Introduction to Machine Learning

A quick refresher course in probability theory

Third lecture, 02.03.2022

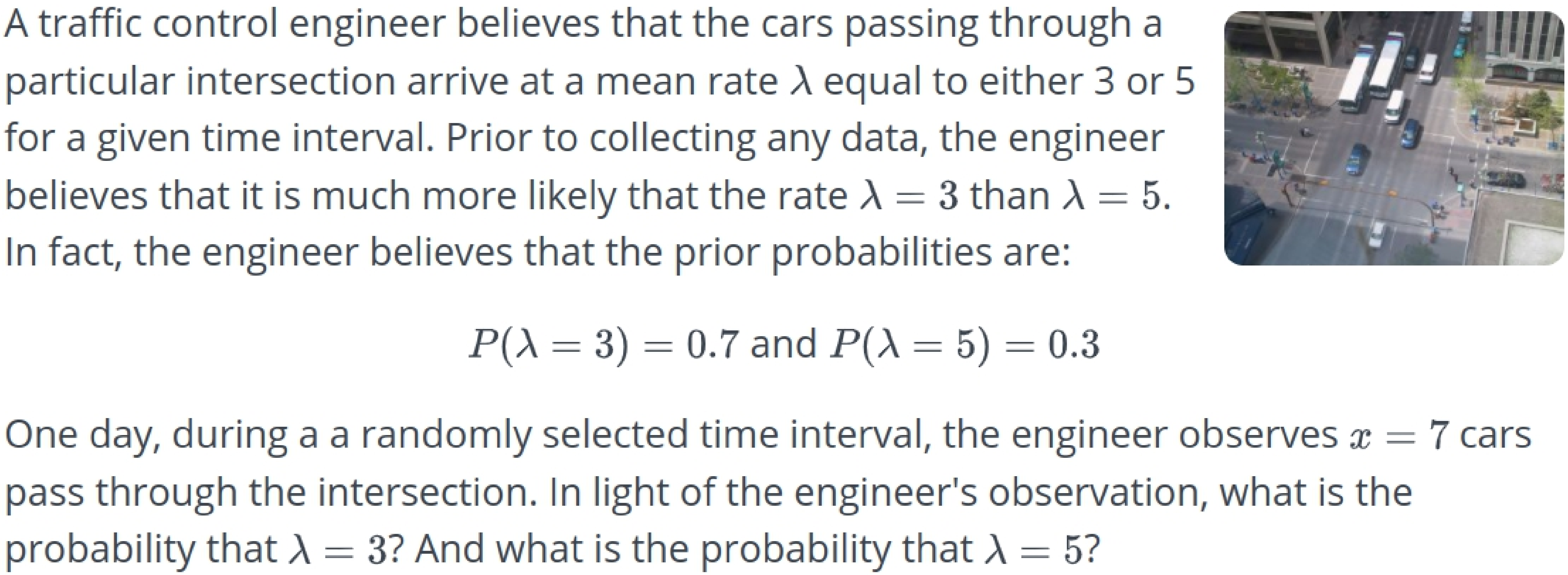
Phuc Loi Luu, PhD
p.luu@garvan.org.au
luu.p.loi@googlemail.com

Roadmap for today

1. *Bayesian Estimation*

2. *Maximum A Posteriori (MAP) Estimation*

Example

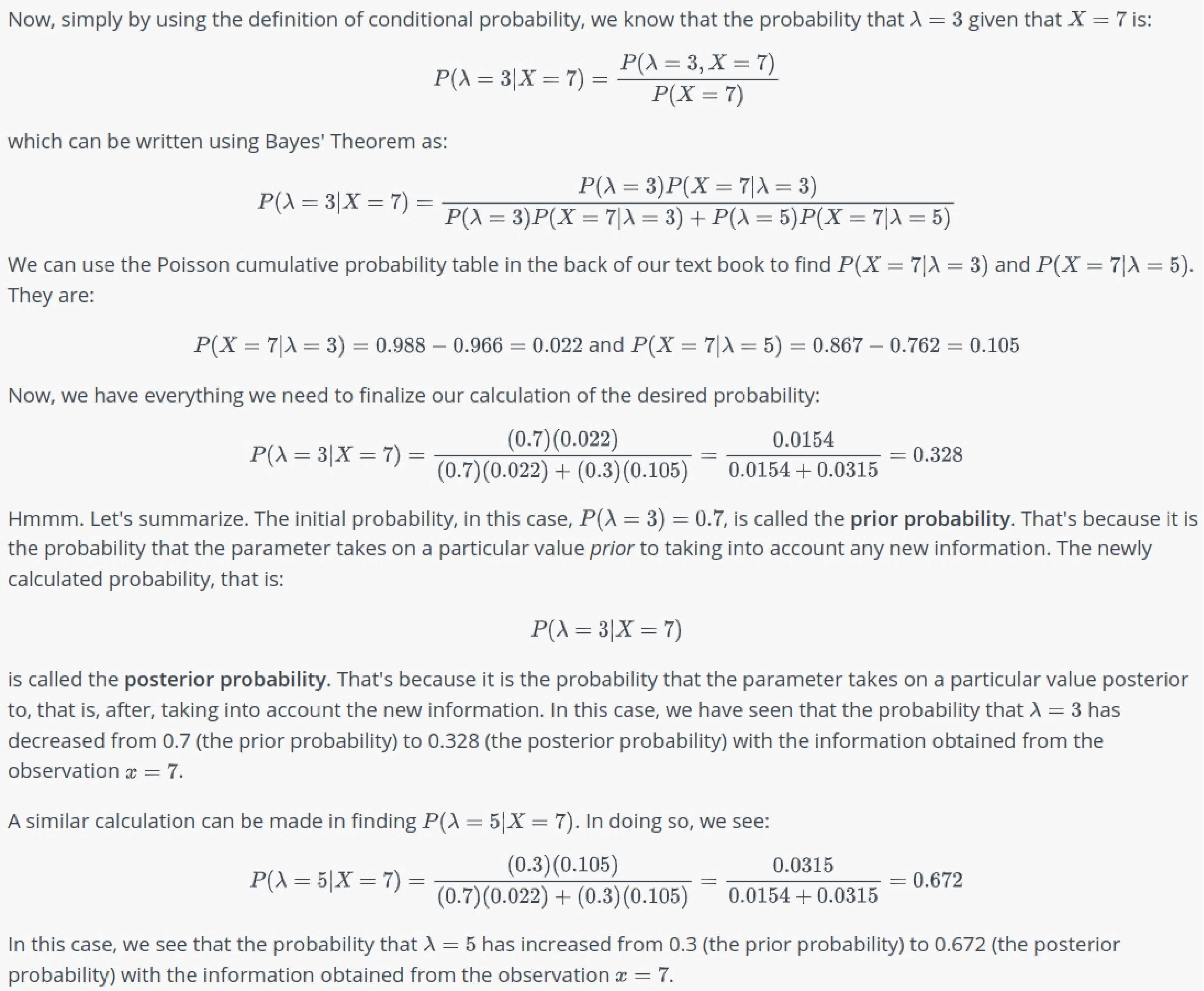


A traffic control engineer believes that the cars passing through a particular intersection arrive at a mean rate λ equal to either 3 or 5 for a given time interval. Prior to collecting any data, the engineer believes that it is much more likely that the rate $\lambda = 3$ than $\lambda = 5$. In fact, the engineer believes that the prior probabilities are:

$$P(\lambda = 3) = 0.7 \text{ and } P(\lambda = 5) = 0.3$$

One day, during a a randomly selected time interval, the engineer observes $x = 7$ cars pass through the intersection. In light of the engineer's observation, what is the probability that $\lambda = 3$? And what is the probability that $\lambda = 5$?

https://online.stat.psu.edu/stat415/lesson/23/23.2



Now, simply by using the definition of conditional probability, we know that the probability that $\lambda = 3$ given that $X = 7$ is:

$$P(\lambda = 3|X = 7) = \frac{P(\lambda = 3, X = 7)}{P(X = 7)}$$

which can be written using Bayes' Theorem as:

$$P(\lambda = 3|X = 7) = \frac{P(\lambda = 3)P(X = 7|\lambda = 3)}{P(\lambda = 3)P(X = 7|\lambda = 3) + P(\lambda = 5)P(X = 7|\lambda = 5)}$$

We can use the Poisson cumulative probability table in the back of our text book to find $P(X = 7|\lambda = 3)$ and $P(X = 7|\lambda = 5)$. They are:

$$P(X = 7|\lambda = 3) = 0.988 - 0.966 = 0.022 \text{ and } P(X = 7|\lambda = 5) = 0.867 - 0.762 = 0.105$$

Now, we have everything we need to finalize our calculation of the desired probability:

$$P(\lambda = 3|X = 7) = \frac{(0.7)(0.022)}{(0.7)(0.022) + (0.3)(0.105)} = \frac{0.0154}{0.0154 + 0.0315} = 0.328$$

Hmmm. Let's summarize. The initial probability, in this case, $P(\lambda = 3) = 0.7$, is called the **prior probability**. That's because it is the probability that the parameter takes on a particular value *prior* to taking into account any new information. The newly calculated probability, that is:

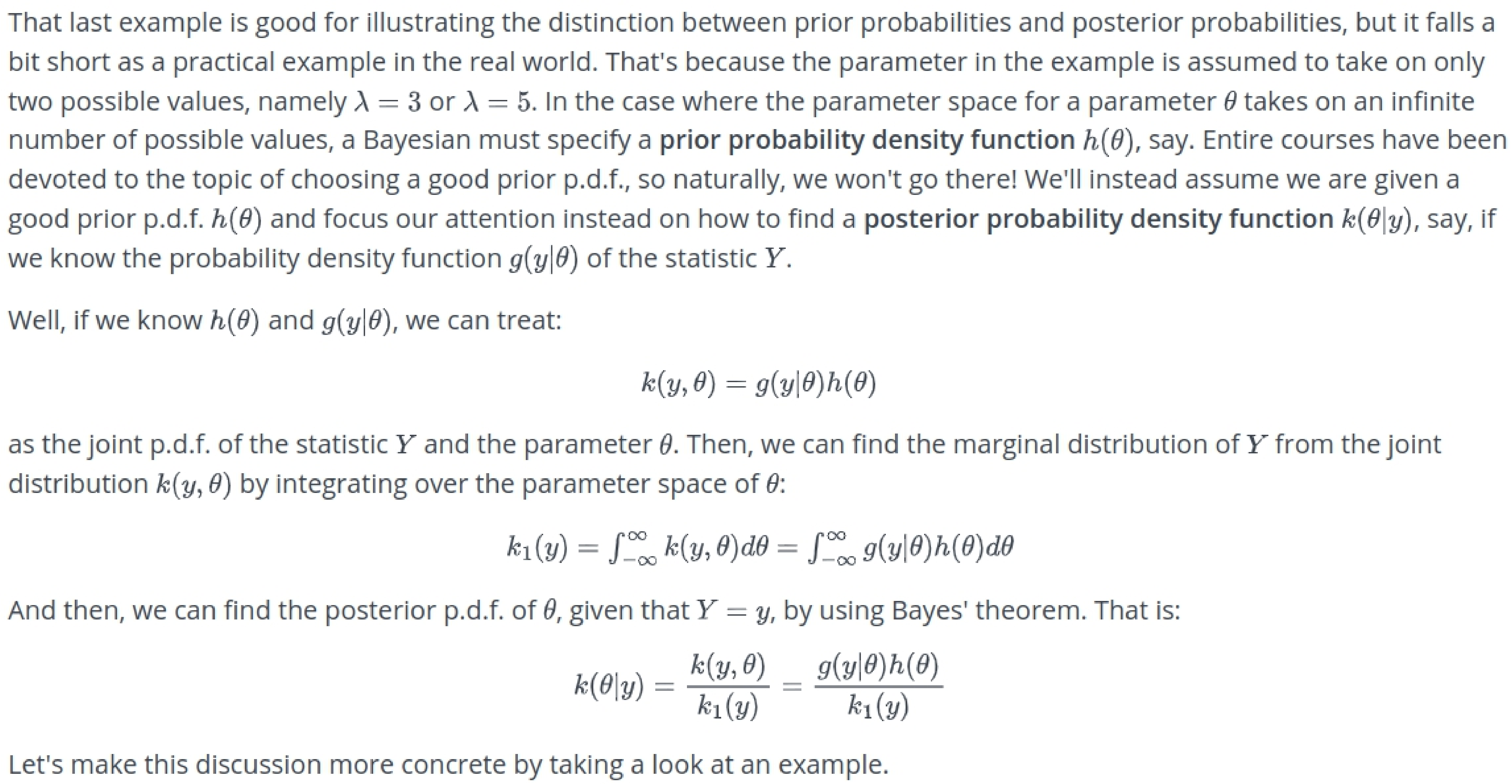
$$P(\lambda = 3|X = 7)$$

is called the **posterior probability**. That's because it is the probability that the parameter takes on a particular value posterior to, that is, after, taking into account the new information. In this case, we have seen that the probability that $\lambda = 3$ has decreased from 0.7 (the prior probability) to 0.328 (the posterior probability) with the information obtained from the observation $x = 7$.

A similar calculation can be made in finding $P(\lambda = 5|X = 7)$. In doing so, we see:

$$P(\lambda = 5|X = 7) = \frac{(0.3)(0.105)}{(0.7)(0.022) + (0.3)(0.105)} = \frac{0.0315}{0.0154 + 0.0315} = 0.672$$

In this case, we see that the probability that $\lambda = 5$ has increased from 0.3 (the prior probability) to 0.672 (the posterior probability) with the information obtained from the observation $x = 7$.



That last example is good for illustrating the distinction between prior probabilities and posterior probabilities, but it falls a bit short as a practical example in the real world. That's because the parameter in the example is assumed to take on only two possible values, namely $\lambda = 3$ or $\lambda = 5$. In the case where the parameter space for a parameter θ takes on an infinite number of possible values, a Bayesian must specify a **prior probability density function** $h(\theta)$, say. Entire courses have been devoted to the topic of choosing a good prior p.d.f., so naturally, we won't go there! We'll instead assume we are given a good prior p.d.f. $h(\theta)$ and focus our attention instead on how to find a **posterior probability density function** $k(\theta|y)$, say, if we know the probability density function $g(y|\theta)$ of the statistic Y.

Well, if we know $h(\theta)$ and $g(y|\theta)$, we can treat:

$$k(y, \theta) = g(y|\theta)h(\theta)$$

as the joint p.d.f. of the statistic Y and the parameter θ. Then, we can find the marginal distribution of Y from the joint distribution $k(y, \theta)$ by integrating over the parameter space of θ:

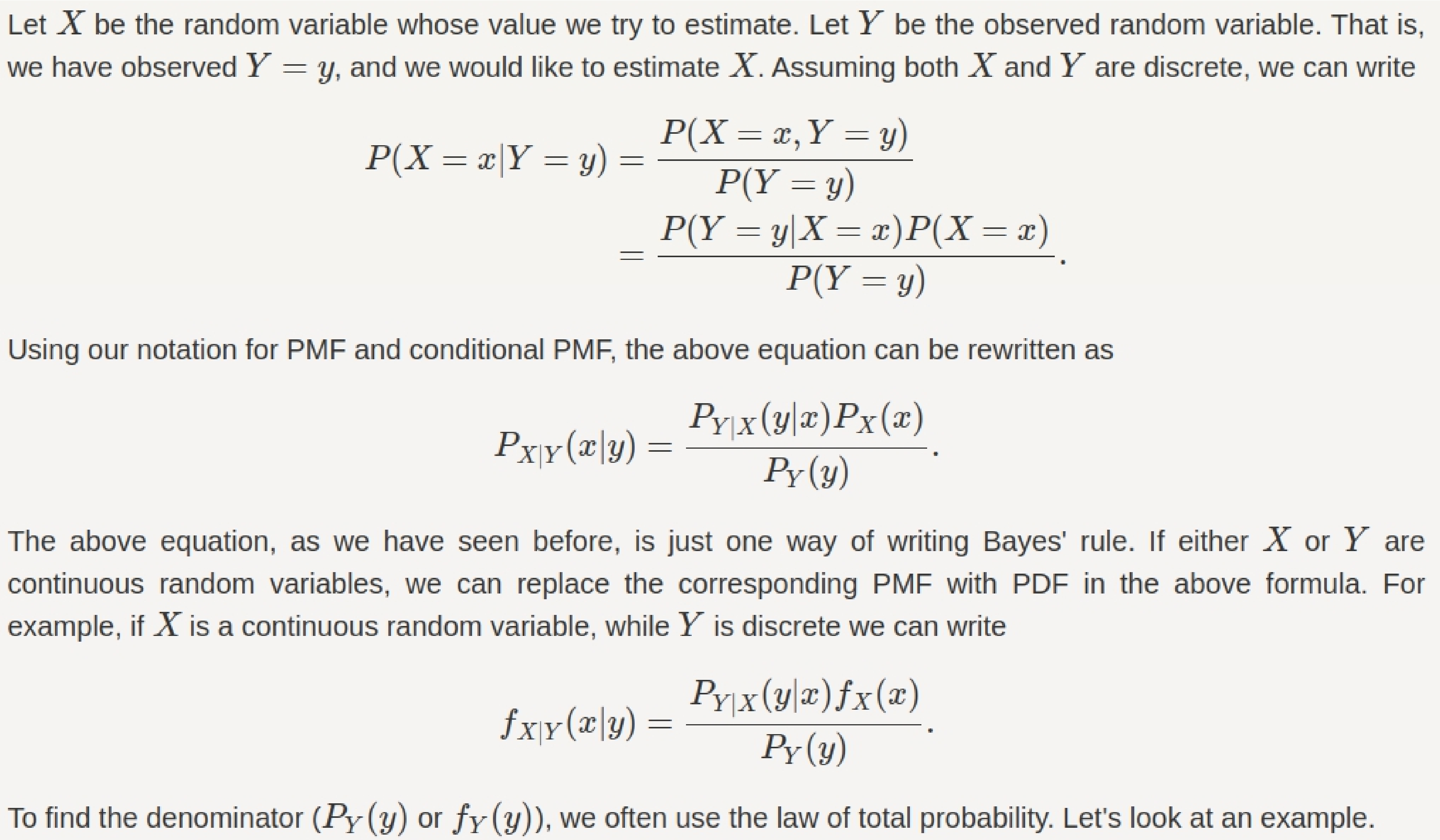
$$k_1(y) = \int_{-\infty}^{\infty} k(y, \theta) d\theta = \int_{-\infty}^{\infty} g(y|\theta)h(\theta) d\theta$$

And then, we can find the posterior p.d.f. of θ, given that $Y = y$, by using Bayes' theorem. That is:

$$k(\theta|y) = \frac{k(y, \theta)}{k_1(y)} = \frac{g(y|\theta)h(\theta)}{k_1(y)}$$

Let's make this discussion more concrete by taking a look at an example.

Prior and Posterior



Let X be the random variable whose value we try to estimate. Let Y be the observed random variable. That is, we have observed $Y = y$, and we would like to estimate X. Assuming both X and Y are discrete, we can write

$$\begin{aligned} P(X=x|Y=y) &= \frac{P(X=x, Y=y)}{P(Y=y)} \\ &= \frac{P(Y=y|X=x)P(X=x)}{P(Y=y)}. \end{aligned}$$

Using our notation for PMF and conditional PMF, the above equation can be rewritten as

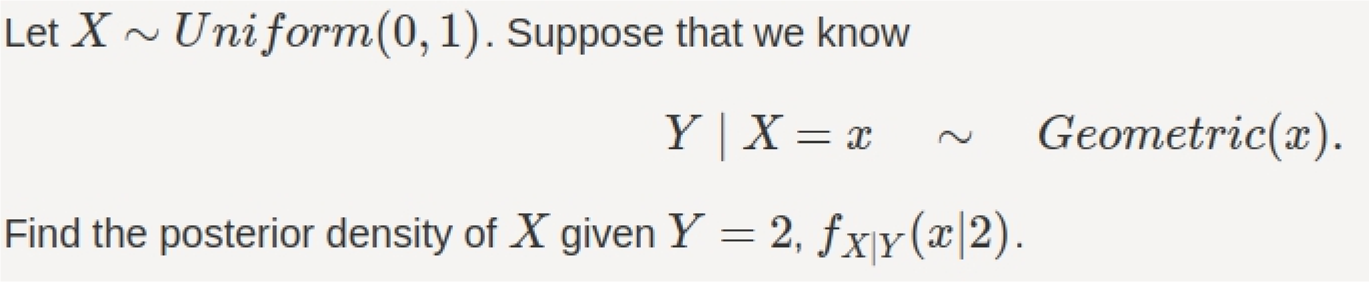
$$P_{X|Y}(x|y) = \frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)}.$$

The above equation, as we have seen before, is just one way of writing Bayes' rule. If either X or Y are continuous random variables, we can replace the corresponding PMF with PDF in the above formula. For example, if X is a continuous random variable, while Y is discrete we can write

$$f_{X|Y}(x|y) = \frac{P_{Y|X}(y|x)f_X(x)}{P_Y(y)}.$$

To find the denominator ($P_Y(y)$ or $f_Y(y)$), we often use the law of total probability. Let's look at an example.

Prior and Posterior, example

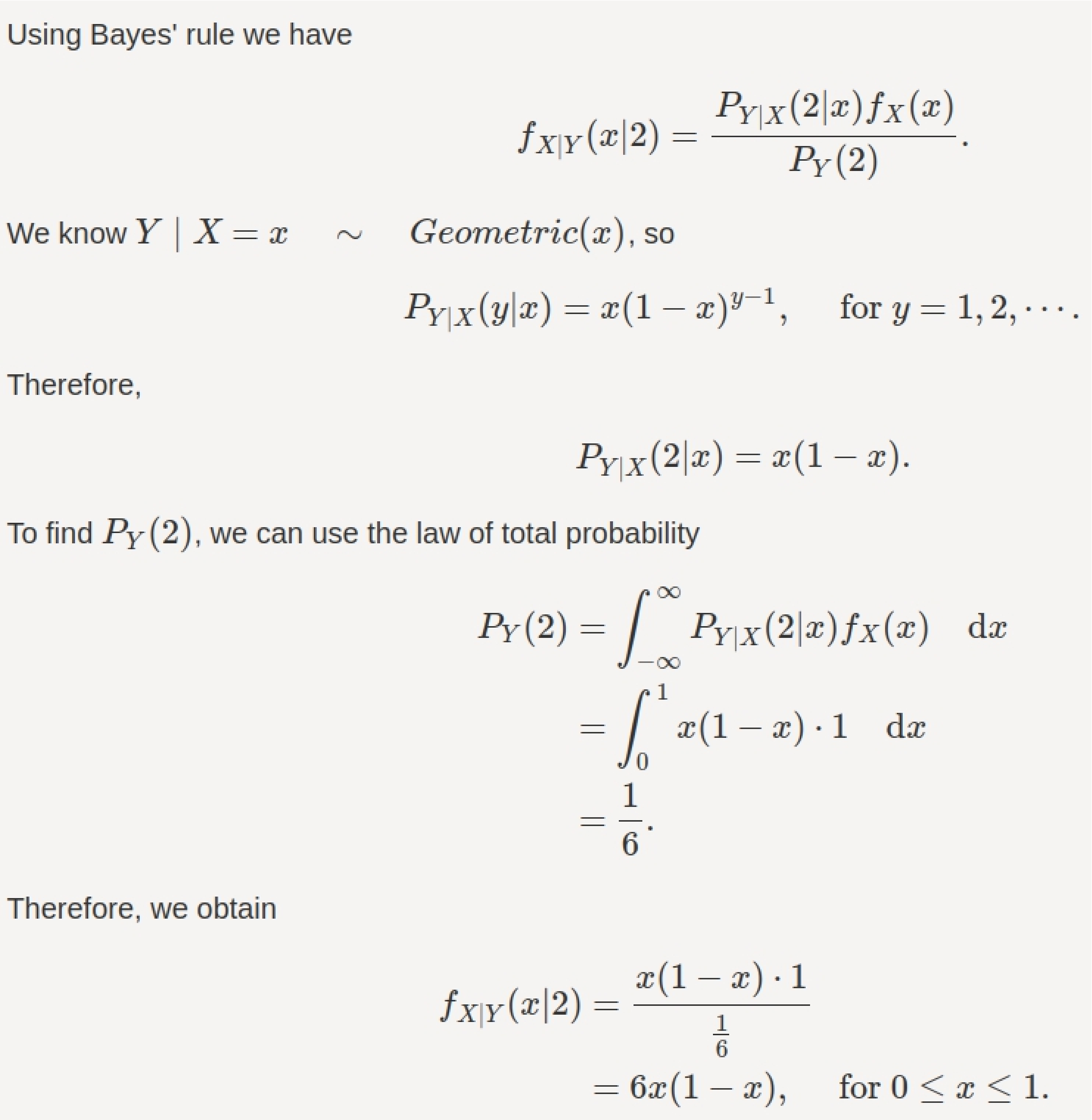


Let $X \sim Uniform(0,1)$. Suppose that we know

$$Y \mid X = x \quad \sim \quad Geometric(x).$$

Find the posterior density of X given $Y = 2$, $f_{X|Y}(x|2)$.

Prior and Posterior, example



Using Bayes' rule we have

$$f_{X|Y}(x|2) = \frac{P_{Y|X}(2|x) f_X(x)}{P_Y(2)}.$$

We know $Y \mid X = x \quad \sim \quad Geometric(x)$, so

$$P_{Y|X}(y|x) = x(1-x)^{y-1}, \quad \text{for } y = 1, 2, \cdots.$$

Therefore,

$$P_{Y|X}(2|x) = x(1-x).$$

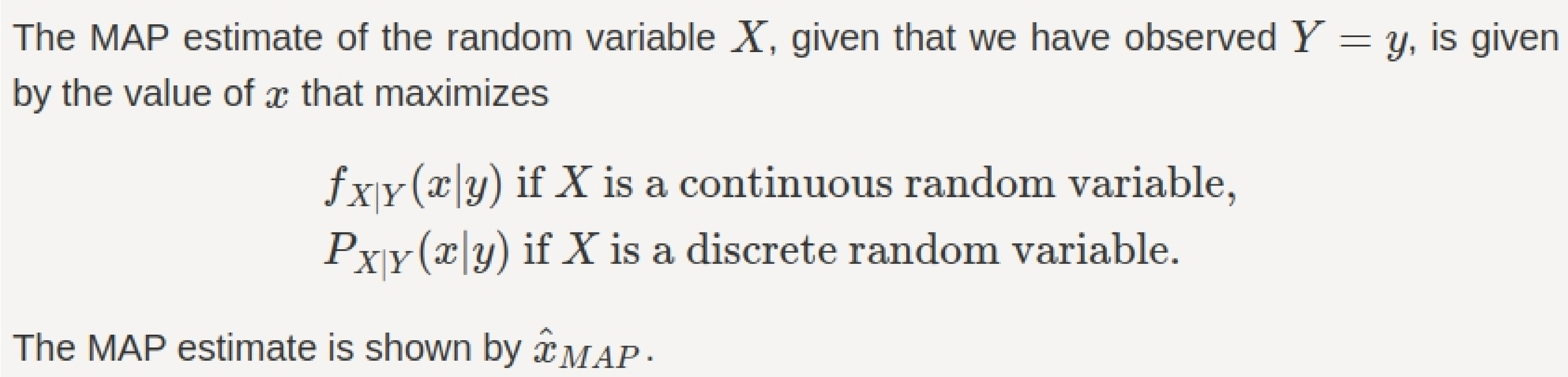
To find $P_Y(2)$, we can use the law of total probability

$$\begin{aligned} P_Y(2) &= \int_{-\infty}^{\infty} P_{Y|X}(2|x) f_X(x) \quad \mathrm{d}x \\ &= \int_0^1 x(1-x) \cdot 1 \quad \mathrm{d}x \\ &= \frac{1}{6}. \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} f_{X|Y}(x|2) &= \frac{x(1-x) \cdot 1}{\frac{1}{6}} \\ &= 6x(1-x), \quad \text{for } 0 \le x \le 1. \end{aligned}$$

Maximum A Posteriori (MAP) Estimation



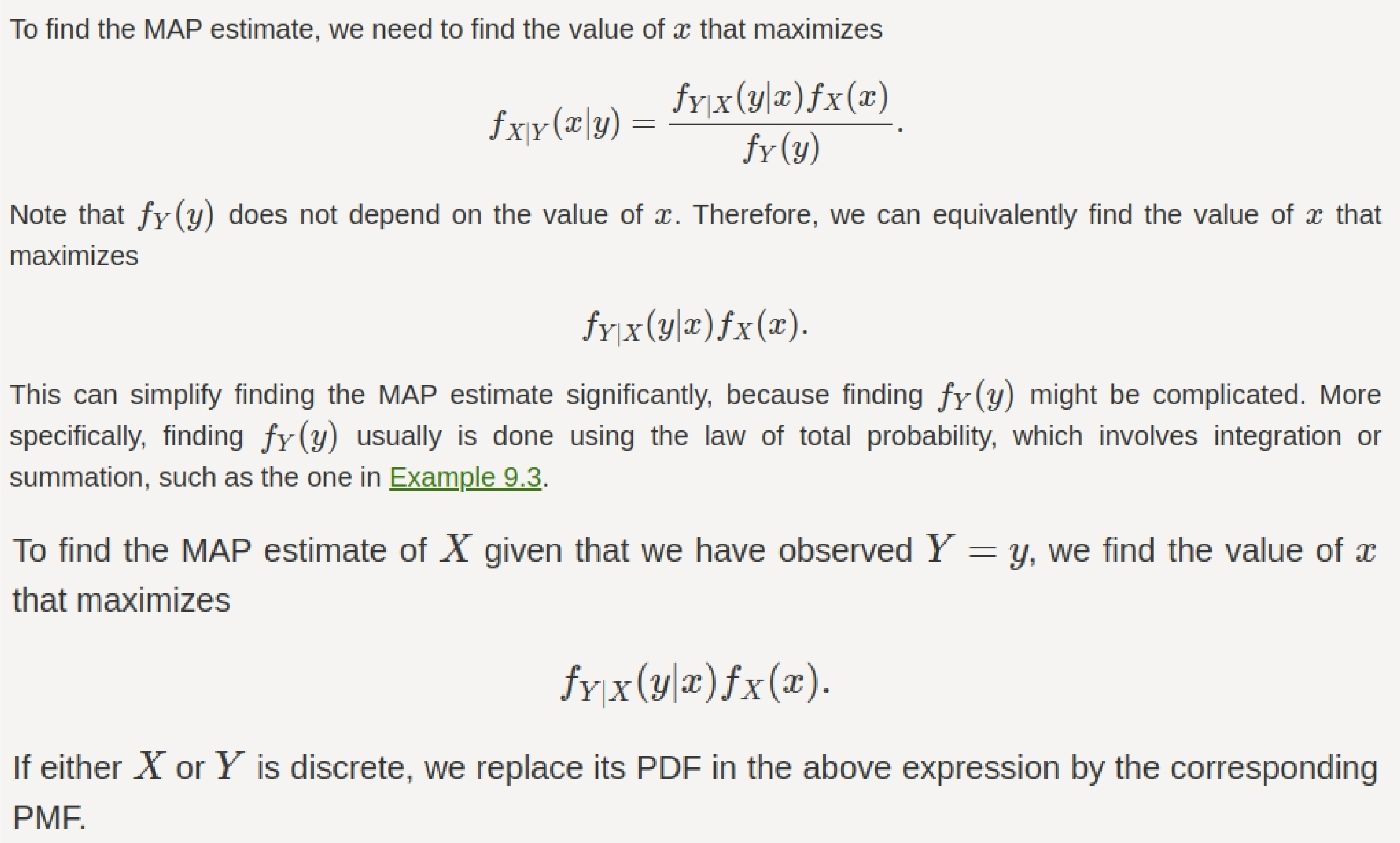
The MAP estimate of the random variable X, given that we have observed $Y = y$, is given by the value of x that maximizes

$$\begin{array}{l} f_{X|Y}(x|y) \text{ if } X \text{ is a continuous random variable,} \\ P_{X|Y}(x|y) \text{ if } X \text{ is a discrete random variable.} \end{array}$$

The MAP estimate is shown by $\hat{x}_{MAP}$.

https://www.probabilitycourse.com/chapter9/9_1_2_MAP_estimation.php

MAP Estimation: simlification



To find the MAP estimate, we need to find the value of x that maximizes

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}.$$

Note that $f_Y(y)$ does not depend on the value of x. Therefore, we can equivalently find the value of x that maximizes

$$f_{Y|X}(y|x) f_X(x).$$

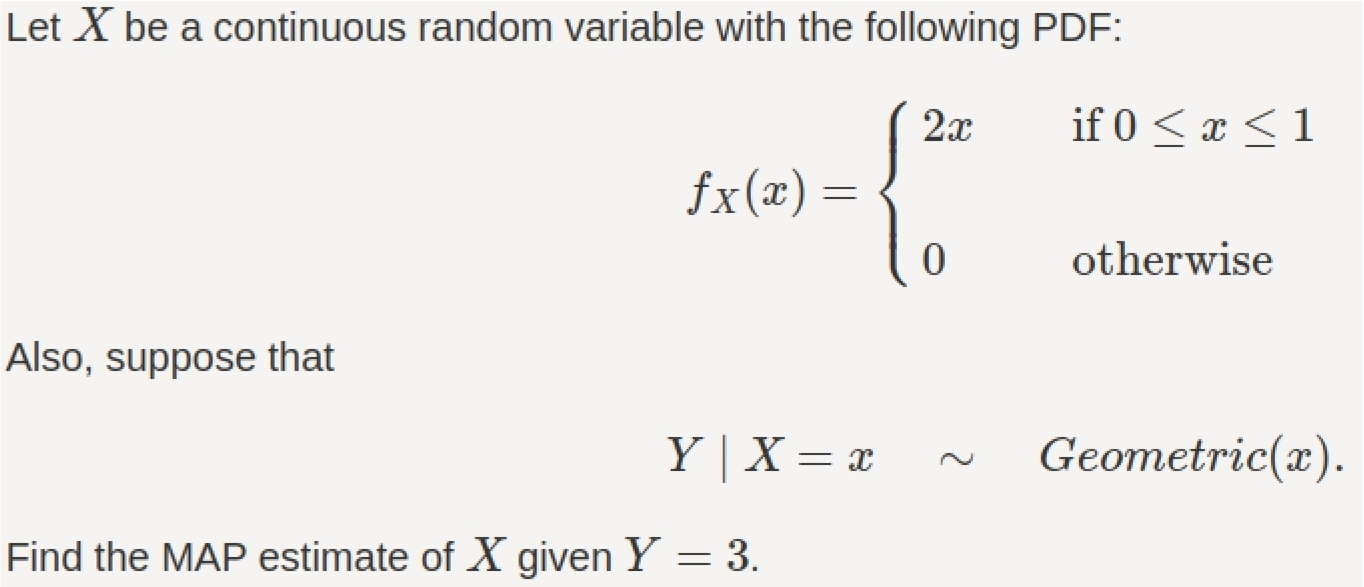
This can simplify finding the MAP estimate significantly, because finding $f_Y(y)$ might be complicated. More specifically, finding $f_Y(y)$ usually is done using the law of total probability, which involves integration or summation, such as the one in Example 9.3.

To find the MAP estimate of X given that we have observed $Y = y$, we find the value of x that maximizes

$$f_{Y|X}(y|x) f_X(x).$$

If either X or Y is discrete, we replace its PDF in the above expression by the corresponding PMF.

MAP Estimation, example



Let X be a continuous random variable with the following PDF:

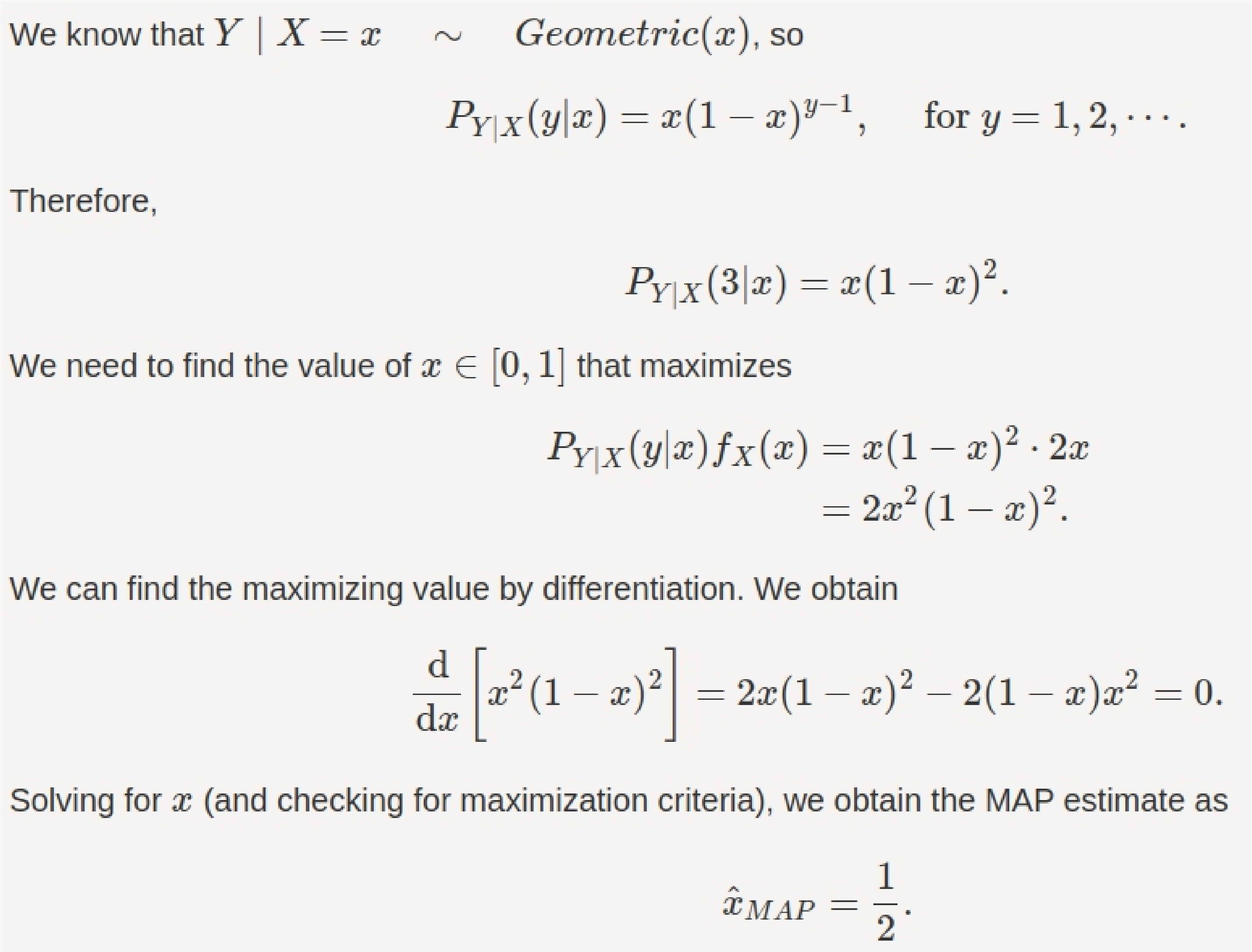
$$f_X(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise} \end{cases}$$

Also, suppose that

$$Y \mid X = x \quad \sim \quad Geometric(x).$$

Find the MAP estimate of X given $Y = 3$.

MAP Estimation, example



We know that $Y \mid X = x \quad \sim \quad Geometric(x)$, so

$$P_{Y|X}(y|x) = x(1-x)^{y-1}, \qquad \text{for } y = 1, 2, \cdots.$$

Therefore,

$$P_{Y|X}(3|x) = x(1-x)^2.$$

We need to find the value of $x \in [0, 1]$ that maximizes

$$\begin{aligned} P_{Y|X}(y|x) f_X(x) &= x(1-x)^2 \cdot 2x \\ &= 2x^2(1-x)^2. \end{aligned}$$

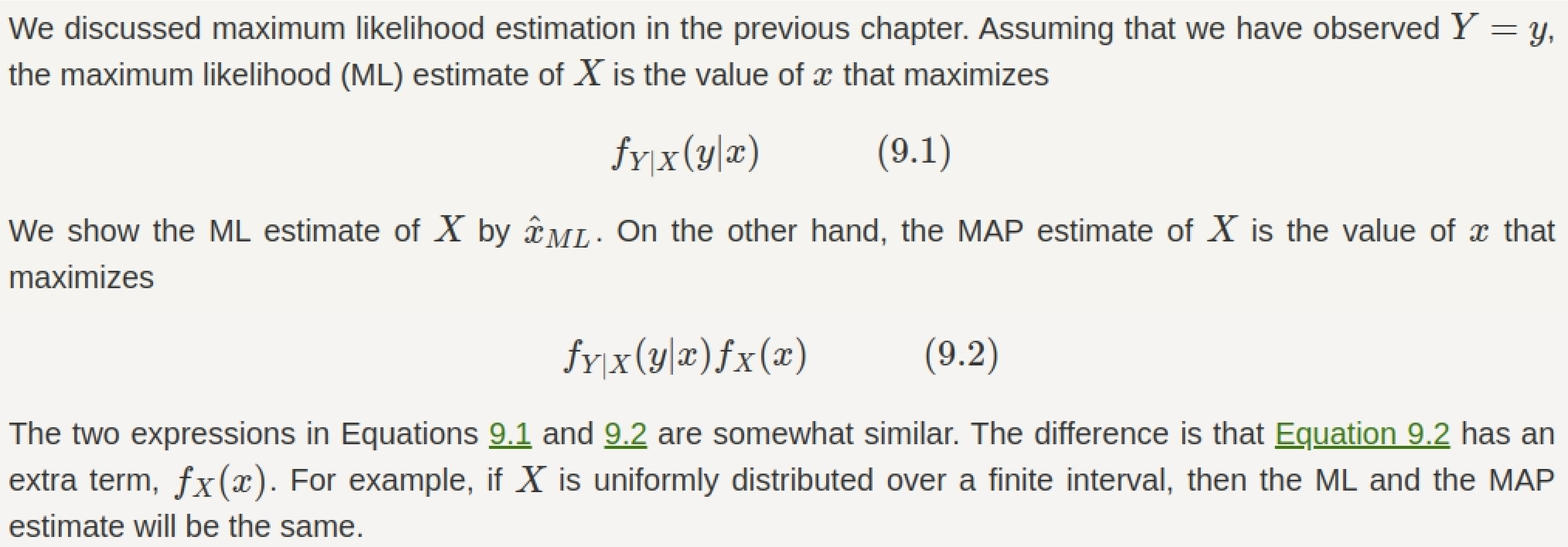
We can find the maximizing value by differentiation. We obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[x^2(1-x)^2\right] = 2x(1-x)^2 - 2(1-x)x^2 = 0.$$

Solving for x (and checking for maximization criteria), we obtain the MAP estimate as

$$\hat{x}_{MAP} = \frac{1}{2}.$$

MAP vs ML Estimation



We discussed maximum likelihood estimation in the previous chapter. Assuming that we have observed $Y = y$, the maximum likelihood (ML) estimate of X is the value of x that maximizes

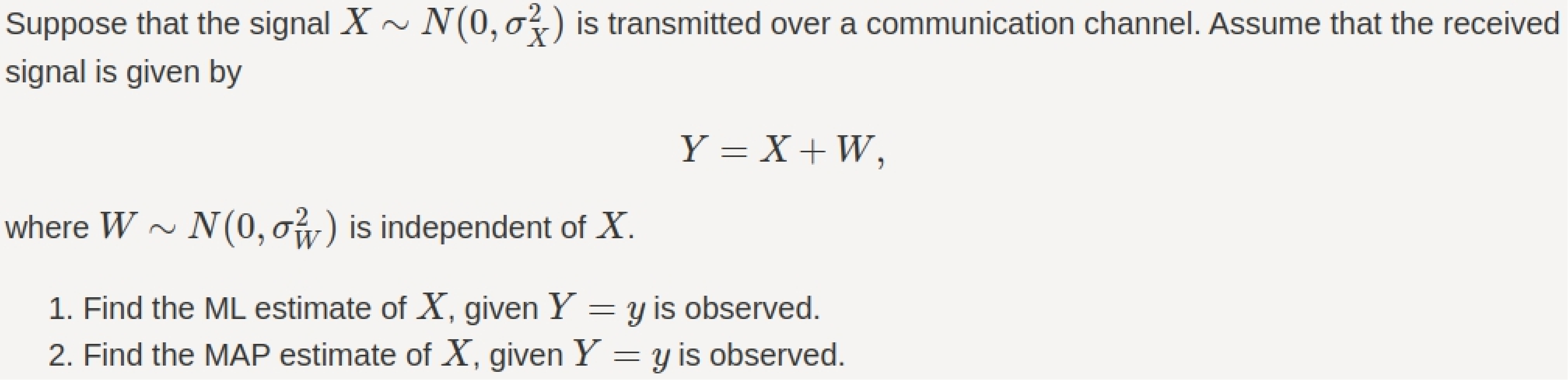
$$f_{Y|X}(y|x) \qquad (9.1)$$

We show the ML estimate of X by $\hat{x}_{ML}$. On the other hand, the MAP estimate of X is the value of x that maximizes

$$f_{Y|X}(y|x)f_X(x) \qquad (9.2)$$

The two expressions in Equations 9.1 and 9.2 are somewhat similar. The difference is that Equation 9.2 has an extra term, $f_X(x)$. For example, if X is uniformly distributed over a finite interval, then the ML and the MAP estimate will be the same.

MAP vs ML Estimation, example



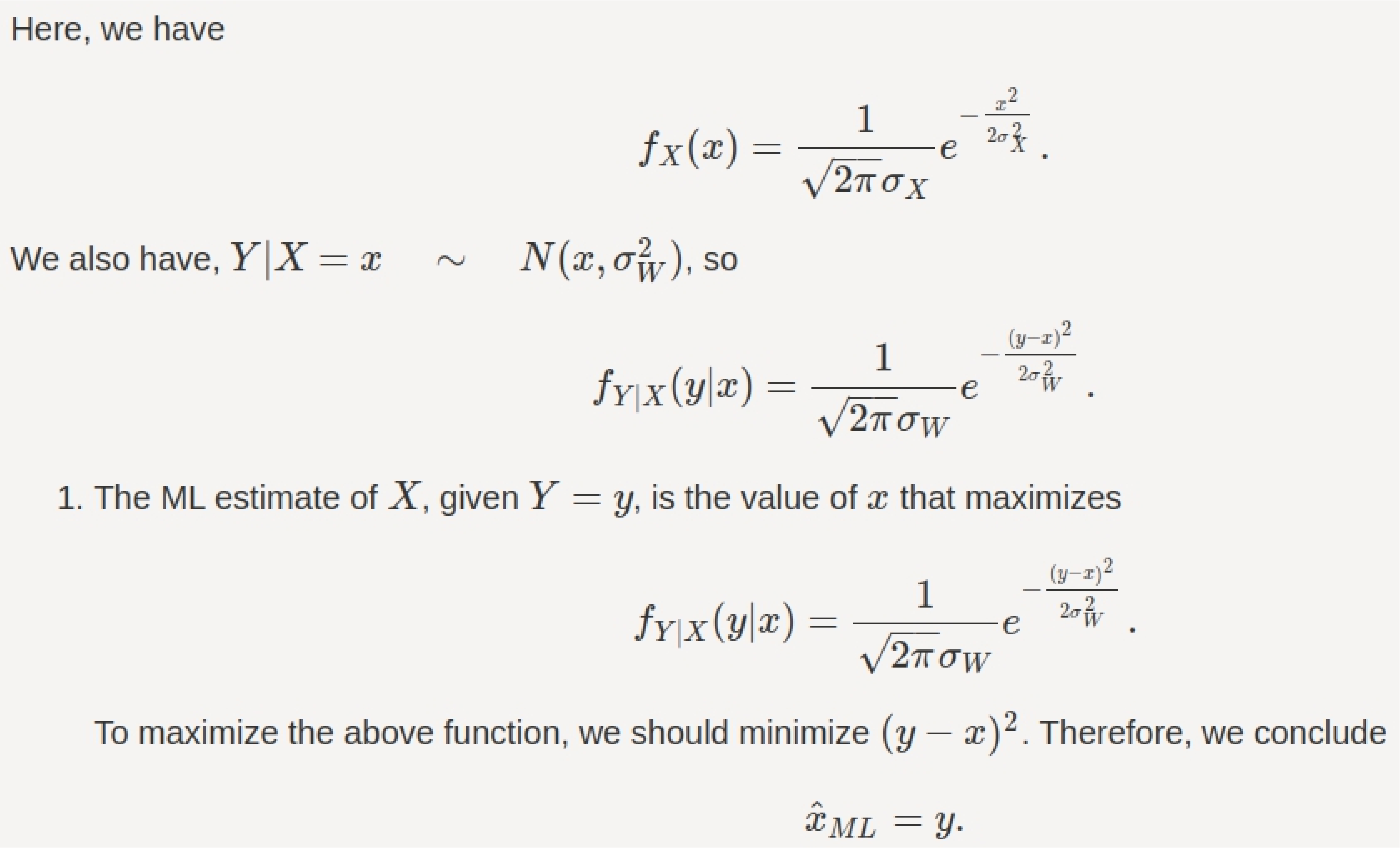
Suppose that the signal $X \sim N(0, \sigma_X^2)$ is transmitted over a communication channel. Assume that the received signal is given by

$$Y = X + W,$$

where $W \sim N(0, \sigma_W^2)$ is independent of X.

1. Find the ML estimate of X, given $Y = y$ is observed.
2. Find the MAP estimate of X, given $Y = y$ is observed.

MAP vs ML Estimation, example



Here, we have

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{x^2}{2\sigma_X^2}}.$$

We also have, $Y|X = x \quad \sim \quad N(x, \sigma_W^2)$, so

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sigma_W} e^{-\frac{(y-x)^2}{2\sigma_W^2}}.$$

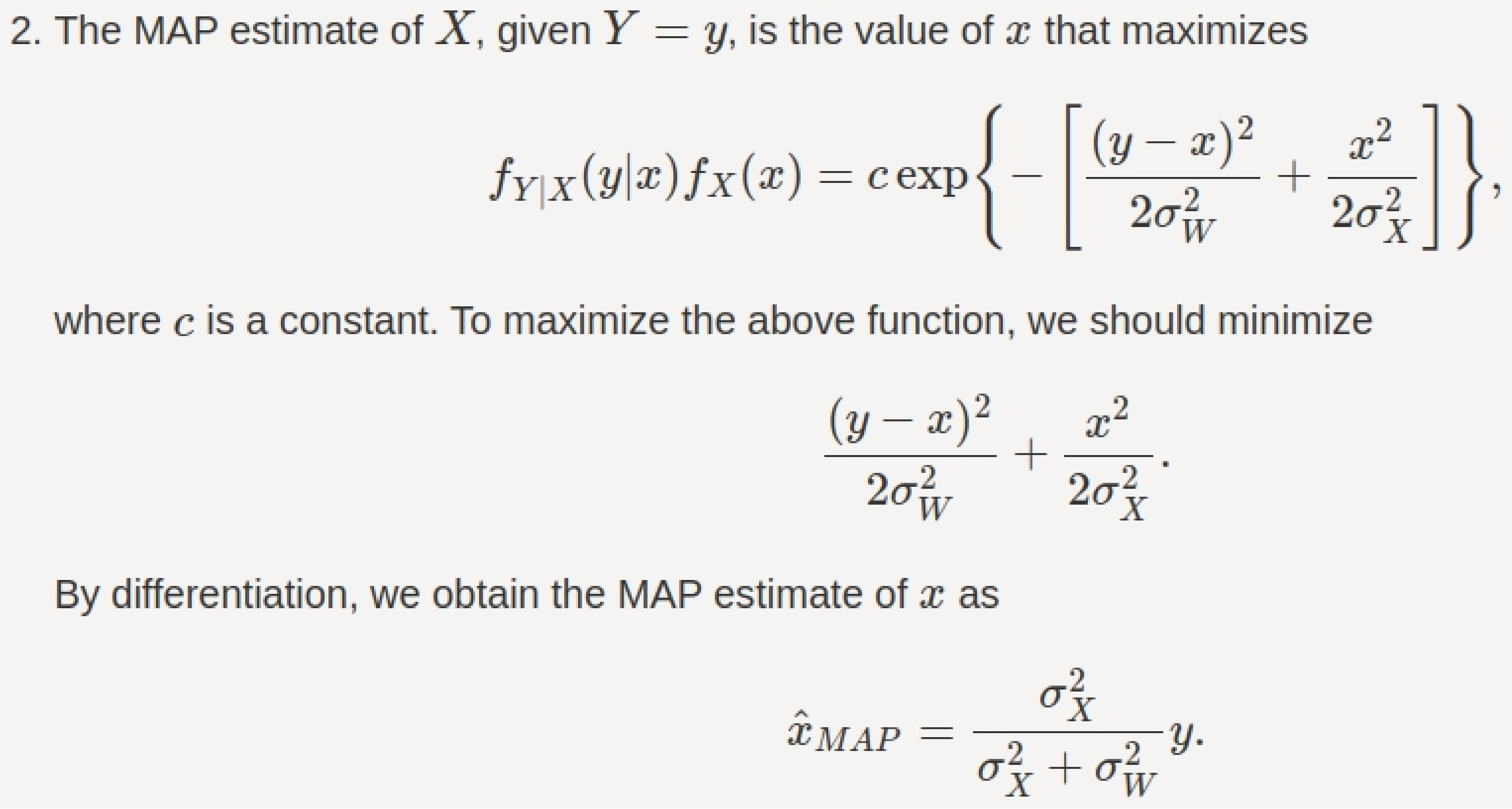
1. The ML estimate of X, given $Y = y$, is the value of x that maximizes

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sigma_W} e^{-\frac{(y-x)^2}{2\sigma_W^2}}.$$

To maximize the above function, we should minimize $(y - x)^2$. Therefore, we conclude

$$\hat{x}_{ML} = y.$$

MAP vs ML Estimation, example



2. The MAP estimate of X, given $Y = y$, is the value of x that maximizes

$$f_{Y|X}(y|x) f_X(x) = c \exp\left\{ - \left[\frac{(y-x)^2}{2\sigma_W^2} + \frac{x^2}{2\sigma_X^2} \right] \right\},$$

where c is a constant. To maximize the above function, we should minimize

$$\frac{(y-x)^2}{2\sigma_W^2} + \frac{x^2}{2\sigma_X^2}.$$

By differentiation, we obtain the MAP estimate of x as

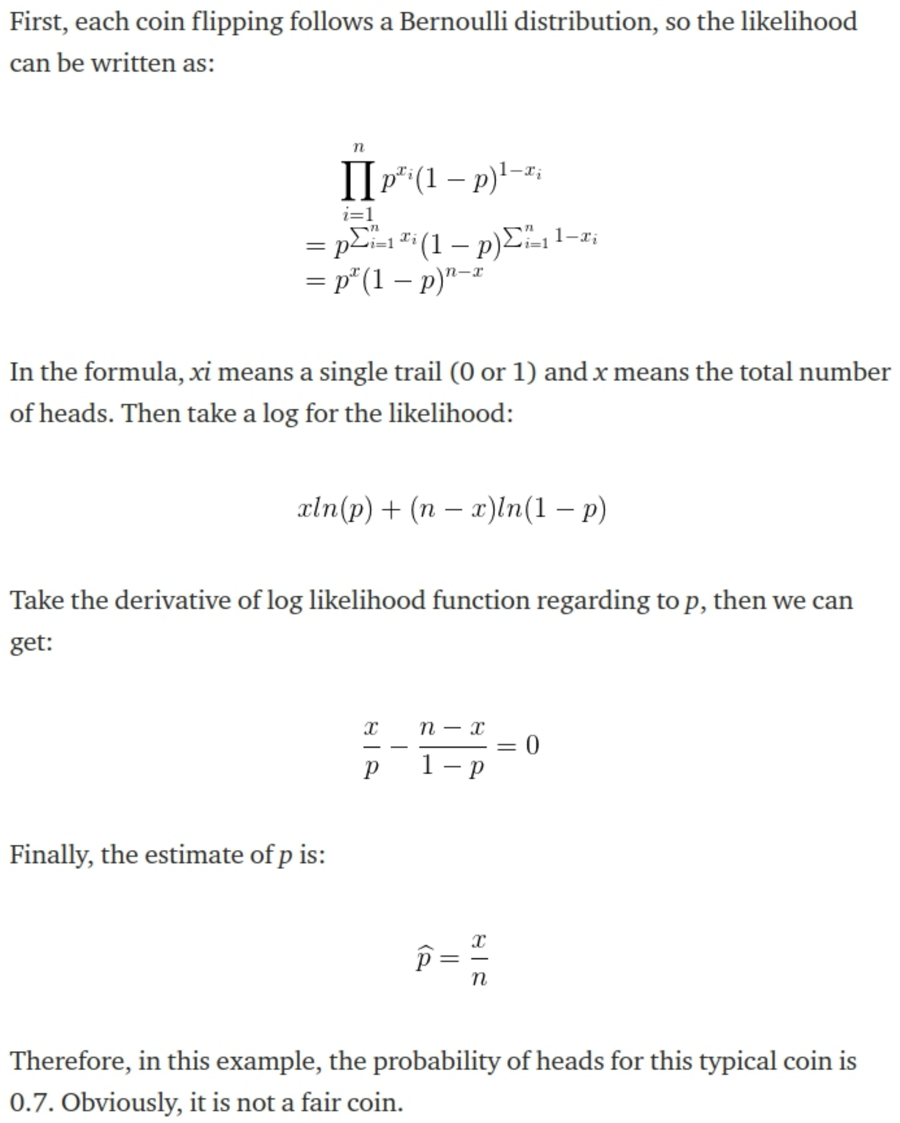
$$\hat{x}_{MAP} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2} y.$$

MAP vs ML Estimation, exercise

You toss this coin n=10 times and there are 7 heads and 3 tails. Each coin flipping follows a Bernoulli distribution.

1. Find the MLE for the observation.
2. Find the MAP for the observation.

MAP vs ML Estimation, exercise



First, each coin flipping follows a Bernoulli distribution, so the likelihood can be written as:

$$\begin{aligned}&\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}\\ &= p^{\sum_{i=1}^{n} x_i}(1-p)^{\sum_{i=1}^{n} 1-x_i}\\ &= p^{x}(1-p)^{n-x}\end{aligned}$$

In the formula, *xi* means a single trail (0 or 1) and *x* means the total number of heads. Then take a log for the likelihood:

$$xln(p) + (n-x)ln(1-p)$$

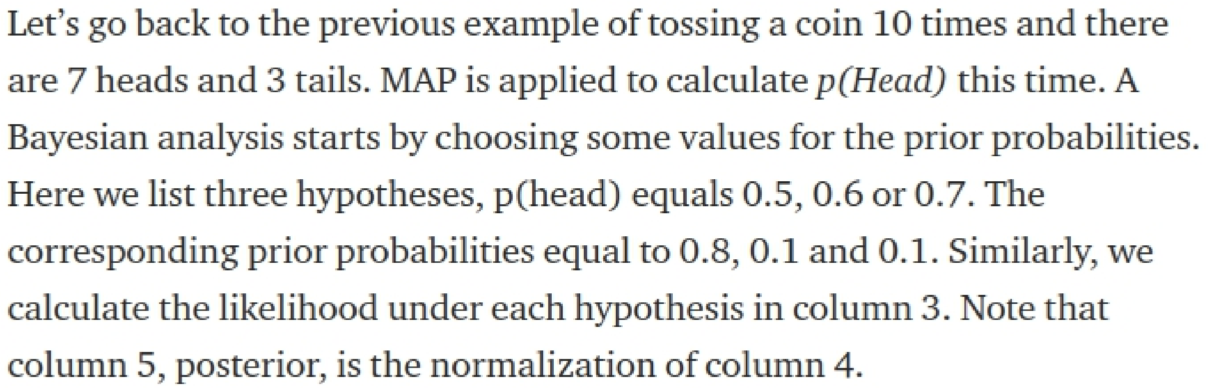
Take the derivative of log likelihood function regarding to *p*, then we can get:

$$\frac{x}{p} - \frac{n-x}{1-p} = 0$$

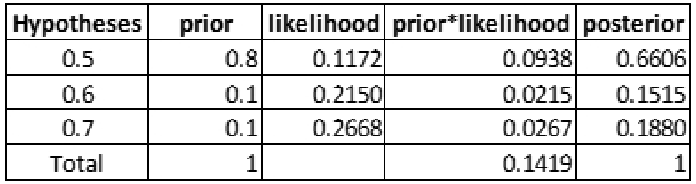
Finally, the estimate of *p* is:

$$\widehat{p} = \frac{x}{n}$$

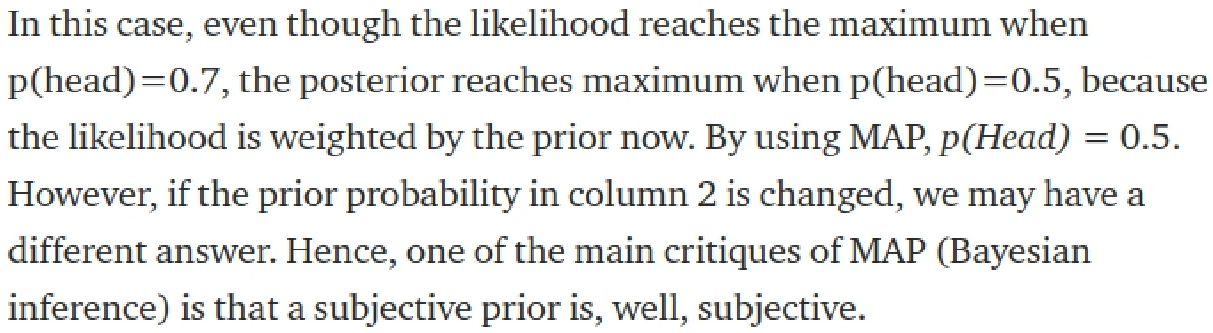
Therefore, in this example, the probability of heads for this typical coin is 0.7. Obviously, it is not a fair coin.



Let's go back to the previous example of tossing a coin 10 times and there are 7 heads and 3 tails. MAP is applied to calculate *p(Head)* this time. A Bayesian analysis starts by choosing some values for the prior probabilities. Here we list three hypotheses, p(head) equals 0.5, 0.6 or 0.7. The corresponding prior probabilities equal to 0.8, 0.1 and 0.1. Similarly, we calculate the likelihood under each hypothesis in column 3. Note that column 5, posterior, is the normalization of column 4.



Hypotheses	prior	likelihood	prior*likelihood	posterior
0.5	0.8	0.1172	0.0938	0.6606
0.6	0.1	0.2150	0.0215	0.1515
0.7	0.1	0.2668	0.0267	0.1880
Total	1		0.1419	1



In this case, even though the likelihood reaches the maximum when p(head)=0.7, the posterior reaches maximum when p(head)=0.5, because the likelihood is weighted by the prior now. By using MAP, *p(Head)* = 0.5. However, if the prior probability in column 2 is changed, we may have a different answer. Hence, one of the main critiques of MAP (Bayesian inference) is that a subjective prior is, well, subjective.