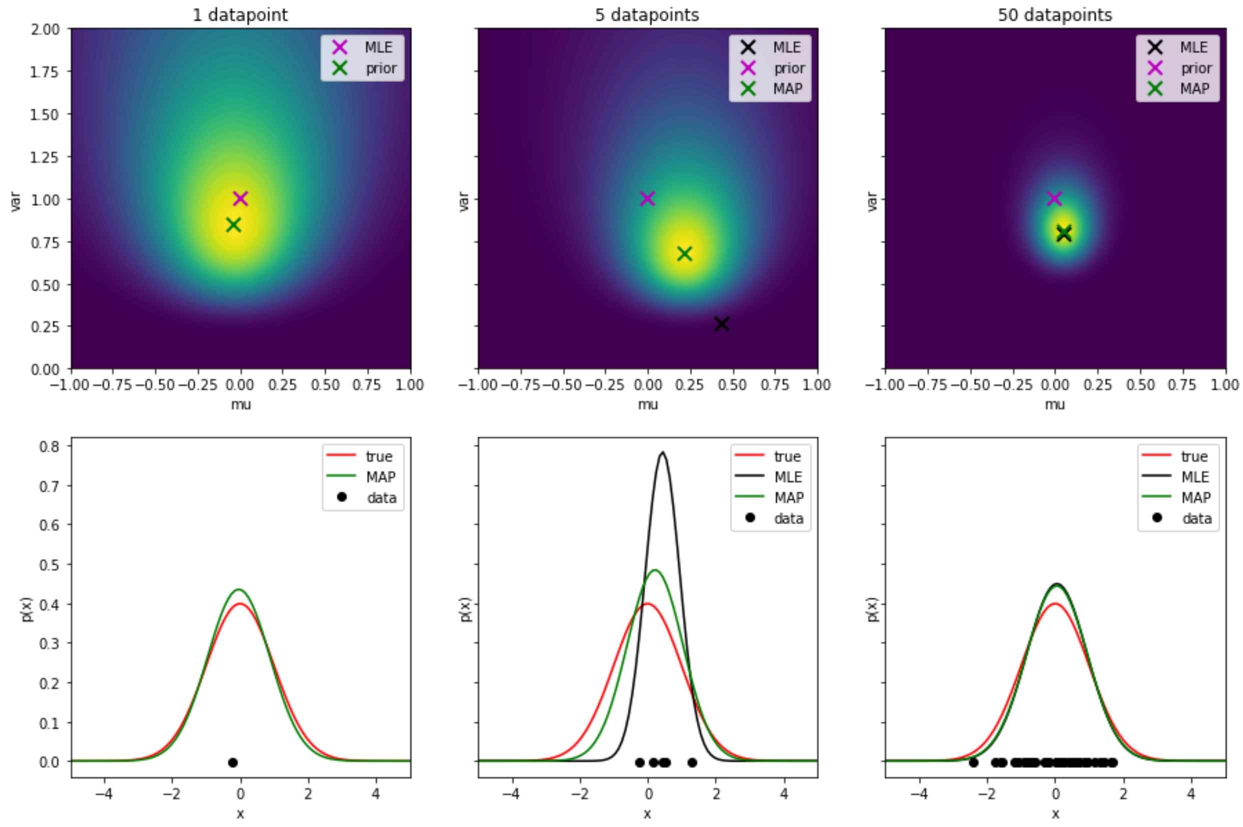
COURSE NAME

MLE vs MAP



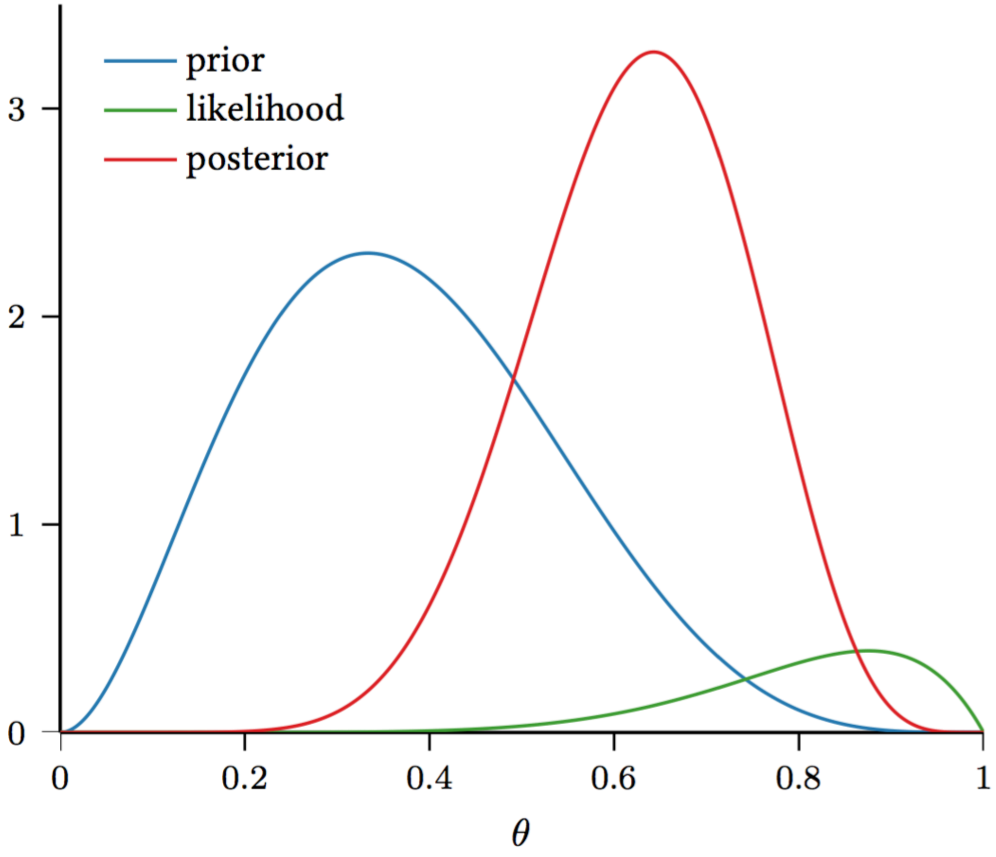
Introduction

Imagine you are evaluating whether a coin is fair or not, so you collect a sequence of heads and tails as your data set. In MLE, we simply look at the data we collected and find the maximum likelihood… **This works well when we have no prior knowledge to leverage (i.e. we have no idea if the coin is fair or not).**

By contrast, in **MAP we take the same likelihood we used in MLE but now multiply by our prior knowledge.** For instance, **we may**

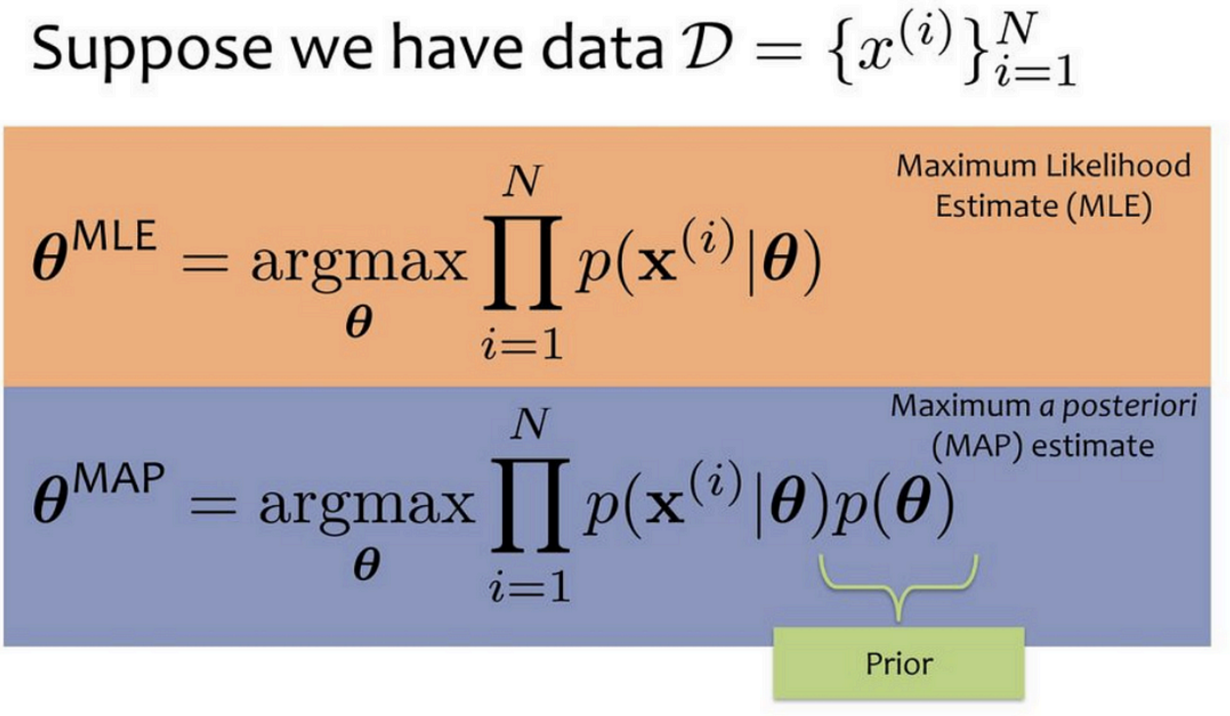
strongly suspect that our coin is biased and so we can influence our estimate with that knowledge through a prior distribution. This new estimate is a mixture of what **we believe (our prior)** and what **we measured (our likelihood).**

Thinking of two extreme cases here would be 1) if we very strongly believe in our prior then we would need to collect a lot of data to influence the resulting estimate away from the prior. Conversely, if we know very little up front (i.e. we have an uninformative prior) then finding the MAP estimate is equivalent to the MLE estimate because our prior did not influence the result.



MLE is finding the maximum of the green curve.

MAP is 1) multiplying the blue curve by the green curve to create the red curve and 2) finding the maximum of the newly created red curve.



Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

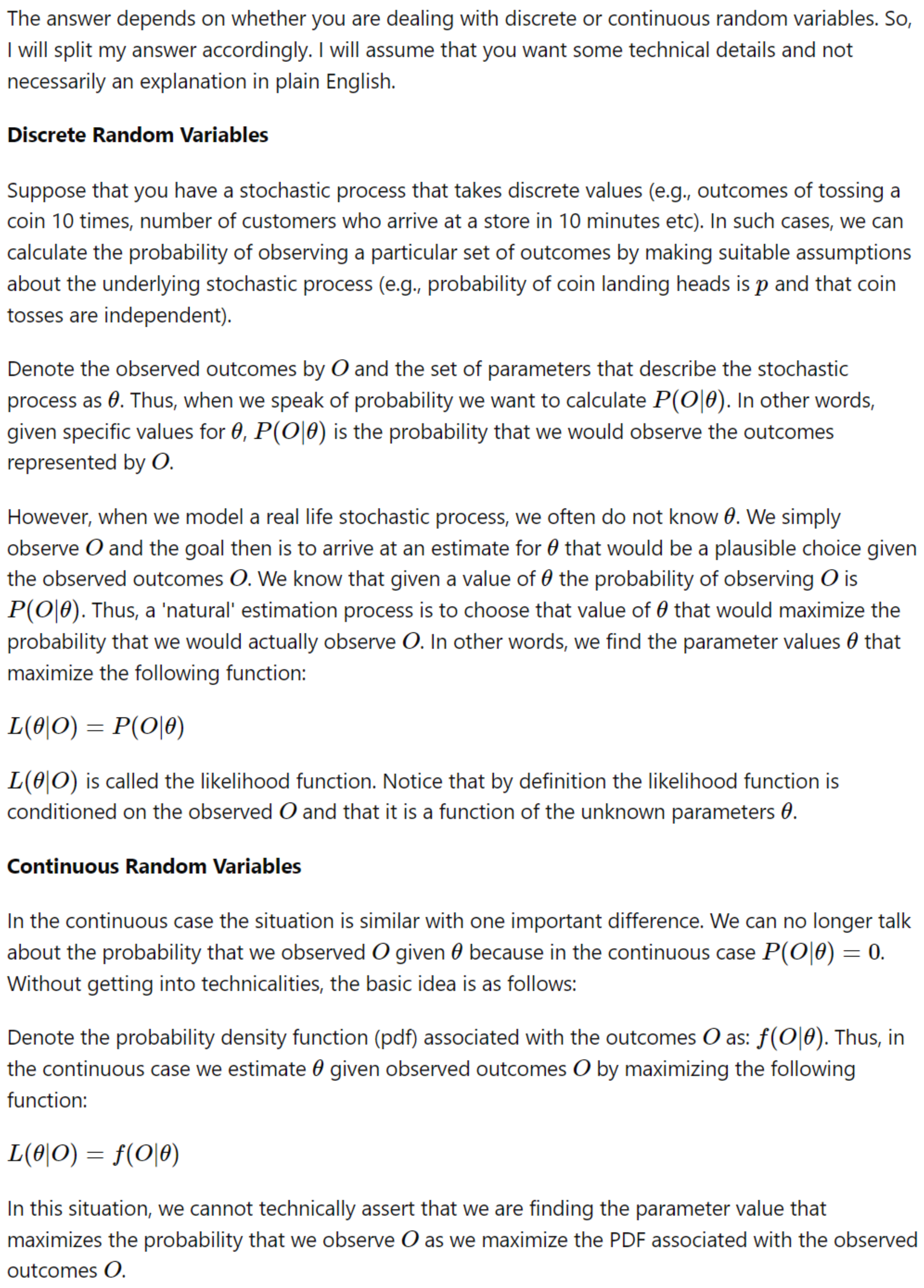
Maximum Likelihood Estimate (MLE)

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum *a posteriori* (MAP) estimate

$$\boldsymbol{\theta}^{\text{MAP}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})\underbrace{p(\boldsymbol{\theta})}_{\text{Prior}}$$

What is the difference between "likelihood" and "probability"?



The answer depends on whether you are dealing with discrete or continuous random variables. So, I will split my answer accordingly. I will assume that you want some technical details and not necessarily an explanation in plain English.

Discrete Random Variables

Suppose that you have a stochastic process that takes discrete values (e.g., outcomes of tossing a coin 10 times, number of customers who arrive at a store in 10 minutes etc). In such cases, we can calculate the probability of observing a particular set of outcomes by making suitable assumptions about the underlying stochastic process (e.g., probability of coin landing heads is p and that coin tosses are independent).

Denote the observed outcomes by O and the set of parameters that describe the stochastic process as θ. Thus, when we speak of probability we want to calculate $P(O|\theta)$. In other words, given specific values for θ, $P(O|\theta)$ is the probability that we would observe the outcomes represented by O.

However, when we model a real life stochastic process, we often do not know θ. We simply observe O and the goal then is to arrive at an estimate for θ that would be a plausible choice given the observed outcomes O. We know that given a value of θ the probability of observing O is $P(O|\theta)$. Thus, a 'natural' estimation process is to choose that value of θ that would maximize the probability that we would actually observe O. In other words, we find the parameter values θ that maximize the following function:

$$L(\theta|O) = P(O|\theta)$$

$L(\theta|O)$ is called the likelihood function. Notice that by definition the likelihood function is conditioned on the observed O and that it is a function of the unknown parameters θ.

Continuous Random Variables

In the continuous case the situation is similar with one important difference. We can no longer talk about the probability that we observed O given θ because in the continuous case $P(O|\theta) = 0$. Without getting into technicalities, the basic idea is as follows:

Denote the probability density function (pdf) associated with the outcomes O as: $f(O|\theta)$. Thus, in the continuous case we estimate θ given observed outcomes O by maximizing the following function:

$$L(\theta|O) = f(O|\theta)$$

In this situation, we cannot technically assert that we are finding the parameter value that maximizes the probability that we observe O as we maximize the PDF associated with the observed outcomes O.

This idea is also called Bayes' rule, which is the core of Bayesian statistics. Here's how it looks:

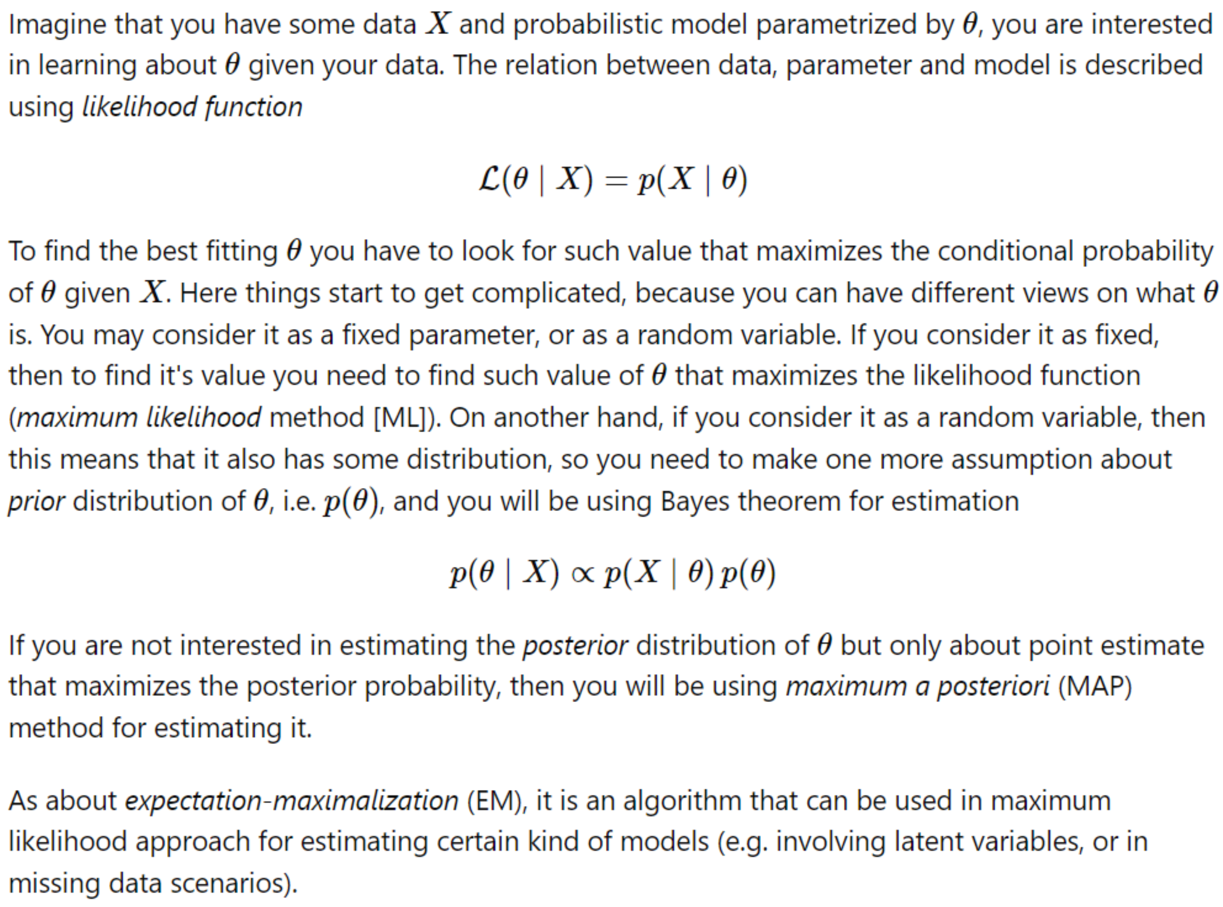
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In order to better explain the various parts of that equation, I will stick to the M&Ms analogy.

- $P(A)$: Like I outlined earlier, we might assume a distribution of the probability that a certain amount of M&Ms have a certain color. This assumption is also called **prior**.
- $P(A|B)$: This is the true distribution of the colors which we can assume based on the M&Ms in our bought package. Or statistically speaking, what is the probability of A given our data B. This part is called the **posterior**.
- $P(B)$: This is the probability for our given data.
- $P(B|A)$: This is the likelihood function. The likelihood function is the probability of data given the parameter, but as a function of the parameter holding the data fixed [3, lecture 3]. I am going to put a focus on this topic later on in the respective chapters of MAP, MLE and ME.

Now we could read our Bayes' rule like so:

$$Posterior \ = \ \frac{Likelihood \ * \ Prior/Believe}{Evidence}$$



Imagine that you have some data X and probabilistic model parametrized by θ, you are interested in learning about θ given your data. The relation between data, parameter and model is described using *likelihood function*

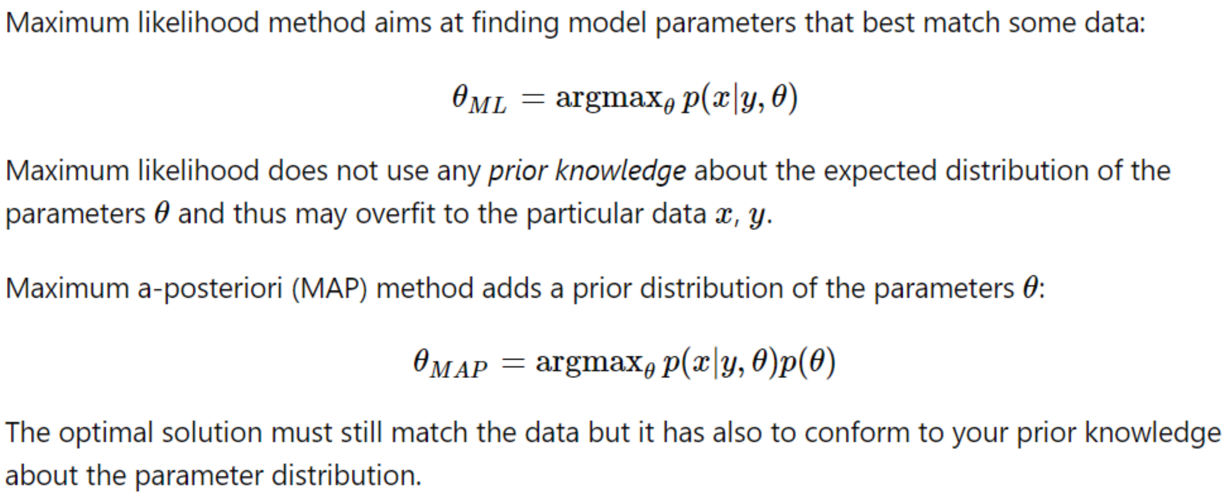
$$\mathcal{L}(\theta \mid X) = p(X \mid \theta)$$

To find the best fitting θ you have to look for such value that maximizes the conditional probability of θ given X. Here things start to get complicated, because you can have different views on what θ is. You may consider it as a fixed parameter, or as a random variable. If you consider it as fixed, then to find it's value you need to find such value of θ that maximizes the likelihood function (*maximum likelihood* method [ML]). On another hand, if you consider it as a random variable, then this means that it also has some distribution, so you need to make one more assumption about *prior* distribution of θ, i.e. $p(\theta)$, and you will be using Bayes theorem for estimation

$$p(\theta \mid X) \propto p(X \mid \theta)\, p(\theta)$$

If you are not interested in estimating the *posterior* distribution of θ but only about point estimate that maximizes the posterior probability, then you will be using *maximum a posteriori* (MAP) method for estimating it.

As about *expectation-maximalization* (EM), it is an algorithm that can be used in maximum likelihood approach for estimating certain kind of models (e.g. involving latent variables, or in missing data scenarios).



Maximum likelihood method aims at finding model parameters that best match some data:

$$\theta_{ML} = \operatorname{argmax}_\theta p(x|y, \theta)$$

Maximum likelihood does not use any *prior knowledge* about the expected distribution of the parameters θ and thus may overfit to the particular data x, y.

Maximum a-posteriori (MAP) method adds a prior distribution of the parameters θ:

$$\theta_{MAP} = \operatorname{argmax}_\theta p(x|y, \theta)p(\theta)$$

The optimal solution must still match the data but it has also to conform to your prior knowledge about the parameter distribution.

Say you have some data. Say you're willing to assume that the data comes from some distribution -- perhaps Gaussian. There are an infinite number of different Gaussians that the data could have come from (which correspond to the combination of the infinite number of means and variances that a Gaussian distribution can have). MLE will pick the Gaussian (i.e., the mean and variance) that is "most consistent" with your data (the precise meaning of *consistent* is explained below).

So, say you've got a data set of $y = \{-1, 3, 7\}$. The most consistent Gaussian from which that data could have come has a mean of 3 and a variance of 16. It could have been sampled from some other Gaussian. But one with a mean of 3 and variance of 16 is most consistent with the data in the following sense: *the probability of getting the particular y values you observed is greater with this choice of mean and variance, than it is with any other choice.*

Moving to regression: instead of the mean being a constant, the mean is a linear function of the data, as specified by the regression equation. So, say you've got data like $x = \{2, 4, 10\}$ along with y from before. The mean of that Gaussian is now the fitted regression model $X'\hat{\beta}$, where $\hat{\beta} = [-1.9, .9]$

Moving to GLMs: replace Gaussian with some other distribution (from the exponential family). The mean is now a linear function of the data, as specified by the regression equation, transformed by the link function. So, it's $g(X'\beta)$, where $g(x) = e^x/(1 + e^x)$ for logit (with binomial data).

MLE is a special case of MAP

Recall the Bayes' rule, we could get the posterior as a product of likelihood and prior:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$\propto P(X|\theta)P(\theta)$$

We ignore the normalizing constant as we are strictly speaking about optimization here, so proportionality should be sufficient.

If we replace the likelihood in the MLE formula above with the posterior, we can get:

$$\theta_{MAP} = \arg\max_{\theta} P(X|\theta)P(\theta)$$

$$= \arg\max_{\theta} \log P(X|\theta) + \log P(\theta)$$

$$= \arg\max_{\theta} \log \prod_{i} P(x_i|\theta) + \log P(\theta)$$

$$= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta) + \log P(\theta)$$

Comparing both MLE and MAP equations, the only thing that differs is the inclusion of prior P(θ) in MAP, otherwise, they are identical. What it means is that the likelihood is now weighted with some weight coming from the prior.

Let's consider what if we use the simplest prior in our MAP estimation, say **uniform prior**. This means we can *assign equal weights everywhere*, on all possible values of the θ. The implication is that the likelihood is equivalently weighted by some constants. Being constant, it could be ignored from our MAP equation, as it will not contribute to the max.

Let's be more concrete, let's say we could assign six possible values into θ. Now, our prior P(θ) is 1/6 everywhere in the distribution. And consequently, we could ignore that constant in our MAP estimation.

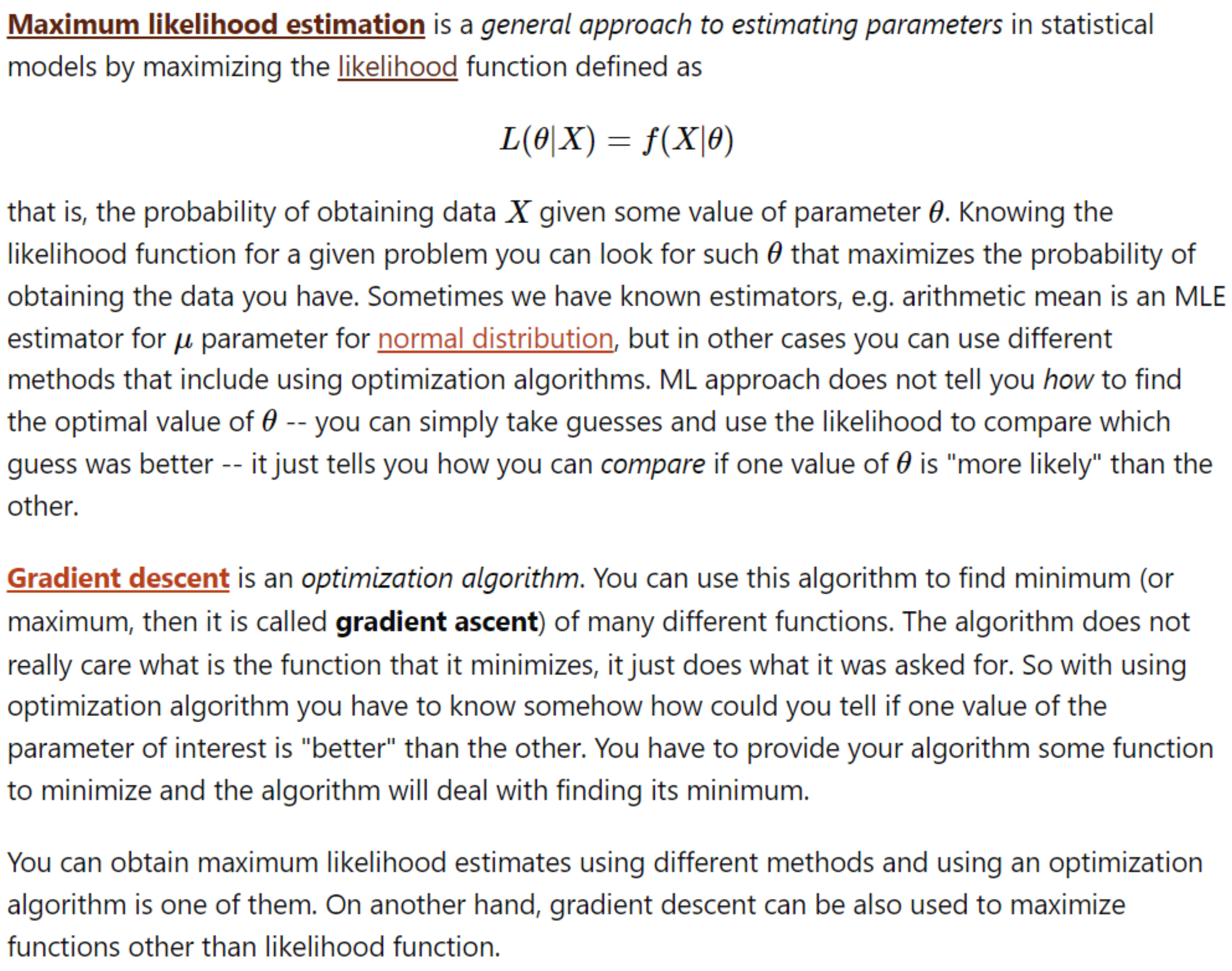
$$\begin{aligned}
\theta_{MAP} &= \arg\max_{\theta} \sum_i \log P(x_i|\theta) + \log P(\theta) \\
&= \arg\max_{\theta} \sum_i \log P(x_i|\theta) + const \\
&= \arg\max_{\theta} \sum_i \log P(x_i|\theta) \\
&= \theta_{MLE}
\end{aligned}$$

As it shows, we are back at the MLE equation!

If we choose a different prior other than uniform, say a Gaussian one instead, then the prior is no longer a constant anymore. The probability is likely to be high or low but won't be the same, as this depends on the region of the distribution.

So clearly we can draw the conclusion: ***MLE is a special case of MAP when the prior is uniform***.

What is the difference between Maximum Likelihood Estimation & Gradient Descent?



Maximum likelihood estimation is a *general approach to estimating parameters* in statistical models by maximizing the likelihood function defined as

$$L(\theta|X) = f(X|\theta)$$

that is, the probability of obtaining data X given some value of parameter θ. Knowing the likelihood function for a given problem you can look for such θ that maximizes the probability of obtaining the data you have. Sometimes we have known estimators, e.g. arithmetic mean is an MLE estimator for μ parameter for normal distribution, but in other cases you can use different methods that include using optimization algorithms. ML approach does not tell you *how* to find the optimal value of θ -- you can simply take guesses and use the likelihood to compare which guess was better -- it just tells you how you can *compare* if one value of θ is "more likely" than the other.

Gradient descent is an *optimization algorithm*. You can use this algorithm to find minimum (or maximum, then it is called **gradient ascent**) of many different functions. The algorithm does not really care what is the function that it minimizes, it just does what it was asked for. So with using optimization algorithm you have to know somehow how could you tell if one value of the parameter of interest is "better" than the other. You have to provide your algorithm some function to minimize and the algorithm will deal with finding its minimum.

You can obtain maximum likelihood estimates using different methods and using an optimization algorithm is one of them. On another hand, gradient descent can be also used to maximize functions other than likelihood function.

Useful Links

1. Maximum Likelihood Estimation and Maximum A Posterior Estimation
2. Numerical example to understand Expectation-Maximization - Cross Validated
3. MLE vs. MAP | Zhiya Zuo
4. Explaining MLE and MAP machine learning principles visually for a newbie | by Sanjeev Kumar