17 Bayesian Statistical Inference

In Chapter 16, we defined an estimator of some unknown quantity, θ, based on experimentally sampled data, X. This estimator, denoted by $\hat{\theta}_{\text{ML}}(X)$, is called a *maximum likelihood (ML) estimator*, because it returns that value of θ that produces the highest likelihood of witnessing the particular sampled data. Specifically,

$$\hat{\theta}_{\text{ML}}(X = x) \equiv \underset{\theta}{\text{argmax}}\, \mathbf{P}\{X = x \mid \theta\}. \tag{17.1}$$

The ML estimator makes a lot of sense in situations where we have no a priori knowledge of θ. However, what do we do in situations where we have some knowledge about θ – for example, we know that θ is likely to be high? $\hat{\theta}_{\text{ML}}(X = x)$ as defined in (17.1) doesn't have any way of incorporating this a priori knowledge.

In this chapter, we therefore introduce a new kind of estimator, called a maximum a posteriori (MAP) estimator. Like the ML estimator, the MAP estimator is again an estimator of an unknown quantity, θ, based on experimentally sampled data, X. However, the MAP estimator starts with a distribution Θ on the possible values of θ, allowing us to specify that some values are more likely than others. The MAP estimator then incorporates the joint distribution of Θ and the sampled data X to estimate θ.

Because it assumes a prior distribution, Θ, the MAP estimator is a **Bayesian estimator**, as compared with the ML estimator which is a classical estimator. We will start with a motivating example that sheds some light on how the MAP estimator and the ML estimator are related.

17.1 A Motivating Example

Example 17.1 (Gold or silver coin?)

In this example, you are given a coin that you can't see. The coin is either gold or silver. If the coin is gold, then it has bias $p = 0.6$ (chance $p = 0.6$ of heads). If the coin is silver, then it has bias $p = 0.4$.

We wish to determine whether $p = 0.6$ or $p = 0.4$. To do this, we flip the coin nine times. Let X denote the number of heads observed.

Question: Define the ML estimator to determine whether $p = 0.6$ or $p = 0.4$.

Answer:

$$\hat{p}_{\text{ML}}(X = x) = \underset{p \in \{0.4, 0.6\}}{\text{argmax}} \; \mathbf{P}\{X = x \mid p\}. \tag{17.2}$$

Question: Consider these two expressions: $\mathbf{P}\{X = x \mid p = 0.4\}$ versus $\mathbf{P}\{X = x \mid p = 0.6\}$. Which is bigger?

Answer: The answer depends on x.

$$\mathbf{P}\{X = x \mid p = 0.4\} = \binom{9}{x}(0.4)^x(0.6)^{9-x}$$
$$\mathbf{P}\{X = x \mid p = 0.6\} = \binom{9}{x}(0.6)^x(0.4)^{9-x}.$$

So $\mathbf{P}\{X = x \mid p = 0.4\}$ is larger if $x < 5$, and $\mathbf{P}\{X = x \mid p = 0.6\}$ is larger if $x \geq 5$.

Thus, we have that

$$\hat{p}_{\text{ML}}(X = x) = \begin{cases} 0.4 & \text{if } x \in \{0, 1, 2, 3, 4\} \\ 0.6 & \text{if } x \in \{5, 6, 7, 8, 9\} \end{cases}. \tag{17.3}$$

Example 17.2 (Gold or silver coin with added information)

Now suppose we are in the same setting as Example 17.1, but we are given the additional information that gold coins are four times more common than silver ones. So, absent any samples, with probability 80% our coin is gold.

To capture this, define a random variable (r.v.) P, where P represents the bias of the coin:

$$P = \text{ bias of coin } = \begin{cases} 0.4 & \text{w/prob } 20\% \\ 0.6 & \text{w/prob } 80\% \end{cases}.$$

Question: How can we incorporate this distributional information about the bias into our ML estimator?

Answer: Our ML estimator, as defined in (17.2), does not have a way of incorporating the distributional information represented by P.

Question: Intuitively, how do you imagine that knowing that the bias is modeled by P might change the result in (17.3)?

Answer: It seems like the output of $p = 0.6$ should be more likely, given the fact that most coins are gold. Thus, even when the sampled data is $X = x < 5$, it may still be true that the best estimate for p is $p = 0.6$.

As an idea for how to incorporate the distributional information embodied by P, consider the weighted ML estimator, given in (17.4). This new estimator starts with the ML estimator given in (17.2), but multiplies the likelihood function by the prior:

$$\hat{p}_{\text{weightedML}}(X = x) = \underset{p \in \{0.4, 0.6\}}{\text{argmax}} \left(\underbrace{\mathbf{P}\{X = x \mid p\}}_{\text{likelihood}} \cdot \underbrace{\mathbf{P}\{P = p\}}_{\text{prior}} \right). \tag{17.4}$$

This "weighted ML" estimator clearly puts more weight on the output $p = 0.6$ as compared to $p = 0.4$. We will soon see that this weighted ML estimator in (17.4) is equivalent to the MAP estimator, which we define next!

17.2 The MAP Estimator

We will first define the MAP estimator in the context of Example 17.2 and then define it more generally a little later.

Definition 17.3 (MAP estimator for Example 17.2) *Our goal is to estimate $p \in \{0.4, 0.6\}$. We are given a* **prior distribution** *on the possible values for p, denoted by r.v. P (we intentionally use the capitalized form of p). We also have experimental data, denoted by r.v. X.*

We say that $\hat{P}_{MAP}(X)$ is the **MAP estimator** *of p. We use a capital $\hat{P}$ to denote that the estimator takes into account both the prior distribution P and the data X to create an estimate of p:*

$$\hat{P}_{MAP}(X = x) = \underset{p \in \{0.4, 0.6\}}{argmax} \ \mathbf{P}\{P = p \mid X = x\}. \tag{17.5}$$

Note that $\hat{P}_{MAP}(X)$ is a function of a r.v. X and thus is a r.v., while $\hat{P}_{MAP}(X = x)$ is a constant.

Let us compare $\hat{P}_{\text{MAP}}(X = x)$ in (17.5) with $\hat{p}_{\text{ML}}(X = x)$ in (17.2). Both of these are estimates of p based on data sample $X = x$. Both involve finding the value of p which maximizes some expression. However, (17.5) uses the prior distribution P and has swapped the order of the conditional as compared to (17.2).

Question: Argue that $\hat{P}_{\text{MAP}}(X = x)$ from (17.5) is equal to $\hat{p}_{\text{weightedML}}(X = x)$ from (17.4).

Answer: Starting with $\hat{P}_{\text{MAP}}(X = x)$, and applying Bayes' Rule, observe that we are looking for the p that maximizes:

$$\mathbf{P}\{P = p \mid X = x\} = \frac{\mathbf{P}\{P = p \ \& \ X = x\}}{\mathbf{P}\{X = x\}} = \frac{\mathbf{P}\{X = x \mid P = p\} \cdot \mathbf{P}\{P = p\}}{\mathbf{P}\{X = x\}}.$$

But the $\mathbf{P}\{X = x\}$ term doesn't affect this maximization, so we're really looking for the p that maximizes

$$\underbrace{\mathbf{P}\{X = x \mid P = p\}}_{\text{likelihood}} \cdot \underbrace{\mathbf{P}\{P = p\}}_{\text{prior}}. \tag{17.6}$$

But this in turn is exactly the expression that we're maximizing in (17.4).

Question: Is there any situation where $\hat{P}_{\text{MAP}} = \hat{p}_{\text{ML}}$?

Answer: Yes, this happens when the prior, P, provides no additional information, in that all possible values of p are equally likely. For our current example, this would mean that the gold and silver coins are equally likely. In the case of a continuous setting, P would follow a Uniform distribution.

We now proceed to evaluate

$$\hat{P}_{\text{MAP}}(X = x) = \underset{p \in \{0.4, 0.6\}}{\text{argmax}} \ \mathbf{P}\{P = p \mid X = x\}.$$

Given that there are only two possible values of p, we simply need to compare the following two expressions:

$$\mathbf{P}\{P = 0.4 \mid X = x\} = \frac{\binom{9}{x} \cdot 0.4^x \cdot 0.6^{9-x} \cdot 20\%}{\mathbf{P}\{X = x\}} \tag{17.7}$$

$$\mathbf{P}\{P = 0.6 \mid X = x\} = \frac{\binom{9}{x} \cdot 0.6^x \cdot 0.4^{9-x} \cdot 80\%}{\mathbf{P}\{X = x\}}. \tag{17.8}$$

Question: How do we determine which of (17.7) and (17.8) is higher?

Answer: It's easiest to look at their ratio and see when the ratio exceeds 1:

$$\frac{\mathbf{P}\{P = 0.6 \mid X = x\}}{\mathbf{P}\{P = 0.4 \mid X = x\}} = 4 \cdot \left(\frac{3}{2}\right)^{2x-9}.$$

But

$$4 \cdot \left(\frac{3}{2}\right)^{2x-9} > 1 \qquad \Longleftrightarrow \qquad x \geq 3.$$

Thus, $p = 0.6$ is the maximizing value when $x \geq 3$. So

$$\hat{P}_{\text{MAP}}(X = x) = \begin{cases} 0.4 & \text{if } x \in \{0, 1, 2\} \\ 0.6 & \text{if } x \in \{3, 4, 5, 6, 7, 8, 9\} \end{cases}. \tag{17.9}$$

Thus,

$$\hat{P}_{\text{MAP}}(X) = \begin{cases} 0.4 & \text{if } X < 3 \\ 0.6 & \text{if } X \geq 3 \end{cases}. \tag{17.10}$$

Intuitively, this makes sense, since we are starting out with a coin that is gold with probability 80%.

We end this section by defining the MAP estimator in general settings, beyond the context of Example 17.2.

Definition 17.4 *Our goal is to estimate some unknown θ. We are given a* **prior distribution** *on the possible values for θ, denoted by r.v. Θ. We also have experimental data, denoted by r.v. X.*
We say that $\hat{\Theta}_{MAP}(X)$ is our **MAP estimator** *of θ. We use a capital $\hat{\Theta}$ in our estimator to denote that the estimator takes into account both the prior distribution Θ and the data X to create an estimate of θ.*

In the case where Θ is a discrete r.v., *the MAP estimator is defined by:*

$$\begin{aligned} \hat{\Theta}_{MAP}(X = x) &= \underset{\theta}{argmax}\, \mathbf{P}\{\Theta = \theta \mid X = x\} \\ &= \begin{cases} \underset{\theta}{argmax}\, \mathbf{P}\{X = x \mid \Theta = \theta\} \cdot \mathbf{P}\{\Theta = \theta\} & \text{if } X \text{ is discrete} \\ \underset{\theta}{argmax}\, f_{X|\Theta=\theta}(x) \cdot \mathbf{P}\{\Theta = \theta\} & \text{if } X \text{ is continuous} \end{cases} \end{aligned}$$

In the case where Θ is a continuous r.v., *the MAP estimator is defined by:*

$$\begin{aligned} \hat{\Theta}_{MAP}(X = x) &= \underset{\theta}{argmax}\, f_{\Theta|X=x}(\theta) \\ &= \begin{cases} \underset{\theta}{argmax}\, \mathbf{P}\{X = x \mid \Theta = \theta\} \cdot f_{\Theta}(\theta) & \text{if } X \text{ is discrete} \\ \underset{\theta}{argmax}\, f_{X|\Theta=\theta}(x) \cdot f_{\Theta}(\theta) & \text{if } X \text{ is continuous} \end{cases} \end{aligned}$$

Note that $\hat{\Theta}_{MAP}(X)$ is a function of a r.v. X and thus is a r.v., while $\hat{\Theta}_{MAP}(X = x)$ is a constant.

Definition 17.5 *While the r.v. Θ represents the prior distribution, the conditional r.v., $[\Theta \mid X = x]$, represents the* **posterior distribution** *since it represents the updated version of the prior distribution, given the value of the data. Likewise, $\mathbf{P}\{\Theta = \theta \mid X = x\}$ is called the* **posterior probability** *(where we write $f_{\Theta|X=x}(\theta)$ for the continuous case). Thus $\hat{\Theta}_{MAP}(X = x)$ represents the value of θ that maximizes the posterior probability.*

Remark: While $\hat{\Theta}_{\text{MAP}}(X)$ in Definition 17.4 depends on both the prior distribution Θ and also on X, we note that $\hat{\Theta}_{\text{MAP}}(X)$ is a function of just X. Specifically, once we specify the value of X, say $X = x$, then $\hat{\Theta}_{\text{MAP}}(X)$ becomes a constant.

17.3 More Examples of MAP Estimators

Example 17.6 (Estimating voting probability)

Suppose we want to estimate the fraction of people who will vote in the next election. Let's call this quantity p. To estimate p, we sample 100 people independently at random. Suppose that 80 of the sampled people say that they plan to vote. This feels high, so we go back to look at prior elections and how many people voted in prior elections. We find that the fraction of people who voted in prior elections is well modeled by the r.v. P, with density function:

$$f_P(p) = 3(1-p)^2, \qquad \text{where } 0 \le p \le 1,$$

shown in Figure 17.1. Given this prior, P, and the sample $X = 80$, how can we estimate the true fraction of people, p, who will actually vote?

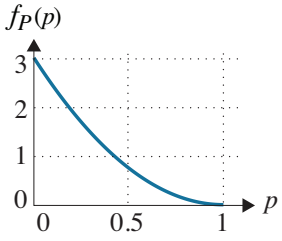

Figure 17.1 *Illustration of $f_P(p)$.*

In order to formulate this question in terms of Definition 17.4, we start with a few questions.

Question: If X denotes the number of people sampled, how is X distributed?

Answer: $X \sim \text{Binomial}(100, p)$.

Question: Which of the cases of Definition 17.4 should we be looking at?

Answer: $\Theta = P$ is continuous, and X is discrete. Thus,

$$\hat{P}_{\text{MAP}}(X = 80) = \underset{p}{\text{argmax}} f_{P|X=80}(p) = \underset{p}{\text{argmax}}\, \mathbf{P}\{X = 80 \mid P = p\} \cdot f_P(p).$$

Since $X \sim \text{Binomial}(100, p)$, we know that

$$\mathbf{P}\{X = 80 \mid P = p\} = \binom{100}{80} p^{80}(1-p)^{20}.$$

Our posterior probability is thus:

$$\begin{aligned}\mathbf{P}\{X = 80 \mid P = p\} \cdot f_P(p) &= \binom{100}{80} p^{80}(1-p)^{20} \cdot 3(1-p)^2 \\ &= \binom{100}{80} \cdot 3p^{80}(1-p)^{22}.\end{aligned}$$

To find the maximizing p, we differentiate the posterior with respect to p, ignoring the constant unrelated to p, and set the derivative equal to 0, yielding:

$$0 = p^{80} \cdot 22 \cdot (1-p)^{21} \cdot (-1) + 80p^{79} \cdot (1-p)^{22}.$$

This in turn is easily solved by dividing both sides by $(1-p)^{21} \cdot p^{79}$, yielding:

$$p = \frac{80}{102}.$$

Thus,

$$\hat{P}_{\text{MAP}}(X = 80) = \frac{80}{102} \approx 78\%.$$

Question: This may still feel off to you. Shouldn't the prior matter more?

Answer: The answer lies in the number of people sampled. The fact that we sampled 100 people (picked uniformly at random) makes the prior distribution not so meaningful. Had we sampled a smaller number of people, then the prior distribution would matter much more.

Question: Repeat the voting example, where now we sample five people, uniformly at random and $X = 4$ report that they will vote. What is our estimate for p now?

Answer: You should get

$$\hat{P}_{\text{MAP}}(X = 4) = \frac{4}{7} \approx 57\%.$$

Observe that the prior distribution has much more of an effect now.

Another example of where estimation comes up has to do with signals that are (partially) corrupted by noise.

Example 17.7 (Deducing original signal in a noisy environment)

When sending a signal, θ, some random noise gets added to the signal, where the noise is represented by r.v. $N \sim \text{Normal}(0, \sigma_N^2)$. What is received is the sum of the original signal, θ, and the random noise, N. We represent the data received by r.v. X, where

$$X = \theta + N. \tag{17.11}$$

Suppose that we receive $X = x$. Based on that, we'd like to estimate the original signal, θ.

We will consider two situations: In the first, we have no prior information about the original signal. In the second, we have a prior distribution on the original signal.

Question: What is $\hat{\theta}_{\text{ML}}(X = x)$?

Since X is continuous, by Definition 16.6 we have that

$$\begin{aligned}\hat{\theta}_{\text{ML}}(X = x) &= \underset{\theta}{\text{argmax}}\, f_{X|\theta}(x) \\ &= \underset{\theta}{\text{argmax}}\, f_N(x - \theta), \qquad \text{by (17.11)}.\end{aligned}$$

Question: Now, where does N have its highest density?

Answer: Since $N \sim \text{Normal}(0, \sigma_N^2)$, we know that it achieves its highest density at 0. Thus, $f_N(x - \theta)$ is highest when $\theta = x$. So

$$\hat{\theta}_{\text{ML}}(X = x) = x. \tag{17.12}$$

Since this holds for all x, we have that $\hat{\theta}_{\text{ML}}(X) = X$.

Question: Why does (17.12) make sense?

Answer: We are trying to estimate the original signal, θ. We know that the noise

is symmetric, meaning that it is equally likely to add or subtract from the original signal. Thus, when we receive x, our best guess for the original signal is x.

Now consider that we have *additional information* in the form of a *prior distribution* on the original signal, represented by r.v. $\Theta \sim \text{Normal}(\mu, \sigma^2)$. Thus we can think of X as a sum of two independent random variables:

$$X = \Theta + N.$$

Again, we are trying to estimate the original signal, θ, given that we have received data $X = x$. To do this, we use a MAP estimator.

Question: What is $\hat{\Theta}_{\text{MAP}}(X = x)$?

Answer: By Definition 17.4,

$$\hat{\Theta}_{\text{MAP}}(X = x) = \underset{\theta}{\text{argmax}}\ f_{X|\Theta=\theta}(x) \cdot f_{\Theta}(\theta).$$

Now, since $X = \Theta + N$ and $\Theta \perp N$, we know that

$$[X \mid \Theta = \theta] \sim \text{Normal}(\theta, \sigma_N^2).$$

Hence,

$$f_{X|\Theta=\theta}(x) = \frac{1}{\sqrt{2\pi}\sigma_N} e^{-\frac{1}{2\sigma_N^2}(x-\theta)^2}. \tag{17.13}$$

So

$$\begin{aligned}
\hat{\Theta}_{\text{MAP}}(X = x) &= \underset{\theta}{\text{argmax}}\ f_{X|\Theta=\theta}(x) \cdot f_{\Theta}(\theta) \\
&= \underset{\theta}{\text{argmax}} \left(\frac{1}{\sqrt{2\pi}\sigma_N} e^{-\frac{1}{2\sigma_N^2}(x-\theta)^2} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(\theta-\mu)^2} \right) \\
&= \underset{\theta}{\text{argmax}} \left(e^{-\frac{1}{2\sigma_N^2}(x-\theta)^2 - \frac{1}{2\sigma^2}(\theta-\mu)^2} \right) \qquad \text{(can ignore constants)} \\
&= \underset{\theta}{\text{argmax}} \left(-\frac{1}{2\sigma_N^2}(x-\theta)^2 - \frac{1}{2\sigma^2}(\theta-\mu)^2 \right),
\end{aligned}$$

where the last line follows since it suffices to maximize the exponent. Let

$$g(\theta) = -\frac{1}{2\sigma_N^2}(x-\theta)^2 - \frac{1}{2\sigma^2}(\theta-\mu)^2.$$

To find the maximizing θ, we take the derivative and set it equal to 0, obtaining

$$0 = g'(\theta) = -\frac{1}{2\sigma_N^2} \cdot (-2)(x-\theta) - \frac{1}{2\sigma^2} \cdot 2(\theta - \mu),$$

which easily solves to

$$\theta = \frac{\frac{x}{\sigma_N^2} + \frac{\mu}{\sigma^2}}{\frac{1}{\sigma_N^2} + \frac{1}{\sigma^2}} = \frac{\sigma^2}{\sigma^2 + \sigma_N^2} \cdot x + \frac{\sigma_N^2}{\sigma^2 + \sigma_N^2} \cdot \mu.$$

Thus,

$$\hat{\Theta}_{\text{MAP}}(X = x) = \frac{\sigma^2}{\sigma^2 + \sigma_N^2} \cdot x + \frac{\sigma_N^2}{\sigma^2 + \sigma_N^2} \cdot \mu. \tag{17.14}$$

Question: What is the meaning behind the fact that the MAP estimate of θ in (17.14) looks like a weighted average?

Answer: Observe that (17.14) represents a weighted average of the received data, x, and the prior mean μ. So the MAP takes into account both the received data and also the prior distribution. Looking at the weights, we see that they depend on the variance of the original signal, σ^2, and also the variance of the noise, σ_N^2. If the variance of the noise is (relatively) low, then we weigh the received data, x, more highly in our estimate. If the variance of the noise is (relatively) high, then we weigh the mean of the prior, μ, more highly in our estimate.

17.4 Minimum Mean Square Error Estimator

This chapter has been devoted to coming up with an estimator, in the case where we have a prior distribution, denoted by r.v. Θ, and also data, denoted by r.v. X. The idea has been to create a *posterior distribution*, denoted by

$$[\Theta \mid X = x].$$

Then, from Definition 17.4,

$$\hat{\Theta}_{\text{MAP}}(X = x) = \underset{\theta}{\text{argmax}}\, \mathbf{P}\{\Theta = \theta \mid X = x\}.$$

We can view $\hat{\Theta}_{\text{MAP}}$ as the **mode of the posterior** distribution. In the case of a discrete distribution, this represents the value, θ, that comes up most frequently in the posterior distribution. In the case of a continuous distribution, this represents the value with highest density.

One could alternatively define a different Bayesian estimator for θ that is the **mean of the posterior** distribution. We do this now.

Definition 17.8 *Our goal is to estimate some unknown θ. We are given a prior distribution Θ on the possible values for θ. We also have experimental data, denoted by r.v. X.*

We say that $\hat{\Theta}_{MMSE}(X)$ is the **minimum mean squared error (MMSE) estimator** *of θ, where*

$$\hat{\Theta}_{MMSE}(X) = \mathbf{E}\left[\Theta \mid X\right].$$

This is shorthand for saying that, for any x,

$$\hat{\Theta}_{MMSE}(X = x) = \mathbf{E}\left[\Theta \mid X = x\right].$$

Note that $\hat{\Theta}_{MMSE}(X)$ is a function of a r.v. X and thus is a r.v., while $\hat{\Theta}_{MMSE}(X = x)$ is a constant.

The estimator $\hat{\Theta}_{\text{MMSE}}(X = x)$ gets its name from the fact that this estimator in fact produces the minimum possible mean squared error of any estimator. We will prove this fact in Theorem 17.12. For now, let's consider a few examples of this new estimator to better understand how it compares with the MAP estimator.

Example 17.9 (Coin with unknown probability: revisited)

We revisit Example 7.14, where there is a coin with some unknown *bias*, where the "bias" of the coin is its probability of coming up heads. We are given that the coin's bias is drawn from distribution $P \sim \text{Uniform}(0, 1)$. We are also given that the coin has resulted in $X = 10$ heads out of the first 10 flips. Based on this, we would like to estimate the coin's bias.

Question: What is $\hat{P}_{\text{MMSE}}(X = 10)$?

Answer:

$$\hat{P}_{\text{MMSE}}(X = 10) = \mathbf{E}\left[P \mid X = 10\right].$$

To derive this, we need to first derive the conditional probability density function (p.d.f.) of P given $X = 10$:

$$f_{P|X=10}(p) = \frac{\mathbf{P}\left\{X = 10 \mid P = p\right\} \cdot f_P(p)}{\mathbf{P}\left\{X = 10\right\}}$$

$$= \begin{cases} \frac{p^{10} \cdot 1}{\mathbf{P}\{X=10\}} & \text{if } 0 \le p \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

Here,

$$\begin{aligned}\mathbf{P}\{X=10\} &= \int_0^1 \mathbf{P}\{X=10 \mid P=p\} \cdot f_P(p)dp \\ &= \int_0^1 p^{10}dp \\ &= \frac{1}{11}.\end{aligned}$$

So,

$$\begin{aligned}f_{P|X=10}(p) &= \frac{\mathbf{P}\{X=10 \mid P=p\} \cdot f_P(p)}{\mathbf{P}\{X=10\}} \\ &= \begin{cases} 11p^{10} & \text{if } 0 \le p \le 1 \\ 0 & \text{otherwise} \end{cases}. \end{aligned} \tag{17.15}$$

Hence,

$$\hat{P}_{\text{MMSE}}(X=10) = \mathbf{E}[P \mid X=10] = \int_0^1 p11p^{10}dp = \frac{11}{12}.$$

Question: How does $\hat{P}_{\text{MMSE}}(X=10)$ compare with $\hat{P}_{\text{MAP}}(X=10)$?

Answer: The prior P is continuous and X is discrete, so using Definition 17.4 and (17.15), we have:

$$\hat{P}_{\text{MAP}}(X=10) = \underset{p}{\text{argmax}}\ f_{P|X=10}(p) = \underset{p}{\text{argmax}}\ \left(11p^{10}\right) = 1.$$

Question: Which is the more believable estimator?

Answer: This is a matter of opinion, but it feels like the MMSE estimator does a better job of capturing the prior distribution than the MAP estimator.

Let's consider one more example comparing the MMSE estimator and the MAP estimator.

Example 17.10 (Supercomputing: estimating the true job size)

In supercomputing centers, users are asked to provide an upper bound on their job's size (running time). The upper bound provided by the user is typically several times larger than the job's actual size [49]. We can think of the upper bound provided as a scalar multiple of the original job size. The relationship between the original job and upper bound provided can be represented by:

$$X = S \cdot \Theta,$$

where Θ is a r.v. denoting the original job size, S is a scalar multiple where $S \geq 1$, and X is the reported upper bound. We will assume that $S \perp \Theta$. Given a value on the upper bound, $X = x$, how do we estimate the original job size, $\Theta = \theta$, from this? Specifically, we will be interested in deriving $\hat{\Theta}_{\text{MAP}}(X = x)$ and $\hat{\Theta}_{\text{MMSE}}(X = x)$.

To keep the computations from getting too messy, we assume: $\Theta \sim \text{Pareto}(\alpha = 3)$ and $S \sim \text{Pareto}(\alpha = 2)$. Hence,

$$f_\Theta(\theta) = 3\theta^{-4}, \qquad \text{if } \theta \geq 1$$
$$f_S(s) = 2s^{-3}, \qquad \text{if } s \geq 1.$$

Both estimators will require deriving $f_{\Theta|X=x}(\theta)$. To get there, we will have to start with the other direction, namely $f_{X|\Theta=\theta}(x)$.

Question: Given that $X = S \cdot \Theta$, what is $f_{X|\Theta=\theta}(x)$?

Hint: Is it $f_S\left(\frac{x}{\theta}\right)$?

Answer: The correct answer is

$$f_{X|\Theta=\theta}(x) = \frac{1}{\theta} \cdot f_S\left(\frac{x}{\theta}\right), \qquad x \geq \theta \geq 1.$$

To see why, recall that we need to make the arguments over probabilities, not densities:

$$\mathbf{P}\{X \leq x \mid \Theta = \theta\} = \mathbf{P}\left\{S \leq \frac{x}{\theta}\right\}$$
$$\int_{t=0}^{t=x} f_{X|\Theta=\theta}(t)dt = \int_{t=0}^{t=\frac{x}{\theta}} f_S(t)dt$$
$$\frac{d}{dx}\int_{t=0}^{t=x} f_{X|\Theta=\theta}(t)dt = \frac{d}{dx}\int_{t=0}^{t=\frac{x}{\theta}} f_S(t)dt$$
$$f_{X|\Theta=\theta}(x) = \frac{1}{\theta} f_S\left(\frac{x}{\theta}\right) \qquad \text{by FTC, see (1.6) and (1.7) .}$$

We use our conditional density to get the joint density as follows:

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta=\theta}(x) \cdot f_\Theta(\theta) = \frac{1}{\theta} \cdot 2\left(\frac{x}{\theta}\right)^{-3} \cdot 3\theta^{-4} = \frac{6}{\theta^2 x^3}.$$

We can integrate the joint density to get $f_X(x)$, as follows:

$$\begin{aligned} f_X(x) &= \int_{\theta=1}^{\theta=x} f_{X,\Theta}(x,\theta)d\theta \\ &= \int_{\theta=1}^{\theta=x} \frac{6}{\theta^2 x^3} d\theta \\ &= 6x^{-3} - 6x^{-4}, \qquad x \geq 1. \end{aligned}$$

We are finally ready to obtain $f_{\Theta|X=x}(\theta)$:

$$\begin{aligned} f_{\Theta|X=x}(\theta) &= \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} \\ &= \frac{\frac{6}{\theta^2 x^3}}{6x^{-3} - 6x^{-4}} \\ &= \frac{x}{\theta^2 x - \theta^2} \\ &= \frac{1}{\theta^2} \cdot \frac{x}{x-1}, \qquad x \geq \theta \geq 1. \end{aligned}$$

Question: So what is $\hat{\Theta}_{\text{MAP}}(X=x)$?

Answer:

$$\hat{\Theta}_{\text{MAP}}(X=x) = \underset{\theta}{\text{argmax}}\, f_{\Theta|X=x}(\theta) = \underset{\theta}{\text{argmax}}\, \frac{1}{\theta^2} \cdot \frac{x}{x-1} = 1.$$

Question: What is $\hat{\Theta}_{\text{MMSE}}(X=x)$?

Answer:

$$\begin{aligned} \hat{\Theta}_{\text{MMSE}}(X=x) &= \mathbf{E}\left[\Theta \mid X=x\right] \\ &= \int_{\theta=1}^{\theta=x} \theta \cdot f_{\Theta|X=x}(\theta)d\theta \\ &= \frac{x}{x-1} \int_{\theta=1}^{\theta=x} \frac{1}{\theta} d\theta \\ &= \frac{x \ln x}{x-1} \\ &= \ln x + \frac{\ln x}{x-1}. \end{aligned}$$

Question: Which is the more believable estimator?

Answer: The MAP estimator is pretty useless, given that it simply returns an

answer of $\theta = 1$. The problem is that the density of the prior is maximized at $\theta = 1$, and somehow this isn't improved when we look at the conditional density.

The MMSE estimator returns a more reasonable answer of $\theta \approx \ln x$. This makes more sense given that the upper bound on job size is x.

Question: You might wonder if the answers change if we make the problem a little more symmetric, where Θ and S have the same distribution. For example, what do you think might happen if $\Theta \sim \text{Pareto}(\alpha = 2)$ and $S \sim \text{Pareto}(\alpha = 2)$?

Answer: We find that, disappointingly, $\hat{\Theta}_{\text{MAP}}(X = x)$ remains at 1. However, now

$$\hat{\Theta}_{\text{MMSE}}(X = x) = \frac{x - 1}{\ln x}.$$

17.5 Measuring Accuracy in Bayesian Estimators

We have seen different estimators, producing different results. It is helpful to have some metrics for evaluating the *accuracy* of our estimators. One common metric for measuring the accuracy of estimators is the mean squared error (MSE).

Recall the MSE as given by Definition 15.4, when we were looking at non-Bayesian estimators. Here, θ was an unknown constant, X represented the sample data, and $\hat{\theta}(X)$ was our estimator for θ. Under this setting we defined:

$$\textbf{MSE}\left(\hat{\theta}(X)\right) = \textbf{E}\left[\left(\hat{\theta}(X) - \theta\right)^2\right]. \tag{17.16}$$

For Bayesian estimators we need an *adaptation* of the definition in (17.16) because θ is no longer a constant, but rather is drawn from a prior distribution, Θ. For Bayesian estimators, we use Definition 17.11 for the MSE.

Definition 17.11 *Let $\hat{\Theta}(X)$ be an estimator where Θ represents the prior distribution and X the sample data. Then the* **mean squared error (MSE)** *of $\hat{\Theta}(X)$ is defined by*

$$\textbf{MSE}\left(\hat{\Theta}(X)\right) = \textbf{E}\left[\left(\hat{\Theta}(X) - \Theta\right)^2\right]. \tag{17.17}$$

Question: How should one interpret Definition 17.11? What is the expectation over?

Answer: Both terms within the expectation in (17.17) are random variables.

The first term is a r.v. which is a function of just X (once a value of X is specified, $\hat{\Theta}(X)$ becomes a constant). The second term is the r.v. Θ. The expectation in (17.17) is over the joint distribution of Θ and X (that is, it's a double sum).

At first, Definition 17.11 may seem a little strange. However, it's actually very similar to our definition in (17.16) except that now the value of θ is picked from the prior distribution. To see this, we condition on θ:

$$\begin{aligned}\mathbf{MSE}\left(\hat{\Theta}(X)\right) &= \mathbf{E}\left[\left(\hat{\Theta}(X)-\Theta\right)^2\right] \\ &= \int_\theta \mathbf{E}\left[\left(\hat{\Theta}(X)-\Theta\right)^2 \,\middle|\, \Theta=\theta\right] f_\Theta(\theta)d\theta \\ &= \int_\theta \mathbf{E}\left[\left(\hat{\Theta}(X)-\theta\right)^2 \,\middle|\, \Theta=\theta\right] f_\Theta(\theta)d\theta.\end{aligned}$$

Observe that the integrand looks very similar to (17.16). The point is, whatever our chosen value, θ, we want to say that our estimator, $\hat{\Theta}(X)$, is close to that value in expectation.

Now recall the estimator $\hat{\Theta}_{\text{MMSE}}(X)$. Theorem 17.12 says that this estimator has the lowest MSE compared to all other estimators.

Theorem 17.12 *$\hat{\Theta}_{MMSE}(X)$ minimizes the MSE over all estimators $\hat{\Theta}(X)$.*

Proof: We start by defining:

$$\mathbf{MSE}\left(\hat{\Theta}(X=x)\right) = \mathbf{E}\left[\left(\hat{\Theta}(X)-\Theta\right)^2 \,\middle|\, X=x\right]. \tag{17.18}$$

We will show that $\hat{\Theta}_{\text{MMSE}}(X=x)$ minimizes $\mathbf{MSE}\left(\hat{\Theta}(X=x)\right)$ for all values of x. It then follows that $\hat{\Theta}_{\text{MMSE}}(X)$ minimizes the MSE over all estimators $\hat{\Theta}(X)$.

$$\begin{aligned}\mathbf{MSE}\left(\hat{\Theta}(X=x)\right) &= \mathbf{E}\left[\left(\hat{\Theta}(X)-\Theta\right)^2 \,\middle|\, X=x\right] \\ &= \mathbf{E}\left[\hat{\Theta}(X)^2 - 2\hat{\Theta}(X)\Theta + \Theta^2 \,\middle|\, X=x\right] \\ &= \hat{\Theta}(X=x)^2 - 2\hat{\Theta}(X=x)\mathbf{E}\left[\Theta \mid X=x\right] \\ &\quad + \mathbf{E}\left[\Theta^2 \mid X=x\right].\end{aligned} \tag{17.19}$$

We now want to find the minimizing $\hat{\Theta}(X=x)$ in (17.19). Recall that $\hat{\Theta}(X=x)$ is a constant function of x. We'll denote this by $c(x)$ and replace $\hat{\Theta}(X=x)$ with

$c(x)$ throughout, obtaining:

$$\begin{aligned}\textbf{MSE}\left(\hat{\Theta}(X=x)\right) &= c(x)^2 - 2c(x)\mathbf{E}\left[\Theta \mid X=x\right] + \mathbf{E}\left[\Theta^2 \mid X=x\right] \\ &= \left(c(x) - \mathbf{E}\left[\Theta \mid X=x\right]\right)^2 + \mathbf{E}\left[\Theta^2 \mid X=x\right] \\ &\quad - \mathbf{E}\left[\Theta \mid X=x\right]^2,\end{aligned}$$

which is clearly minimized when the first term is 0, namely when

$$c(x) = \mathbf{E}\left[\Theta \mid X=x\right] = \hat{\Theta}_{\text{MMSE}}(X=x). \qquad \blacksquare$$

17.6 Exercises

17.1 **Deducing original signal in a noisy environment**
We have an original signal, represented by r.v. Θ, where $\Theta \sim \text{Normal}(0,1)$. We also have noise, represented by r.v. N, where $N \sim \text{Normal}(0,1)$. The received signal, represented by r.v. X, is then:

$$X = \Theta + N.$$

Derive the MMSE estimator, $\hat{\Theta}_{\text{MMSE}}(X=x)$. How does your answer compare to $\hat{\Theta}_{\text{MAP}}(X=x)$ under the same setting?

17.2 **Mean squared error of the MMSE estimator**
In Theorem 17.12, we saw that $\hat{\Theta}_{\text{MMSE}}(X)$ minimizes the MSE. But what exactly is this error? Prove that $\textbf{MSE}\left(\hat{\Theta}_{\text{MMSE}}(X=x)\right)$ is the variance of the posterior distribution.

17.3 **MMSE estimator for gold vs. silver coin problem**
For the Bayesian coin problem from Example 17.2, derive the MMSE estimator, $\hat{P}_{\text{MMSE}}(X)$.

17.4 **Hypothesis testing for COVID: MLE vs. MAP**
To determine whether you have COVID, you take an antigen self-test. Rather than outputting "yes" or "no," the test outputs a number, L, from the set $\{0,1,2,3\}$, where L indicates the level of antigen detected. The level L is not a perfect indicator. Table 17.1, called a "likelihood matrix," shows the probability distribution over the level output by the test, depending on whether you have COVID or not. For example, if you don't have COVID, then the test outputs $L=0$ with probability 0.6 and $L=1$ with probability 0.3, etc. By contrast, if you have COVID, the probability distribution is more biased toward higher levels.

	$L=0$	$L=1$	$L=2$	$L=3$
H_0: Don't have COVID	0.6	0.3	0.1	0.0
H_1: Have COVID	0.1	0.2	0.3	0.4

Table 17.1 *Likelihood matrix.*

Consider two hypotheses: H_0 that you don't have COVID and H_1 that you do.

(a) For each possible reading of L, determine which hypothesis is returned by the MLE, which returns the hypothesis with highest likelihood.

(b) For each possible reading of L, determine which hypothesis is returned by the MAP decision rule. Assume that $\mathbf{P}\{H_0\} = 0.8$ and $\mathbf{P}\{H_1\} = 0.2$.

17.5 **Estimating the minimum: MLE vs. MAP**

You observe 10 i.i.d. data samples, $X_1, X_2, \ldots, X_{10} \sim \text{Uniform}(a, 1)$. You know that $a \geq 0$ but not the exact value of a. Your goal is to estimate a.

(a) Determine $\hat{a}_{\text{ML}}(X_1, X_2, \ldots, X_{10})$, the ML estimator of a.

(b) Suppose that we have a prior on a, denoted by r.v. A, with p.d.f.:

$$f_A(a) = \begin{cases} \frac{20e^{-20a}}{1-e^{-20}} & \text{if } 0 \leq a \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

Determine $\hat{A}_{\text{MAP}}(X_1, X_2, \ldots, X_{10})$, the MAP estimator of a.

17.6 **Interaction graph**

Annie, Ben, and Caroline are three CMU students. CMU has only two clubs: PnC club and Buggy club. Each student must join one and only one club. Suppose that you (as an outsider) know that Annie has joined the PnC club, but you cannot see which clubs Ben and Caroline join. However, you can see the interaction graph in Figure 17.2. The interaction graph tells us something about which students at CMU interact with other students. But

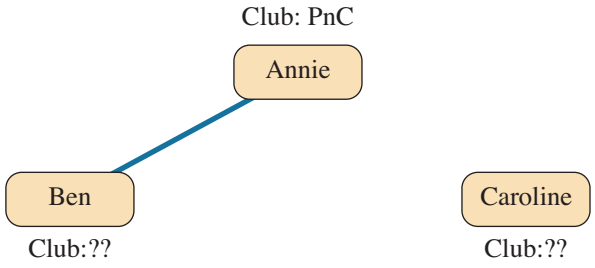

Figure 17.2 *Interaction graph for Exercise 17.6.*

the interaction graph is not perfect: An edge between two people exists with probability $\frac{1}{2}$ if the two people are in the same club and exists with probability $\frac{1}{6}$ if the two people are in different clubs.

(a) What is your ML estimate of the clubs Ben and Caroline each joins?

(b) Suppose that you know Ben and Caroline well enough to have the following prior: Ben joins PnC with probability $\frac{3}{8}$ and joins Buggy with probability $\frac{5}{8}$. Caroline joins PnC with probability $\frac{7}{8}$ and joins Buggy with probability $\frac{1}{8}$. They make their choices independently. What is your MAP estimate of the clubs Ben and Caroline each joins?

17.7 Error correcting codes

Suppose you want to transmit a message to your friend through a wireless channel. Your message, denoted as M, has three possible values with this distribution:

$$\mathbf{P}\{M=0\}=\frac{1}{2}, \qquad \mathbf{P}\{M=1\}=\frac{7}{16}, \qquad \mathbf{P}\{M=2\}=\frac{1}{16}.$$

You have decided to use a 5-bit string, $U = U_1U_2U_3U_4U_5$ to encode message M as follows:

$$\begin{aligned} M=0 &\Longrightarrow U=00000,\\ M=1 &\Longrightarrow U=11110,\\ M=2 &\Longrightarrow U=10101. \end{aligned}$$

Here, the leftmost two bits, U_1, U_2, are used to differentiate among the values of M, and the remaining three bits, U_3, U_4, U_5 are redundant bits for error correcting – that is, the remaining bits reinforce the information in the first two bits. This coding scheme sets $U_3 = U_1$, $U_4 = U_2$, and $U_5 = U_1 + U_2 \mod 2$.

When you transmit the string U, each bit U_i gets flipped with probability $\epsilon = 0.2$, and $U_1, U_2, \ldots, U_5$ get flipped independently. Let $X = X_1X_2X_3X_4X_5$ denote the string that your friend receives. Your friend must estimate the value of M based on the received string. For two binary strings with the same length, the **Hamming distance** between the strings, denoted by $d_H(\cdot,\cdot)$, is defined to be the number of bits on which the two strings differ.

(a) Suppose your friend decodes X by comparing X with the three strings $\{00000, 11110, 10101\}$ and selecting the string that has the smallest Hamming distance to X. Then she declares that the value of M that corresponds to the selected string is the value transmitted. When there is a tie, she declares the smaller value for M. For example, if she receives $X = 10100$, then 10101 is the string from $\{00000, 11110, 10101\}$ that is the closest to X. So she declares that $M = 2$ is the value transmitted. If she receives $X = 11000$, then $d_H(X, 00000) = 2$, $d_H(X, 11110) = 2$, and $d_H(X, 10101) = 3$. So she breaks the tie and declares that $M = 0$.

(i) What type of estimation is your friend doing?

(ii) Suppose that the received string is $X = k \triangleq k_1k_2k_3k_4k_5$. How does your friend determine whether M equals 0 or 1 or 2? (Write down the probabilities involved using expressions involving $d_H(k, \cdot)$.).

(b) If your friend uses a MAP decoder to estimate M, what will she declare when she receives $X = 10100$?

17.8 **MMSE estimator of temperature given noise**

There is a heat source with temperature $T \sim \text{Normal}(100, 16)$. You want to know the value of the temperature, $T = t$, but you cannot directly access the source. You are, however, able to approximately measure the temperatures at two nearby locations: Let X_A denote your measurement of the temperature at location A, which is 1 mile away and known to have temperature $\frac{T}{2}$. Let X_B denote your measurement at location B, which is 2 miles away with temperature $\frac{T}{4}$. Unfortunately, X_A and X_B are both affected by noise, and hence what you actually read is:

$$X_A = \frac{T}{2} + W_A, \qquad W_A \sim \text{Normal}(0, 1),$$
$$X_B = \frac{T}{4} + W_B, \qquad W_B \sim \text{Normal}(0, 1),$$

where the noises W_A, W_B, and T are independent.

(a) What is the conditional p.d.f. of T given that you observe $X_A = x_A$ and $X_B = x_B$?

(b) What distribution does T follow given you observe $X_A = x_A$ and $X_B = x_B$?

(c) What is $\hat{T}_{\text{MMSE}}(X_A, X_B)$?

[Hint: If a r.v. Y has a p.d.f. of the form $f_Y(y) = C \cdot e^{-\frac{1}{2}(ay^2+by+c)}$, where C, a, b, c are constants, independent of y, then $Y \sim \text{Normal}\left(-\frac{b}{2a}, \frac{1}{a}\right)$.]

17.9 **The MMSE is an unbiased estimator**

Prove that the MMSE estimator is unbiased. That is, $\mathbf{E}\left[\hat{\Theta}_{\text{MMSE}}(X)\right] = \mathbf{E}[\Theta]$.

17.7 Acknowledgment

This chapter was written in collaboration with Weina Wang, who was a major contributor to the chapter contents and the exercises.