ELE8088: Control & Estimation Theory **QUB, 2021**

Handout 13: Bayesian Estimation

Lecturer: Pantelis Sopasakis *Date:* _____

Topics: Bayesian Estimation ∘ Conjugate Priors.

13.1 Preliminaries: Parametric Distributions

Firstly we need to introduce some standard parametric distributions which are popular in probability theory and statistics.

13.1.1 Bernoulli

A discrete random variable X supported on $\{0, 1\}$ is said to follow the *Bernoulli distribution* with parameter θ if

$$\mathsf{P}[X = 1] = \theta, \tag{13.1}$$

where $\theta \in [0, 1]$, which implies that $\mathsf{P}[X = 0] = 1 - \theta$. We can write concisely

$$\mathsf{P}[X = k] = \theta^k (1 - \theta)^{1-k}, \tag{13.2}$$

for $k \in \{0, 1\}$. We write $X \sim \text{Ber}(\theta)$.

The Bernoulli distribution is use to model experiments that have only two outcomes (1: success, 0: failure). A coin can be modelled using the Bernoulli distribution.

It is easy to confirm that if $X \sim \text{Ber}(\theta)$, then $\mathbb{E}[X] = \theta$ and $\text{Var}[X] = \theta(1 - \theta)$.

13.1.2 Binomial

Consider an experiment where we toss a coin n times and we count the number of heads we observe — this will be an integer from 0 to n. The number of heads is a random variable which follows the *binomial distribution*.

More specifically, suppose $X_1, X_2, \ldots, X_n$ are independent iid Bernoulli random variables ($X_i \sim \text{Ber}(\theta)$); define $S_n = X_1 + X_2 + \ldots + X_n$. Then S_n follows the binomial distribution with parameters n (number of trials) and θ (probability of success). We write $S_n \sim \text{Binom}(n, \theta)$.

The probability mass function of the binomial distribution is

$$\mathsf{P}[S_n = k] = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \tag{13.3}$$

for $k \in \{0, 1, \ldots, n\}$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \tag{13.4}$$

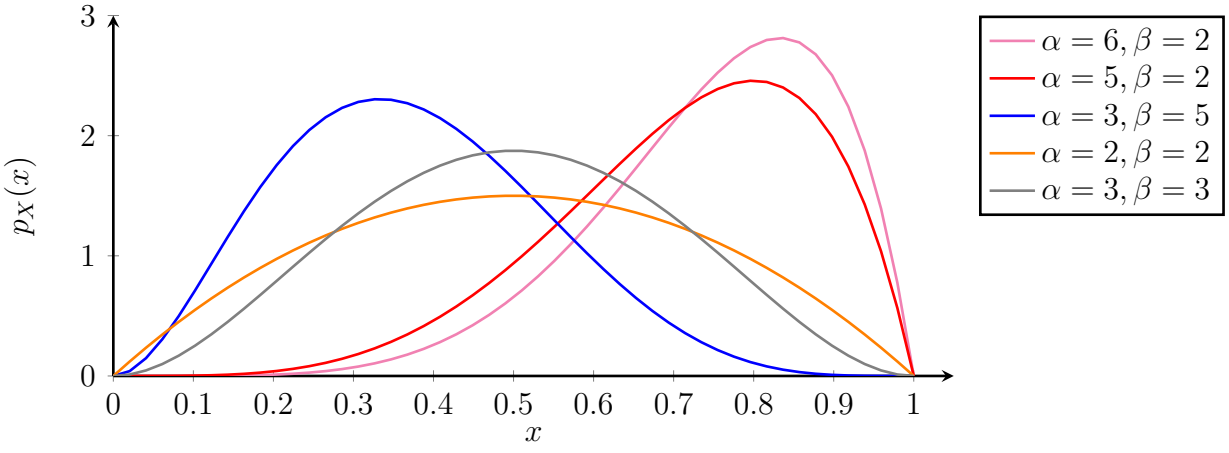
is the *binomial coefficient* (read: "n-choose-k").

13.1.3 Beta

A random variable X, supported on $[0, 1]$, is said to follow the Beta distribution with shape parameters $\alpha, \beta > 0$ — denoted as $X \sim \text{Beta}(\alpha, \beta)$ — if its pdf is

$$p_X(x) \propto x^{\alpha-1}(1-x)^{\beta-1}, \tag{13.5}$$

for $x \in [0, 1]$. In particular, the proportionality coefficient is $\frac{1}{B(\alpha,\beta)}$ where B is the Beta function[1]. The effect of α and β on the shape of the pdf is illustrated below.



Note that Beta$(1, 1)$ is the uniform distribution on $[0, 1]$.

[1]The Beta function is defined as $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t$ and has the property $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$.

If $X \sim \text{Beta}(\alpha, \beta)$, then

$$\mathbb{E}[X] = \frac{\alpha}{\alpha + \beta}, \tag{13.6}$$

and the mode of X is

$$\text{mode}(X) = \underset{x \in [0,1]}{\arg\max}\, p_X(x) = \frac{\alpha - 1}{\alpha + \beta - 2}, \tag{13.7}$$

for $\alpha, \beta > 1$.

13.1.4 Poisson

A random variable X, supported on $\mathbb{N}$, is said to follow the Poisson distribution with parameter $\lambda > 0$ if

$$\mathsf{P}[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{13.8}$$

for $k \in \mathbb{N}$. For the Poisson distribution we have $\mathbb{E}[X] = \lambda$ and $\text{Var}[X] = \lambda$.

The Poisson distribution is used to model the number of occurences of events in a certain time period, if the events happen at a constant mean rate (events/time) and the probability of the occurence of an event is not conditioned by the occurence of another event. Examples where the Poisson distribution can be used include (i) the number of meteorites hitting the Earth in a year, (ii) the number of clients arriving in a shop in a day, and (iii) the number of goals in a football match.

13.1.5 Gamma

The gamma distribution is a parametric continuous distribution, supported on $(0, \infty)$, with two parameters: (i) $\alpha > 0$, which is referred to as the *shape parameter* and (ii) $\beta > 0$, which is known as the *rate parameter*. The pdf of the gamma distribution, $\Gamma(\alpha, \beta)$ is

$$p_X(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\beta x}, \tag{13.9}$$

for $x > 0$. We can simply write

$$p_X(x; \alpha, \beta) \propto x^{\alpha-1} e^{-\beta x}. \tag{13.10}$$

Exercise 1 (☕). Let $X \sim \Gamma(\alpha, \beta)$. Show that

$$\mathbb{E}[X] = \frac{\alpha}{\beta}, \tag{13.11a}$$

$$\mathrm{Var}[X] = \frac{\alpha}{\beta^2}. \tag{13.11b}$$

Exercise 2 (☕). Let $X \sim \Gamma(\alpha, \beta)$ with $\alpha \geq 1$. Show that the mode of X is

$$\mathrm{mode}(X) = \arg\max_{x>0} p_X(x) = \frac{\alpha - 1}{\beta}. \tag{13.12}$$

13.1.6 Normal

We say that a real-valued random variable follows the normal distribution with mean μ and variance σ^2 if its pdf is

$$p_X(x) \propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{13.13}$$

We denote this by $X \sim \mathcal{N}(\mu, \sigma^2)$ and $\mathbb{E}[X] = \mu$, $\mathrm{Var}[X] = \sigma^2$.

We introduced the multivariate normal distribution, $\mathcal{N}(\mu, \Sigma)$, in Handout 11. Let us just recall that the multivariate normal pdf is

$$p_X(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{n/2}} \exp\left[-\tfrac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)\right], \tag{13.14}$$

that is,

$$p_X(x) \propto \exp\left[-\tfrac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)\right]. \tag{13.15}$$

13.2 Bayesian estimation

In the Bayesian estimation framework, we use some prior information about an unknown parameter $\theta \in \Theta$ — which we treat as a random variable — and some measurements $x_1, \ldots, x_N$, to update our prior knowledge and produce a *posterior* distribution of θ.

In particular, we assume that $X_1, \ldots, X_N$ are independent random variables and identicially distributed with a pdf (or pmf) that depends parametrically on an unknown parameter θ, i.e., $p(x_i \mid \theta)$. We assume that θ follows a prior distribution, $p(\theta)$, which reflects our prior knowledge. The *likelihood* of the data is the pdf $p(x_1, \ldots, x_N \mid \theta)$ and because of the independence assumption we have

$$p(x_1, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta). \tag{13.16}$$

By Bayes' theorem, the posterior distribution of θ given the observations $x_1, \ldots, x_N$ is

$$p(\theta \mid x_1, \ldots, x_N) \propto p(\theta) \prod_{i=1}^{N} p(x_i \mid \theta). \tag{13.17}$$

Then we can do a lot with the posterior distribution... For example, we can estimate θ by extracting a value from the posterior distribution. If we take the *mode* of the posterior, we will have precisely a *maximum a posteriori* (MAP) estimate, that is,

$$\widehat{\theta}_{\text{map}}(x_1, \ldots, x_N) \in \arg\max_{\theta \in \Theta} p(\theta \mid x_1, \ldots, x_N). \tag{13.18}$$

But the Bayesian estimation framework does not only give us an *estimate* (a point), but a probability distribution, which is a lot more informative.

In summary, the Bayesian estimation approach consists in:

1. Treating $\theta \in \Theta$ as a *random variable*, which follows a certain *prior* distribution, $p(\theta)$
2. Obtaining some *independent* observations, x_i, which follow a distribution with pdf (or pmf) $p(x_i \mid \theta)$

3. Using Bayes' formula in Equation (13.17) to determine the posterior — yes, it is just a multiplication!

How do we know the prior? Examples: (i) we have a regular coin; it is reasonable to assume that it is described by $\text{Ber}(\theta)$ where θ is more likely to be close to 0.5, rather than 0.95 or 0.05. We can use something like $\theta \sim \text{Beta}(2, 2)$, or, even better, we can ask the mint whether they have any idea about θ. (ii) we want to find the percentage of men in a population; it makes sense to use some prior that encodes the fact that according to experience and previous studies, this percentage is close to 50%. (iii) we want to know our position on the map and we have some GPS measurements (which, of course, are corrupted by noise); we can use some prior information that encodes the fact that we are standing in the street and we are not inside a building. The choice of the prior is a matter of intuition, availability of prior information and convenience. Indeed, some priors lend themselves to easier derivations — these are known as *conjugate priors* (see Section 13.3).

What if I have no prior information? Then, don't use the Bayesian approach[2].

Spread of the posterior. Lastly, having observed some data $x_1, \ldots, x_N$, we are interested in the dispersion of the posterior distribution of θ; it will be interesting to look at the (conditional) variance of θ (given $x_1, \ldots, x_N$), which of course depends on the prior. We will also be interested in determining regions (sets), R_α, such that $\mathsf{P}[\theta \in R_\alpha \mid x_1, \ldots, x_N] = 1-\alpha$, where $\alpha \in (0, 1)$. Such sets are known as *credible regions* (not "confidence regions"). If θ is a scalar parameter, we can define a credible interval, $[a, b]$, as one for which

$$\int_a^b p(\theta \mid x_1, \ldots, x_N)\mathrm{d}\theta = 1 - \alpha. \tag{13.19}$$

Note that credible intervals are defined for a particular collection of observations, $x_1, \ldots, x_N$.

[2]Have a look at this interesting discussion: https://stats.stackexchange.com/q/326484

13.3 Conjugate Priors

13.3.1 Bernoulli and Binomial Distributions

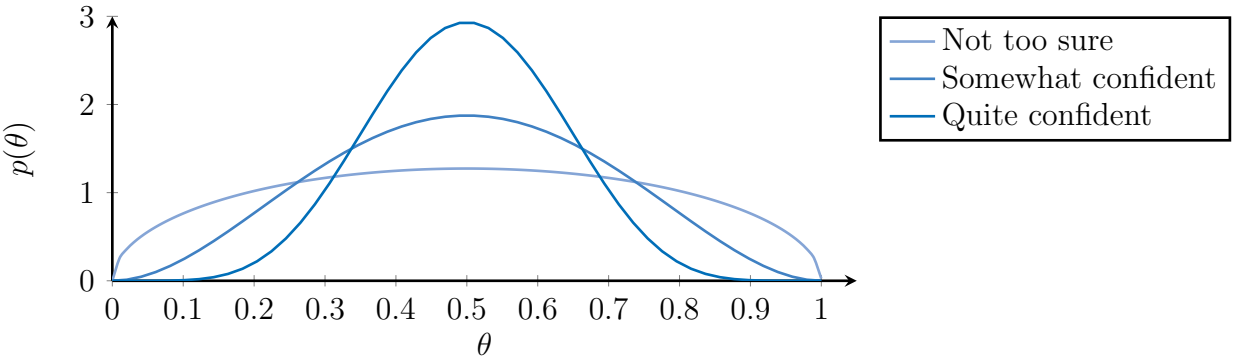
Example 1 (Estimation of Bernoulli parameter with MAP with Beta prior). We want to determine whether a coin is fair; to that we obtain a set of N *independent* measurements (where heads is 1 and tails is 0), $X_1, X_2, \ldots, X_N$. Every coin toss is a Bernoulli random variable with parameter θ, that is $X_i \sim \text{Ber}(\theta)$, i.e.,

$$\mathsf{P}[\text{heads}] = \mathsf{P}[X = 1] = \theta, \tag{13.20}$$

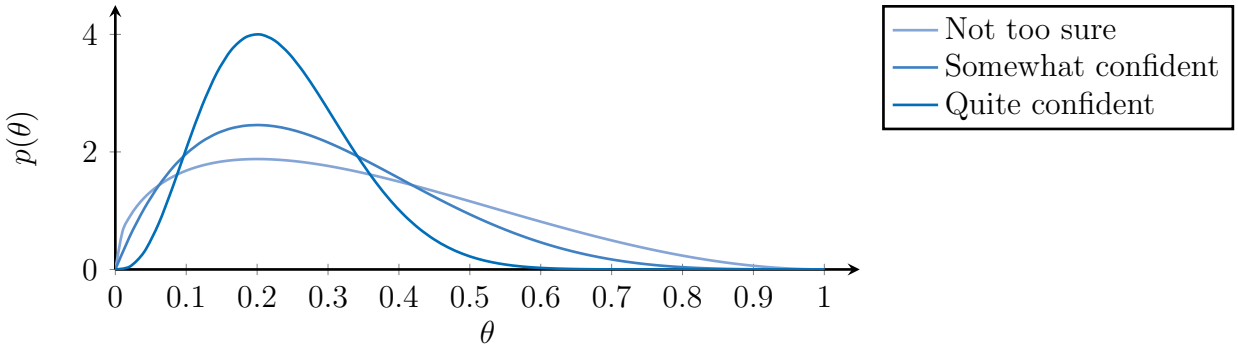
or, what is the same,

$$\mathsf{P}[X = k] = \theta^k (1-\theta)^{1-k}, \ k \in \{0, 1\}. \tag{13.21}$$

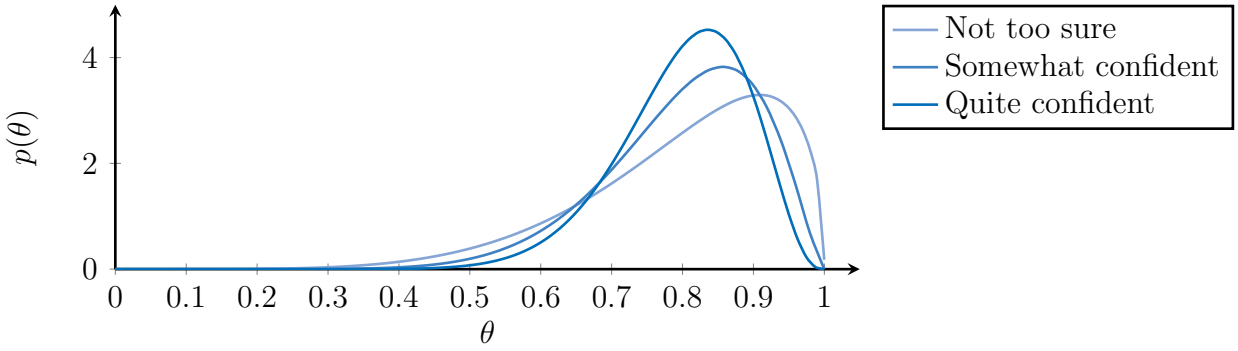
Suppose also that we have some prior information about θ. For example, we may believe that θ is about 0.5; we can describe this by using a prior distribution for θ that can look like this:



Or we may have reasons to belive (to a greater or lesser extent) that θ follows a distribution with mode 0.2 in which case it may make sense to choose a prior like this:



Or we may believe (to a greater or lesser extent) that the expected value of θ is 0.8 in which case the pdf of θ can look like this:



We need to choose a parametric distribution for our prior which is flexible enough, but will not lead to an overly complex MAP estimation problem. In particular, the posterior distribution, that is, $p(\theta \mid x_1, \ldots, x_N) \propto p(x_1, \ldots, x_N \mid \theta)p(\theta)$, should be of a "convenient" form.

In this case, it turns out that the Beta distribution is an appropriate prior. Firstly, we can choose α and β to shape $p(\theta)$ as shown above. Secondly, if we assume that $\theta \sim \text{Beta}(\alpha, \beta)$, i.e., $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, then the MAP estimate of θ becomes

$$\widehat{\theta}_{\text{map}}(x_1, \ldots, x_N) = \arg\max_{\theta \in [0,1]} p(x_1, \ldots, x_N \mid \theta)p(\theta), \tag{13.22}$$

where

$$\begin{aligned} p(\theta \mid x_1, \ldots, x_N) &\propto p(x_1, \ldots, x_N \mid \theta)p(\theta) \\ &\propto \prod_{i=1}^{N} p(x_i \mid \theta) \cdot p(\theta) \end{aligned}$$

$$
\begin{aligned}
&\propto \prod_{i=1}^{N} \underbrace{\theta^{x_i}(1-\theta)^{1-x_i}}_{\text{Ber}(\theta)} \cdot \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{Beta}(\alpha,\beta)} \\
&= \theta^{\sum_i x_i+\alpha-1}(1-\theta)^{\sum_i(1-x_i)+\beta-1} \\
&= \theta^{\sum_i x_i+\alpha-1}(1-\theta)^{N-\sum_i x_i+\beta-1} \\
&= \text{Beta}\left(\sum_i x_i+\alpha, N-\sum_i x_i+\beta\right). \qquad (13.23)
\end{aligned}
$$

We see that if we choose the prior to be a Beta distribution, the posterior is also a Beta distribution! We say that the Beta distribution is a **conjugate prior** of the Bernoulli distribution for the parameter θ. In brief, we showed that

Theorem 13.1 (Bernoulli with Beta prior) *Suppose* $X_1, \ldots, X_N \overset{iid}{\sim} \text{Ber}(\theta)$ *and* $\theta \sim \text{Beta}(\alpha, \beta)$. *Then,*

$$\theta \mid x_1, \ldots, x_N \sim \text{Beta}\left(\sum_i x_i+\alpha, N-\sum_i x_i+\beta\right), \qquad (13.24)$$

and the MAP estimate of θ *given* $x_1, \ldots, x_N$ *is*

$$\widehat{\theta}_{\text{map}}(x_1, \ldots, x_N) = \frac{\sum_i x_i+\alpha-1}{\alpha+\beta+N-2}. \qquad (13.25)$$

With regards to the MAP estimate, from Equation (13.7) we have that if both parameters of the Beta distribution in Equation (13.23) are larger than 1, the MAP estimate of θ given $x_1, x_2, \ldots, x_N$ is

$$\widehat{\theta}_{\text{map}}(x_1, \ldots, x_N) = \frac{\sum_i x_i+\alpha-1}{\cancel{\sum_i x_i}+\alpha+N-\cancel{\sum_i x_i}+\beta-2} = \frac{\sum_i x_i+\alpha-1}{\alpha+\beta+N-2} \qquad (13.26)$$

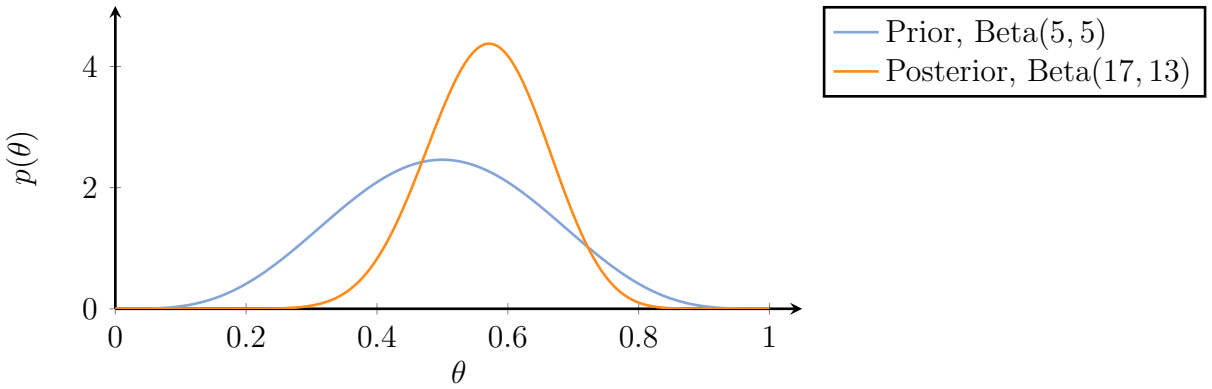
In addition, the fact that the posterior is of the same type as the prior — both are beta distributions — allows us to compare them. We see that

$$\text{Beta}(\alpha, \beta) \xrightarrow{x_1,\ldots,x_N} \text{Beta}\left(\sum_i x_i+\alpha, N-\sum_i x_i+\beta\right). \qquad (13.27)$$

Example 2 (Prior to Posterior). We have a coin which we believe to be fair. In particular, we believe that its outcome, X, follows a Bernoulli distribution with parameter θ which follows the Beta distribution, Beta$(5, 5)$. We tossed the coin 20 times and we observed 12 heads. The posterior distribution of θ given the observations is

$$\theta \mid x_1, \ldots, x_{20} \sim \text{Beta}(12 + 5, 20 - 12 + 5) = \text{Beta}(17, 13). \tag{13.28}$$

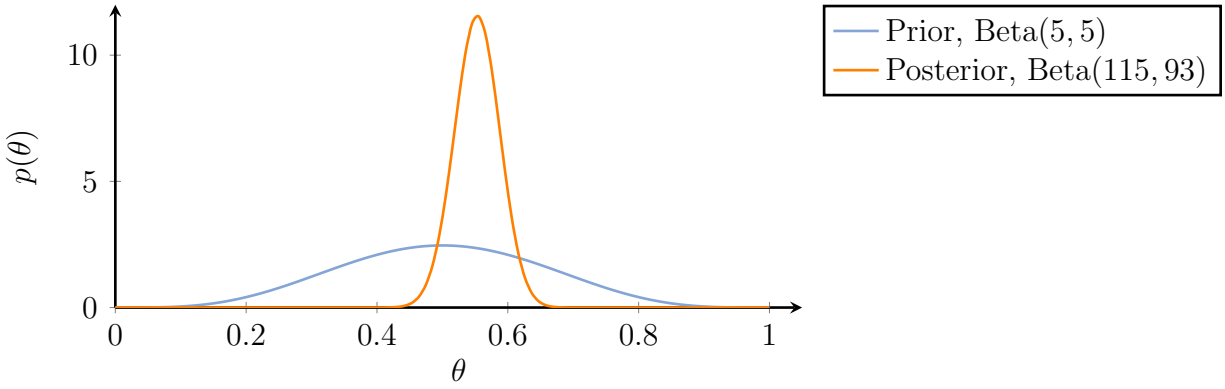
The posterior is shown below and the MAP estimate of θ is $\widehat{\theta}_{\text{map}}(x_1, \ldots, x_{20}) = 0.5714$.



Now suppose that we flip the coin $N = 200$ times and we observe 110 heads. Then the posterior is

$$\theta \mid x_1, \ldots, x_{200} \sim \text{Beta}(110 + 5, 200 - 112 + 5) = \text{Beta}(115, 93), \tag{13.29}$$

which is shown below.



We observe that as we accumulate more data, the posterior becomes more narrow. For comparison,

N	Posterior	MAP Estimate	Posterior Variance
20	$\text{Beta}(17, 13)$	0.5714	0.00792
200	$\text{Beta}(115, 93)$	0.5534	0.00118

Example 3 (Binomial with Beta prior). Suppose that $X \sim \text{Binom}(n, \theta)$ where n is known and fixed and we need to estimate θ from independent observations. Let us assume that we have some prior knowledge about θ, which is described by the prior $\theta \sim \text{Beta}(\alpha, \beta)$. We will show that $\theta \mid x_1, \ldots, x_N$ follows a Beta distribution. The reader can follow the same procedure as above to show that

Theorem 13.2 (Binomial with Beta prior) *Suppose that* $X_1, \ldots, X_N \overset{iid}{\sim} \text{Binom}(n, \theta)$, *where* n *is known and* $\theta \sim \text{Beta}(\alpha, \beta)$. *Then,*

$$\theta \mid x_1, \ldots, x_N \sim \text{Beta}\left(\sum_i x_i + \alpha, nN - \sum_i x_i + \beta\right), \tag{13.30}$$

Exercise 3 (☕). What is the maximum *a posteriori* estimate of θ? Use Equation (13.30).

Exercise 4 (☕). Suppose that $(x_i)_{i=1}^N$ are independent binomial random variables with $x_i \sim \text{Binom}(n_i, \theta)$, where n_i are known and $\theta \sim \text{Beta}(\alpha, \beta)$. Determine the posterior distribution, $p(\theta \mid x_1, \ldots, x_N)$.

13.3.2 Poisson with gamma prior

Suppose that $X_1, X_2, \ldots, X_N$ are independently identically distributed and follow the Poisson distribution with an unknown parameter $\lambda > 0$. Suppose that we have the prior infromation

$\lambda \sim \Gamma(\alpha, \beta)$ with $\alpha, \beta > 0$. We will show that the posterior is another Gamma distribution. In other words, the Gamma distribution is a *conjugate prior* of Poisson for the parameter $\lambda > 0$.

Theorem 13.3 (Poisson with gamma prior) *Suppose that* $X_1, \ldots, X_N \overset{iid}{\sim} \text{Poisson}(\lambda)$, *where* $\lambda \sim \Gamma(\alpha, \beta)$. *Then,*

$$\lambda \mid x_1, \ldots, x_N \sim \Gamma\left(\sum_{i=1}^{N} x_i + \alpha, N + \beta\right). \tag{13.31}$$

Proof: The posterior distribution is

$$\begin{aligned} p(\lambda \mid x_1, \ldots, x_N) &\propto p(x_1, \ldots, x_N \mid \lambda) p(\lambda) \\ &\propto \prod_{i=1}^{N} p(x_i \mid \lambda) \cdot p(\lambda) \\ &\propto \prod_{i=1}^{N} \underbrace{e^{-\lambda} \lambda^{x_i}}_{\text{Poisson}(\lambda)} \cdot \underbrace{\lambda^{\alpha-1} e^{-\beta\lambda}}_{\Gamma(\alpha,\beta)} \\ &= e^{-N\lambda} \lambda^{\sum_i x_i} \lambda^{\alpha-1} e^{-\beta\lambda} \\ &= e^{-N\lambda-\beta\lambda} \lambda^{\sum_i x_i + \alpha - 1} = \Gamma\left(\sum_i x_i + \alpha, N + \beta\right), \end{aligned} \tag{13.32}$$

which completes the proof. ■

Note that the variance of the posterior, according to Equation (13.11b), is

$$\text{Var}[\lambda \mid x_1, \ldots, x_N] = \frac{\sum_{i=1}^{N} x_i + \alpha}{(N + \beta)^2}, \tag{13.33}$$

which goes to 0 as $N \to \infty$[3].

Exercise 5 (☕). What is the MAP estimate of the Poisson parameter λ given the independent observations $x_1, \ldots, x_N$ and using the prior $\lambda \sim \Gamma(\alpha, \beta)$?

[3] This statement is not rigorous, but we will skip the details.

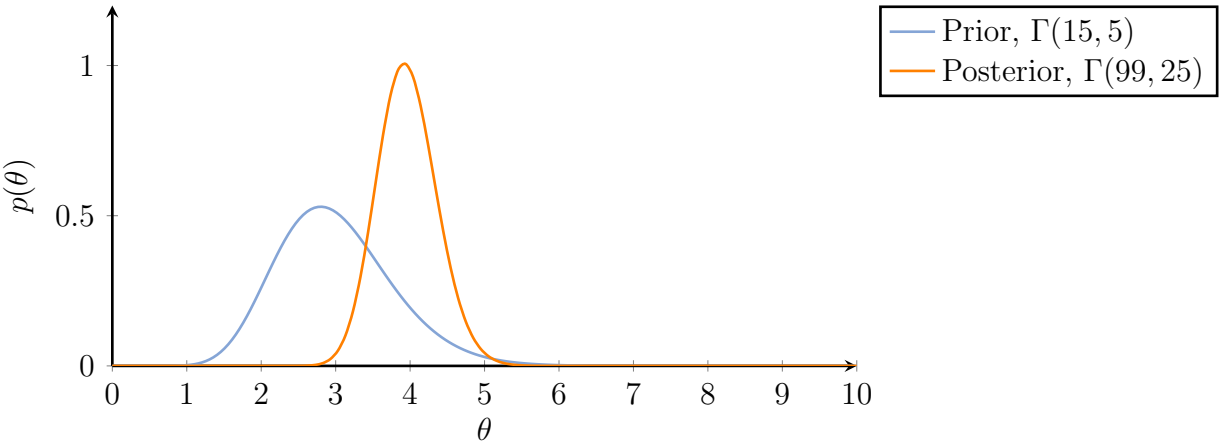
Example 4 (Poisson with Gamma prior). We have obtained the following twenty measurements from a Poisson distribution with an unknown parameter $\lambda > 0$:

$$4, 7, 0, 5, 3, 4, 4, 3, 3, 411, 2, 2, 8, 2, 7, 6, 5, 3, 1. \tag{13.34}$$

Suppose we have the prior infromation $\lambda \sim \Gamma(15, 5)$. Following Theorem 13.3, the posterior distribution of λ given $x_1, \ldots, x_{20}$ is

$$\lambda \mid x_1, \ldots, x_{20} \sim \Gamma\left(\sum_{i=1}^{N} x_i + \alpha, N + \beta\right) = \Gamma(99, 25). \tag{13.35}$$

The prior and posterior distributions are shown below.



The MAP estimate of θ given the above measurements is the mode of $\Gamma(99, 25)$, the determination of which is left to the reader as an exercise.

Exercise 6 (Gamma with gamma prior for β). Suppose that $X \sim \Gamma(\alpha, \beta)$ where $\alpha > 0$ is known and $\beta \sim \Gamma(\alpha_0, \beta_0)$. Show that β follows a gamma distribution and determine its parameters.

Exercise 7 (Exponential with gamma prior). The pdf of an exponential distribution with parameter $\lambda > 0$ ($\text{Exp}(\lambda)$) is $p_X(x) = \lambda e^{-\lambda x}$, defined for $x \geq 0$. Suppose that $X_1, \ldots, X_N$ are independent samples from $\text{Exp}(\lambda)$ and we assume that $\lambda \sim \Gamma(\alpha, \beta)$. Determine a MAP estimate for λ.

13.3.3 Normal with normal prior

Suppose that $X_1, \ldots, X_N$ are independent samples that follow $\mathcal{N}(\mu, \sigma^2)$ with a known variance σ^2 and unknown mean μ. We will assume that $\mu \sim \mathcal{N}(\mu_0, \sigma_\mu^2)$. Then, the posterior distribution of μ given observations $x_1, \ldots, x_N$ is

$$
\begin{aligned}
p(\mu \mid x_1, \ldots, x_N) &\propto p(\mu) \prod_{i=1}^{N} p(x_i \mid \mu) \\
&\propto \underbrace{\exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_\mu^2}\right]}_{p(\mu)} \underbrace{\prod_{i=1}^{N} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]}_{p(x_1,\ldots,x_N|\mu)} \\
&= \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_\mu^2} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&\propto \exp\left[-\frac{\mu^2 - 2\mu_0\mu}{2\sigma_\mu^2} - \sum_{i=1}^{N} \frac{\mu^2 - 2x_i\mu}{2\sigma^2}\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\sigma_\mu^{-2}\mu^2 - 2\sigma_\mu^{-2}\mu_0\mu + \sigma^{-2}N\mu^2 - 2\sigma^{-2}\mu\sum_i x_i\right)\right] \\
&= \exp\left[-\frac{1}{2}\left((\sigma_\mu^{-2} + \sigma^{-2}N)\mu^2 - 2\left(\sigma_\mu^{-2}\mu_0 + \sigma^{-2}\sum_i x_i\right)\mu\right)\right] \\
&= \exp\left[-\frac{1}{2}(\sigma_\mu^{-2} + \sigma^{-2}N)\left(\mu^2 - 2\frac{\sigma_\mu^{-2}\mu_0 + \sigma^{-2}\sum_i x_i}{\sigma_\mu^{-2} + \sigma^{-2}N}\mu\right)\right] \\
&\propto \exp\left[-\frac{1}{2}(\sigma_\mu^{-2} + \sigma^{-2}N)\left(\mu - \frac{\sigma_\mu^{-2}\mu_0 + \sigma^{-2}\sum_i x_i}{\sigma_\mu^{-2} + \sigma^{-2}N}\right)^2\right] \\
&= \mathcal{N}\left(\frac{\sigma_\mu^{-2}\mu_0 + \sigma^{-2}\sum_i x_i}{\sigma_\mu^{-2} + \sigma^{-2}N}, \frac{1}{\sigma_\mu^{-2} + \sigma^{-2}N}\right)
\end{aligned}
$$

It is interesting to note that the posterior is a normal distribution and that the variance of the posterior is $\mathcal{O}(1/N)$. To put it simply, as we accumulate more data, we become increasingly more certain about our estimate. The reader can confirm that the maximum a posteriori estimate of μ is

$$
\widehat{\mu}_{\text{map}}(x_1, \ldots, x_N) = \frac{\sigma^2\mu_0 + \sigma_\mu^2 \sum_{i=1}^{N} x_i}{N\sigma_\mu^2 + \sigma^2}. \tag{13.36}
$$

Note that $\widehat{\mu}_{\text{map}}$ is a weighted average of the prior μ_0 and the observed sampled mean $\bar{x}_N = \sum_{i=1}^N x_i/N$. It is

$$\widehat{\mu}_{\text{map}}(x_1, \ldots, x_N) = \frac{\sigma^2}{N\sigma_\mu^2 + \sigma^2}\mu_0 + \frac{N\sigma_\mu^2}{N\sigma_\mu^2 + \sigma^2}\bar{x}_N. \tag{13.37}$$

Note also that if N is large,

$$\widehat{\mu}_{\text{map}}(x_1, \ldots, x_N) \approx \bar{x}_N. \tag{13.38}$$

Similar results can be obtained for the multivariate case. In particular,

Theorem 13.4 (Normal with a normal prior) *Suppose* $X_1, \ldots, X_N \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$ *with known* Σ *and* $\mu \sim \mathcal{N}(\mu_0, \Sigma_\mu)$. *Then,*

$$\mu \mid x_1, \ldots, x_N \sim \mathcal{N}\left((\Sigma_\mu^{-1} + N\Sigma^{-1})^{-1}(\Sigma_\mu^{-1}\mu_0 + \Sigma^{-1}S_N), (\Sigma_\mu^{-1} + N\Sigma^{-1})^{-1}\right), \tag{13.39}$$

where $S_N = \sum_{i=1}^N x_i$.

We see that the variance of the posterior converges to 0 at a rate of $\mathcal{O}(1/N)$.