Maximum Likelihood Estimation

Maximum Likelihood Estimation

- For different pattern recognition tasks, we often find ourselves having to estimate parameters of a given model
- We have some samples of the data, and want to use them to estimate parameters of the model
- This happens in many pattern recognition applications, e.g.,
  - Regression analysis
  - Modeling Biometric score distributions
  - Logistic Regression
  - Time series analysis

Maximum Likelihood Estimation

- Assume we have data samples $Y_1, Y_2, \ldots, Y_n$ which are assumed to be independent and identically distributed (iid)
- Let ϴ be the parameter which we seek to estimate
- Since $Y_1, Y_2, \ldots, Y_n$ are iid, the joint distribution of the entire sample can be expressed as:

$$p(y_1, y_2, \ldots, y_n|\Theta) = p(y_1|\Theta) \times p(y_2|\Theta) \times \cdots \times p(y_n|\Theta)$$

- The function $p(y_1, y_2, \ldots, y_n|\Theta)$ is called the likelihood function
- Given some observed data (e.g., $y_1 = 5, y_2 = 6, \ldots, y_n = 4$), maximum likelihood estimation leverages this function to find the most likely value of ϴ

Likelihood Function

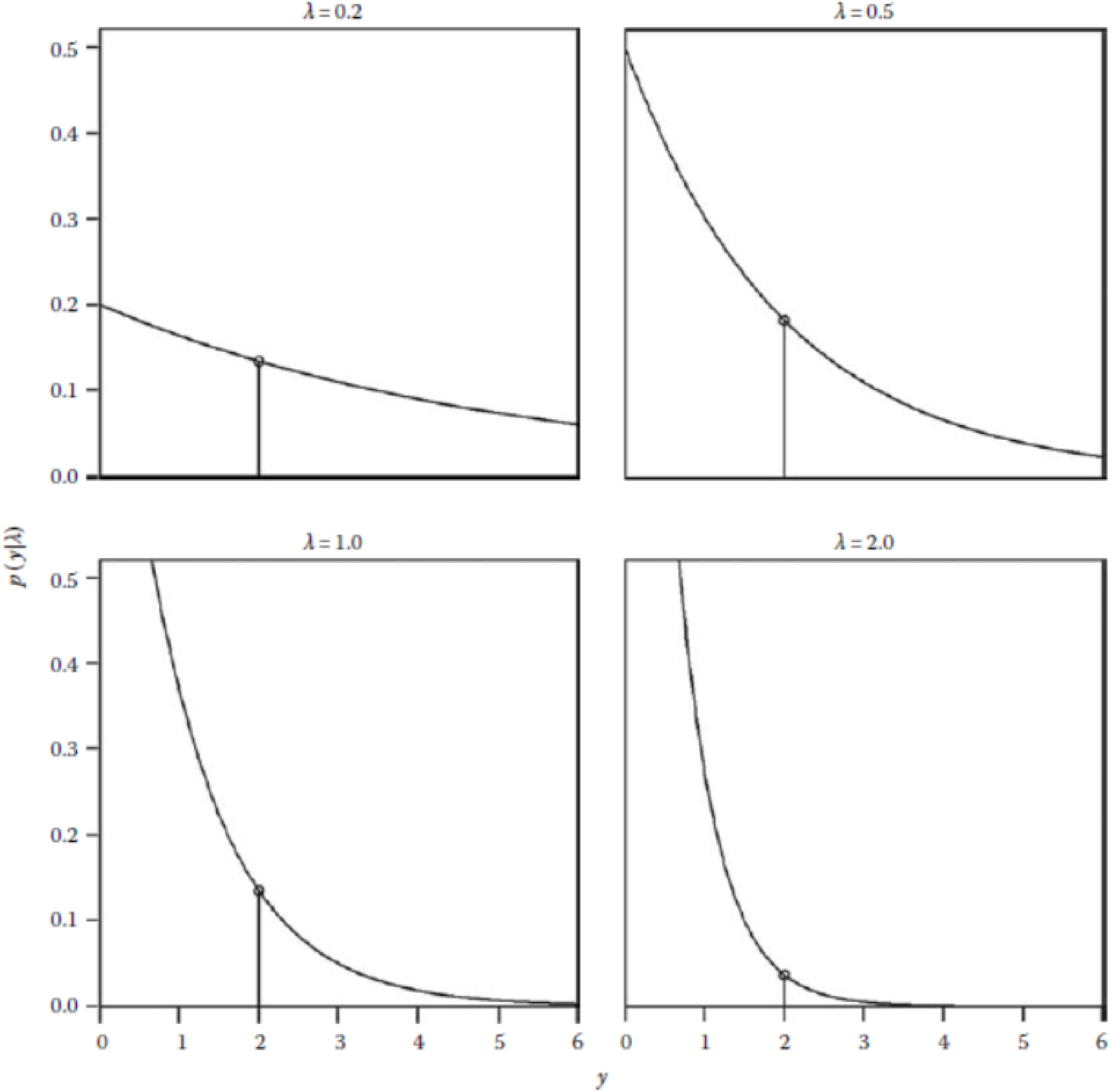
- $L(\Theta | y_1,\ y_2,\ \ldots, y_n)$= $p(y_1, y_2, \ldots, y_n|\Theta)$

is identical to a probability density function except that it is a function of the parameter Θ for the fixed values of $y_1, y_2, \ldots, y_n$ (a pdf is on the other hand a function of $y_1, y_2, \ldots, y_n$ for a fixed value of Θ

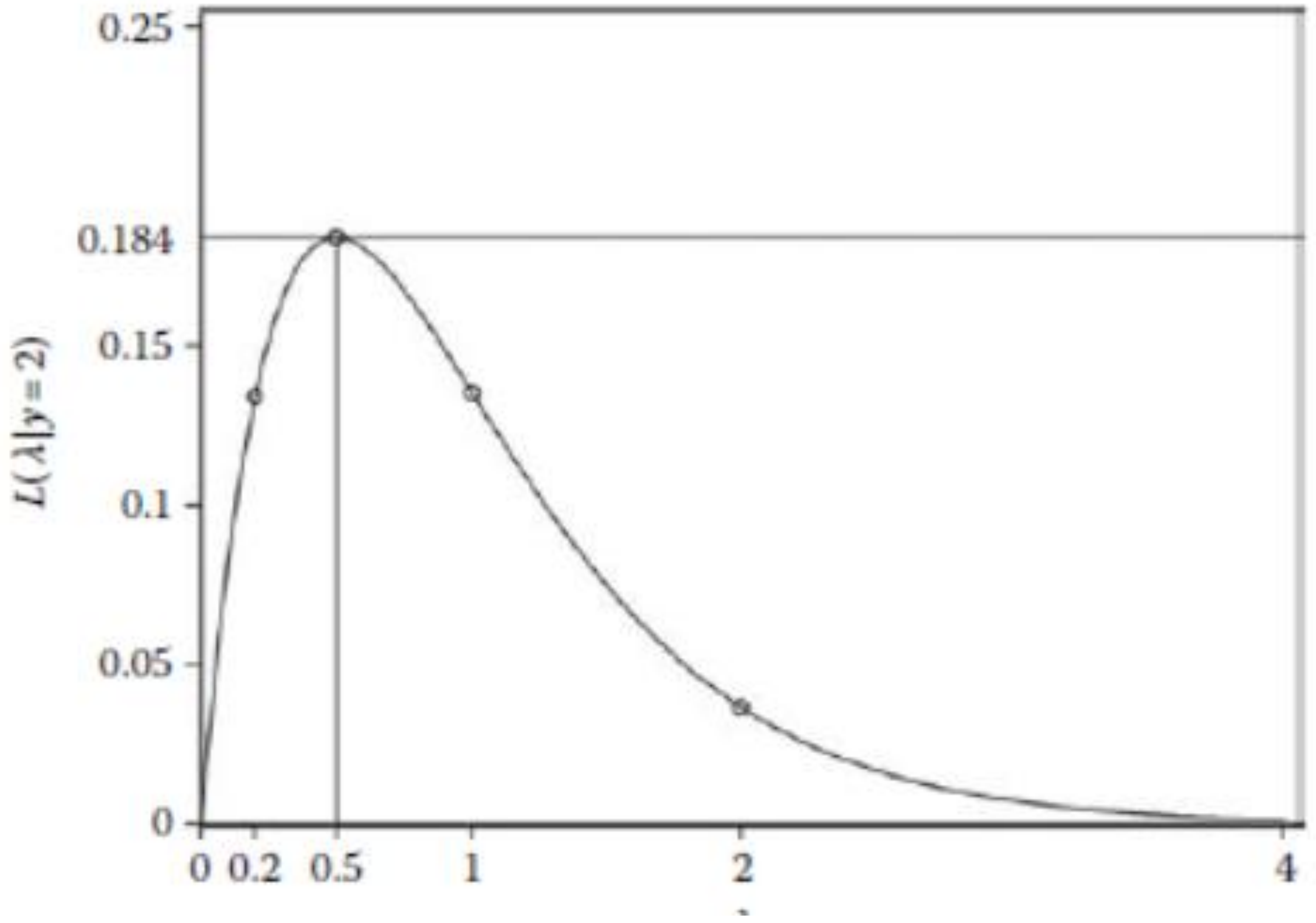
Example

- Recall the exponential distribution, $p(y|\lambda) = \lambda e^{-\lambda y}$ for y>0
- Suppose we hypothesize that the customer service waiting time at a call center follows this distribution
- Suppose we get one observation y_1=2.0 from a single customer. We can attempt to try to find the value of theta using this data
- We have: $L(\lambda|y = 2)$=$\lambda e^{-2\lambda}$

Example: How a single data point reduces our uncertainty about the parameters of $p(y|\lambda) = \lambda e^{-\lambda y}$



Likelihood function based on a single observation, y=2.0



What if we had more than just one sample

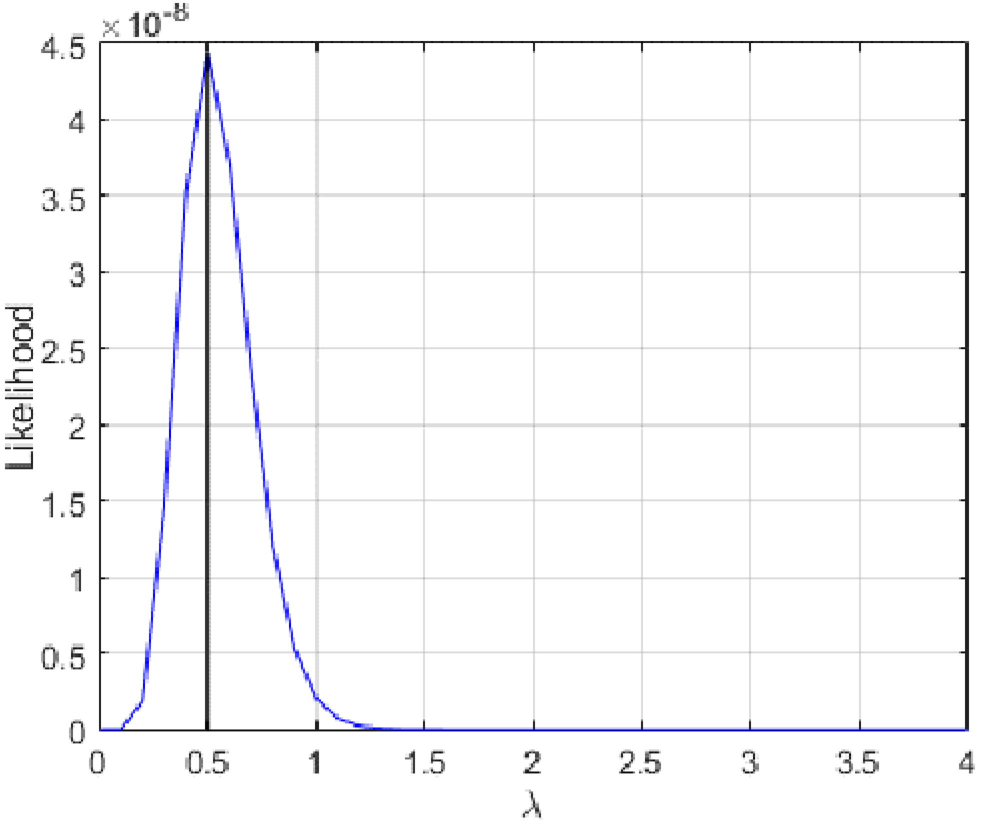
- Assume n=10; and the samples iid.

Data point	Value
y_1	2.0
y_2	1.2
y_3	4.8
y_4	1.0
y_5	3.8
y_6	0.7
y_7	0.3
y_8	0.2
y_9	4.5
y_{10}	1.5
$\bar{y} = \mathbf{2.0}$	

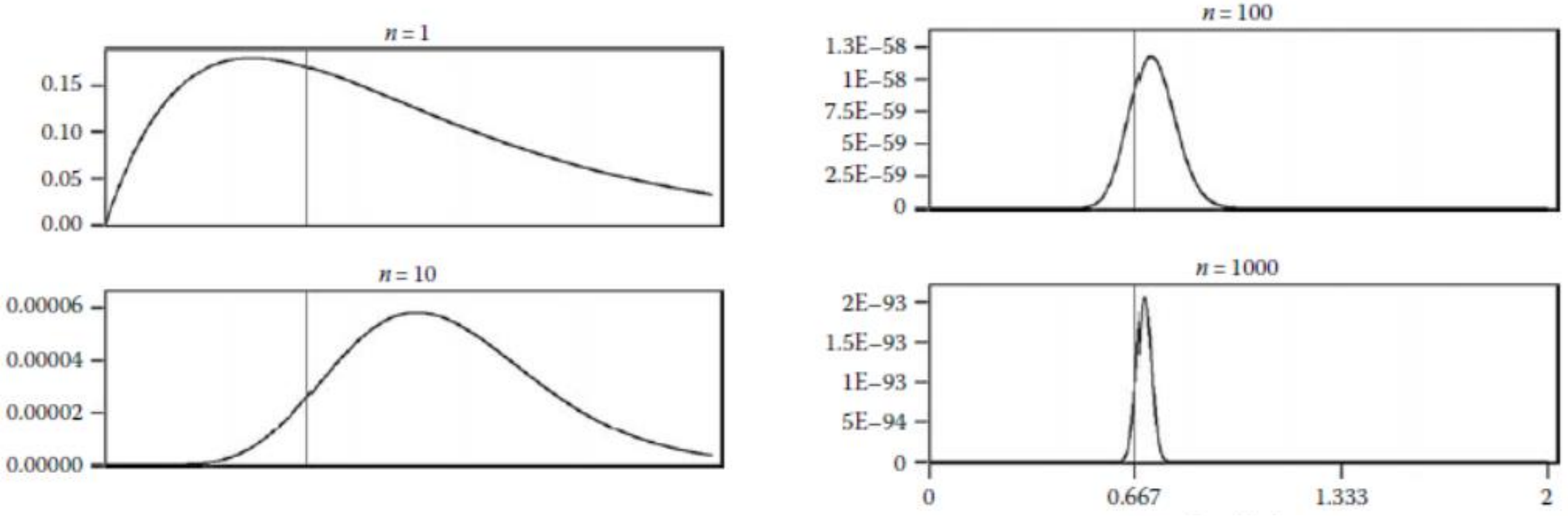
$L(\lambda | y_1, y_2, \ldots, y_{10})$
$= p(y_1, y_2, \ldots, y_{10} | \lambda)$
$= p(y_1|\lambda) \times p(y_2|\lambda) \times \cdots \times p(y_{10}|\lambda)$
$= \lambda e^{-\lambda y_1} \times \lambda e^{-\lambda y_2} \times \ldots \times \lambda e^{-\lambda y_{10}}$
$= \lambda^{10} e^{-\lambda y_1 - \lambda y_2 - \cdots - \lambda y_{10}}$
$= \lambda^{10} e^{-\lambda \sum_{i=1}^{10} y_i}$

But $\bar{y} = \frac{1}{10} \sum_{i=1}^{10} y_i \Rightarrow 10\bar{y} = \sum_{i=1}^{10} y_i \Rightarrow$
$L(\lambda | y_1, y_2, \ldots, y_{10}) = \lambda^{10} e^{-10\lambda\bar{y}}$

$L(\lambda | \bar{y} = 2) = \lambda^{10} e^{-20\lambda}$



Likelihood estimates with more samples



- ✓ Vertical axis is likelihood
- ✓ Horizontal axis is λ

Computing MLE

- Previous plots helped us visualize the behavior of the likelihood function as sample sizes increased.
- However in practice, we may not be able to graph the function that easily – often one has to deal with lots of parameters (both in terms of numbers and variety)
- **Options:**
  - Sometimes it is possible to use calculus to find the parameter value(s) which maximize the likelihood function
  - Numerical methods can also be used to find the parameter values

Computing MLE

- To find the maxima, we could use $\frac{\partial L(\Theta \mid y_1, y_2, \ldots, y_n)}{\partial \Theta} = 0$
- However, often its much simpler to maximize the logarithm of the likelihood function instead.
  - **Log-likelihood function:** $\boldsymbol{LL(\Theta \mid y_1, y_2, \ldots, y_n) = \ln(L(\Theta \mid y_1, y_2, \ldots, y_n))}$
- Reasons?
  - Density functions often complex -- have exponential terms
  - Log of product of likelihoods is a sum – easier to deal with
  - Likelihood values tend to be too small –logarithm helps make them bigger and reduces the risk precision loss.

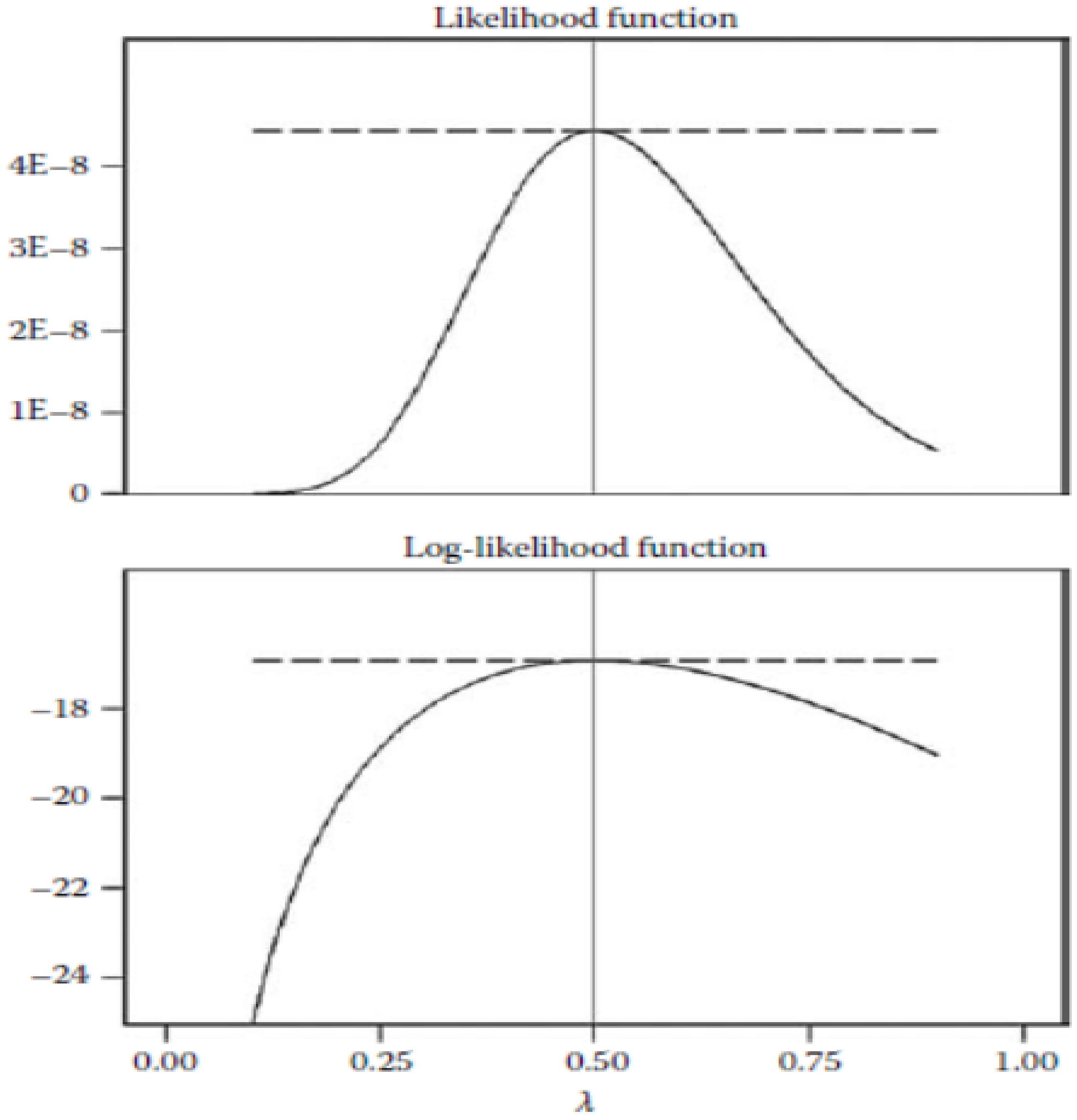
Does the Log-likelihood provide the same MLE?

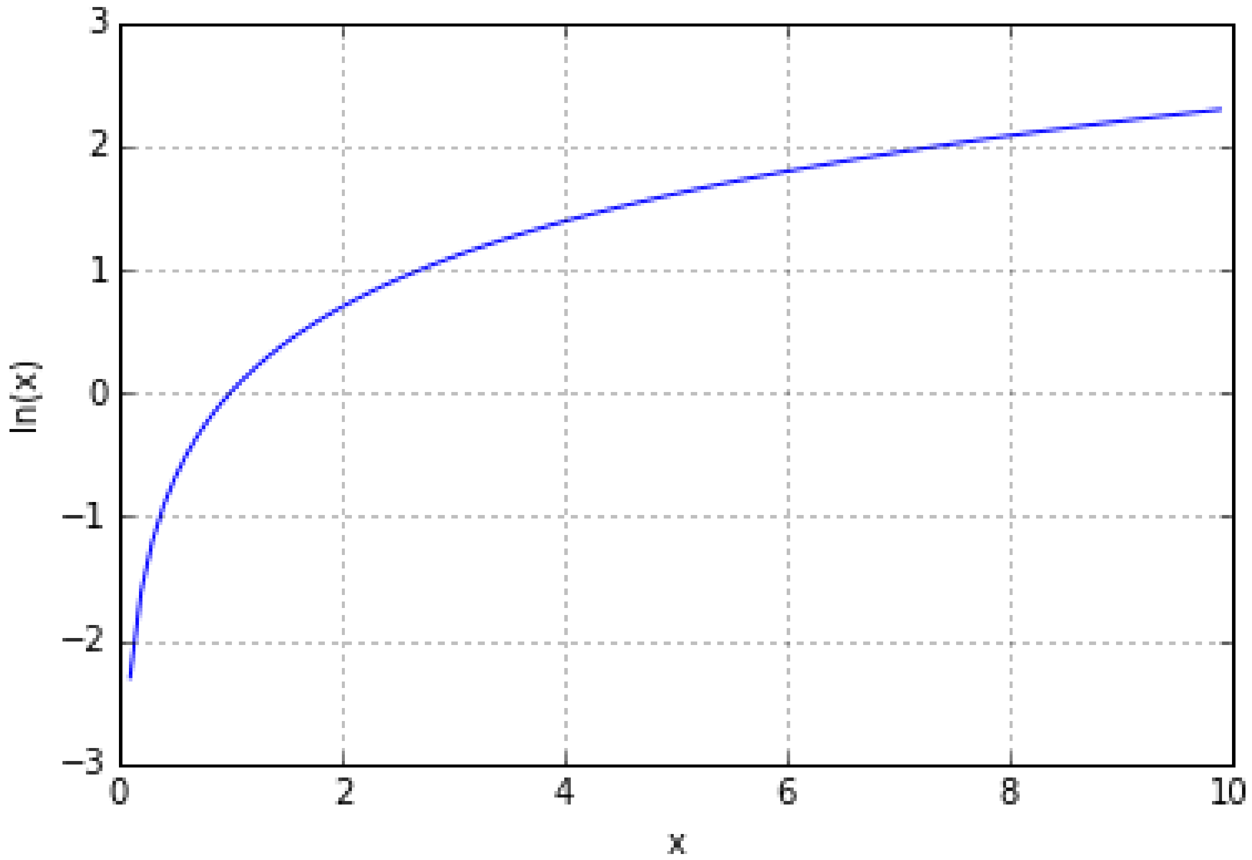
- Natural log is a monotonically increasing function of its argument, so if $a_1 > a_2$ then $\ln(a_1) > \ln(a_2)$
- Thus: $L(\theta_1 | y_1, y_2, \ldots, y_n) > L(\theta_2 | y_1, y_2, \ldots, y_n)$

  is equivalent to

  $ln(L(\theta_1 | y_1, y_2, \ldots, y_n)) > ln(L(\theta_2 | y_1, y_2, \ldots, y_n))$
- So we are free to maximize the log-likelihood function

Does Log-likelihood provide the same MLE?





Coin Toss Example

- Problem: We have a coin, and want to estimate its bias – whats the probability it lands on heads/tails?
- Let $P(Heads) = \theta$ and $P(Tails) = 1 - \theta$. Assume we toss the coin 12 times and obtain: $HHHHHHHTHTHH$.
- $L(\theta) = \theta^{10}(1-\theta)^2 = \theta^{10}(1 - 2\theta + \theta^2) = \theta^{10} - 2\theta^{11} + \theta^{12}$
- If we maximize the likelihood directly:

$$\frac{d}{d\theta} L(\theta) = 10\theta^9 - 22\theta^{10} + 12\theta^{11}$$

$$\frac{d}{d\theta} L(\theta) = 0 \Rightarrow 10\theta^9 - 22\theta^{10} + 12\theta^{11} = 0 \Rightarrow \theta^9(10 - 22\theta + 12\theta^2) = 0$$

- $\Rightarrow \theta = 0, \theta = 1, \theta = \frac{10}{12}$. Further evaluation of each turning point confirms $\theta = \frac{10}{12}$

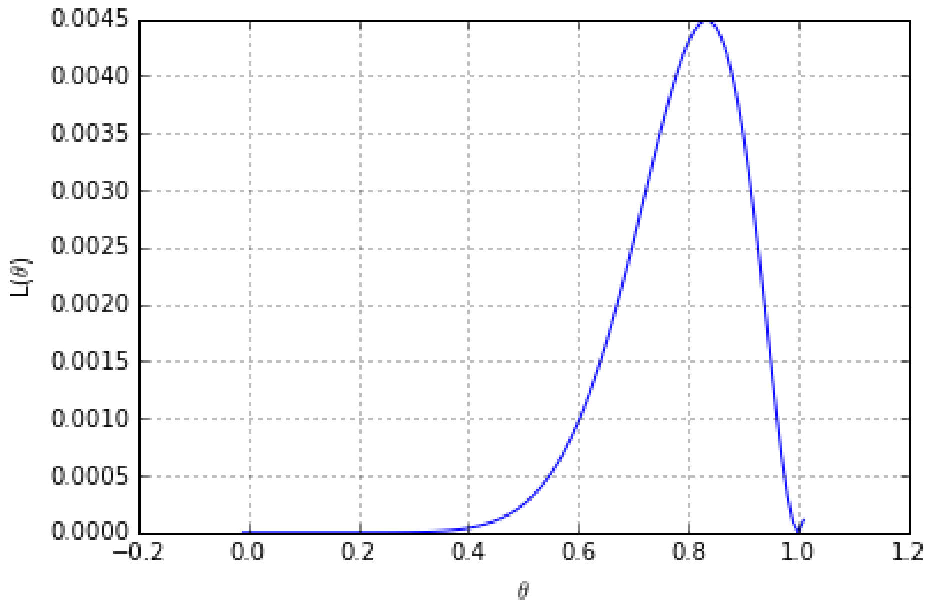
Example: Coin Toss

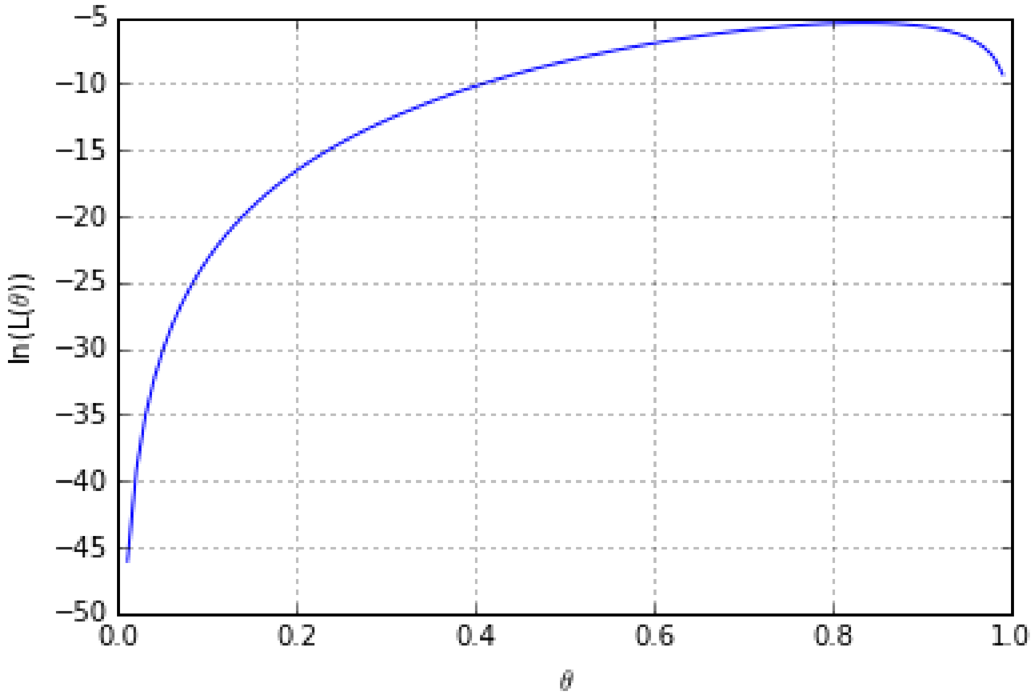
$$L(L(\theta)) = \ln(L(\theta)) = \ln(\theta^{10}(1-\theta)^2)$$
$$= 10ln\theta + 2ln(1-\theta)$$

$$\frac{d(L(L(\theta))}{\theta} = \frac{10}{\theta} - \frac{2}{1-\theta}$$

$$\frac{d(L(L(\theta))}{d\theta} = 0$$

$$\Rightarrow 10(1-\theta) - 2\theta = 0 \Rightarrow \theta = \frac{10}{12}$$





General Case of binary valued rv

- Assume a binary valued r.v X having: $P(X = 1) = \theta \; and \; P(X = 0) = 1 - \theta$
- $L(\theta) = \theta^x(1 - \theta)^y$ if we observe x ones and y zeros.
- $ln(L(\theta)) = xln(\theta) + yln(1 - \theta)$
- $\frac{\partial L(\theta)}{\partial \theta} = \frac{x}{\theta} + \frac{-1.y.}{(1- \theta)}$
- $\frac{\partial L(\theta)}{\partial \theta} = 0 \Rightarrow \frac{x}{\theta} + \frac{-1.y.}{(1- \theta)} = 0 \Rightarrow x(1 - \theta) - y\theta = 0$
- $\Rightarrow x - \theta(x + y) = 0 \Rightarrow \theta = \frac{x}{x+y}$

MLE for Univariate Gaussian

- $Gaussian\ pdf$: $P(x|\,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- If we have received N samples $x_1, x_2, \ldots, x_N$

$$L(\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \times \ \ldots \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_N-\mu)^2}{2\sigma^2}}$$

$$= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

MLE for Univariate Gaussian

$$ln\big(L(\mu,\sigma)\big) = \ln\left(\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) = \sum_{i=1}^{N} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$= K + \sum_{i=1}^{N} \left(-\log\sigma - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\frac{\partial}{\partial\mu} ln\big(L(\mu,\sigma)\big) = \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu)$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) \ = 0 \Rightarrow \sum_{i=1}^{N}(x_i-\mu) \ = 0 \Rightarrow -N\mu + \sum_{i=1}^{N}(x_i) = 0$$

MLE for Univariate Gaussian

$$-N\mu + \sum_{i=1}^{N}(x_i) = 0 \Rightarrow N\mu = \sum_{i=1}^{N} x_i$$

$$\Rightarrow \mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$$

- Recall: $ln\big(L(\mu,\sigma)\big) = K + \sum_{i=1}^{N}(-\log\sigma - \frac{(x_i-\mu)^2}{2\sigma^2})$
- $\frac{\partial}{\partial\sigma} ln\big(L(\mu,\sigma)\big) = \sum_{i=1}^{N}(-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}) = 0$
- $\Rightarrow \frac{-N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i - \mu)^2 = 0 \Rightarrow N\sigma^2 = \sum_{i=1}^{N}(x_i - \mu)^2$

$$\sigma_{ML}^2 = \frac{\sum_{i=1}^{N}(x_i - \mu_{ML})^2}{N}$$

Biasness of Gaussian MLE Estimators -- Mean

- **μ_{ML} is unbiased if E(μ_{ML})=μ**
- Recall: $\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$
- $E(\mu_{ML}) = E(\bar{x}) = E\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \frac{1}{N} E\left(\sum_{i=1}^{N} x_i\right)$
- $= \frac{1}{N}\left(\sum_{i=1}^{N} E(x_i)\right)$
- For iid samples, $E(x_i) = E(x)$
- $\Rightarrow E(\mu_{ML}) = \frac{1}{N}\left(\sum_{i=1}^{N}(E(x)\right) = \frac{1}{N}.N.E(x)$
- $= E(x) = \mu$

Biasness of Gaussian MLE Estimators -- Variance

- **σ^2_{ML} is unbiased if E(σ^2_{ML})=σ^2;**

$$E(\sigma^2_{ML}) = E\left(\frac{\sum_{i=1}^{N}(x_i-\mu_{ML})^2}{N}\right) = \frac{1}{N}E\left(\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)$$

- Recall μ_{ML}=sample mean, $\bar{x}$.

$$\Rightarrow \sum_{i=1}^{N}(x_i-\bar{x})^2 = \sum_{i=1}^{N}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) = N\bar{x}^2 - 2N\bar{x}\,\bar{x} + \sum_{i=1}^{N} x_i^2$$

Previous expression is after manipulating two terms in the summation using $\sum_{i=1}^{N}\bar{x}^2 = N\,\bar{x}^2$ and $N\bar{x} = \sum_{i=1}^{N} x_i$. Substituting the expression back into $E(\sigma^2_{ML})$:

$$\Rightarrow E(\sigma^2_{ML}) = \frac{1}{N}E\left(N\bar{x}^2 - 2N\bar{x}^2 + \sum_{i=1}^{N} x_i^2\right)$$

Biasness of Gaussian MLE Estimators

$$E(\sigma_{ML}^2) = \frac{1}{N} E\left(N\bar{x}^2 - 2N\bar{x}^2 + \sum_{i=1}^{N} x_i^2\right) = \frac{1}{N} E\left(\sum_{i=1}^{N} x_i^2\right) - E(\bar{x}^2)$$

$$\Rightarrow E(\sigma_{ML}^2) = \frac{1}{N} \sum_{i=1}^{N} E(x_i^2) - E(\bar{x}^2)$$

Recall standard formulae for variance:

$\sigma^2 = E(x^2) - \mu^2$ and $E(\bar{x}^2) - \mu^2 = Var(\bar{x}) = \frac{\sigma^2}{N}$

$$\Rightarrow E(\sigma_{ML}^2) = \frac{1}{N} (\sum_{i=1}^{N} (\sigma^2 + \mu^2)) - \left(\frac{\sigma^2}{N} + \mu^2\right) = \frac{1}{N} (N\sigma^2 + N\mu^2) - \frac{\sigma^2}{N} - \mu^2$$

Biasness of Gaussian MLE Estimators

- $E(\sigma_{ML}^2) = \frac{1}{N}(N\sigma^2 + N\mu^2) - \frac{\sigma^2}{N} - \mu^2 = \sigma^2 + \mu^2 - \frac{\sigma^2}{N} - \mu^2$
- $E(\sigma_{ML}^2) = \sigma^2 - \frac{\sigma^2}{N} = \sigma^2 \frac{(N-1)}{N}$
- Any observations on biasness of the MLE estimator for variance ?

Gaussian MLE Estimators – Correcting the Bias in σ_{ML}^2

$Since\ E(\sigma_{ML}^2) = \sigma^2 - \frac{\sigma^2}{N} = \sigma^2 \frac{(N-1)}{N},$

$\Rightarrow E(\frac{N}{N-1}.\sigma_{ML}^2) = \sigma^2 \frac{(N-1)}{N}.\frac{N}{N-1} = \sigma^2$

Recall $\sigma_{ML}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2$

$$\frac{N}{N-1}.\sigma_{ML}^2 = \frac{N}{N-1}.\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(x_i - \mu_{ML})^2$$