

# Optimality of the Maximum Likelihood Method

Hossein Falaki  
Computer Science Department,  
University of California, Los Angeles  
falaki@cs.ucla.edu

## 1 Introduction

In this note we prove the consistency, asymptotic normality, and optimality of Maximum Likelihood.

## 2 Method of Moments

The method of moments matches moments of the model to the moments of the sample to get the required number of equations that can be solved for the missing parameters. Therefore the equations are:

$$\frac{1}{n} \sum_{i=1}^n (X_i^k - m_k(\Theta)) = 0$$

We can generalize this to:

$$\frac{1}{n} \sum_{i=1}^n h(X_i, \Theta) = 0$$

And as we can see for MoM:

$$\forall \Theta E_{\Theta}(h(X_i, \Theta)) = 0 \tag{1}$$

This condition is referred to as *consistency*.

## 3 Maximum Likelihood

For MLE the function  $h$  is:

$$h(X_i, \Theta) = \frac{\partial}{\partial \Theta} \log p(X_i, \Theta) \tag{2}$$

### 3.1 Consistency

To prove the consistency of MLE we need to prove that:

$$E_{\Theta}(MLE) = E_{\Theta}\left(\frac{\partial}{\partial\Theta}\log p(X_i, \Theta)\right) = 0 \quad (3)$$

dff

$$\begin{aligned} E_{\Theta}(MLE) &= E_{\Theta}\left(\frac{1}{p(X, \Theta)} \frac{\partial}{\partial\Theta} p(X_i, \Theta)\right) = 0 \\ &= \int \left[\frac{1}{p(x, \Theta)} \frac{\text{partial}}{\text{partial}\Theta} p(x, \Theta)\right] \cdot p(x, \Theta) dx \\ &= \int \frac{\partial}{\partial\Theta} p(x, \Theta) dx \\ &= \frac{\partial}{\partial\Theta} \int p(x, \Theta) dx = \frac{\partial}{\partial\Theta} 1 = 0 \end{aligned}$$

Therefore we can conclude that the MLE equation is a consistent equation.

### 3.2 Asymptotic Normality

We will prove that the difference between the MLE estimation,  $\hat{\Theta}$  and the true value,  $\Theta_{true}$ , follows a zero-mean normal distribution.

The first order Taylor expansion of the MLE equation around  $\Theta_{true}$  gives:

$$\frac{1}{n} \sum_{i=1}^n h(X_i, \Theta) \doteq \underbrace{\frac{1}{n} \sum_{i=1}^n h(X_i, \Theta_{true})}_{\text{intercept}} + \underbrace{\frac{1}{n} \sum_{i=1}^n h'(X_i, \Theta_{true})(\Theta - \Theta_{true})}_{\text{slope}} \quad (4)$$

Therefore:

$$\underbrace{\hat{\Theta} - \Theta_{true}}_{\text{displacement}} = - \frac{\overbrace{\frac{1}{n} \sum_{i=1}^n h(X_i, \Theta_{true})}_{\text{intercept}}}{\underbrace{\frac{1}{n} \sum_{i=1}^n h'(X_i, \Theta_{true})}_{\text{slope}}}$$

We already know that as  $n \rightarrow \infty$ :

$$\begin{aligned} \text{slope} &\rightarrow E_{\Theta_{true}}(h'(X, \Theta_{true})) \\ \text{intercept} &\rightarrow E_{\Theta_{true}}(h(X, \Theta_{true})) = 0 \end{aligned}$$

To find the optimal  $h$  we zoom in (put displacement under microscope):

$$\sqrt{n}(\hat{\Theta} - \Theta_{true}) = -\frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n h(X_i, \Theta_{true})}{\frac{1}{n} \sum_{i=1}^n h'(X_i, \Theta_{true})} \Rightarrow \frac{N(0, Var_{\Theta_{true}}(h(X, \Theta_{true})))}{E_{\Theta_{true}}(h'(X, \Theta_{true}))}$$

According to the Central Limit Theorem, as  $n \rightarrow \infty$ :

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n h(X_i, \Theta_{true}) \rightarrow N(0, Var_{\Theta_{true}}(h(X, \Theta_{true})))$$

This proves that the displacement follows a zero-mean normal distribution.

### 3.3 Optimality

The optimal  $h$  is the one consistent  $h$  that has the lowest variance. We need to minimize  $Var_h$  over all possible consistent  $h$

$$\forall \Theta : \int h(x, \Theta)p(x, \Theta)dx = 0$$

If we take partial derivatives:

$$\int \left[ \underbrace{h'(x, \Theta)p(x, \Theta)}_{E(h'(X, \Theta))} + h(x, \Theta) \overbrace{\frac{\partial}{\partial \Theta} \log p(x, \Theta)p(x, \Theta)}^{p'(x, \Theta)} \right] dx = 0$$

Therefore:

$$\begin{aligned} E(h'(X, \Theta)) &= - \int [h(x, \Theta) \frac{\partial}{\partial \Theta} \log p(x, \Theta)] p(x, \Theta) dx \\ &= -E(h(X, \Theta) \frac{\partial}{\partial \Theta} \log p(X, \Theta)) \\ &= Cov(h(X, \Theta), \frac{\partial}{\partial \Theta} \log p(X, \Theta)) \end{aligned}$$

As we know:

$$Cov^2(X, Y) \leq Var(X) \times Var(Y)$$

The equality happens when  $X$  and  $Y$  are aligned. Therefore the Variance is minimized when:

$$h(X, \Theta) \propto \frac{\partial}{\partial \Theta} \log p(X, \Theta) \quad (5)$$

Which is the MLE function.

## 4 Acknowledgment

This note is based on Prof. Ying Nian Wu's lectures on Theoretical Statistics at UCLA. Figures 1 and 2 are from the Internet.