Week 5: Bayesian Inference

Bayesian Statistics for Machine Learning

Dr Daniel Worrall

AMLab, University of Amsterdam

October 3, 2019

Probability and Statistics

We are mostly concerned with models which look like

$$p(\mathbf{x}|\boldsymbol{\theta}).$$

In many case $\mathbf{x}$ refers to an *observation* and $\boldsymbol{\theta}$ refers to a set of *parameters*.

Probability

$$\begin{tikzcd}
\boldsymbol{\theta} \arrow[d, "p(\mathbf{x}|\boldsymbol{\theta})"] \\
\{\mathbf{x}_1, \mathbf{x}_2, ...\}
\end{tikzcd}$$

Statistics

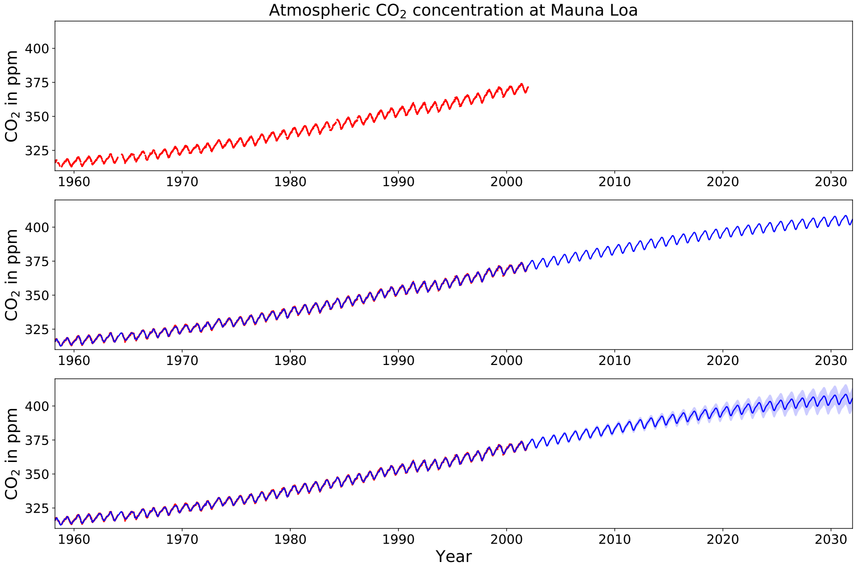
$$\begin{tikzcd}
\{\mathbf{x}_1, \mathbf{x}_2, ...\} \arrow[d, "p(\mathbf{x}|\boldsymbol{\theta})\text{or}p(\boldsymbol{\theta}|\mathbf{x})"] \\
\boldsymbol{\theta}
\end{tikzcd}$$

Machine Learning

$$\begin{tikzcd}
\{\mathbf{x}_1, \mathbf{x}_2, ...\} \arrow[d, "p(\mathbf{x}|\boldsymbol{\theta})\text{or}p(\boldsymbol{\theta}|\mathbf{x})"] \\
\boldsymbol{\theta} \arrow[d, "p(\mathbf{x}_*|\boldsymbol{\theta})"] \\
\{\mathbf{x}_*\}
\end{tikzcd}$$

Machine learning

In machine learning, we use past data to make predictions about the future.

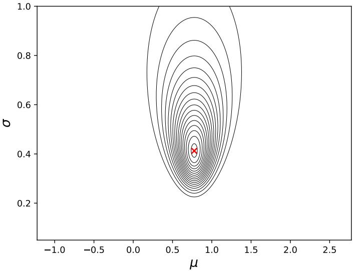


I: Bayesian Inference

Bayesian Inference

We previously learnt about maximum likelihood, where we solved

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} p(\mathcal{D}|\boldsymbol{\theta}).$$



- Is this the best we can do?
- It is almost certainly wrong $P(\boldsymbol{\theta}_{\text{ML}} = \boldsymbol{\theta}_{\text{true}}) = 0$.
- In reality $\mathcal{D} \nsim p(\mathcal{D}|\boldsymbol{\theta})$ but $\mathcal{D} \sim p_{\text{true}}(\mathcal{D})$.

Why do we want $\boldsymbol{\theta}_{\text{ML}}$ anyway?

Options

1. We are actually interested in knowing $\boldsymbol{\theta}$
2. We don't care: actually want to generate samples $\{\mathbf{x}_*^{(1)}, \mathbf{x}_*^{(2)}, ..\}$ from the data generating distribution $p_{\text{true}}(\mathbf{x})$.

Posterior distributions

All we need is the *posterior distribution*

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

Read, *"the probability of the parameters, given the data"*.

This is way more descriptive than a single point estimate $\boldsymbol{\theta}_{\mathsf{ML}}$. It is a probability distribution over the entire space of $\boldsymbol{\theta}$, telling us how much each one explains the data.

Posterior distributions

We compute the posterior using Bayes' theorem

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}.$$

These terms each have names:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

Later on we shall dissect this equation in agonising detail, but for now, let's just use it and get a feel for how it works.

NOTE: the evidence also goes by the name of *marginal likelihood*.

Two TAs

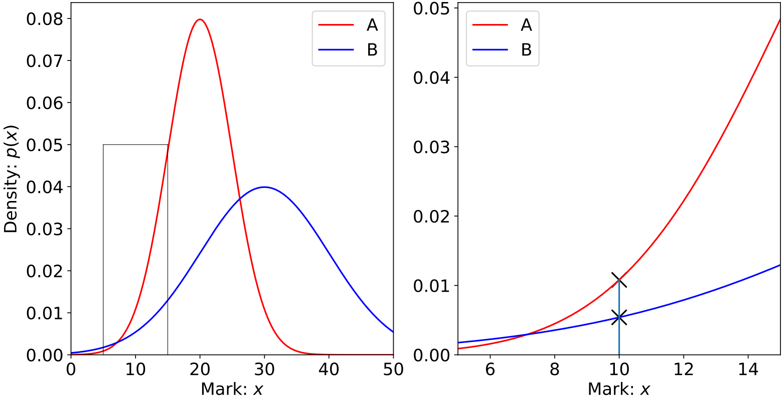
e.g. Two TAs are marking exams. If x denotes the marks awarded by a TA, the distributions of the two markers is

$$p(x|\mathsf{TA_A}) = \mathcal{N}(x; 30, 10^2)$$
$$p(x|\mathsf{TA_B}) = \mathcal{N}(x; 20, 5^2).$$

You flunk your course receiving $x = 10$ marks. Unethically, you wish to seek out who marked your exam paper. Find the posterior probability that $\mathsf{TA_A}$ marked your manuscipt, assuming you initially expect $\mathsf{TA_A}$ marked your paper with probability $\pi = 0.5$.

$$p(\mathsf{TA_A}|\mathcal{D} = \{10\}) = \frac{p(\mathcal{D}|\mathsf{TA_A})p(\mathsf{TA_A})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mathsf{TA_A})p(\mathsf{TA_A})}{\sum_{i\in\{A,B\}} p(\mathcal{D}|\mathsf{TA_i})p(\mathsf{TA_i})}$$

Two TAs



Density: $p(x)$
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50
Mark: x
A
B
0.05
0.04
0.03
0.02
0.01
0.00
6
8
10
12
14
Mark: x
A
B

Two TAs

$$
\begin{aligned}
p(\mathsf{TA_A}|\mathcal{D}=\{10\},\pi) &= \frac{p(\mathcal{D}|\mathsf{TA_A})p(\mathsf{TA_A})}{p(\mathcal{D}|\mathsf{TA_A})p(\mathsf{TA_A})+p(\mathcal{D}|\mathsf{TA_B})p(\mathsf{TA_B})} \\
&= \frac{\mathcal{N}(10;30,10^2)\pi}{\mathcal{N}(10;30,10^2)\pi+\mathcal{N}(10;20,5^2)(1-\pi)} \\
&= \frac{\pi}{\pi+\frac{\mathcal{N}(10;20,5^2)}{\mathcal{N}(10;30,10^2)}(1-\pi)}
\end{aligned}
$$

$$
\frac{\mathcal{N}(10;20,5^2)}{\mathcal{N}(10;30,10^2)} = \frac{\frac{1}{5\sqrt{2\pi}}\exp\{-\frac{(10-20)^2}{2\cdot 5^2}\}}{\frac{1}{10\sqrt{2\pi}}\exp\{-\frac{(10-30)^2}{2\cdot 10^2}\}} = \frac{10e^{-2}}{5e^{-2}} = 2
$$

$$
p(\mathsf{TA_A}|\mathcal{D}=\{10\},\pi) = \frac{\pi}{\pi+2(1-\pi)}
$$

If $\pi = 0.5$ then $p(\mathsf{TA_A}|\mathcal{D}=\{10\},\pi=0.5)=\frac{1}{3}$.

Two TAs

Say you get a tip off that $\mathsf{TA_A}$ marked 80% of the manuscripts. What now is your posterior belief that $\mathsf{TA_A}$ marked your paper?

Easy peasy lemon squeezy. We just set the prior to

$$p(\mathsf{TA_A}) = \pi = 0.8$$

So

$$p(\mathsf{TA_A} | \mathcal{D} = \{10\}, \pi = 0.8) = \frac{0.8}{0.8 + 2(1 - 0.8)} = \frac{2}{3}.$$

The Bayesian framework allows us to incorporate *prior knowledge* into our inferences. How would we achieve this in the maximum likelihood setting?

The Bent Coin

e.g. This is the original inference problem studied by Thomas Bayes in 1763.

You are given a bent coin. You flip it N times. It lands `heads` H times. If we denote the probability of the coin landing `heads` as π, what is the ML solution and the posterior distribution $p(\pi|\mathcal{D})$?

ML solution If `heads` is 1, we have $p(x|\pi) = \pi^x(1-\pi)^{1-x}$, so

$$\begin{aligned}\mathcal{L}(\pi) &= \sum_{i=1}^{N} x_i \log \pi + (1 - x_i)\log(1-\pi) \\ &= H \log \pi + (N - H)\log(1-\pi)\end{aligned}$$

which has a maximum at

$$\pi_{\mathsf{ML}} = \frac{H}{N}.$$

The Bent Coin cont'd

Bayesian solution Always start with Bayes' Theorem

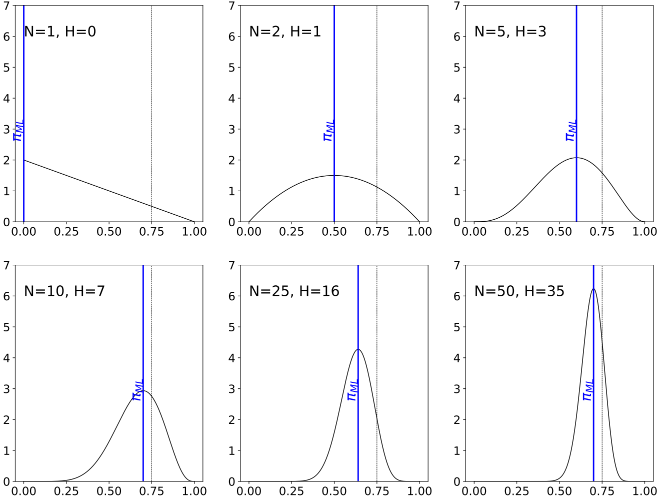
$$p(\pi|\mathcal{D}) = \frac{p(\mathcal{D}|\pi)p(\pi)}{p(\mathcal{D})} = \frac{\left[\prod_{i=1}^{N} p(x_i|\pi)\right] p(\pi)}{p(\mathcal{D})}$$

We need a prior, let's pick $p(\pi) = \text{Uniform}(\pi; 0, 1) = \mathbb{I}[\pi \in [0,1]]$ for now.

$$\begin{aligned} p(\pi|\mathcal{D}) &= \frac{\left[\prod_{i=1}^{N} \pi^{x_i}(1-\pi)^{1-x_i}\right] \mathbb{I}[\pi \in [0,1]]}{p(\mathcal{D})} \\ &= \frac{\left[\pi^{H}(1-\pi)^{N-H}\right] \mathbb{I}[\pi \in [0,1]]}{p(\mathcal{D})} \\ &= \frac{1}{Z}\pi^{H}(1-\pi)^{N-H} \end{aligned}$$

The Bent Coin cont'd

D = [0 1 0 1 1 1 0 1 1 1 1 0 0 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 0 1
1 1 1 1 1 0 0 0 1 0 0 1 1]



The Bent Coin cont'd

Notice how the posterior is 'less temperamental' than the likelihood function.

Next we need to figure out the *marginal likelihood*

$$Z = p(\mathcal{D}) = \int p(\mathcal{D}, \pi) \, \mathrm{d}\pi = \int \underbrace{p(\mathcal{D}|\pi)}_{\text{likelihood}} \underbrace{p(\pi)}_{\text{prior}} \, \mathrm{d}\pi.$$

The marginal likelihood is an instance of the famous Beta integral[1]

$$p(\mathcal{D}) = \int_0^1 \pi^H (1-\pi)^{N-H} \, \mathrm{d}\pi = B(H+1, N-H+1) = \frac{H!(N-H)!}{(N+1)!}$$

$$\boxed{p(\pi|\mathcal{D}) = \frac{(N+1)!}{H!(N-H)!} \pi^H (1-\pi)^{N-H}}$$

Don't worry if this integral scares you. It frightens me too! Resources such as Wolfram Alpha, Wikipedia, the Bishop book, and the MacKay book are handy.

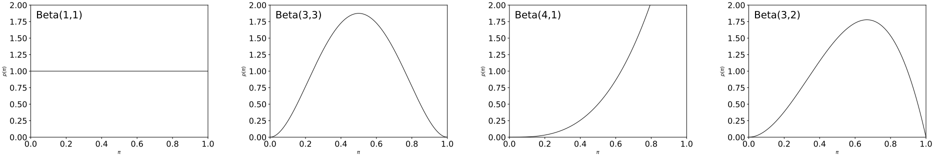
[1] $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1} \, \mathrm{d}t$

The Bent Coin cont'd

The posteror has the form of a *Beta distribution*

$$\text{Beta}(\pi|\alpha,\beta) = \frac{1}{Z(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}, \qquad Z(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- The Beta distribution is a probability distribution over probabilities.
- The two parameters α and β control the shape of the distribution.
- The *Gamma* function[2] satisfies $\Gamma(\alpha) = (\alpha-1)!$ and $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.
- The mean of the distribution is $\mathbb{E}[\pi] = \frac{\alpha}{\alpha+\beta}$.



[2] $\Gamma(\alpha) := \int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x$

The Bent Coin cont'd

So, we see that the posterior distribution gets sharper and sharper as N increases. In fact it becomes a Dirac delta peak in the limit of infinite data.

Its maximum is called the *maximum a posteriori* or *MAP*. In this example

$$\pi_{\mathsf{MAP}} = \frac{H}{N}$$

which coincides with π_{ML}. This is not true in general.

The Bent Coin cont'd

We can use the posterior to make predictions. We can compute the probability of the next coin toss as $p(x_* = \text{head}|\mathcal{D})$. We can find this by expanding a marginal

$$p(x_* = \text{head}|\mathcal{D}) = \int p(x_* = \text{head}, \pi|\mathcal{D})\,\mathrm{d}\pi = \int \underbrace{p(x_* = \text{head}|\pi, \cancel{\mathcal{D}})}_{\text{forward likelihood}} \underbrace{p(\pi|\mathcal{D})}_{\text{posterior}}\,\mathrm{d}\pi$$

$$= \int_0^1 \pi \cdot \frac{\pi^H (1-\pi)^{N-H}}{p(\mathcal{D})}\,\mathrm{d}\pi = \mathbb{E}_\pi\left[\text{Beta}(\pi|H+1, N-H+1)\right].$$

This is the mean of the Beta distribution, which is $\mathbb{E}_\pi \text{Beta}(\pi|a,b) = \frac{a}{a+b}$ so

$$p(x_* = \text{head}|\mathcal{D}) = \frac{H+1}{N+2}$$

which is called *Laplace's rule of succession*.

The Bent Coin cont'd

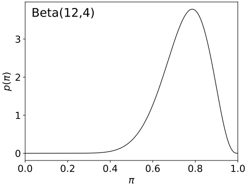
What just happened? We computed a quantity called the *posterior predictive distribution*. It is

$$\boxed{\underbrace{p(x_*|\mathcal{D})}_{\text{posterior predictive}} = \int_{\boldsymbol{\Theta}} \underbrace{p(x_*|\boldsymbol{\theta})}_{\text{forward likelihood}} \underbrace{p(\boldsymbol{\theta}|\mathcal{D})}_{\text{posterior}} \, \mathrm{d}\boldsymbol{\theta}}$$

If we interpret $p(x_*|\boldsymbol{\theta})$ as a particular model of x_* given the parameters θ, then $p(x_*|\mathcal{D})$ is a weighted average of *all possible models*, where the weights are determined by the posterior i.e. the training data.

The Bent Coin cont'd

Say someone had given us a sneaky hint beforehand that the coin's bias is about 0.7-0.8. We would choose a Beta prior with this mean, say



$$p(\pi) = \text{Beta}(\pi|\alpha = 12, \beta = 4)$$

The posterior is

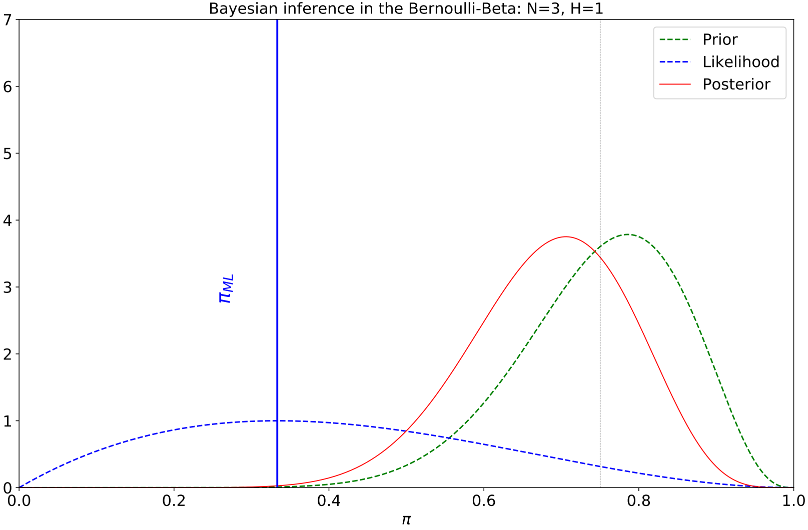
$$p(\pi|\mathcal{D}) = \frac{\overbrace{\left[\pi^{H}(1-\pi)^{N-H}\right]}^{\text{likelihood}} \cdot \overbrace{\pi^{\alpha-1}(1-\pi)^{\beta-1}}^{\text{prior}}}{p(\mathcal{D})} = \frac{\pi^{H+\alpha-1}(1-\pi)^{N-H+\beta-1}}{Z(H+\alpha, N-H+\beta)}$$

But this is is just the Beta distribution, with different parameters!

$$p(\pi|\mathcal{D}) = \text{Beta}(\pi; H+\alpha, N-H+\beta).$$

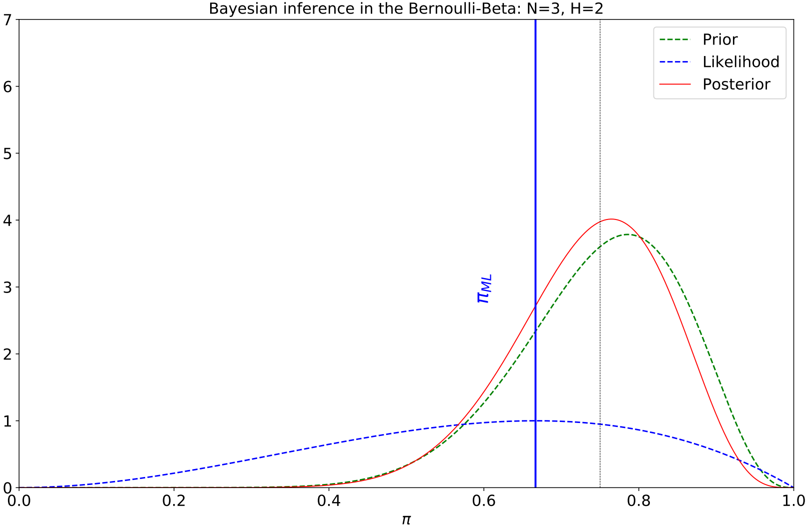
The Bent Coin cont'd

Bayesian solution



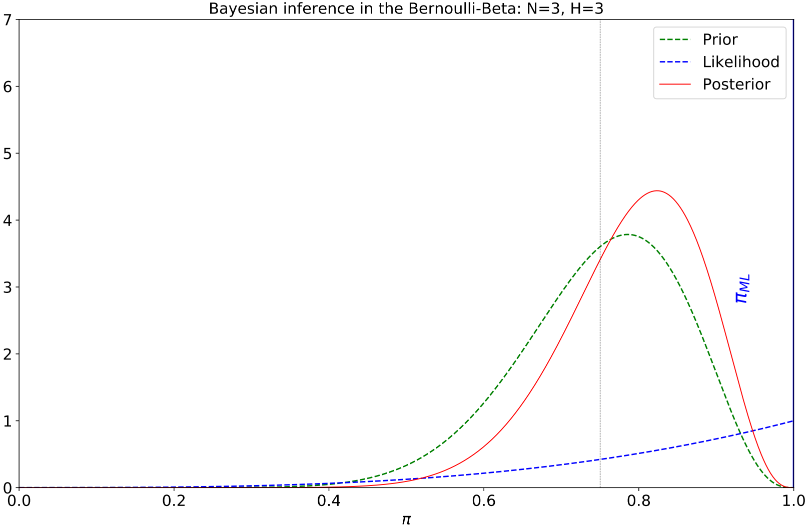
The Bent Coin cont'd

Bayesian solution

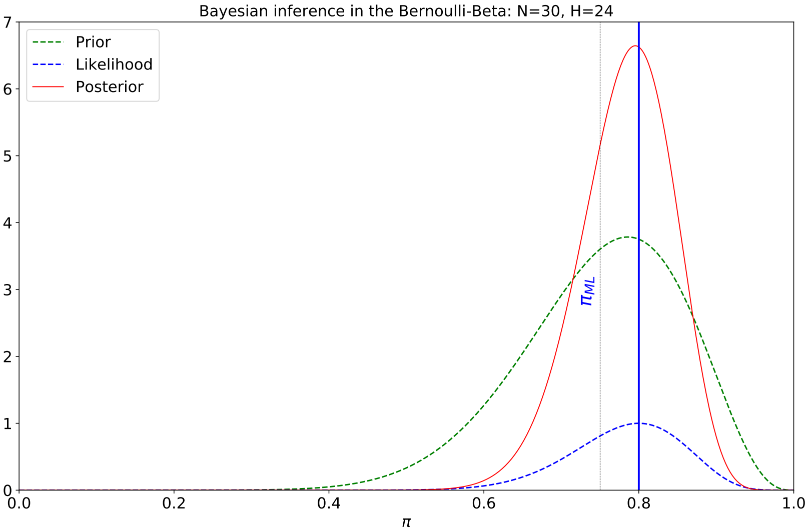


The Bent Coin cont'd

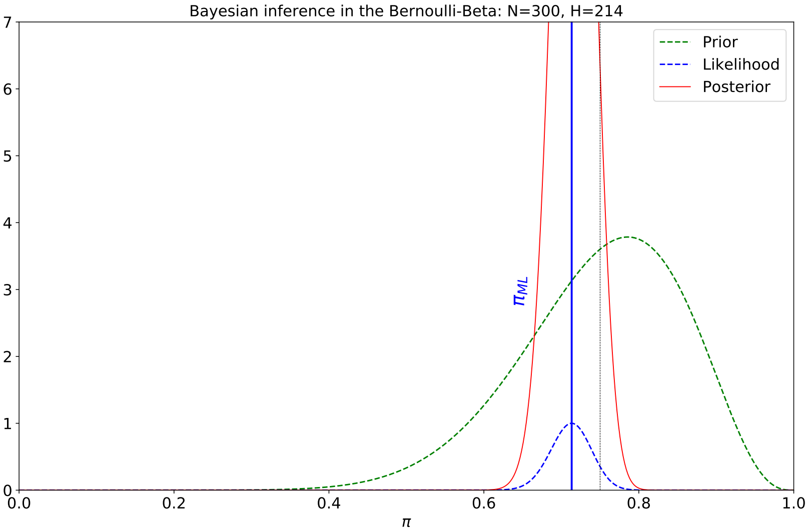
Bayesian solution



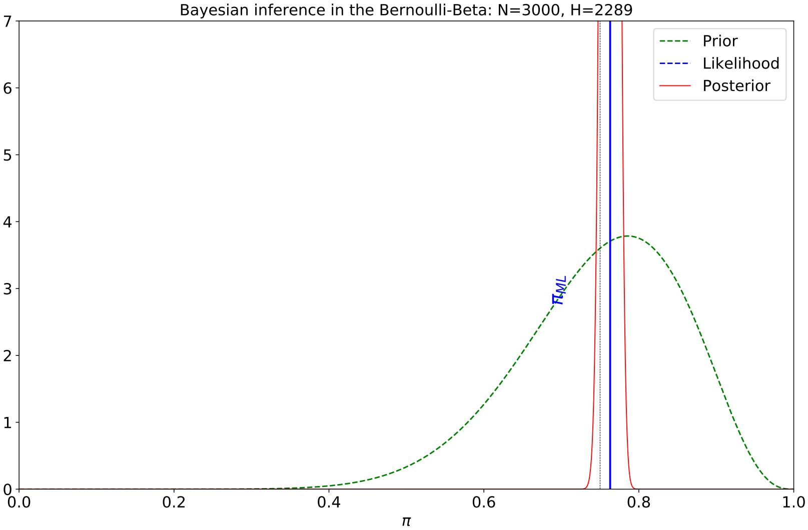
The Bent Coin cont'd



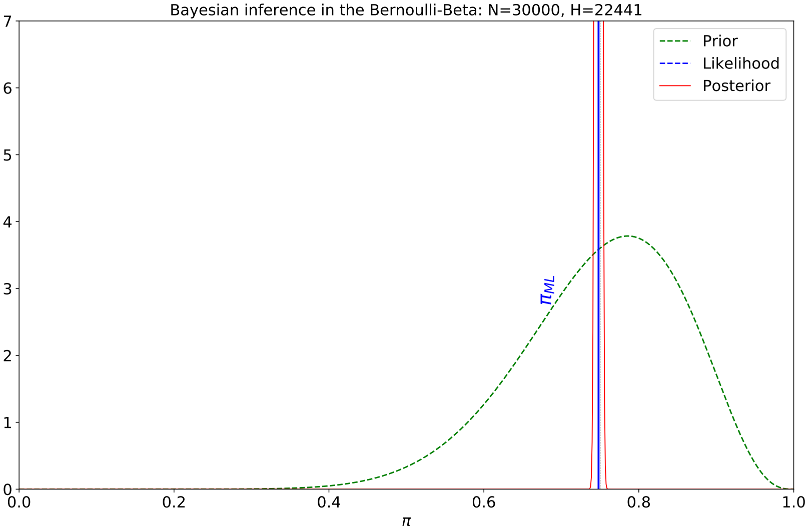
The Bent Coin cont'd



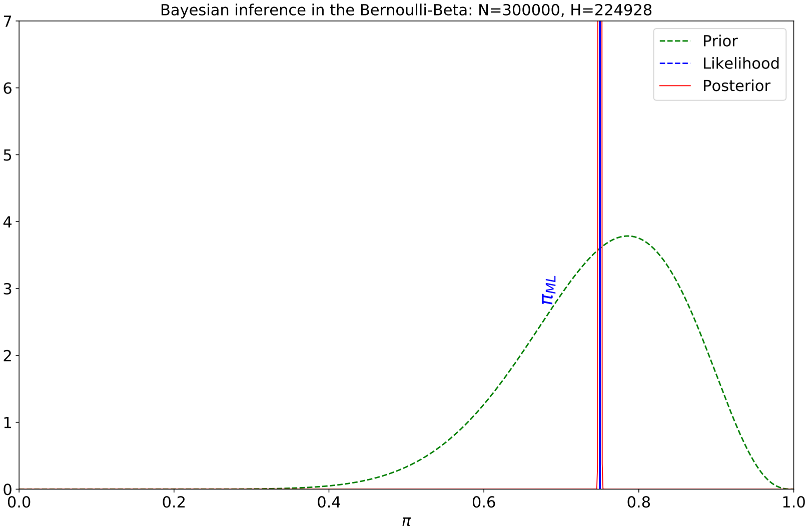
The Bent Coin cont'd



The Bent Coin cont'd

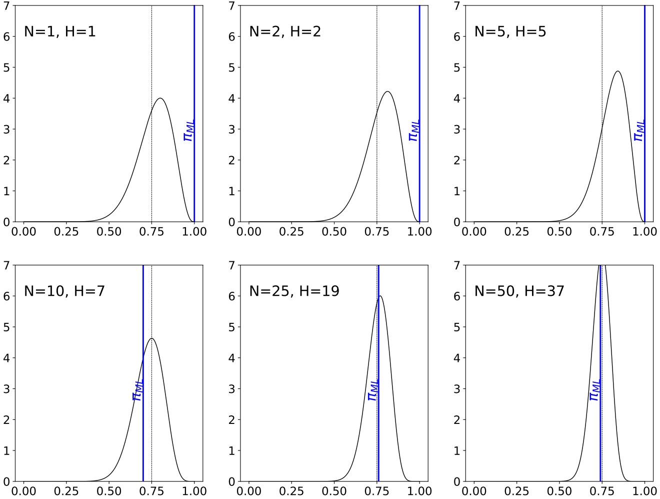


The Bent Coin cont'd



The Bent Coin cont'd

D = [1 1 1 1 1 0 1 0 1 0 1 1 1 1 0 1 0 1 1 1 1 1 1 1 0 1 1 1 0 0 1 0 1 1 0 1 1
1 1 1 1 1 1 1 1 1 1 0 0 0]



The Bent Coin cont'd

$$\boxed{p(\pi|\mathcal{D}) = \frac{\pi^{H+\alpha-1}(1-\pi)^{N-H+\beta-1}}{Z(H+\alpha, N-H+\beta)}}$$

A couple of things to notice:

- The posterior was a Beta distribution, just like the prior. This property is called *conjugacy*. (More on this later)
- The *MAP* (peak of the posterior) is at

  $$\pi_{\mathsf{MAP}} = \frac{H+\alpha-1}{N+\alpha+\beta-2} = \frac{\frac{H}{N}+\frac{\alpha-1}{N}}{1+\frac{\alpha+\beta-2}{N}} = \frac{\pi_{\mathsf{ML}}+\frac{\alpha-1}{N}}{1+\frac{\alpha+\beta-2}{N}}$$

  If you now take the limit $N \to \infty$, $\lim_{N\to\infty} \pi_{\mathsf{MAP}} = \pi_{\mathsf{ML}}$. So ML and MAP match in the limit. Some would argue that maximum likelihood is a limit case of Bayesian inference.
- The parameters α and β acts like extra data! They are called *pseudocounts*, with *effective sample size* $\alpha + \beta$.

The Bent Coin cont'd

In general, the probability of the next toss is

$$p(x_* = \mathsf{head}|\mathcal{D}, \alpha, \beta) = \frac{H + \alpha + 1}{N + \alpha + \beta + 2} = \frac{\pi_{\mathsf{ML}} + \frac{\alpha+1}{N}}{1 + \frac{\alpha+\beta+2}{N}}$$

Compare this with the prediction from the maximum likelihood π_{ML}:

$$p(x_* = \mathsf{head}|\pi_{\mathsf{ML}}) = \frac{H}{N}$$

Or maximum *a posteriori* π_{MAP}:

$$p(x_* = \mathsf{head}|\pi_{\mathsf{MAP}}) = \frac{\pi_{\mathsf{ML}} + \frac{\alpha-1}{N}}{1 + \frac{\alpha+\beta-2}{N}}$$

We have 3 different solutions, which should we prefer?

Conjugate priors

For a Bernoulli likelihood we can have a Beta prior and posterior. This property is known as *conjugacy*.

Not all likelihoods admit conjugate priors: only distributions in the *exponential family*. Here is a list of likelihoods and their conjugate priors

Distribution	Parameter	Prior
Gaussian	Mean	Gaussian
Gaussian	Variance	Inverse-Gamma
Bernoulli	Mean	Beta
Poisson	Rate (λ)	Gamma
Exponential	Rate (λ)	Gamma
Uniform	Upper limit	Pareto

We like distributions which admit a conjugate prior, because we can find the posterior parameters very efficiently given simple statistics about the input data. Furthermore the posterior distributions are tractable.

Gaussian example

Consider a Gaussian likelihood over N datapoints. To make life easier, we shall reparameterize using the precision $\lambda = 1/\sigma^2$, this can be written

$$\begin{aligned}\prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \lambda^{-1}) &= \prod_{i=1}^{N} \frac{\lambda}{\sqrt{2\pi}} \exp\left\{-\frac{\lambda(x_i-\mu)^2}{2}\right\} \\ &= \left(\sqrt{\frac{\lambda}{2\pi}}\right)^{N} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}\end{aligned}$$

Let's compute the posterior distribution over the precision, given a conjugate Gamma prior and *fixed mean*

$$p(\lambda|\alpha, \beta) = \frac{1}{Z(\alpha, \beta)} \lambda^{\alpha-1} \exp\left\{-\beta\lambda\right\}$$

Gaussian example

$$
\begin{aligned}
p(\lambda|\alpha,\beta,\mu,\{x_i\}) &\propto \left[\prod_{i=1}^{N}\mathcal{N}(x_i|\mu,\lambda^{-1})\right]p(\lambda|\alpha,\beta) \\
&= \left(\frac{\lambda}{2\pi}\right)^{\frac{N}{2}}\exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}\frac{1}{Z(\alpha,\beta)}\lambda^{\alpha-1}\exp\left\{-\beta\lambda\right\} \\
&\propto \lambda^{\frac{N}{2}+\alpha-1}\exp\left\{-\lambda\left(\beta+\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right)\right\}
\end{aligned}
$$

This is a new Gamma distribution $p(\lambda|\alpha',\beta')$ with

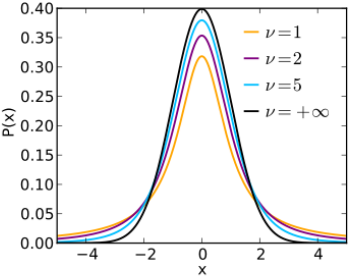
$$\alpha' = \frac{N}{2}+\alpha \qquad \beta' = \beta + \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2$$

Gaussian example

Let's look at the posterior predictive distribution.

$$p(x_*|\alpha, \beta, \mu, \{x_i\}) = \int \mathcal{N}(x_*|\mu, \lambda^{-1}) p(\lambda|\alpha', \beta') \,\mathrm{d}\lambda$$

This is in fact the definition of the Student-t distribution. Note that the posterior predictive is not generally the same form as the prior or the forward likelihood.



Models of this form are called Gaussian scale mixtures. They are useful in modelling because they have significantly heavier tails than the Gaussian distribution, which make them robust to outliers.

Categorical-Dirichlet example

The categorical distribution is the D-way generalization of the Bernoulli distribution. It has the form

$$p(\mathbf{x}|\boldsymbol{\pi}) = \pi_1^{x_1}\pi_2^{x_2}\cdots\pi_N^{x_D} = \prod_{d=1}^{D} \pi_d^{x_d}, \qquad \sum_{d=1}^{D} \pi_d = 1$$

where $\mathbf{x}$ is *one-hot* i.e. the d^{th} entry is 1, the rest are 0. So $p(x_d|\boldsymbol{\pi}) = \pi_d$. This is typically used for modelling the distribution of words in sentences, where each word is associated with a label x_d.

The conjugate prior to the categorical is the *Dirichlet distribution*

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})}\prod_{d=1}^{D} \pi_d^{\alpha_d - 1}.$$

The *hyperparameter* $\boldsymbol{\alpha}$ controls the shape of the Dirichlet. Each element of $\boldsymbol{\alpha}$ must satisfy $\alpha_d > 0$

Categorical-Dirichlet example

If for each category we get counts $\mathbf{N} = [N_1, N_2, ..., N_D]^\top$, where $N = \sum_{d=1}^{D} N_d$, the likelihood is

$$\prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\pi}) = \prod_{i=1}^{N} \left(\prod_{d=1}^{D} \pi_d^{x_{d,i}} \right) = \prod_{d=1}^{D} \pi_d^{N_d}$$

The posterior distribution is then

$$p(\boldsymbol{\pi}|\{\mathbf{x}_i\}, \boldsymbol{\alpha}) \propto \left[\prod_{d=1}^{D} \pi_d^{N_d} \right] \frac{1}{Z(\boldsymbol{\alpha})} \prod_{d=1}^{D} \pi_d^{\alpha_d - 1} \propto \prod_{d=1}^{D} \pi_d^{N_d} \pi_d^{\alpha_d - 1} = \prod_{d=1}^{D} \pi_d^{N_d + \alpha_d - 1}$$

which has the form of a Dirichlet distribution

$$p(\boldsymbol{\pi}|\{\mathbf{x}_i\}, \boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\mathbf{N} + \boldsymbol{\alpha})$$

Categorical-Dirichlet example

The posterior predictive is fiddly to work out. It is useful to know that the mean of the d^{th} dimension is $\alpha_d / \sum_{d=1}^{D} \alpha_d$

$$\begin{aligned}
p(x_{*,d} | \{\mathbf{x}_i\}, \boldsymbol{\alpha}) &= \int_{\Delta} p(x_{*,d} | \boldsymbol{\pi}) p(\boldsymbol{\pi} | \{\mathbf{x}_i\}, \boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\pi} \\
&= \int_{\Delta} \pi_{*,d} \text{Dirichlet}(\boldsymbol{\pi} | \mathbf{N} + \boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\pi} \\
&= \mathbb{E}_{\pi_d} [\text{Dirichlet}(\boldsymbol{\pi} | \mathbf{N} + \boldsymbol{\alpha})] \\
&= \frac{N_d + \alpha_d}{\sum_{d=1}^{D} N_d + \alpha_d}
\end{aligned}$$

We see the role of the prior is to add *pseudocounts* to shift the posterior predictive estimates away from the maximum likelihood solution.

Bayesian Inference Summary

We wanted to compute two quantities

1. The posterior distribution $p(\pi|\mathcal{D})$
   - Use Bayes' theorem for $p(\pi|\mathcal{D}) = \frac{p(\mathcal{D}|\pi)p(\pi)}{p(\mathcal{D})}$
   - Write out the numerator $p(\mathcal{D}|\pi)p(\pi)$
   - Isolate terms in π and match to the form of the prior
   - Optionally normalize by $p(\mathcal{D})$, which is the marginal of the numerator $p(\mathcal{D}) = \int p(\mathcal{D}|\pi)p(\pi)\,\mathrm{d}\pi$
2. The probability distribution of a test sample $\mathbf{x}_*$
   - This is $p(\mathbf{x}_*|\mathcal{D})$
   - We expand it as $p(\mathbf{x}_*|\mathcal{D}) = \int p(\mathbf{x}_*|\pi)p(\pi|\mathcal{D})\,\mathrm{d}\pi$
   - This is called the *posterior predictive distribution*

The Exponential Family*

Which distributions have analytical maximum likelihood solutions?

Most of the distributions we have looked at are fairly similar.

They have three main components:

$$\boxed{p(x|\boldsymbol{\theta}) = \underbrace{\frac{1}{Z(\boldsymbol{\theta})}}_{\text{normalizer}} \cdot \overbrace{b(x)}^{\text{fnc of } x} \cdot \underbrace{\exp\left\{\boldsymbol{\theta}^\top \mathbf{t}(x)\right\}}_{\text{exp of linear fnc of } \boldsymbol{\theta}}}$$

$\boldsymbol{\theta}$ is called the *natural parameters*, and $\mathbf{t}(x)$ is called the *sufficient statistics* of the distribution.

Conjugate Exponential Priors*

A natural choice of prior for the exponential family is

$$\boxed{p(\boldsymbol{\theta}|\boldsymbol{\tau},\nu) = \underbrace{\frac{1}{Z(\boldsymbol{\tau},\nu)}}_{\text{normalizer}} \cdot \overbrace{\left(\frac{1}{Z(\boldsymbol{\theta})}\right)^{\nu}}^{\text{not a normalizer}} \cdot \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{\tau}\right\}}$$

$\boldsymbol{\tau}$ and ν are hyperparameters called the *pseudo-observations* and *scale* respectively.

Conjugate exponential posterior Let's start by looking at the likelihood

$$\begin{aligned}\prod_{i=1}^N p(x_i|\boldsymbol{\theta}) &= \prod_{i=1}^N \frac{1}{Z(\boldsymbol{\theta})} b(x) \exp\left\{\boldsymbol{\theta}^\top \mathbf{t}(x)\right\} \\ &= \left(\frac{1}{Z(\boldsymbol{\theta})}\right)^N \left[\prod_{i=1}^N b(x_i)\right] \exp\left\{\boldsymbol{\theta}^\top \sum_{i=1}^N \mathbf{t}(x_i)\right\}\end{aligned}$$

Conjugate Exponential Priors*

So the posterior is

$$
\begin{aligned}
p(\boldsymbol{\theta}|\{x_i\}, \boldsymbol{\tau}, \nu) &\propto \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\tau}, \nu) \\
&= \left(\frac{1}{Z(\boldsymbol{\theta})}\right)^{N} \left[\prod_{i=1}^{N} b(x_i)\right] \exp\left\{\boldsymbol{\theta}^\top \sum_{i=1}^{N} \mathbf{t}(x_i)\right\} \cdot \frac{1}{Z(\boldsymbol{\tau}, \nu)} \left(\frac{1}{Z(\boldsymbol{\theta})}\right)^{\nu} \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{\tau}\right\} \\
&\propto \left(\frac{1}{Z(\boldsymbol{\theta})}\right)^{N+\nu} \exp\left\{\boldsymbol{\theta}^\top \left(\boldsymbol{\tau} + \sum_{i=1}^{N} \mathbf{t}(x_i)\right)\right\}
\end{aligned}
$$

This is of the same form as the prior, so we know the normalizer is $Z(\boldsymbol{\tau} + \sum_{i=1}^{N} \mathbf{t}(x_i), \nu + N)$, thus

$$
\boxed{p(\boldsymbol{\theta}|\{x_i\}, \boldsymbol{\tau}, \nu) = \frac{Z(\boldsymbol{\theta})^{-(N+\nu)}}{Z(\boldsymbol{\tau} + \sum_{i=1}^{N} \mathbf{t}(x_i), \nu + N)} \exp\left\{\boldsymbol{\theta}^\top \left(\boldsymbol{\tau} + \sum_{i=1}^{N} \mathbf{t}(x_i)\right)\right\}}
$$