior pdf is related to the prior pdf and likelihood function via:

$$f(\theta|x) = \frac{p(x|\theta)\, f(\theta)}{\int_a^b p(x|\theta) f(\theta)\, d\theta} = \frac{p(x|\theta)\, f(\theta)}{p(x)}$$

Regarding both sides as functions of θ, we can again express Bayes' theorem in the form:

$$f(\theta|x) \propto p(x|\theta) \cdot f(\theta)$$

6. The use of $\theta \pm d\theta/2$ for hypotheses gets a little tedious. At times we may allow ourselves to be a bit sloppy and just write θ, but we will still mean $\theta \pm d\theta/2$

Example 4. *Coin with flat prior.* Suppose we have a bent coin with unknown probability θ of heads. Also suppose we flip the coin once and get heads. Starting from a flat prior pdf compute the posterior pdf for θ.

answer: By a flat prior we mean $f(\theta) = 1$ on $[0, 1]$; that is, we assume the true probability of heads is equally likely to be any probability. As we usually do with coin flips we let $x = 1$ for heads. In this case the definition of θ says the likelihood $p(x|\theta) = \theta$. We get the following table:

hypothesis	prior	likelihood	unnormalized posterior	posterior
$\theta \pm d\theta/2$	$1 \cdot d\theta$	θ	$\theta\, d\theta$	$2\theta\, d\theta$
total	$\int_a^b f(\theta)\, d\theta = 1$		$T = \int_0^1 \theta\, d\theta = 1/2$	1

Therefore the posterior pdf (after seeing 1 heads) is $f(\theta|x) = 2\theta$.

5.1 From discrete to continuous Bayesian updating

To develop intuition for the transition from discrete to continuous Bayesian updating, we'll walk a familiar road from calculus. Namely we will:

i) approximate the continuous range of hypotheses by a finite number.

ii) create the discrete updating table for the finite number of hypotheses.

iii) consider how the table changes as the number of hypotheses goes to infinity.

In this way, will see the prior and posterior pmf's converge to the prior and posterior pdf's.

Example 5. To keep things concrete, we will work with the 'bent' coin in Example 4.

We start by slicing $[0, 1]$ into 4 equal intervals: $[0, 1/4]$, $[1/4, 1/2]$, $[1/2, 3/4]$, $[3/4, 1]$.

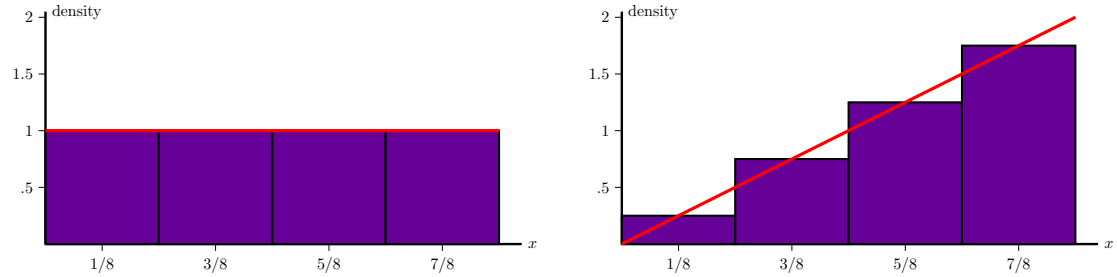
Each slice has width $\Delta\theta = 1/4$. We put our 4 hypotheses θ_i at the centers of the four slices:

$$\theta_1\text{: } \text{`}\theta = 1/8\text{'}, \quad \theta_2\text{: } \text{`}\theta = 3/8\text{'}, \quad \theta_3\text{: } \text{`}\theta = 5/8\text{'}, \quad \theta_4\text{: } \text{`}\theta = 7/8\text{'}.$$

The flat prior gives each hypothesis a probability of $1/4 = 1 \cdot \Delta\theta$. We have the table:

hypothesis	prior	likelihood	un. posterior	posterior
$\theta = 1/8$	$1/4$	$1/8$	$(1/4) \times (1/8)$	$1/16$
$\theta = 3/8$	$1/4$	$3/8$	$(1/4) \times (3/8)$	$3/16$
$\theta = 5/8$	$1/4$	$5/8$	$(1/4) \times (5/8)$	$5/16$
$\theta = 7/8$	$1/4$	$7/8$	$(1/4) \times (7/8)$	$7/16$
Total	1	–	$\sum_{i=1}^{n} \theta_i \, \Delta\theta$	1

Here are the density histograms of the prior and posterior pmf. The prior and posterior pdfs from Example 4 are superimposed on the histograms in red.

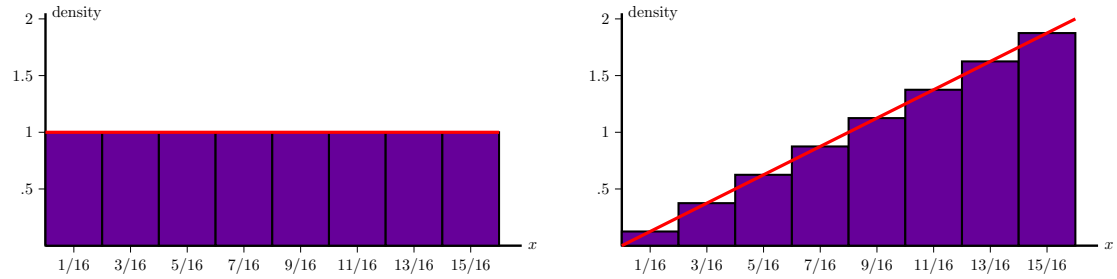


Next we slice [0,1] into 8 intervals each of width $\Delta\theta = 1/8$ and use the center of each slice for our 8 hypotheses θ_i.

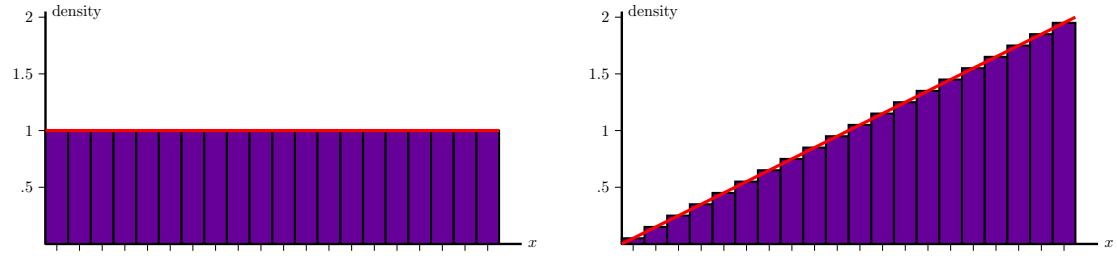
θ_1: '$\theta = 1/16$', θ_2: '$\theta = 3/16$', θ_3: '$\theta = 5/16$', θ_4: '$\theta = 7/16$'
θ_5: '$\theta = 9/16$', θ_6: '$\theta = 11/16$', θ_7: '$\theta = 13/16$', θ_8: '$\theta = 15/16$'

The flat prior gives each hypothesis the probablility $1/8 = 1 \cdot \Delta\theta$. Here are the table and density histograms.

hypothesis	prior	likelihood	un. posterior	posterior
$\theta = 1/16$	$1/8$	$1/16$	$(1/8) \times (1/16)$	$1/64$
$\theta = 3/16$	$1/8$	$3/16$	$(1/8) \times (3/16)$	$3/64$
$\theta = 5/16$	$1/8$	$5/16$	$(1/8) \times (5/16)$	$5/64$
$\theta = 7/16$	$1/8$	$7/16$	$(1/8) \times (7/16)$	$7/64$
$\theta = 9/16$	$1/8$	$9/16$	$(1/8) \times (9/16)$	$9/64$
$\theta = 11/16$	$1/8$	$11/16$	$(1/8) \times (11/16)$	$11/64$
$\theta = 13/16$	$1/8$	$13/16$	$(1/8) \times (13/16)$	$13/64$
$\theta = 15/16$	$1/8$	$15/16$	$(1/8) \times (15/16)$	$15/64$
Total	1	–	$\sum_{i=1}^{n} \theta_i \, \Delta\theta$	1



Finally we slice [0,1] into 20 pieces. This is essentially identical to the previous two cases. Let's skip right to the density histograms.



Looking at the sequence of plots we see how the prior and posterior density histograms converge to the prior and posterior probability density functions.

5.2 Using the posterior pdf

Example 6. In Example 4, after observing one heads, what is the (posterior) probability that the coin is biased towards heads?

answer: Since the parameter θ is the probability the coin lands heads, the problem asks for $P(\theta > .5 \,|\, x)$. This is easily computed from the posterior pdf.

$$P(\theta > .5|x) = \int_{.5}^{1} f(\theta|x)\, d\theta = \int_{.5}^{1} 2\theta\, d\theta = \theta^2\big|_{.5}^{1} = \frac{3}{4}.$$

This can be compared with the prior probability that the coin is biased towards heads:

$$P(\theta > .5) = \int_{.5}^{1} f(\theta)\, d\theta = \int_{.5}^{1} 1 \cdot d\theta = \theta\big|_{.5}^{1} = \frac{1}{2}.$$

We see that observing one heads has increased the probability that the coin is biased towards heads from 1/2 to 3/4.

6 The law of total probability

Recall that for discrete hypotheses $\mathcal{H}_1, \mathcal{H}_2, \ldots \mathcal{H}_n$ the law of total probability says the *prior probability of data* $\mathcal{D}$ is

$$P(\mathcal{D}) = \sum_{i=1}^{n} P(\mathcal{D}|\mathcal{H}_i)P(\mathcal{H}_i).$$

This is the *prior* probability of $\mathcal{D}$ because we used the prior probability $P(\mathcal{H}_i)$ If instead we use the values, $\theta_1, \theta_2, \ldots, \theta_n$ for hypotheses and x for data then this is written

$$p(x) = \sum_{i=1}^{n} p(x|\theta_i)p(\theta_i).$$

We call this the *prior predictive probability* of x to distinguish it from the prior probability of θ.

Suppose we then collect data x_2. Assuming data x_1 and x_2 are conditionally independent (i.e., they are independent if we condition on a hypothesis) then we can replace the prior probability for θ by the posterior $p(\theta|x_1)$ to get the *posterior predictive probability* of x_2 given x_1:

$$p(x_2|x_1) = \sum_{i=1}^{n} p(x_2|\theta_i)p(\theta_i|x_1).$$

If x_1 and x_2 are not conditionally independent then the probability of x_2 also needs to be conditioned on x_1. We get the somewhat more complicated formula for the posterior predictive probability of x_2 given x_1: (*This is not something we will see in 18.05.*)

$$p(x_2|x_1) = \sum_{i=1}^{n} p(x_2|\theta_i, x_1)p(\theta_i|x_1).$$

Likewise for continuous priors and posteriors over the range $[a, b]$ and discrete data we have

$$p(x) = \int_a^b p(x|\theta) f(\theta)\, d\theta$$

and (assuming x_1 and x_2 are conditionally independent)

$$p(x_2|x_1) = \int_a^b p(x_2|\theta) f(\theta|x_1)\, d\theta.$$

Example 7. In Example 4, compute the prior predictive and posterior predictive probability of heads on the second toss (i.e., prior and posterior to taking into account that the first toss was heads).

answer: The prior predictive probability of $x_2 = 1$ is

$$p(x_2) = \int_0^1 p(x_2|\theta)\, f(\theta)\, d\theta = \int_0^1 \theta \cdot 1 \cdot d\theta = \int_0^1 d\theta = \frac{1}{2}.$$

The posterior predictive probability of $x_2 = 1$ given $x_1 = 1$ is

$$p(x_2|x_1) = \int_0^1 p(x_2|\theta)\, f(\theta|x_1)\, d\theta = \int_0^1 \theta \cdot 2\theta\, d\theta = \int_0^1 2\theta^2\, d\theta = \frac{2}{3}\theta^3\Big|_0^1 = \frac{2}{3}.$$

We see that observing a heads on the first toss has increased the probability of heads on the second toss from 1/2 to 2/3.

18.05 Introduction to Probability and Statistics

Spring 2014