# 5. Maximum Likelihood Estimator

## Review: Maximizing composite functions

1/1 point (graded)

The **arguments of the minima** (*resp.* **arguments of the maxima** ) of a function $f(x)$, denoted by $\textrm{argmin} f(x)$ (*resp.* $\textrm{argmax}\, f(x)$), is the value(s) of $x$ at which $f(x)$ is minimum (*resp.* maximum). We can also restrict to a subset $S$ of the domain of $f$, and denote by $\textrm{argmin}_{x\in S}\, f(x)$ (*resp.* $\textrm{argmax}_{x\in S}\, f(x)$) the value(s) of $x \in S$ at which $f(x)$ is minimum (*resp.* maximum) over $S$.

Let $f(x) > 0$ be continuous **positive** function with $\max_x f(x) = 1.$ (Note that $\max_x f(x)$ is the maximum value of the function, which is different from $\textrm{argmax} f(x)$, the value of the argument $x$ at which the function is maximum.)

Which of the following functions of $f(x)$ has the same $\textrm{argmax}$ as $f(x)$? In other words, which of the following attain their maxima at the same $x$-value(s) as $f(x)$?
(Choose all that apply.)

- [x] $f(x)^2$
- [x] $\sqrt{f(x)}$
- [x] $\ln(f(x))$
- [x] $-\ln\left(\frac{1}{f(x)}\right)$
- [ ] $\cos(f(x))$
- [x] $-\cos(2f(x))$

**Solution:**

We go through the choices in order.

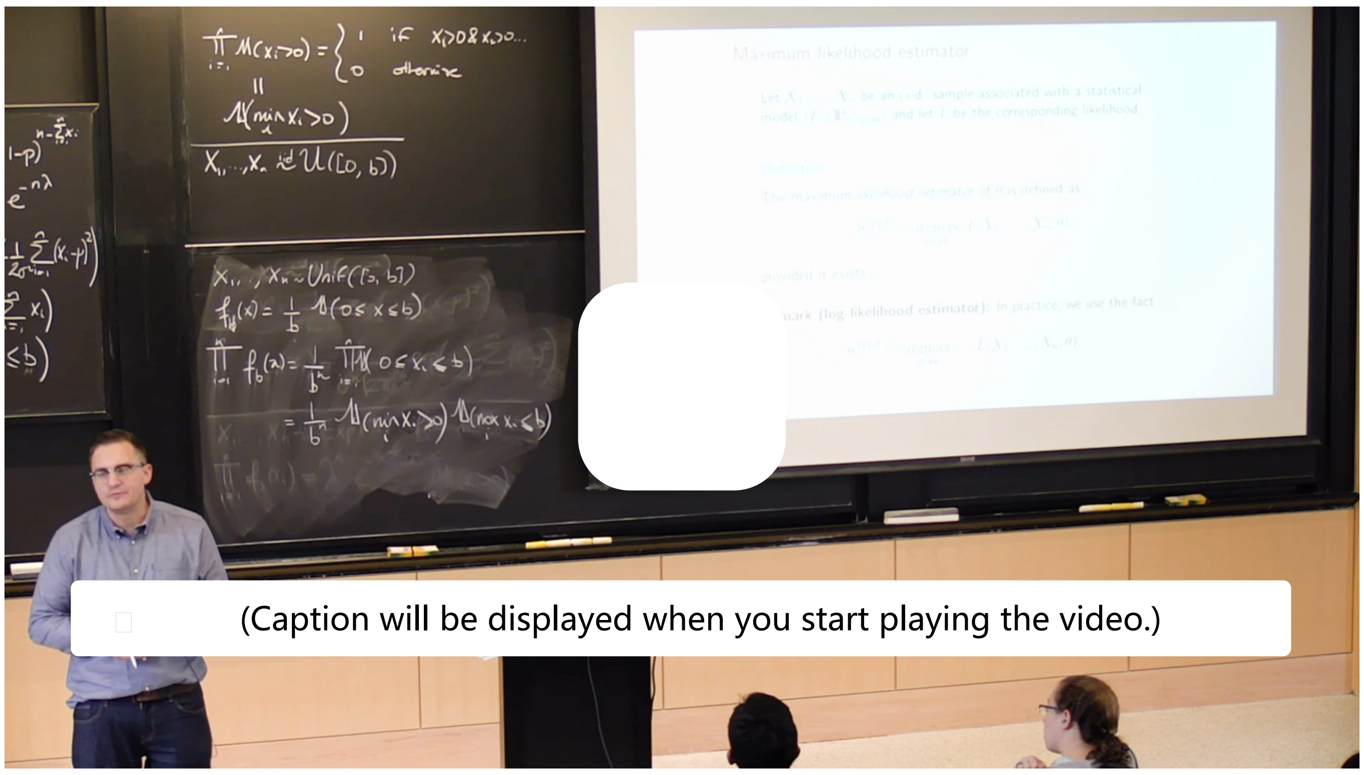
- Since $y^2$, $\sqrt{y}$, $\ln(y) = -\ln\left(\frac{1}{y}\right)$ are all **strictly increasing** functions, their value increases as $y$ increases. Hence, the functions $f(x)^2$, $\sqrt{f(x)}$, $\ln(f(x))$, $-\ln\left(\frac{1}{f(x)}\right)$ attain their maxima when $f(x)$ attain its maximum, which is at $x = \textrm{argmax}\, f(x)$.
- The cosine function is strictly decreasing in $(0, \pi)$. Given $\max_x f(x) = 1 < \pi$, $\cos(f(x))$ is in fact minimum when $f(x)$ is maximum.
- On the other hand, $-\cos(2y)$ is strictly increasing for $0 < 2y < \pi$. Since $\max_x 2f(x) = 2 < \pi$, we conclude that $-\cos(2f(x))$ is maximum again when $f(x)$ is maximum, at $x = \textrm{argmax}\, f(x)$.

提交 你已经尝试了2次（总共可以尝试2次）

Answers are displayed within the problem

# Definition of Maximum Likelihood Estimator and Log Likelihood

Start of transcript. Skip to the end.




So now that I've written a bunch of likelihoods,

I would like to be able to use them to compute an estimator.

And remember what we did, we said that minimum estimated kl

the same as maximizing likelihood.

So now I'm just left with a question, which

is how do I maximize those functions as functions

**视频**
下载视频文件

**字幕**
下载 SubRip (.srt) file
下载 Text (.txt) file

## Concept Check: Interpreting the Maximum Likelihood Estimator

1/1 point (graded)

Let $X_1,\ldots,X_n \overset{iid}{\sim} \mathbf{P}_{\theta^*}$ be discrete random variables. We construct a statistical model $(E, \{\mathbf{P}_\theta\}_{\theta\in\mathbb{R}})$ where $\mathbf{P}_\theta$ has pmf $p_\theta$. We observe our sample to be $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. The **maximum likelihood estimator** for $\theta^*$ is defined to be

$$\hat{\theta}_n^{MLE} = \mathrm{argmax}_{\theta\in\mathbb{R}} \left(\prod_{i=1}^{n} p_\theta\left(X_i\right)\right).$$

Which of the following is a correct interpretation of the maximum likelihood estimator (MLE) when applied to the sample $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$?
(Choose all that apply.)

- [x] The value of $\theta$ that maximizes the probability that $\mathbf{P}_\theta$ generates the data set $(x_1,\ldots,x_n)$.
- [x] The value of $\theta$ that minimizes an estimator of the KL divergence between $\mathbf{P}_\theta$ and the true distribution $\mathbf{P}_{\theta^*}$.
- [ ] It is the true parameter $\theta^*$

**Solution:**

- "The value of $\theta$ that maximizes the probability that $\mathbf{P}_\theta$ generates the data set $(x_1,\ldots,x_n)$." is correct. Since the likelihood is the joint density of $n$ iid samples from $\mathbf{P}_\theta$,

  $$\mathbf{P}_\theta\left[X_1 = x_1,\ldots,X_n = x_n\right] = L_n\left(x_1,\ldots,x_n,\theta\right).$$

  Hence, the MLE finds $\hat{\theta}_n$ that maximizes the probability that $x_1,\ldots,x_n$ were sampled from $P_{\hat{\theta}_n}$.

- "The value of $\theta$ that minimizes the KL divergence between $\mathbf{P}_\theta$ and the true distribution $\mathbf{P}_{\theta^*}$." is correct. In fact, this is how the MLE was derived from KL divergence. See the third section "Parameter Estimation via KL Divergence" of this lecture to review this fact.

- "It is the true parameter $\theta^*$ " is incorrect. The MLE is an estimator– it is constructed from the finite amount of data $x_1, \ldots, x_n$ that we are given– so we can't hope for it to exactly recover the true parameter.

**Remark:** Under some technical conditions the MLE is a **weakly consistent estimator** for $\theta^*$, meaning that the MLE will converge to $\theta^*$ in probability under these conditions. However, there are examples of statistical models where the maximum likelihood estimator will **not** converge to the true parameter.

提交 你已经尝试了1次（总共可以尝试2次）

Answers are displayed within the problem

## 讨论

显示讨论

**主题：** Unit 3 Methods of Estimation:Lecture 9: Introduction to Maximum Likelihood Estimation / 5. Maximum Likelihood Estimator