Chapter 3

Bayesian Parameter Estimation

3.1 From Prior to Posterior

In the Bayesian philosophy, unknown parameters are viewed as random. So, our knowledge about the parameter can be encoded as a distribution. The distribution representing our belief before observing the data is called the **prior distribution**. After we observe the data, our belief changes, resulting in the **posterior distribution**.

Specifically, the components of a Bayesian estimation problem are:

- Data $\boldsymbol{x}$: The data is a realization of a random variable $\boldsymbol{X}$. The distribution of $\boldsymbol{X}$ depends on a parameter Θ.
- Parameter Θ: The parameter of the distribution of $\boldsymbol{X}$, which is unknown, and hence a random variable in the Bayesian framework.
- Joint and marginal distributions p: A joint distribution $p_{\boldsymbol{X},\Theta}$ and its marginals $p_{\boldsymbol{X}}$ and p_{Θ}.

The steps of Bayesian estimation of a parameter θ are:

1. Identifying the **prior** distribution, $p_{\Theta}(\theta)$. This is called the prior because it encodes our beliefs about Θ before seeing any data.
2. Collecting **data** $\boldsymbol{x}$ and forming the likelihood: $p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)$
3. Finding the **posterior** distribution $p_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ as

$$p_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) = \frac{p_{\Theta}(\theta)p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)}{p_{\boldsymbol{X}}(\boldsymbol{x})}, \tag{3.1}$$

The distribution $p_{\Theta|\boldsymbol{X}}$ is called the posterior distribution since it encodes our knowledge about the parameter after observing the data. Usually, since the distribution is clear from the argument, we drop the subscripts of p, writing the above equation as

$$p(\theta|\boldsymbol{x}) = \frac{p(\theta)p(\boldsymbol{x}|\theta)}{p(\boldsymbol{x})}, \tag{3.2}$$

Normalizing distributions. Finding the posterior distribution requires computing the integral $p(\boldsymbol{x}) = \int_{\theta} p(\theta)p(\boldsymbol{x}|\theta)d\theta$. Since we have to compute an integral anyway, we might as well drop all multiplicative terms that are constant in θ and then normalize the final distribution. In particular, $p(\boldsymbol{x})$ is one such term. So we often first find a function *proportional* to $p(\theta|\boldsymbol{x})$ as

$$p(\theta|\boldsymbol{x}) \propto p(\theta)p(\boldsymbol{x}|\theta), \tag{3.3}$$

where we can also drop constant terms in θ from $p(\theta)$ and $p(\boldsymbol{x}|\theta)$. We can then normalize the result by integration. This is often difficult to do. Sometimes, given this function, we can identify the distribution. More generally, we can use computational methods, such as Markov Chain Monte Carlo, or approximation methods, such as variational inference, as we will see later. Finally, in certain cases, we can find what we need without any integration. For example, if our goal is to find the value of θ maximizing $p(\theta|\boldsymbol{x})$.

Example 3.1. Let Θ denote the unknown parameter of a geometric random variable X, where

$$p_{X|\Theta}(x|\theta) = \theta(1-\theta)^{x-1}.$$

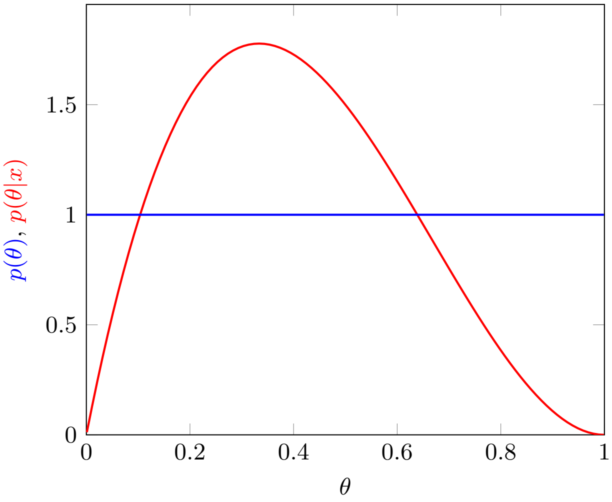
Suppose we observe $X = x$. We would like to estimate Θ based on this observation. If all possible values of Θ are equally likely, we may choose $\Theta \sim \text{Uni}(0,1)$. We then have

$$p(\theta) = 1 \tag{3.4}$$

$$p(x|\theta) = \theta(1-\theta)^{x-1} \tag{3.5}$$

$$p(\theta|x) \propto p(\theta)p(x|\theta) \propto \theta(1-\theta)^{x-1} \tag{3.6}$$

The expression $\theta(1-\theta)^{x-1}$ as a function of x is the geometric distribution. But as a function of θ, it is proportional to $\text{Beta}(2,x)$. As an example, if $x = 3$, then $\Theta|x \sim \text{Beta}(2,3)$:



△

Exercise 3.2. The probability of 1 (success) in a Bernoulli experiment (e.g., flipping a coin, a system working or not working, etc) is Θ, which we would like to estimate. Suppose that the experiment is performed once and the outcome x is observed to be $x = 1$. Assuming a uniform prior, find the posterior distribution of Θ, i.e., $p_{\Theta|X}(\theta|1)$. △

Example 3.3. The probability of success in a Bernoulli experiment is Θ, which we would like to estimate. We show success in the ith trial with $y_i = 1$ and failure by $y_i = 0$.

- Prior distribution: Assuming that a priori we do not know anything about Θ, it is appropriate to choose $p_\Theta \sim \text{Uni}[0,1]$, i.e., $p(\theta) = 1$ in the interval $[0,1]$.
- Likelihood: We then perform the experiment n times. Suppose that we observe s successes and f failures. Let us denote this observation as $\boldsymbol{x} = (s, f)$. The likelihood is

$$p(\boldsymbol{x}|\theta) = \binom{n}{s}\theta^s(1-\theta)^f \tag{3.7}$$

- The posterior distribution:
$$p(\theta|\boldsymbol{x}) \propto 1 \cdot \theta^s(1-\theta)^f = \theta^s(1-\theta)^f \tag{3.8}$$
We observe that this distribution is of the form of a beta distribution, $\text{Beta}(y;\alpha,\beta) \sim y^{\alpha-1}(1-y)^{\beta-1}$. Hence,
$$p(\theta|\boldsymbol{x}) \sim \text{Beta}(s+1, f+1). \tag{3.9}$$

$\triangle$

Note that since we are interested in Θ, we can drop multiplicative terms that are constant with respect to θ, such as $\binom{n}{s}$, in the above example.

Now that we have the posterior distribution, we can answer questions about the parameter, for example, What is the probability that $0.4 < \Theta < 0.6$?

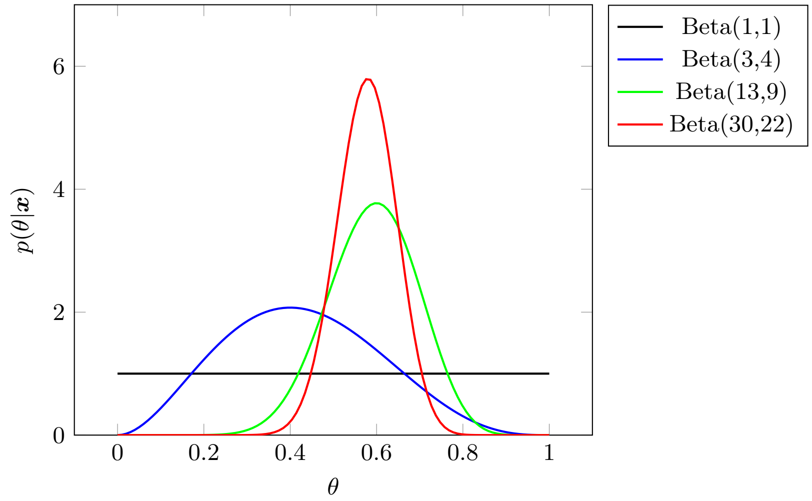
$$\int_{0.4}^{0.6} p(\theta|\boldsymbol{x})d\theta \tag{3.10}$$

Example 3.4 (Consecutive Bayesian updating)**.** Continuing the previous example, suppose that we collect more data $\boldsymbol{x}' = (s', f')$, consisting of s' successes and f' failures. Our prior distribution now is the posterior of the previous example, $p(\theta) \propto \theta^s(1-\theta)^f$. We have

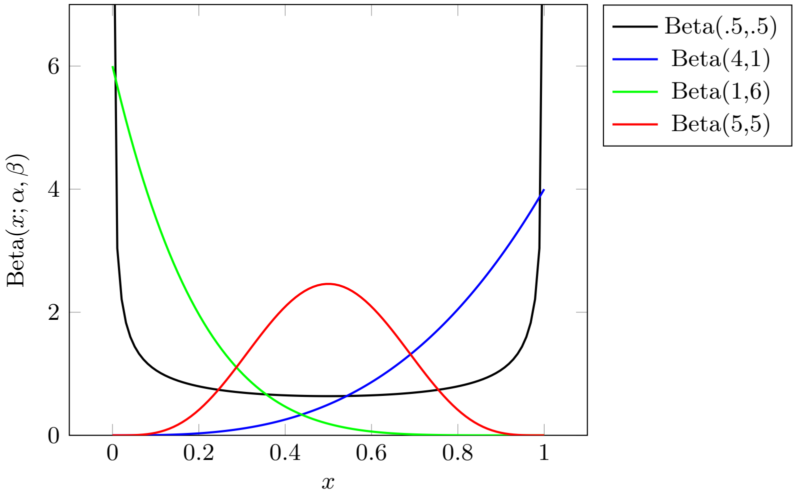
$$\begin{aligned} p(\boldsymbol{x}'|\theta) &= \binom{s'+f'}{s'}\theta^{s'}(1-\theta)^{f'} \\ p(\theta|\boldsymbol{x}') &\propto \theta^s(1-\theta)^f\theta^{s'}(1-\theta)^{f'} \\ &= \theta^{s+s'}(1-\theta)^{f+f'} \\ \Theta|\boldsymbol{x}' &\sim \text{Beta}(s+s'+1, f+f'+1). \end{aligned} \tag{3.11}$$

Equivalently, we can update our uniform prior $p(\theta) \propto 1$ with data $(s+s', f+f')$ to obtain $p(\theta|(s+s', f+f')) \sim \text{Beta}(s+s'+1, f+f'+1)$. As we can see, the Bayesian approach provides a way to update our belief in a consistent manner.

The figure below provides an example of the posterior with 0, 5, 20, and 50 samples. It can be observed that the posterior becomes sharper as more data is collected. $\triangle$



Example 3.5. Beta is a common prior for the probability of Bernoulli experiments. Based on the discussion above, one way to interpret a Beta prior with parameters $\alpha \geq 1, \beta \geq 1$ is to imagine that, starting with the uniform prior, we have already collected $\alpha+\beta-2$ samples, with $\alpha-1$ successes. The following plot shows the Beta distribution with different parameters to give a sense of the range of possible priors. $\triangle$



Example 3.6. Suppose that Y has a Poisson distribution with paramter Λ. That is, $p_{Y|\Lambda}(y|\lambda) = \text{Poi}(y; \lambda)$. Hence,

$$p(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y \in \{0, 1, \dots\}$$

We intend to estimate Λ based on n iid samples $y_1^n = (y_1, \dots, y_n)$ of Y.

We assume that the prior for Λ is given as $p(\lambda) = \text{Gamma}(\lambda; \alpha, \beta) \propto \lambda^{\alpha-1} e^{-\beta\lambda}$. We have

$$p(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda} \tag{3.12}$$

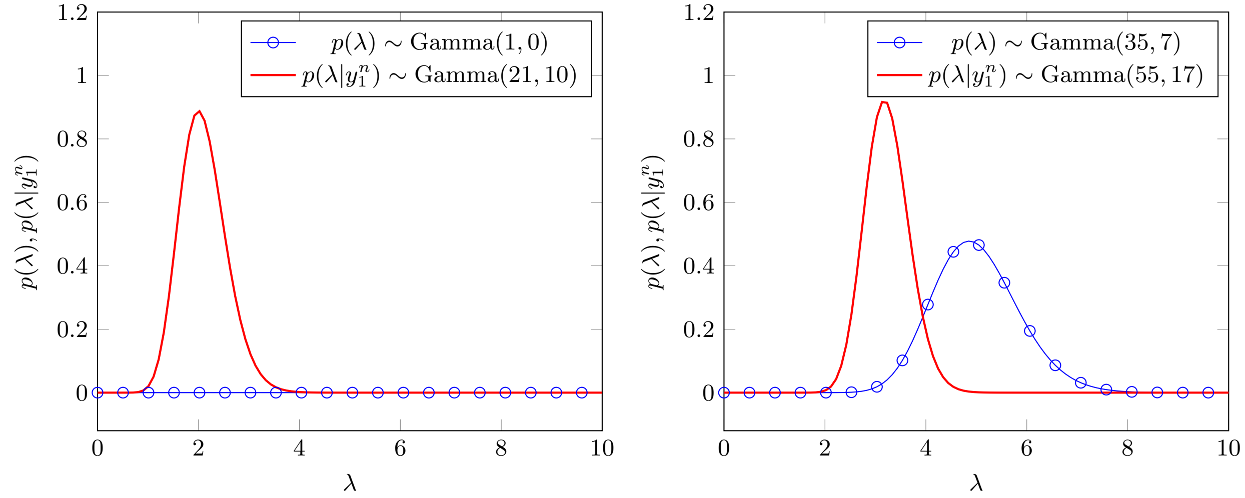
$$p(y_1^n|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \propto \prod_{i=1}^{n} \lambda^{y_i} e^{-\lambda} = e^{-n\lambda} \lambda^{n\bar{y}}, \tag{3.13}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$. Note that while $p(y_1^n|\lambda)$ is a distribution in y_1^n, we still dropped the $y_i!$ from its expression since our final goal is to find a distribution in λ and for this purpose terms that are independent of λ can be viewed as constant. The posterior is

$$p(\lambda|y_1^n) \propto \lambda^{\alpha-1} e^{-\beta\lambda} e^{-n\lambda} \lambda^{n\bar{y}} = \lambda^{\alpha+n\bar{y}-1} e^{-\lambda(n+\beta)} \propto \text{Gamma}(\lambda; \alpha + n\bar{y}, n + \beta). \tag{3.14}$$

If we choose $\alpha = 1, \beta = 0$, then the Gamma prior is flat, giving all possible values the same prior probability. But this is not a proper distribution. However, as long as the final posterior is a proper distribution, an **improper prior** is deemed acceptable.

Suppose that $n = 10$ and $\bar{y} = 2$. The figure below shows the posterior distribution with different priors. The prior on the left is called a **non-informative prior** because it is flat and the one on the right is an **informative prior** given that it represents a prior belief that certain values have a higher probability.



△

3.2 Bayesian Point Estimates

Having the complete distribution for $p_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ is useful since it provides the probability for different values for θ. But sometimes we want to estimate Θ with a single value $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$ as a function of the data, similar to maximum likelihood. It is very important to note that in the Bayesian framework, the data is given and the estimate is known (and not random). The best choice for $\hat{\theta}$ then depends on how we characterize the estimation error:

Average Error	Optimal Estimator
$\mathbb{E}[(\Theta - \hat{\theta})^2 \vert \boldsymbol{x}]$	$\hat{\theta} = \mathbb{E}[\Theta \vert \boldsymbol{x}]$ (**mean**)
$\mathbb{E}[\vert\Theta - \hat{\theta}\vert \,\vert \boldsymbol{x}]$	$\hat{\theta} =$ **median** of $p(\theta \vert \boldsymbol{x})$
$\Pr(\Theta \neq \hat{\theta} \vert \boldsymbol{x}) = \mathbb{E}[I(\Theta \neq \hat{\theta}) \vert \boldsymbol{x}]$	$\hat{\theta} = \arg\max_\theta p(\theta \vert \boldsymbol{x})$ (**mode**)

In the table, $I(\textit{condition})$ is 1 if the condition is satisfied and is 0 otherwise.

We prove the first case in the table. Let $\bar{\theta} = \mathbb{E}[\Theta|\boldsymbol{x}]$. We have

$$\begin{aligned}
\mathbb{E}[(\hat{\theta} - \Theta)^2|\boldsymbol{x}] &= \mathbb{E}[((\hat{\theta} - \bar{\theta}) - (\Theta - \bar{\theta}))^2|\boldsymbol{x}] && (3.15)\\
&= \mathbb{E}[(\hat{\theta} - \bar{\theta})^2 - 2(\hat{\theta} - \bar{\theta})(\Theta - \bar{\theta}) + (\Theta - \bar{\theta})^2|\boldsymbol{x}] && (3.16)\\
&= (\hat{\theta} - \bar{\theta})^2 - 2(\hat{\theta} - \bar{\theta})\,\mathbb{E}[(\Theta - \bar{\theta})|\boldsymbol{x}] + \mathbb{E}[(\Theta - \bar{\theta})^2|\boldsymbol{x}] && (3.17)\\
&= (\hat{\theta} - \bar{\theta})^2 + \mathbb{E}[(\Theta - \bar{\theta})^2|\boldsymbol{x}] && (3.18)\\
&= (\hat{\theta} - \bar{\theta})^2 + \text{Var}(\Theta|\boldsymbol{x}) && (3.19)\\
&\geq \text{Var}(\Theta|\boldsymbol{x}), && (3.20)
\end{aligned}$$

and the lower bound on the error is achieved when $\hat{\theta} = \bar{\theta}$.

Example 3.7. Generalizing Example 3.3 by assuming $p(\theta) = \text{Beta}(\theta; \alpha, \beta)$, we obtain $p(\theta|\boldsymbol{x}) = \text{Beta}(\theta; \alpha +$

$s, \beta + f)$ (for Uniform, $\alpha = \beta = 1$). We have

$$\text{Mean} = \frac{s+\alpha}{s+f+\alpha+\beta}, \tag{3.21}$$

$$\text{Median} \simeq \frac{s+\alpha-1/3}{s+f+\alpha+\beta-2/3}, \tag{3.22}$$

$$\text{Mode} = \frac{s+\alpha-1}{s+f+\alpha+\beta-2}. \tag{3.23}$$

Generally speaking, Bayesian point estimates are between what is suggested only using the prior and what would be obtained using only the likelihood. For example, the mean of the prior is $\frac{\alpha}{\alpha+\beta}$ and the maximum likelihood solution is $\frac{s}{s+f}$. The mean of the posterior, $\frac{s+\alpha}{s+f+\alpha+\beta}$, is between these two. △

3.3 Posterior Predictive Distribution

Given n iid samples, $y_1^n = (y_1, \ldots, y_n)$, we are often interested in the distribution of the next (unobserved) value, $p_{Y_{n+1}|Y_1^n}(y_{n+1}|y_1^n)$. This distribution is referred to as *predictive posterior*. We have

$$\begin{aligned} p(y_{n+1}|y_1^n) &= \int p(y_{n+1}, \theta|y_1^n) d\theta && (3.24)\\ &= \int p(\theta|y_1^n) p(y_{n+1}|\theta, y_1^n) d\theta && (3.25)\\ &= \int p(\theta|y_1^n) p(y_{n+1}|\theta) d\theta, && (3.26) \end{aligned}$$

where we have used the fact that $Y_{n+1} \perp\!\!\!\perp Y_1^n|\Theta$. We have thus written the predictive posterior in terms of two known distributions.

Example 3.8. Continuing Example 3.3, let success in the $n+1$st experiment be denoted by $Y_{n+1} = 1$ and failure by $Y_{n+1} = 0$. We have

$$p_{Y_{n+1}|Y_1^n}(1|y_1^n) = \int \theta p(\Theta|y_1^n) = \mathbb{E}[\Theta|y_1^n] = \frac{s+1}{s+f+2}, \tag{3.27}$$

where we have used the facts that $p_{Y_{n+1}|\Theta}(1|\theta) = \theta$ and that the mean of Beta$(s+1, f+1)$ is $\frac{s+1}{s+f+2}$. △

We may also ask about the expected value of Y_{n+1} given y_1^n, i.e., $\mathbb{E}[Y_{n+1}|y_1^n]$. We can find this by first finding $p(y_{n+1}|y_1^n)$ explicitly. But it is often easier to use the law of iterated expectations, since y_1^n influences Y_{n+1} through Θ. Recall that

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y], \qquad \mathbb{E}[\mathbb{E}[Y|X, Z=z]|Z=z] = \mathbb{E}[Y|Z=z]. \tag{3.28}$$

Hence,

$$\mathbb{E}[Y_{n+1}|y_1^n] = \mathbb{E}[\mathbb{E}[Y_{n+1}|\Theta, y_1^n]|y_1^n] = \mathbb{E}[\mathbb{E}[Y_{n+1}|\Theta]|y_1^n], \tag{3.29}$$

where the last step follows from the fact that $Y_{n+1} \perp\!\!\!\perp Y_1^n|\Theta$, implying that $\mathbb{E}[Y_{n+1}|\Theta, y_1^n] = \mathbb{E}[Y_{n+1}|\Theta]$.

Example 3.9. Let's find $\mathbb{E}[Y_{n+1}|y_1^n]$ in Example 3.6. First, observe that $\mathbb{E}[Y_{n+1}|\Lambda] = \Lambda$. We hence need to find $\mathbb{E}[\Lambda|y_1^n]$. We know from before that $\Lambda|y_1^n$ is distributed according to Gamma$(\alpha + n\bar{y}, \beta + n)$. Therefore, $\mathbb{E}[Y_{n+1}|y_1^n] = \mathbb{E}[\Lambda|y_1^n] = \frac{\alpha+n\bar{y}}{\beta+n}$. △

3.4 Gaussian Prior and Likelihood

Suppose that we want to estimate the mean of a Gaussian distribution with known variance,

$$p(y_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta)^2}{2\sigma^2}} \tag{3.30}$$

given iid data $\{y_1, \ldots, y_n\}$.

Improper priors. Assuming that we have no information about this mean, it makes sense to choose the prior

$$p(\theta) \propto 1. \tag{3.31}$$

But since the integral $\int_{-\infty}^{\infty} 1 d\theta = \infty$, this does not lead to a valid distribution. Nevertheless, such a choice is acceptable, if the posterior is a valid distribution. Such priors are called *improper priors*. An improper prior does not necessarily have to be uniform.

Example 3.10. Consider the above likelihood and prior and let $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. We have

$$\begin{align}
p(\theta|y_1^n) &\propto p(y_1^n|\theta) \cdot 1 \tag{3.32}\\
&\propto \exp\left(-\frac{\sum_{i=1}^{n}(y_i-\theta)^2}{2\sigma^2}\right) \tag{3.33}\\
&\propto \exp\left(-\frac{\sum_{i=1}^{n}(\theta^2-2y_i\theta+y_i^2)}{2\sigma^2}\right) \tag{3.34}\\
&\propto \exp\left(-\frac{\theta^2-2\bar{y}\theta}{2\sigma^2/n}\right) \tag{3.35}\\
&\propto \exp\left(-\frac{(\theta-\bar{y})^2}{2\sigma^2/n}\right) \tag{3.36}\\
\Theta|y_1^n &\sim \mathcal{N}(\bar{y}, \sigma^2/n). \tag{3.37}
\end{align}$$

For the expected value of the next sample, we have

$$\mathbb{E}[Y_{n+1}|y_1^n] = \mathbb{E}[\mathbb{E}[Y_{n+1}|\Theta]|y_1^n] = \mathbb{E}[\Theta|y_1^n] = \bar{y}. \tag{3.38}$$

We can see more explicitly as well,

$$\begin{align}
\mathbb{E}[Y_{n+1}|y_1^n] &= \int y_{n+1} p(y_{n+1}|y_1^n) dy_{n+1} \tag{3.39}\\
&= \int y_{n+1} \int p(y_{n+1}, \theta|y_1^n) d\theta dy_{n+1} \tag{3.40}\\
&= \int y_{n+1} \int p(y_{n+1}|\theta) p(\theta|y_1^n) d\theta dy_{n+1} \tag{3.41}\\
&= \int p(\theta|y_1^n) \int y_{n+1} p(y_{n+1}|\theta) dy_{n+1} d\theta \tag{3.42}\\
&= \int \theta p(\theta|y_1^n) d\theta \tag{3.43}\\
&= \mathbb{E}[\Theta|y_1^n] \tag{3.44}\\
&= \bar{y}. \tag{3.45}
\end{align}$$

△

For the posterior predictive variance, we have

$$\mathrm{Var}(Y_{n+1}|y_1^n) = \sigma^2 + \sigma^2/n. \tag{3.46}$$

It can be shown that $Y_{n+1}|y_1^n$ has a Gaussian distribution, and as we know its conditional mean and variance, we have

$$Y_{n+1}|y_1^n \sim \mathcal{N}(\bar{y}, \sigma^2 + \sigma^2/n) \tag{3.47}$$

From the variance, we can see that there are two sources of uncertainty. One is the inherent randomness in Y, quantified by σ^2 and the other is the result of the uncertainty of our estimate of the mean, quantified by σ^2/n.

Example 3.11 (†)**.** Let us prove that $\mathrm{Var}(Y_{n+1}|y_1^n) = \sigma^2 + \sigma^2/n$:

$$\begin{aligned}\mathrm{Var}(Y_{n+1}|y_1^n) &= \mathbb{E}\Big[(Y_{n+1}-\bar{y})^2|y_1^n\Big]\\ &= \mathbb{E}\Big[\mathbb{E}\Big[(Y_{n+1}-\bar{y})^2|\Theta, y_1^n\Big]|y_1^n\Big]\\ &= \mathbb{E}\Big[\sigma^2 + (\Theta-\bar{y})^2|y_1^n\Big]\\ &= \sigma^2 + \mathbb{E}\Big[(\Theta-\bar{y})^2|y_1^n\Big]\\ &= \sigma^2 + \sigma^2/n\end{aligned}$$

△

We now consider the same problem with a proper Gaussian prior. Note that below as $\tau_0 \to \infty$, the proper prior below tends to the improper prior $p(\theta) \propto 1$.

Example 3.12. We would like to estimate the mean Θ of normally distributed independent values $y_1^n = (y_1, \dots, y_n)$. Let $\bar{y} = \sum y_i/n$. We assume

$$\Theta \sim \mathcal{N}\big(\theta_0, \tau_0^2\big) \tag{3.48}$$

$$Y_i|\theta \sim \mathcal{N}\big(\theta, \sigma^2\big) \tag{3.49}$$

where θ_0 and τ_0^2 are the prior mean and variance, respectively, and σ^2 is known. We have

$$p(\theta|y_1^n) \propto p(\theta)p(y_1^n|\theta) \tag{3.50}$$

$$\propto \frac{1}{\sigma\tau_0}\exp\left(-\frac{\sum_{i=1}^n (y_i-\theta)^2}{2\sigma^2} - \frac{(\theta-\theta_0)^2}{2\tau_0^2}\right) \tag{3.51}$$

With some algebra, it can be shown that conditioned on y_1^n, Θ is normally distributed,

$$\Theta|y_1^n \sim \mathcal{N}\left(\frac{\frac{n\bar{y}}{\sigma^2} + \frac{\theta_0}{\tau_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}\right). \tag{3.52}$$

△

Example 3.13. (†) Let us prove (3.52) using (3.51). We start with the following **claim**: If $p_X(x) \propto e^{-f(x)}$, where $f(x) = ax^2 - bx + c$ with $a > 0$, then $X \sim \mathcal{N}\big(\frac{b}{2a}, \frac{1}{2a}\big)$. Observe that

$$ax^2 - bx + c = \frac{x^2 - bx/a + c/a}{1/a} = \frac{\big(x - \frac{b}{2a}\big)^2 - (\frac{b}{2a})^2 + \frac{c}{a}}{2(1/(2a))} = \frac{(x - b/(2a))^2}{2(1/(2a))} + C, \tag{3.53}$$

where C is a constant independent from x. Hence,

$$p_X(x) \propto \exp\left(\frac{(x-b/(2a))^2}{2(1/(2a))}\right), \tag{3.54}$$

proving the claim. Then, (3.52) can be proven by setting

$$a = \frac{n}{2\sigma^2} + \frac{1}{2\tau_0^2}, \quad b = \frac{n\bar{y}}{\sigma^2} + \frac{\theta_0}{\tau_0^2}. \tag{3.55}$$

△

Example 3.14 (Bias-variance trade-off for a Bayesian point estimator)**.** Suppose that the prior for Θ is $\Theta \sim \mathcal{N}\big(0, \tau_0^2\big)$. Then, from (3.52), the mean (also the mode and median) Bayesian point estimator for Θ is

$$\hat{\theta}_B = \bar{y}\left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n}\right), \tag{3.56}$$

while the maximum-likelihood estimator is $\hat{\theta}_{mle} = \bar{y}$. We can evaluate both estimators in the frequentist framework, finding their MSE.

Note that the frequentist framework requires us to assume a true value θ^* and view $\hat{\theta}_B$ and $\hat{\theta}_{mlue}$ as functions of random data Y_1^n. So they are random variables (however, we won't switch to capital letters to represent them here). First, as the MLE is unbiased,

$$\mathrm{MSE}(\hat{\theta}_{mle}) = \mathrm{Var}(\hat{\theta}_{mle}) = \mathrm{Var}(\bar{Y}) = \sigma^2/n, \tag{3.57}$$

and by CRLB, this is the best unbiased estimator.

For the Bayesian estimator, we have

$$\mathrm{Bias}(\hat{\theta}_B) = \mathbb{E}[\hat{\theta}_B] - \theta^* = \theta^*\left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n}\right) - \theta^* = -\theta^*\left(\frac{\sigma^2/n}{\tau_0^2 + \sigma^2/n}\right) \tag{3.58}$$

$$\mathrm{Var}(\hat{\theta}_B) = \frac{\sigma^2}{n}\left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n}\right)^2 \tag{3.59}$$

$$\mathrm{MSE}(\hat{\theta}_B) = (\theta^*)^2\left(\frac{\sigma^2/n}{\tau_0^2 + \sigma^2/n}\right)^2 + \frac{\sigma^2}{n}\left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n}\right)^2 \tag{3.60}$$

We can see that the bias term is decreasing in τ_0^2 while the variance term is increasing. So, there is a trade-off between the two types of error. Smaller values of τ_0 mean we have a strong prior, thus leading to bias. A strong prior is also less sensitive to data, thus leading to a smaller variance.

In particular, for $\tau_0^2 = (\theta^*)^2$, we have

$$\mathrm{MSE}(\hat{\theta}_B) = \frac{(\theta^*)^2\sigma^2/n}{(\theta^*)^2 + \sigma^2/n} < \sigma^2/n = \mathrm{MSE}(\hat{\theta}_{mle}). \tag{3.61}$$

So, with the right prior, $\hat{\theta}_B$ has lower MSE than the maximum-likelihood estimator. Of course, this requires knowledge of (θ^*), which is not available. However, a good prior found based on experience or intuition can provide good results.

△

3.5 Conjugate Priors

Given a likelihood function, the *conjugate prior* is a distribution that leads to a posterior that is from the same family as the prior. Several examples are given below.

- Bernoulli/Beta: $(y = \sum_{i=1}^n y_i)$

$$p(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i} \qquad \mathrm{Ber}(\theta) \tag{3.62}$$

$$p(y_1^n|\theta) = \theta^{y}(1-\theta)^{n-y} \tag{3.63}$$

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \mathrm{Beta}(\alpha,\beta) \tag{3.64}$$

$$p(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \qquad \mathrm{Beta}(y+\alpha, n-y+\beta) \tag{3.65}$$

- Exponential/Gamma: ($y = \sum_{i=1}^n y_i$)

$$p(y_i|\theta) = \theta \exp(-\theta y_i) \qquad \text{Exp}(\theta) = \text{Gamma}(1, \theta) \tag{3.66}$$

$$p(y_1^n|\theta) = \theta^n \exp(-\theta y) \tag{3.67}$$

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta) \qquad \text{Gamma}(\alpha, \beta) \tag{3.68}$$

$$p(\theta|y_1^n) \propto \theta^{n+\alpha-1} \exp(-(y+\beta)\theta) \qquad \text{Gamma}(n+\alpha, y+\beta) \tag{3.69}$$

- Gaussian/Gaussian (with known σ^2): ($\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$)

$$p(y_i|\theta) \propto \exp\left(\frac{(y_i-\theta)^2}{2\sigma^2}\right) \qquad \mathcal{N}(\theta, \sigma^2) \tag{3.70}$$

$$p(y_1^n|\theta) \propto \exp\left(\frac{\sum_{i=1}^n (y_i-\theta)^2}{2\sigma^2}\right) \tag{3.71}$$

$$p(\theta) \propto \exp\left(\frac{(\theta-\mu_0)^2}{2\tau_0^2}\right) \qquad \mathcal{N}(\mu_0, \tau_0^2) \tag{3.72}$$

$$p(\theta|y_1^n) \propto \exp\left(\frac{(\theta-\mu_1)^2}{2\tau_1^2}\right) \qquad \mathcal{N}(\mu_1, \tau_1^2), \tag{3.73}$$

where

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2/n}\bar{y}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2/n}}, \tag{3.74}$$

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2/n}. \tag{3.75}$$

Note that if a prior is conjugate for the likelihood of a single observation, it is also conjugate for the likelihood of many iid observations. One way to see this is to note that updating the distribution using n iid observations is equivalent to updating the distribution n times using single observations consecutively.

Conjugate priors provide a way to fully determine the posterior distribution without the need to integrate to find the missing constants.

3.6 The Exponential Family (EF)

For a random variable Y with parameter Θ, $p(y|\theta)$ is said to be from the exponential family if it has the following form

$$p(y|\theta) = \exp\big(a(y)^T b(\theta) + f(y) + g(\theta)\big), \tag{3.76}$$

where a, b, y, θ can be vectors and f, g are scalar functions. $b(\theta)$ is referred to as the *natural parameter*.

The exponential family includes many common distributions such as Gaussian, Beta, Gamma, Binomial, etc. For likelihoods in this family, we can identify the conjugate prior, thus simplifying Bayesian estimation. Furthermore, for these distributions all information in the data can be summarized in the *sufficient statistics* described below.

Maximum Likelihood. Suppose that we have n iid observation, leading to the likelihood function

$$p(y_1^n|\theta) \propto \exp\left(\sum_{i=1}^n a(y_i)^T b(\theta) + n g(\theta)\right), \tag{3.77}$$

Define the *sufficient statistics* for this likelihood as $t(y_1^n) = \sum_{i=1}^n a(y_i)$. We then have

$$p(y_1^n|\theta) \propto \exp\big(t(y_1^n)^T b(\theta) + n g(\theta)\big). \tag{3.78}$$

So for finding the maximum likelihood solution, we can summarize all our data as $t(y_1^n)$ and the rest of the information in y_1^n is irrelevant. This is also true for Bayesian estimation. Note that the size of $t(y_1^n)$ is independent of n.

Bayesian Estimation with Conjugate Priors. In this case, we have the general form of the conjugate prior

$$\begin{aligned} p(y_i|\theta) &\propto \exp\big(a(y_i)^T b(\theta) + g(\theta)\big) && (3.79)\\ p(y_1^n|\theta) &\propto \exp\big(t(y_1^n)^T b(\theta) + ng(\theta)\big) && (3.80)\\ p(\theta) &\propto \exp\big(\nu^T b(\theta) + mg(\theta)\big) \qquad Dist(\nu, m) && (3.81)\\ p(\theta|y_1^n) &\propto \exp\big((\nu + t(y_1^n))^T b(\theta) + (m+n)g(\theta)\big) \qquad Dist(\nu + t(y_1^n), m+n), && (3.82) \end{aligned}$$

where $Dist$ refers to a specific type distribution.

Pseudo-observations. The parameters in conjugate priors can be interpreted as representing pseudo-observations by comparing the forms of $p(y_1^n|\theta)$ and $p(\theta)$. In particular, ν plays the same role as $t(y_1^n)$ and m represents the number of pseudo-observations.

Example 3.15. The likelihood for a Bernoulli observation is

$$\begin{aligned} p(y_i|\theta) &= \theta^{y_i}(1-\theta)^{1-y_i} && (3.83)\\ &= \exp(y_i \ln\theta + (1-y_i)\ln(1-\theta)) && (3.84)\\ &= \exp\left(y_i \ln\frac{\theta}{1-\theta} + \ln(1-\theta)\right). && (3.85) \end{aligned}$$

We thus let $a(y_i) = y_i$, $b(\theta) = \ln\frac{\theta}{1-\theta}$, and $g(\theta) = \ln(1-\theta)$. Furthermore, let $y = t(y_1^n) = \sum_{i=1}^n a(y_i) = \sum_{i=1}^n y_i$. Then,

$$\begin{aligned} p(y_1^n|\theta) &= \exp\left(y \ln\frac{\theta}{1-\theta} + n\ln(1-\theta)\right) && (3.86)\\ p(\theta) &= \exp\left(\nu \ln\frac{\theta}{1-\theta} + m\ln(1-\theta)\right) && (3.87)\\ &= \theta^{\nu}(1-\theta)^{m-\nu}, \qquad \text{Beta}(\nu+1, m-\nu+1) && (3.88)\\ p(\theta|y_1^n) &= \exp\left((\nu+y)\ln\frac{\theta}{1-\theta} + (m+n)\ln(1-\theta)\right), && (3.89)\\ &\text{Beta}(\nu+y+1, m+n-\nu-y+1) && (3.90) \end{aligned}$$

△