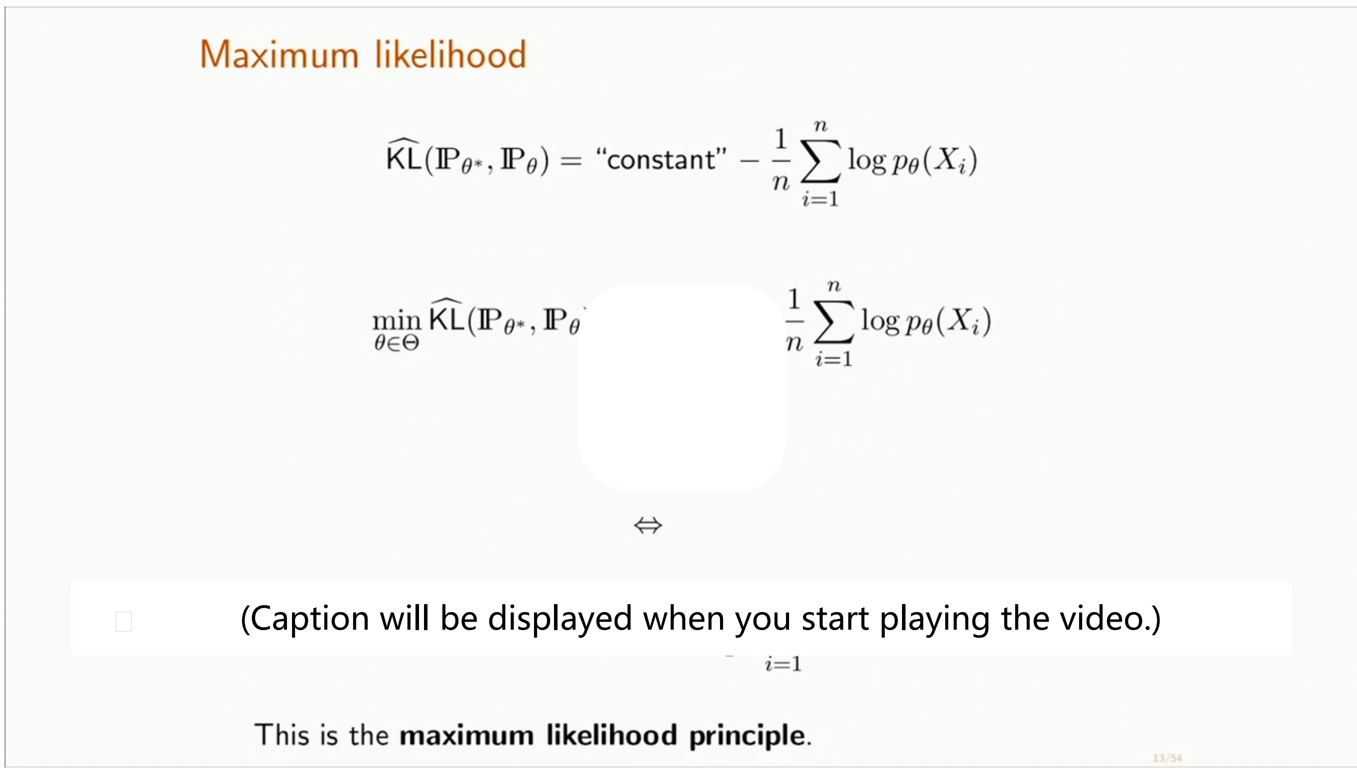
13. Parameter Estimation via KL Divergence

Deriving the Maximum Likelihood Estimator

Start of transcript. Skip to the end.



OK, so what is the maximum likelihood estimator?

Well, I have this KL, I just reproduce the same thing.

So now I am just trying to minimize the KL.

Now, what those equivalence signs

means, what you need to pay attention

is that this equivalence is not an equality.

视频

下载视频文件

字幕

下载 SubRip (.srt) file

下载 Text (.txt) file

The next four problems concern the following statistical set-up.

You observe discrete random variables

$$X_1, \ldots, X_n \stackrel{iid}{\sim} P_{\theta^*}$$

where θ^* is the true parameter. You construct an associated statistical model $(E, \{P_\theta\}_{\theta\in\mathbb{R}})$. The sample space E is discrete.

Intuitively, your goal is to find an estimator $\hat{\theta}_n \in \mathbb{R}$ so that the distributions $P_{\hat{\theta}_n}$ and P_{θ^*} are close. Precisely, you want to find an estimator $\hat{\theta}_n \in \mathbb{R}$ so that the quantity

$$\text{KL}\left(P_{\theta^*}, P_{\hat{\theta}_n}\right)$$

is as small as possible.

This approach will naturally lead to the construction of the **maximum likelihood estimator** .

Finding a Minimizer of KL Divergence

0/1 point (graded)

Consider the optimization problem in which we minimize the KL divergence between P_{θ^*}, the true distribution, and P_θ. Formally, we want to solve

$$\min_{\theta \in \mathbb{R}} \mathrm{KL}\left(P_{\theta^*}, P_\theta\right).$$

We are not so much interested in the minimum value attained by the objective function $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$, but rather the value of θ where the minimum is attained. We refer to such a θ as a **minimizer** .

Let's suppose that there is a unique minimizer for the above optimization problem– *i.e.*, if m is the minimum value of $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$, there is only one point $\theta_{\min}$ such that

$$m = \mathrm{KL}\left(P_{\theta^*}, P_{\theta_{\min}}\right).$$

For which θ is the minimum value of $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$ attained? (Equivalently, what is $\theta_{\min}$?)

- θ^*
- θ
- 0
- None of the above.

Solution:

The KL divergence is nonnegative, so $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right) \geq 0$. The right-hand side is achieved if we set $\theta = \theta^*$: $\mathrm{KL}\left(P_{\theta^*}, P_{\theta^*}\right) = 0$. Since the minimizer is unique by assumption, we conclude that the minimum value is attained at $\theta = \theta^*$.

Remark: The assumption that there is a unique minimizer holds if we are given that the parameter θ is identified. Here is why: since KL divergence is definite, $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right) = 0$ if and only if P_{θ^*} and P_θ are the same distribution. And if θ is identified, this implies that $\theta = \theta^*$
.

提交 你已经尝试了2次（总共可以尝试2次）

Answers are displayed within the problem

Can we Minimize KL Divergence Directly?

2/2 points (graded)
Let's use the same statistical set-up as above. Recall that you have access to the iid samples $X_1, \ldots, X_n$. You use these samples to build an estimator $\hat{\theta}_n$. Can you compute

$$\mathrm{KL}\left(P_{\hat{\theta}_n}, P_{1/2}\right)$$

without knowing θ^*, the true parameter?

- Yes
- No

Can you compute

$$\mathrm{KL}\left(P_{\theta^*}, P_{1/2}\right)$$

without knowing θ^*?

- ○ Yes
- ◉ No

Solution:

In general, we can compute $\mathrm{KL}\left(\mathbf{P},\mathbf{Q}\right)$ if and only if we know both distributions $\mathbf{P}$ and Q. Moreover, by our statistical model, we can compute P_θ if and only if we know the real number θ. Putting these last two facts together, we can compute

$$\mathrm{KL}\left(P_{\hat{\theta}_n}, P_{1/2}\right)$$

because $\hat{\theta}_n$ is known–it is an estimator so its expression does not depend on θ^*, the true parameter. However, regardless of how many samples we take, we cannot compute $\mathrm{KL}\left(P_{\theta^*}, P_{1/2}\right)$ exactly because the distribution P_{θ^*} is unknown.

Remark: Since we cannot even compute the function $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$ for general θ, this implies that the optimization problem

$$\min_{\theta\in\mathbb{R}} \mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$$

cannot be solved exactly, regardless of the number of samples we have. So to estimate the minimizer of this optimization problem (which is the true parameter θ^*) we will have to consider an approximation for $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$.

提交 你已经尝试了1次（总共可以尝试2次）

Answers are displayed within the problem

Finding the Minimizer for an Approximation of KL Divergence

1/1 point (graded)

We use the same statistical set-up as above. Recall that $X_1,\ldots,X_n \stackrel{iid}{\sim} P_{\theta^*}$. Let p_θ be the pmf of P_θ.

Which of the following is a (weakly) **consistent** estimator for

$$\mathbb{E}_{\theta^*}\left[\ln p_\theta\left(X\right)\right] = \sum_{x\in E} p_{\theta^*} \ln p_\theta\left(x\right)\ ?$$

- ○ $\frac{1}{n}\sum_{i=1}^{n} X_i$
- ◉ $\frac{1}{n}\sum_{i=1}^{n} \ln\left(p_\theta\left(X_i\right)\right)$
- ○ $\frac{1}{n}\sum_{i=1}^{n} \ln\left(p_{\theta^*}\left(X_i\right)\right) - \frac{1}{n}\sum_{i=1}^{n} \ln\left(p_\theta\left(X_i\right)\right)$
- ○ $\theta^* - \mathbb{E}_{\theta^*}\left[\ln p_{\theta^*}\right]$

Solution:

By the law of large numbers, $\frac{1}{n}\sum_{i=1}^{n}\ln\left(p_\theta\left(X_i\right)\right) \to \mathbb{E}_{\theta^*}\left[\ln p_\theta\right]$ in probability. Hence, the second choice is correct.

Remark 1: The KL divergence between P_{θ^*} and P_θ can be written

$$\mathrm{KL}\left(P_{\theta^*}, P_\theta\right) = \sum_{x\in E} p_{\theta^*}\ln p_{\theta^*}\left(x\right) - \sum_{x\in E} p_{\theta^*}\ln p_\theta\left(x\right) = \mathbb{E}_{\theta^*}\left[\ln p_{\theta^*}\left(X\right)\right] - \mathbb{E}_{\theta^*}\left[\ln p_\theta\left(X\right)\right]$$

where $X \sim P_{\theta^*}$.

Remark 2: While we can't find θ that minimizes $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$, we can find θ that minimizes

$$\widehat{\mathrm{KL}}\left(P_{\theta^*}, P_\theta\right) := \mathbb{E}_{\theta^*}\left[\ln p_{\theta^*}\right] - \frac{1}{n}\sum_{i=1}^{n}\ln\left(p_\theta\left(X_i\right)\right).$$

Here's why: the first term on the RHS, $\mathbb{E}_{\theta^*}\left[\ln p_{\theta^*}\right]$, does not depend on θ. Hence, the θ that minimizes $\widehat{\mathrm{KL}}\left(P_{\theta^*}, P_\theta\right)$ is the same as the θ that minimizes $-\frac{1}{n}\sum_{i=1}^{n}\ln\left(p_\theta\left(X_i\right)\right)$.

提交 你已经尝试了1次（总共可以尝试2次）

Answers are displayed within the problem

Deriving the Maximum Likelihood Estimator

1/1 point (graded)

We use the same statistical set-up as above. Recall that p_θ is the pmf of P_θ and $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta^*}$.

Suppose that $\theta_{\min}$ is a minimizer for the function

$$f\left(\theta\right) := -\frac{1}{n}\sum_{i=1}^{n}\ln\left(p_\theta\left(X_i\right)\right)$$

Which of the following functions is also minimized at $\theta_{\min}$?

- $g_1\left(\theta\right) = -\prod_{i=1}^{n} p_\theta\left(X_i\right)$
- $g_2\left(\theta\right) = 25 - \prod_{i=1}^{n} p_\theta\left(X_i\right)$
- $g_3\left(\theta\right) = h\left(\theta^*\right) - \prod_{i=1}^{n} p_\theta\left(X_i\right)$ where h is a function of θ^* that does **not** depend on θ.
- $g_4\left(\theta\right) = \theta^* - \prod_{i=1}^{n} p_\theta\left(X_i\right)$
- All of the above

Solution:

Observe that rescaling by n does not change where the minimum of a function is attained. Hence, $f\left(\theta\right)$ and $nf\left(\theta\right)$ have the same minimizer. Next, by the addition property of logarithms,

$$n f(\theta) = \sum_{i=1}^{n} \ln(p_\theta(X_i)) = \ln\left(\prod_{i=1}^{n} p_\theta(X_i)\right).$$

Since $\ln$ is an increasing function, the function

$$\theta \mapsto \prod_{i=1}^{n} p_\theta(X_i)$$

has the same minimizer as $\ln\left(\prod_{i=1}^{n} p_\theta(X_i)\right)$. Thus the first choice is correct.

Moreover, the second and third choices are also correct. Whenever we have an optimization problem

$$\min_{\theta\in\mathbb{R}} C + g(\theta)$$

where C does not depend on θ, then the above will have the same minimizer as the optimization problem

$$\min_{\theta\in\mathbb{R}} g(\theta).$$

In the second choice, $C = 25$ (which is independent of θ), and in the third choice, $C = h(\theta^*)$ (which by assumption is independent of θ).

Remark 1: The quantity

$$\hat{\theta}_n := \text{maximizer of } \prod_{i=1}^{n} p_\theta(X_i)$$

is referred to as the **maximum likelihood estimator** . Note that this is the same as the estimator

$$\hat{\theta}_n := \text{minimizer of } \quad -\frac{1}{n}\sum_{i=1}^{n} \ln(p_\theta(X_i))$$

considered in Remark 2 in the solution of the previous problem.

Remark 2: Under certain technical conditions, the maximum likelihood estimator is guaranteed to (weakly) converge to the true parameter θ^*.

提交 你已经尝试了1次（总共可以尝试3次）

Answers are displayed within the problem

讨论

显示讨论

主题： Unit 3 Methods of Estimation:Lecture 8: Distance measures between distributions / 13. Parameter Estimation via KL Divergence