



# Lecture 4 Addendum: Maximum Likelihood Estimation (MLE)


July 3, 2019

Lecturer: Hunter Schafer

At the end of lecture, we discussed a new topic that will serve as an under-pinning for a lot of what we will talk about next week. This topic is generally pretty tedious with the math necessary to do all the calculations. In our class, we don't really care that you know the exact derivation, but do care you get the big picture idea. In each section, the big idea will be put in a gray box like below to emphasize what we want you to know for that section. We include the complete derivations and their explanations for completeness, but we don't expect you to know every detail from those.

 **Key Idea:** *This is a key idea you should know! If a section has a  in its name, that means that entire section is optional!*

## 1 Math Review

 **Key Idea:** *Unless otherwise noted in this section, it is probably good to know these concepts.*

### 1.1 Logarithm Rules

Some useful properties of the logarithm:

- $\log(a^b) = b \log(a)$
- $\log(ab) = \log(a) + \log(b)$
- The more general form of this rule is:

$$\log \left( \prod_{i=1}^n x_i \right) = \sum_{i=1}^n \log(x_i)$$

Where  $\prod_{i=1}^n x_i = x_1 \cdot x_2 \cdot \dots \cdot x_n$

### 1.2 Probability

This section will have a basic probability review.

If events  $A$  and  $B$  are independent, then the probability of both  $A$  and  $B$  happening will be  $P(AB) = P(A)P(B)$ . Equivalently, this can be written as  $P(B|A) = P(B)$  where  $P(B|A)$  is the probability of  $B$  happening given knowledge that event  $A$  has happened.

### 1.3 Normal Distribution

The normal distribution (Gaussian distribution) is a continuous distribution that describes a normal bell curve centered at some mean with some standard deviation. Because the normal distribution is continuous, we talk about it having a "probability density function" (PDF) that indicates where the probability density lives. The PDF for normally distributed variable  $X \sim \mathcal{N}(\mu, \sigma^2)$  (mean  $\mu$  and variance  $\sigma^2$ ) is:

$$PDF_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A normal distribution PDF looks like the following

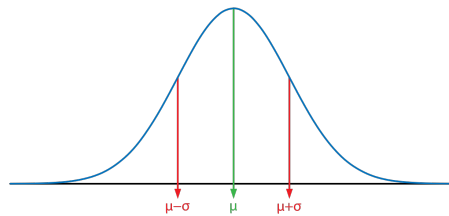


Figure 1: PDF of  $X \sim \mathcal{N}(\mu, \sigma^2)$

### 1.4 Optimization

For many problems, we end up trying to find the value  $x$  that maximizes or minimizes some function  $f(x)$ . Sometimes, if the function is simple, we are able to solve the problem analytically using calculus. Briefly, analytical solutions generally look like taking the derivative of the function and finding the value that makes the derivative 0. We will not ask you to solve any problems analytically in this class. Many times, we generally rely on using an optimization algorithm like gradient descent or ascent that attempts to approximate the optimal value iteratively.

## 2 Estimating Probability of Heads

Suppose someone comes up to you with a coin and asks you to tell you the probability that it will end up heads when flipped. You ask them to flip the coin five times, and they tell you they saw HTHHT. You nicely try to end the conversation quickly and just tell them the probability of a heads  $P(H) = 3/5$ .

However, this person has different plans for you today and they then ask you WHY you told them the probability is  $3/5$ . You ask them to flip the coin 50 times and when they get 29 heads, you tell them your new estimate for  $P(H) = 29/50$ . Again, they ask why so you go to tell them about **maximum likelihood estimation**.

## 2.1 Maximum Likelihood Estimation (MLE)

For MLE, we are given a dataset  $D$ , this case a series of coin flips  $D = \{HTHHT\dots\}$  with  $n$  flips,  $k$  of which are heads. We make the assumption that coin flips are independent and identically distributed (i.i.d). This means that the coin flips all come from the same, but unknown distribution, and the outcome of one flip does not impact the other.

We will have a hypothesis that each coin flip comes from a Bernoulli distribution. This means that there is some  $\theta \in [0, 1]$  such that  $P(H) = \theta$  and  $P(T) = 1 - \theta$ . The trick here is  $\theta$  is unknown to us so we must estimate it using data.

In our case of 5 coin flips, this means the probability of seeing this particular series of flips HTHHT for some parameter  $\theta$  is

$$\begin{aligned} P(HTHHT) &= P(H)P(T)P(H)P(H)P(T) \\ &= \theta \cdot (1 - \theta) \cdot \theta \cdot \theta \cdot (1 - \theta) \cdot \theta \\ &= \theta^3(1 - \theta)^2 \end{aligned}$$

We will use this idea of computing the probability of seeing our data for some setting of  $\theta$  to help us identify the most likely setting of the parameter.

**Key Idea:** In general for MLE, we compute  $P(D|\theta)$  to help us figure out which setting of  $\theta$  is the most likely. In the general case for our coin flip example with  $k$  heads out of  $n$  flips the likelihood of a parameter  $\theta$  is:

$$P(D|\theta) = \theta^k(1 - \theta)^{n-k}$$

This formula tells us which values of  $\theta$  are more likely and which are less. For example, this could be visualized in Figure 2.

The goal of Maximum Likelihood Estimation (MLE) is to find the value  $\hat{\theta}_{MLE}$  that yields the highest value of  $P(D|\hat{\theta}_{MLE})$ . In other words, MLE finds

$$\hat{\theta}_{MLE} = \max_{\theta} P(D|\theta)$$

This corresponds to the value of  $\theta$  in Figure 2 that has the highest value of  $P(D|\theta)$ .

As a technical note, we generally find the value with the maximum log-likelihood since it tends to be easier to work with analytically and results in numerically stable computations for approximation methods like gradient ascent. It's the same idea, but instead we say

$$\hat{\theta}_{MLE} = \max_{\theta} \log(P(D|\theta))$$

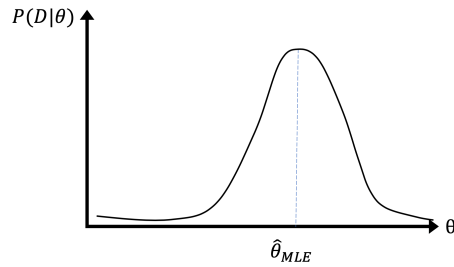


Figure 2: PDF of  $X \sim \mathcal{N}(\mu, \sigma^2)$

## 2.2 $\triangle$ MLE Derivation

In this section, we do the actually step-by-step computation to find the maximum likelihood estimate of the probability of flipping a heads in this scenario  $\theta$ .

$$\begin{aligned}
 \hat{\theta}_{MLE} &= \max_{\theta} P(D|\theta) \\
 &= \max_{\theta} \log P(D|\theta) \\
 &= \max_{\theta} \log (\theta^k (1 - \theta)^{n-k}) \\
 &= \max_{\theta} k \log \theta + (n - k) \log(1 - \theta)
 \end{aligned}$$

This is something that can be solved analytically, so we take the derivative with respect to  $\theta$  and then set that to 0 to find the maximum point.

$$\begin{aligned}
 \frac{d}{d\theta} \log P(D|\theta) &= \frac{d}{d\theta} (k \log \theta + (n - k) \log(1 - \theta)) \\
 &= \frac{k}{\theta} - \frac{n - k}{1 - \theta}
 \end{aligned}$$

We know that optimal value  $\hat{\theta}_{MLE}$  will occur when the derivative is 0 so we set this to 0 when  $\theta$  is  $\hat{\theta}_{MLE}$  and rearrange to solve the  $\hat{\theta}_{MLE}$ . Technically we also need to do what is called a second derivative test to make sure the point is a maxima and not a minima, but we leave that out for this exercise.

$$\begin{aligned} \frac{k}{\hat{\theta}_{MLE}} - \frac{n-k}{1-\hat{\theta}_{MLE}} &= 0 \\ \frac{k}{\hat{\theta}_{MLE}} &= \frac{n-k}{1-\hat{\theta}_{MLE}} \\ k(1-\hat{\theta}_{MLE}) &= (n-k)\hat{\theta}_{MLE} \\ k - k\hat{\theta}_{MLE} &= n\hat{\theta}_{MLE} - k\hat{\theta}_{MLE} \\ \hat{\theta}_{MLE} &= \frac{k}{n} \end{aligned}$$

### 3 MLE: Linear Regression

Remember that in the linear regression model, we were given a dataset  $\{(x_i, y_i)\}_{i=1}^n$  where we assumed  $y_i = w^T x_i + \epsilon_i$  for some unknown coefficients  $w$ . We made a special assumption that the noise  $\epsilon_i$  were drawn i.i.d. from  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ . This forms a probability distribution over the values  $y_i \sim \mathcal{N}(w^T x_i, \sigma^2)$

$$P(y_i|x_i, w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}}$$

**Key Idea:** Under these assumptions, the most likely estimate of  $\hat{w}_{MLE}$  will be

$$\hat{w}_{MLE} = \min_w \sum_{i=1}^n (y_i - w^T x_i)^2$$

Which is exactly the Ordinary Least Squares solution we saw last week! This means the reason we used RSS in the first place was because it was actually finding the MLE under the assumptions we made!

### 3.1 MLE Derivation

In this section, we derive the MLE for the linear regression setup.

$$\begin{aligned}\hat{w}_{MLE} &= \max_w P(D|w) \\ &= \max_w \log P(D|w) \\ &= \max_w \log \left( \prod_{i=1}^n P(y_i|x_i, w) \right) \\ &= \max_w \sum_{i=1}^n \log (P(y_i|x_i, w)) \\ &= \max_w \sum_{i=1}^n \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}} \right) \\ &= \max_w \sum_{i=1}^n \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(y_i - w^T x_i)^2}{2\sigma^2} \log(e) \\ &= \max_w \sum_{i=1}^n -(y_i - w^T x_i)^2 \\ &= \min_w \sum_{i=1}^n (y_i - w^T x_i)^2\end{aligned}$$

The third to last line is true by our properties of logarithms. The second to last line comes from the fact that we can drop constant terms and multiples in our maximization problem without changing the result

$$\max_x a + bf(x) = \max_x f(x)$$

The last line is true because

$$\max_x -f(x) = \min_x f(x)$$

This means that  $\hat{w}_{MLE}$  is the  $w$  that minimizes the residual sum of squares, which was exactly the quality metric we used before!

We do not show the derivation to find what the formula for  $\hat{w}_{MLE}$  is as the calculations are much more tedious. There is technically an analytical solution for the least squares objective, but we have not discussed it since we have just used gradient descent to solve this minimization problem.