Chapter 3

Bayesian Parameter Estimation

3.1 From Prior to Posterior

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

Specifically, the steps of Bayesian estimation of a parameter θ are:

1. Identifying the prior distribution, $p(\theta)$
2. Collecting data and forming the likelihood: $p(\mathcal{D}|\theta)$
3. Finding the posterior distribution $p(\theta|\mathcal{D})$ as

$$p(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})} \tag{3.1}$$

Normalizing distributions. Finding the posterior distribution requires computing the integral $p(\mathcal{D}) = \int_\theta p(\theta)p(\mathcal{D}|\theta)d\theta$. 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, $p(\mathcal{D})$ is one such term. So we often first find a function proportional to $p(\theta|\mathcal{D})$ as

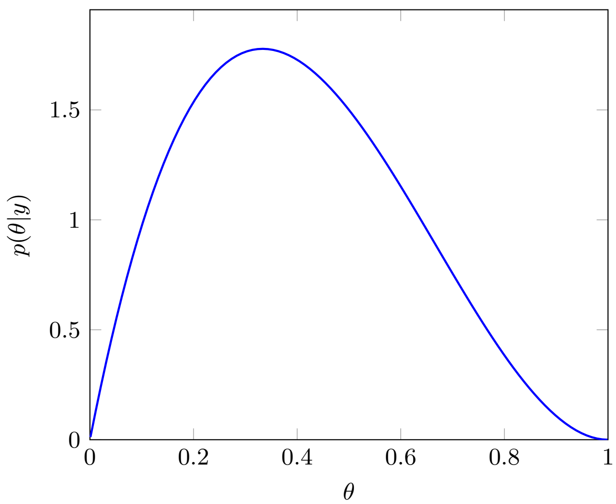
$$p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta),$$

where we can also drop constant terms in θ from $p(\theta)$ and $p(\mathcal{D}|\theta)$. 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, as we will see later. Finally, in certain cases, we can find what we are after without any integration. For example if our goal is to find the value of θ maximizing $p(\theta|\mathcal{D})$.

Example 53. Let θ denote the unknown parameter of a geometric random variable y, where $p(y) = \theta(1-\theta)^{y-1}$. Suppose that we observe y. 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

$$p(\theta) = 1$$
$$p(y|\theta) = \theta(1-\theta)^{y-1}$$
$$p(\theta|y) \propto \theta(1-\theta)^{y-1}$$

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



Exercise 54. 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 y is observed to be $y = 1$. Assuming a uniform prior, find the posterior distribution of θ, i.e., $\theta|y = 1$.

Example 55. 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 $\mathcal{D} = (s, f)$. The likelihood is

$$p(\mathcal{D}|\theta) = \binom{n}{s}\theta^s(1-\theta)^f \tag{3.2}$$

- The posterior distribution:

$$p(\theta|\mathcal{D}) \propto 1 \cdot \theta^s(1-\theta)^f = \theta^s(1-\theta)^f \tag{3.3}$$

  We observe that this distribution is of the form of a beta distribution, $\text{Beta}(x;\alpha,\beta) \sim x^{\alpha-1}(1-x)^{\beta-1}$. Hence,

$$p(\theta|\mathcal{D}) \sim \text{Beta}(s+1, f+1).$$

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 example above.

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

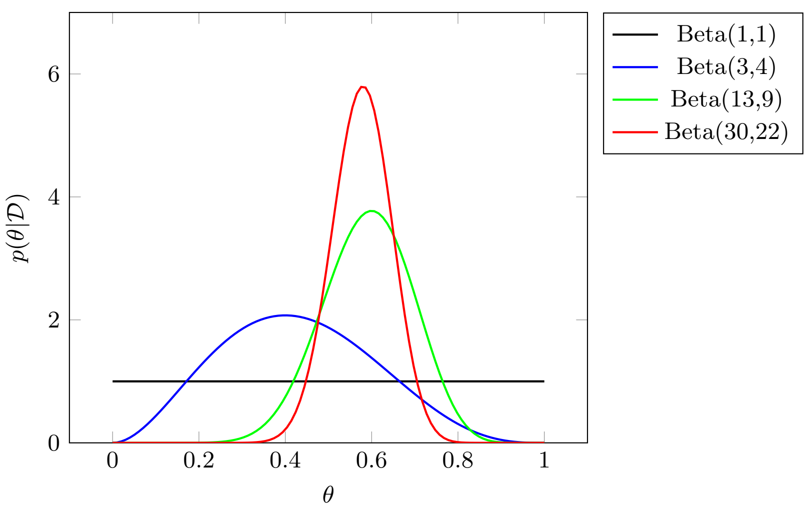
$$\int_{0.4}^{0.6} p(\theta|\mathcal{D})d\theta \tag{3.4}$$

Example 56. Continuing the previous example, suppose that we collect more data $\mathcal{D}' = (s', f')$, 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

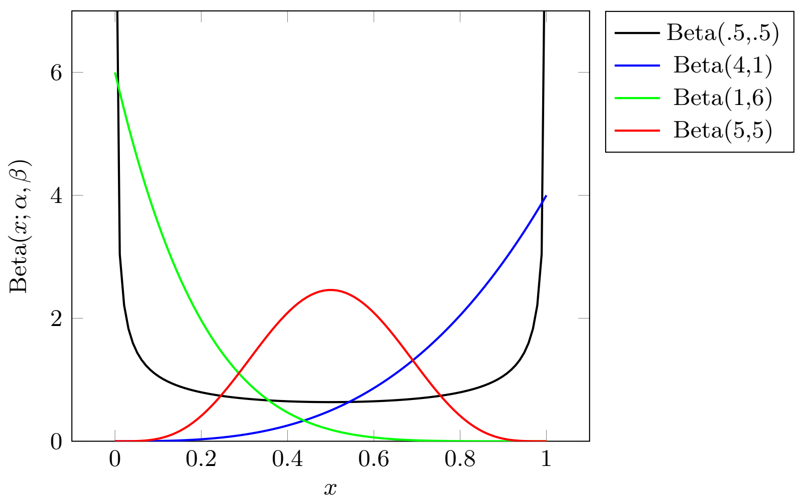
$$\begin{aligned} p(\mathcal{D}'|\theta) &= \binom{s'+f'}{s'}\theta^{s'}(1-\theta)^{f'} \\ p(\theta|\mathcal{D}') &\propto \theta^s(1-\theta)^f\theta^{s'}(1-\theta)^{f'} \\ &= \theta^{s+s'}(1-\theta)^{f+f'} \\ \theta|\mathcal{D}' &\sim \text{Beta}(s+s'+1, f+f'+1). \end{aligned} \tag{3.5}$$

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.



Example 57. 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.



Example 58. Suppose that $y \sim \text{Poi}(\lambda)$ and we intend to estimate λ based on n iid samples $y_1^n = (y_1, \ldots, y_n)$. We assume that the prior for λ is given as $p(\lambda) \sim \text{Gamma}(\lambda; \alpha, \beta) \propto \lambda^{\alpha-1}e^{-\beta\lambda}$. We have

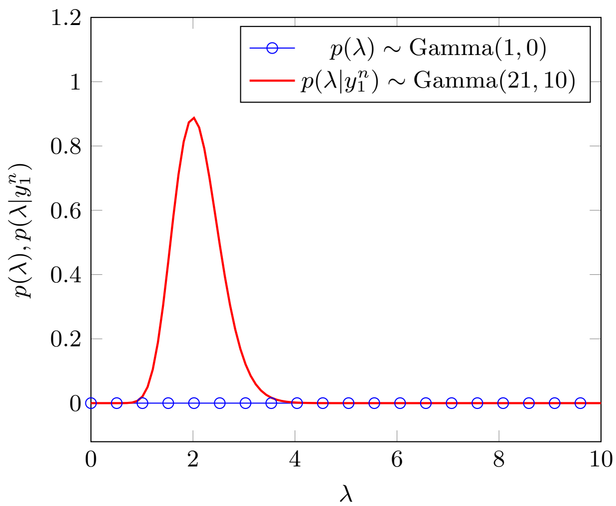
$$p(\lambda) \propto \lambda^{\alpha-1}e^{-\beta\lambda}$$
$$p(y_1^n|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i}e^{-\lambda}}{y_i!} \propto \prod_{i=1}^{n} \lambda^{y_i}e^{-\lambda} = e^{-n\lambda}\lambda^{n\bar{y}},$$

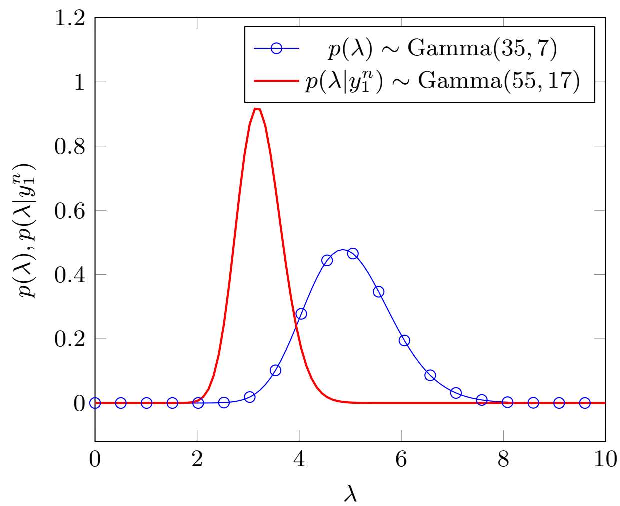
where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Note that while $p(y_1^n|\lambda)$ is a distribution in y_1^n, we still dropped the $y_i!$ from its expression since our final goal is to find a distribution in λ and for this purpose terms that are independent of λ can be viewed as constant. The posterior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

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

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 $p(\theta|\mathcal{D})$ is useful since it provides the probability for different values for θ. But sometimes we want to estimate θ with a single value $\hat{\theta} = \hat{\theta}(\mathcal{D})$ as a function of data, similar to maximum likelihood. The best choice for $\hat{\theta}$ then depends on how we characterize the estimation error:

Average Error	Optimal Estimator
$\mathbb{E}[(\theta - \hat{\theta})^2|\mathcal{D}]$	$\hat{\theta} = \mathbb{E}[\theta|\mathcal{D}]$ (**mean**)
$\mathbb{E}[|\theta - \hat{\theta}||\mathcal{D}]$	$\hat{\theta} =$ **median** of $p(\theta|\mathcal{D})$
$\mathbb{E}[I(\theta \neq \hat{\theta})|\mathcal{D}] = p(\theta \neq \hat{\theta}|\mathcal{D})$	$\hat{\theta} = \arg\max_\theta p(\theta|\mathcal{D})$ (**mode**)

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

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

$$\begin{aligned}
\mathbb{E}[(\hat{\theta} - \theta)^2|\mathcal{D}] &= \mathbb{E}[((\hat{\theta} - \bar{\theta}) + (\bar{\theta} - \theta))^2|\mathcal{D}] \\
&= \mathbb{E}[(\hat{\theta} - \bar{\theta})^2 + 2(\hat{\theta} - \bar{\theta})(\bar{\theta} - \theta) + (\bar{\theta} - \theta)^2|\mathcal{D}] \\
&= \mathbb{E}[(\hat{\theta} - \bar{\theta})^2|\mathcal{D}] + \mathbb{E}[(\bar{\theta} - \theta)^2|\mathcal{D}] \\
&= \mathbb{E}[(\hat{\theta} - \bar{\theta})^2|\mathcal{D}] + \text{Var}(\theta|\mathcal{D}) \\
&\geq \text{Var}(\theta|\mathcal{D}),
\end{aligned}$$

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

Example 59. Generalizing Example 55 by assuming $p(\theta) \sim \text{Beta}(\alpha, \beta)$, we obtain $p(\theta|\mathcal{D}) \sim \text{Beta}(\alpha+s, \beta+f)$ (for Uniform, $\alpha = \beta = 1$). We have

$$\begin{aligned}
\text{Mean} &= \frac{s + \alpha}{s + f + \alpha + \beta}, \\
\text{Median} &\simeq \frac{s + \alpha - 1/3}{s + f + \alpha + \beta - 2/3}, \\
\text{Mode} &= \frac{s + \alpha - 1}{s + f + \alpha + \beta - 2}.
\end{aligned}$$

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

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

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 60. Continuing Example 55, let success in the $n+1$st experiment be denoted by $y_{n+1} = 1$ and failure by $y_{n+1} = 0$. We have

$$p(y_{n+1} = 1|y_1^n) = \int \theta p(\theta|y_1^n) = \mathbb{E}[\theta|y_1^n] = \frac{s+1}{s+f+2},$$

where we have used the facts that $p(y_{n+1} = 1|\theta) = \theta$ 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, given that y_1^n influences y_{n+1} through θ. Recall that This can be found easier using the law of iterated expectations, i.e.,

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

Thus,

$$\mathbb{E}[y_{n+1}|y_1^n] = \mathbb{E}[\mathbb{E}[y_{n+1}|\theta, y_1^n]|y_1^n] = \mathbb{E}[\mathbb{E}[y_{n+1}|\theta]|y_1^n], \tag{3.6}$$

where the last step follows from the fact that $y_{n+1} \perp\!\!\!\perp y_1^n|\theta$.

Exercise 61. Find $\mathbb{E}[y_{n+1}|y_1^n]$ in Example 58.

3.4 Gaussian Prior and Likelihood

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

$$p(y_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta)^2}{2\sigma^2}} \tag{3.7}$$

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

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

$$p(\theta) \propto 1.$$

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

$$\begin{aligned}
p(\theta|y_1^n) &\propto p(y_1^n|\theta)\cdot 1\\
&\propto \exp\left(-\frac{\sum_{i=1}^n (y_i-\theta)^2}{2\sigma^2}\right)\\
&\propto \exp\left(-\frac{\sum_{i=1}^n (\theta^2 - 2y_i\theta + y_i^2)}{2\sigma^2}\right)\\
&\propto \exp\left(-\frac{\theta^2 - 2\bar{y}\theta}{2\sigma^2/n}\right)\\
&\propto \exp\left(-\frac{(\theta-\bar{y})^2}{2\sigma^2/n}\right)\\
\theta|y_1^n &\sim \mathcal{N}(\bar{y}, \sigma^2/n).
\end{aligned}$$

For the expected value of the next sample,we have

$$\mathbb{E}[y_{n+1}|y_1^n] = \mathbb{E}[\mathbb{E}[y_{n+1}|\theta]|y_1^n] = \mathbb{E}[\mathbb{E}[\theta]|y_1^n] = \bar{y}.$$

We can see more explicitly as well,

$$\begin{aligned}
\mathbb{E}[y_{n+1}|y_1^n] &= \int y_{n+1} p(y_{n+1}|y_1^n) dy_{n+1}\\
&= \int y_{n+1} \int p(y_{n+1}, \theta|y_1^n) d\theta dy_{n+1}\\
&= \int y_{n+1} \int p(y_{n+1}|\theta) p(\theta|y_1^n) d\theta dy_{n+1}\\
&= \int p(\theta|y_1^n) \int y_{n+1} p(y_{n+1}|\theta) dy_{n+1} d\theta\\
&= \int \theta p(\theta|y_1^n) d\theta\\
&= \mathbb{E}[\theta|y_1^n]\\
&= \bar{y}.
\end{aligned}$$

Exercise 63. Prove that

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

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

$$\mu \sim \mathcal{N}\big(\mu_0, \tau_0^2\big)$$

$$y_i \sim \mathcal{N}\big(\mu, \sigma^2\big)$$

where μ_0 and τ_0^2 are the prior mean and variance, respectively, and σ^2 is known. We have

$$\begin{aligned} p(\mu|y_1^n) &\propto p(\mu)p(y_1^n|\mu) \\ &\propto \frac{1}{\sigma\tau_0}\exp\left(-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\tau_0^2}\right) \end{aligned}$$

The following claim will be useful.

Claim: If $p(x) \propto e^{-f(x)}$, where $f(x) = ax^2 - bx + c$ with $a > 0$, then $x \sim \mathcal{N}\left(\frac{b}{2a}, \frac{1}{2a}\right)$.

Proof: Since

$$ax^2 - bx + c = \frac{x^2 - bx/a + c/a}{1/a} = \frac{\left(x - \frac{b}{2a}\right)^2 - (\frac{b}{2a})^2 + \frac{c}{a}}{2(1/(2a))},$$

we have

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

proving the claim.

Returning to our problem:

$$a = \frac{n}{2\sigma^2} + \frac{1}{2\tau_0^2}, \quad b = \frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2}.$$

Hence

$$\mu|y_1^n \sim \mathcal{N}\left(\frac{\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}\right).$$

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: $(y = \sum_{i=1}^n y_i)$

$$\begin{aligned} p(y_i|\theta) &= \theta^{y_i}(1-\theta)^{1-y_i} && \mathrm{Ber}(\theta) \\ p(y_1^n|\theta) &= \theta^{y}(1-\theta)^{n-y} && \\ p(\theta) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} && \mathrm{Beta}(\alpha,\beta) \\ p(\theta|y) &\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} && \mathrm{Beta}(y+\alpha, n-y+\beta) \end{aligned}$$

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

$$\begin{aligned} p(y_i|\theta) &= \theta\exp(-\theta y_i) && \mathrm{Exp}(\theta) = \mathrm{Gamma}(1,\theta) \\ p(y_1^n|\theta) &= \theta^n \exp(-\theta y) && \\ p(\theta) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta) && \mathrm{Gamma}(\alpha,\beta) \\ p(\theta|y_1^n) &\propto \theta^{n+\alpha-1}\exp(-(y+\beta)\theta) && \mathrm{Gamma}(n+\alpha, y+\beta) \end{aligned}$$

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

$$
\begin{aligned}
p(y_i|\theta) &\propto \exp\left(\frac{(y_i-\theta)^2}{2\sigma^2}\right) && \mathcal{N}(\theta,\sigma^2)\\
p(y_1^n|\theta) &\propto \exp\left(\frac{\sum_{i=1}^n (y_i-\theta)^2}{2\sigma^2}\right) &&\\
p(\theta) &\propto \exp\left(\frac{(\theta-\mu_0)^2}{2\tau_0^2}\right) && \mathcal{N}(\mu_0,\tau_0^2)\\
p(\theta|y_1^n) &\propto \exp\left(\frac{(\theta-\mu_1)^2}{2\tau_1^2}\right) && \mathcal{N}(\mu_1,\tau_1^2),
\end{aligned}
$$

where

$$
\begin{aligned}
\mu_1 &= \frac{\frac{1}{\tau_0}\mu_0 + \frac{1}{\sigma^2/n}\bar{y}}{\frac{1}{\tau_0}+\frac{1}{\sigma^2/n}},\\
\frac{1}{\tau_1^2} &= \frac{1}{\tau_0^2} + \frac{1}{\sigma^2/n}.
\end{aligned}
$$

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 x with parameter θ, $p(x|\theta)$ is said to be from the exponential family if it has the following form

$$p(x|\theta) = \exp\big(a(x)^T b(\theta) + f(x) + g(\theta)\big),$$

where a, b, x, 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

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

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

$$p(y_1^n|\theta) \propto \exp\big(t(y_1^n)^T b(\theta) + ng(\theta)\big).$$

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

$$\begin{aligned}
p(y_i|\theta) &\propto \exp\bigl(a(y_i)^T b(\theta) + g(\theta)\bigr) \\
p(y_1^n|\theta) &\propto \exp\bigl(t(y_1^n)^T b(\theta) + ng(\theta)\bigr) \\
p(\theta) &\propto \exp\bigl(\nu^T b(\theta) + mg(\theta)\bigr) && Dist(\nu, m) \\
p(\theta|y_1^n) &\propto \exp\bigl((\nu + t(y_1^n))^T b(\theta) + (m+n)g(\theta)\bigr) && Dist(\nu + t(y_1^n), m+n),
\end{aligned}$$

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 65. The likelihood for a Bernoulli observation is

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

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,

$$\begin{aligned}
p(y_1^n|\theta) &= \exp\left(y \ln\frac{\theta}{1-\theta} + n\ln(1-\theta)\right) \\
p(\theta) &= \exp\left(\nu \ln\frac{\theta}{1-\theta} + m\ln(1-\theta)\right) = \theta^{\nu}(1-\theta)^{m-\nu} && \mathrm{Beta}(\nu+1, m-\nu+1) \\
p(\theta|y_1^n) &= \exp\left((\nu+y) \ln\frac{\theta}{1-\theta} + (m+n)\ln(1-\theta)\right) && \mathrm{Beta}(\nu+y+1, m+n-\nu-y+1)
\end{aligned}$$