October 17, 2017

CS 361: Probability & Statistics

Inference

Maximum likelihood: drawbacks

A couple of things might trip up max likelihood estimation:

1) Finding the maximum of some functions can be quite hard

2) If we don't have a large amount of data, we might incorrectly estimate certain model parameters

For example, the MLE for p_i in a multinomial distribution is n_i/N if we have observed n_i instances of face i in N rolls of a die. If we have observed zero 3s in 8 rolls of a die, is it always safe to assume $p\_3=0$?

Bayesian inference

Bayesian inference

An alternative method for doing parameter estimation — figuring out a good theta, given the data — that has a different set of strengths and weaknesses than maximum likelihood estimation is called Bayesian inference

With MLE, we tried to find a theta that maximized the likelihood function

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta)$$

With Bayesian inference, we maximize a different function of theta, by treating theta as a random variable

$$P(\theta|\mathcal{D})$$

The value of theta that maximizes this function is called the **maximum a posteriori estimate** or **MAP estimate**

The prior

From Bayes' rule, we know that we can express our function of interest as

$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{Prior}}}{P(\mathcal{D})}$$

The right hand side contains the likelihood, which we've been working with. Also in the numerator is the so-called **prior probability** of θ

Bayesian inference is useful because it allows us to incorporate prior beliefs we have about the value of θ

The prior

From Bayes' rule, we know that we can express our function of interest as

$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{Prior}}}{P(\mathcal{D})}$$

In principle we can use any distribution we want for the prior, if we wanted to stubbornly insist that the true value of θ must not be between 0.25 and 0.75 no matter what the data tells us, for instance, we can just have a prior that has a probability of 0 for those values of θ and our MAP estimate will never be in that range

The prior

From Bayes' rule, we know that we can express our function of interest as

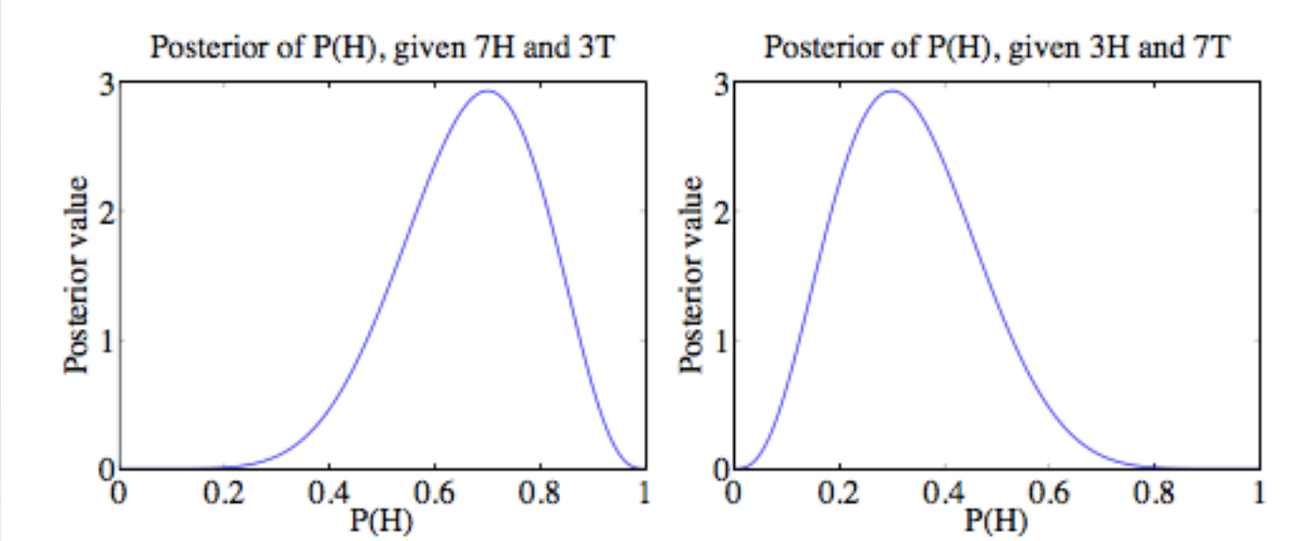
$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{Prior}}}{P(\mathcal{D})}$$

If we had a uniform prior, on the other hand, we are saying we have no particular beliefs about the true value of θ

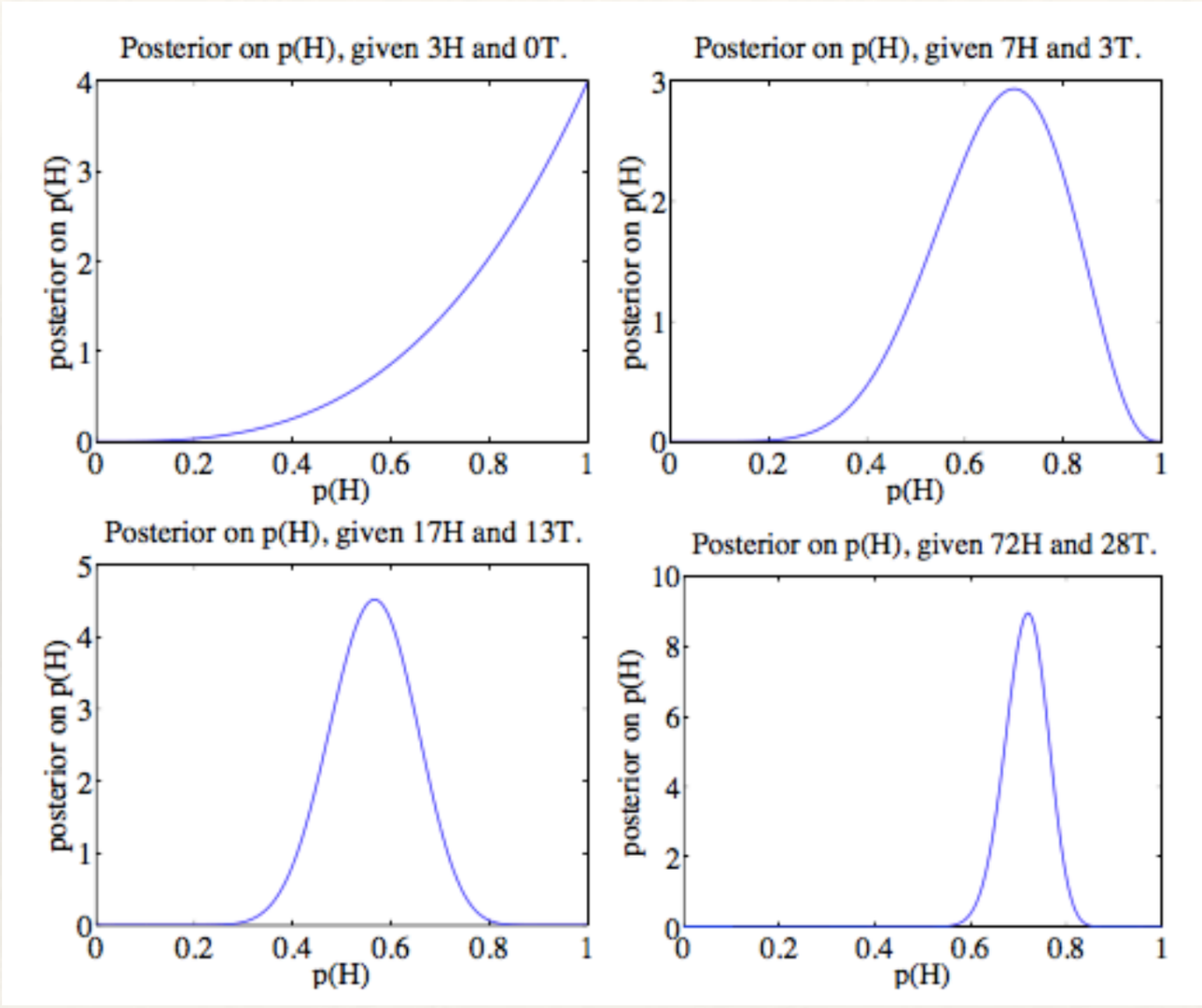
In that case, choosing a theta to maximize the left hand side (the MAP estimate) is the same as choosing a theta to maximize the likelihood (the MLE estimate) since only the likelihood on the RHS depends on theta

Example

Suppose we have a coin with probability θ of heads coming up. We make no assumptions about the prior probability of θ (i.e. assume a uniform prior). We flip the coin 10 times and see 7 heads and 3 tails. Plot a function proportional to p(θ | 7 heads and 3 tails) and for 3 heads, 7 tails.



Example


Posterior on p(H), given 3H and 0T.
posterior on p(H)
p(H)
Posterior on p(H), given 7H and 3T.
posterior on p(H)
p(H)
Posterior on p(H), given 17H and 13T.
posterior on p(H)
p(H)
Posterior on p(H), given 72H and 28T.
posterior on p(H)
p(H)

Which prior?

So we are in this interesting situation where we are considering the probability of a probability in some sense. The probability that a coin is fair or that it comes up heads 90% of the time.

In order to have a good prior for the theta in a coin flipping Binomial model we need a function that is a probability density over the range [0,1]

When we are doing a MAP estimate, we are trying to maximize the product

$$P(\mathcal{D}|\theta)P(\theta)$$

We want this product to be well-behaved enough that we can optimize it to get our MAP estimate, and multiplying two different probability distributions together might produce a really ugly result

Which prior?

Likelihood Prior

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

Posterior

A particular kind of good behavior we might insist upon is the following:

1) For a given problem setup, the likelihood function is largely out of our control. E.g. we suppose that our data is from a normal distribution, the likelihood function is going to be normal
2) So the prior is our only degree of freedom
3) We choose a prior that is expressive enough that we can encode arbitrary beliefs about the prior probability of theta — the unknown parameters in our model
4) But choose a prior such that when it is multiplied by the likelihood function, we get a posterior that is of the same random variable type as the prior

A prior satisfying 4) above is called a **conjugate prior** of the likelihood function

Which prior, binomial

The binomial family of distributions is conjugate to the **beta** family of distributions

A beta random variable is a continuous random variable defined on $0 \leq x \leq 1$ with parameters alpha > 0 and beta > 0 whose density has the following form

$$p(x;\alpha,\beta) = (\text{constant})x^{\alpha-1}(1-x)^{\beta-1}$$

The constant is in terms of a special function called the **gamma function** which is a generalization of the factorial function to positive real values rather than just non-negative integers. Details can be found in the first chapter of the book

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

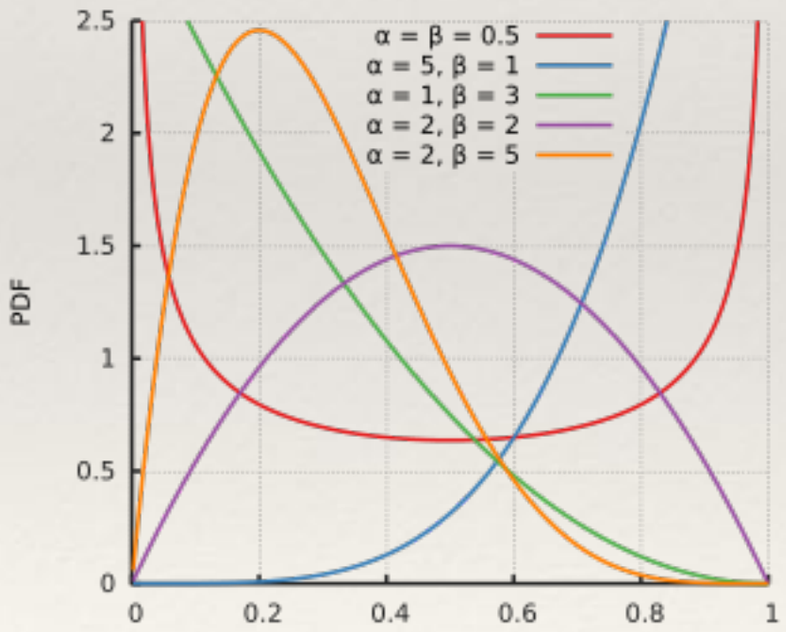
Beta distribution

Useful Facts: 6.6 *Beta distribution*

For a Beta distribution with parameters α, β

1. The mean is $\frac{\alpha}{\alpha+\beta}$.
2. The variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

$$p(x;\alpha,\beta) = (\text{constant})x^{\alpha-1}(1-x)^{\beta-1}$$



The beta distribution is very expressive

Having alpha=beta=1 would give a uniform prior

Binomial likelihood, beta prior

So if we want to do Bayesian inference against a binomial problem setup. Our likelihood be a binomial distribution. Let's see what happens if we pair that likelihood with a beta prior

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

If our data in this case is that we observed h heads in N flips, we have

$$p(\theta|\mathcal{D}) \propto \binom{N}{h}\theta^h(1-\theta)^{N-h}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Likelihood

Prior

Just focusing on theta, we get

$$p(\theta|\mathcal{D}) \propto \theta^{\alpha+h-1}(1-\theta)^{\beta+N-h-1}$$

Binomial likelihood, beta prior

If we do some clever things to make this a density, we get

$$P(\theta|\mathcal{D}) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+h)\Gamma(\beta+N-h)}\theta^{(\alpha+h)-1}(1-\theta)^{(\beta+N-h)-1}$$

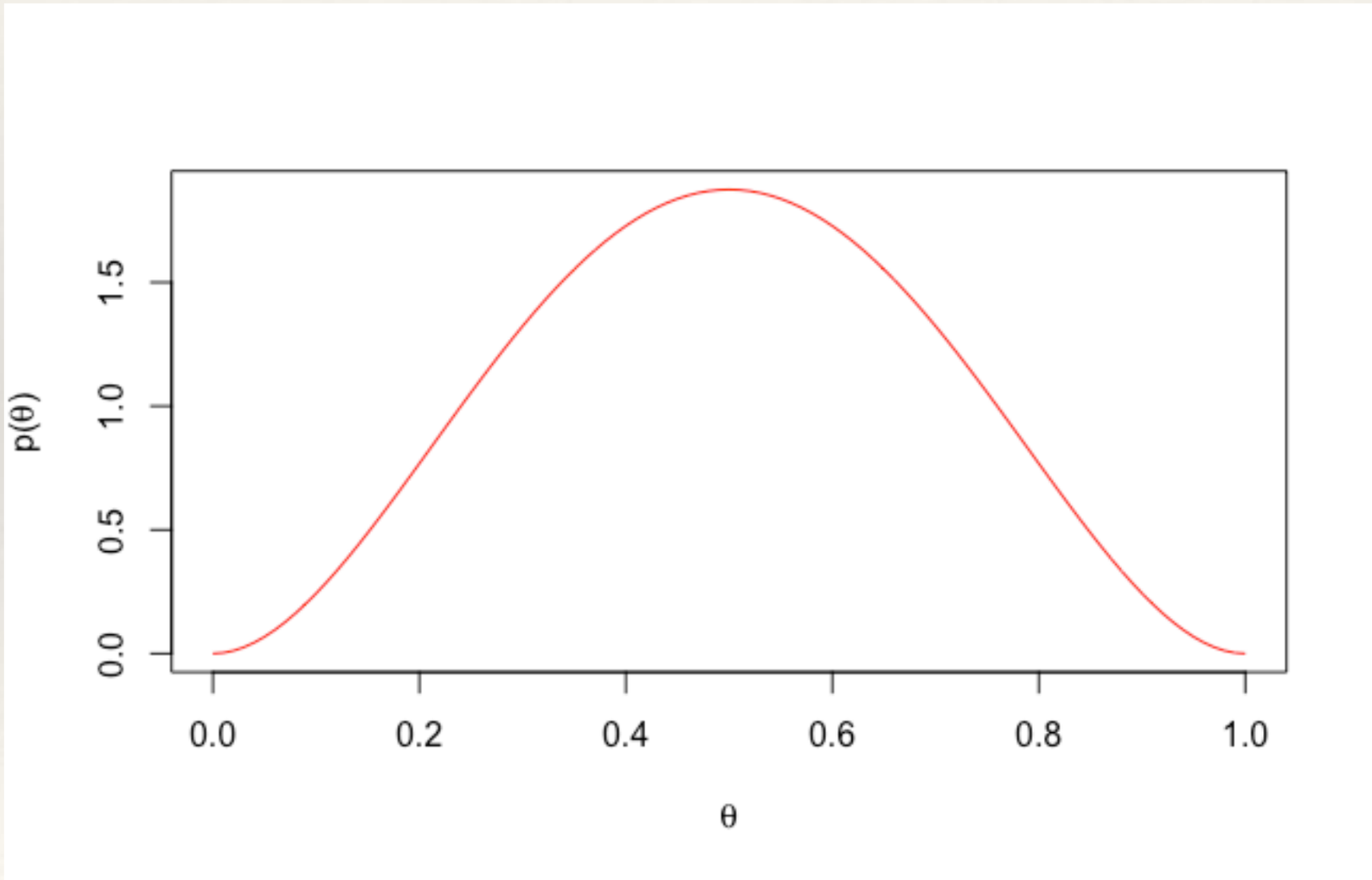
If we pattern match that with our definition of a Beta distribution

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

which is a beta distribution with parameters $\alpha+h$ and $\beta+(N-h)$

Example: updating

Suppose we have a prior belief that the probability of heads for a coin is governed by a beta distribution with parameters $\alpha = 3, \beta = 3$ this is what the density of theta would look like



Example: updating

Now, if we observed 10 coin flips and saw 7 heads, our posterior would be given by this ugly formula

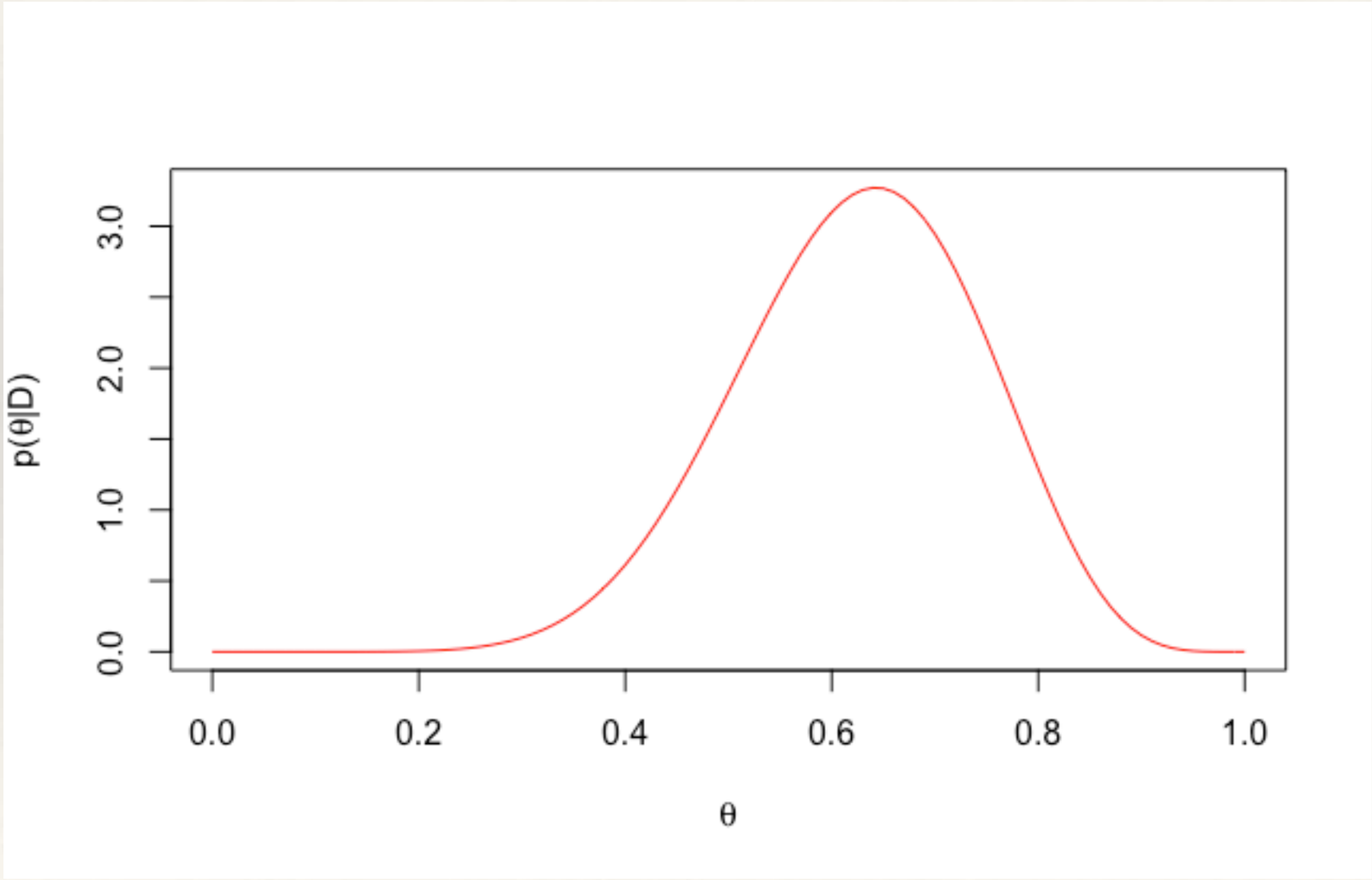
$$P(\theta|\mathcal{D}) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+h)\Gamma(\beta+N-h)}\theta^{(\alpha+h)-1}(1-\theta)^{(\beta+N-h)-1}$$

But we can ignore the ugliness and just realize that starting with $\alpha = 3, \beta = 3$ and then observing 7 heads and 3 tails, we will have a beta distribution with $\alpha = 10, \beta = 6$

We always just add the number of heads we saw to alpha and the number of tails we saw to beta. We don't have to actually think about multiplying the prior and likelihood every time since the beta distribution is a conjugate prior for the binomial!

Example: updating

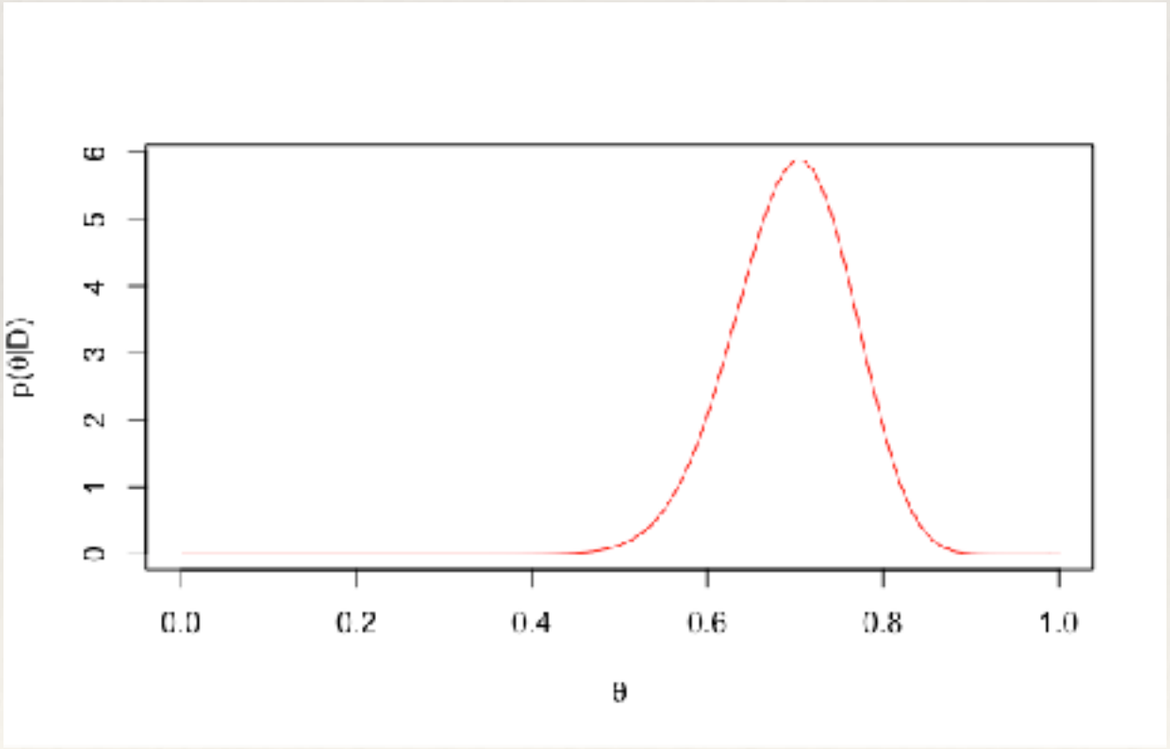
This is what our posterior distribution now looks like



Example: updating

Now here is the real power of conjugate priors. What if we had the same coin and after observing the last 10 flips we got to observe 30 more and this time we saw 22 heads?

We can use the beta posterior from the last slide as the beta prior in this slide. Doing so would give a new posterior that's a beta with $\alpha = 32, \beta = 14$



Beta prior

Since multiplying a likelihood by a conjugate prior gives a posterior that's in the same family as the prior, this makes it easier to have a probabilistic model that we just update with new data as it comes in instead of re-calculating the posterior from scratch every time we have new data

If we have new data we use our old posterior as the prior for our new estimate

Furthermore our example showed that the parameters in a beta distribution just wind up counting heads and tails.

If we started off with a prior with values alpha and beta, we observe a bunch of data in perhaps separate updates, the final posterior we will have will just be a beta distribution with parameters alpha + number of heads and beta + number of tails we have seen so far

Conjugate priors

The beta is also a conjugate prior for geometric and Bernoulli random variables

The exponential and Poisson distributions have another distribution called the **gamma distribution** as their conjugate prior which we will consider next

The normal distribution is conjugate to itself which will not show in depth

Poisson's prior

We liked the Beta distribution because it assigned probabilities in the range [0,1] and when multiplied by the binomial likelihood gave another Beta distribution.

The free parameter in a Poisson distribution is λ and it corresponds to the rate or intensity of the number of events we expect to observe per interval of time or space

Our prior for a Poisson distribution, then, will need to be able to assign a probability to any valid value of λ . In other words it will need to be defined for $x \geq 0$

Gamma distribution

It turns out that the family of distributions with parameters alpha > 0 and beta > 0 called the gamma distribution is the conjugate prior to the Poisson. Its density is

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

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

And we have

Useful Facts: 6.7 *Gamma distribution*

For a Gamma distribution with parameters α, β

1. The mean is $\frac{\alpha}{\beta}$.
2. The variance is $\frac{\alpha}{\beta^2}$.

Example

Suppose we are watching a speech by a politician who swears a lot. We model the politician's swearing with a Poisson distribution. Here are how many swear words we hear in the first 10 intervals in the politician's speech

no. of swear words	no. of intervals
0	5
1	2
2	2
3	1
4	0

For our likelihood we get

$$P(\mathcal{D}|\theta) = \left(\frac{\theta^0 e^{-\theta}}{0!}\right)^5 \left(\frac{\theta^1 e^{-\theta}}{1!}\right)^2 \left(\frac{\theta^2 e^{-\theta}}{2!}\right)^2 \left(\frac{\theta^3 e^{-\theta}}{3!}\right)^1$$

Or

total number of swears observed

number of intervals observed

$$P(\mathcal{D}|\theta) \propto \theta^{9} e^{-10\theta}$$

Example

$$P(\mathcal{D}|\theta) \propto \theta^9 e^{-10\theta}$$

If we multiplied this likelihood by a gamma prior with parameters alpha and beta, we would get

$P(\mathcal{D}|\theta)P(\theta) \propto \theta^9 e^{-10\theta}\theta^{\alpha-1}e^{-\beta\theta}$ or $P(\theta|\mathcal{D}) \propto \theta^{(\alpha+9)-1}e^{-(\beta+10)\theta}$

Any distribution that's proportional to $\theta^{\alpha-1}e^{-\beta\theta}$ is a gamma distribution

Which is to say our posterior would be a gamma with parameters $\alpha + 9$ and $\beta + 10$

Poisson & gamma

To generalize, then, for a Poisson likelihood, if we begin with a prior that is a gamma with parameters alpha and beta

And then we observe *N* intervals with a total of *k* events

The posterior probability of the Poisson parameter is a gamma random variable with parameters alpha + k and beta + N

MAP inference

We've done all of this so that if we encounter a distribution, we know a good prior to choose so that we can write down $P(\theta|\mathcal{D})$ after observing some data

Recall that the point estimate task involves choosing a value for theta. The so-called MAP (maximum a-posteriori) estimate is

$$\hat{\theta} = \arg\max_{\theta} P(\theta|\mathcal{D})$$

Binomial MAP

If we have observed some binomial data, say we have seen *h* heads in *N* coin flips and we have a beta prior with parameters alpha and beta, then we have

$$P(\theta|\mathcal{D}) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+h)\Gamma(\beta+N-h)}\theta^{(\alpha+h)-1}(1-\theta)^{(\beta+N-h)-1}$$

If we differentiate with respect to theta and set equal to 0, we will get a MAP estimate of

$$\hat{\theta} = \frac{\alpha-1+h}{\alpha+\beta-2+N}$$

Note that if we had started off with a prior of alpha=beta=1, we would have a uniform prior and our MAP estimate would be the same as the MLE estimate we derived before

Poisson MAP

If we have observed some Poisson data, say we have seen *k* total events in *N* total intervals and we have a gamma prior with parameters alpha and beta, then we have

$$P(\theta|\mathcal{D}) = \frac{(\beta+N)^{(\alpha+k)}}{\Gamma(\alpha+k)}\theta^{(\alpha+k)-1}e^{-(\beta+N)\theta}$$

Differentiating with respect to theta, setting equal to zero and solving for theta gives a MAP estimate of

$$\hat{\theta} = \frac{\alpha-1+k}{\beta+N}$$

MAP caveats

As we see more and more data, the prior tends to matter less and less. Which is to say our MAP estimate is very close to the MLE estimate for a large number of data items. Hence it might not be worth the trouble thinking about priors unless we have a small amount of data

Justifying the choice of prior can be hard. We chose ones that were mathematically convenient—that gave us a nice posterior—but does nature use conjugate priors? Maybe not for whichever problem you're considering