Consistency of maximum likelihood estimates

Patrick Breheny

October 16, 2024

Introduction

- Today we will begin to prove the important asymptotic properties of maximum likelihood estimates
- We begin with consistency: $\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{P}} \boldsymbol{\theta}^*$ (this is weak consistency; MLEs are also strongly consistent under the same conditions, but we'll only concern ourselves with proving the weak case)
- Broadly speaking, we'll break this up into two cases: where the likelihood is unimodal and where it may not be (the latter case being considerably more complicated as there could be many local maxima, only one of which being the actual MLE)

An inconsistent MLE

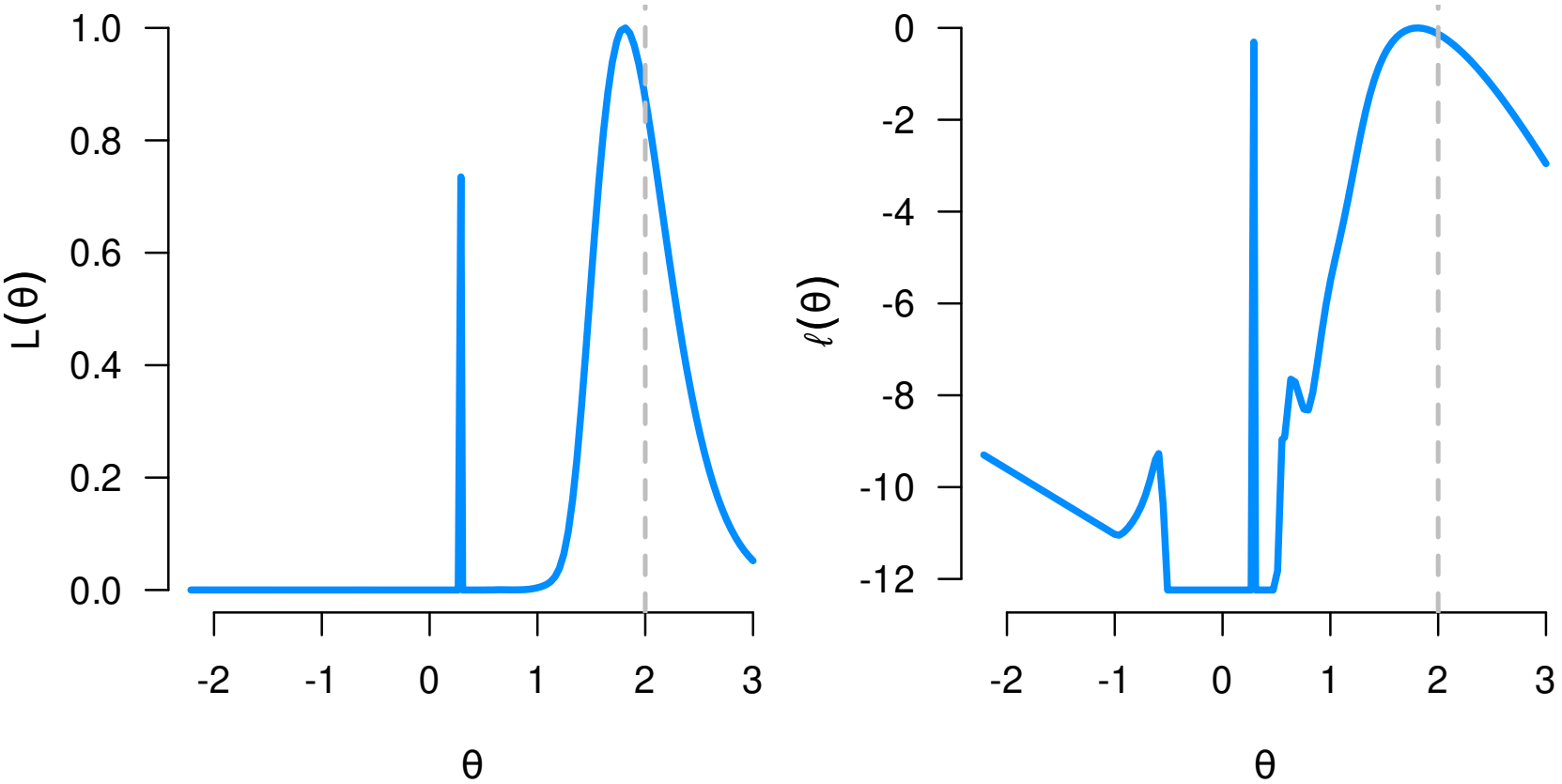
- To get a sense of the problems that arise when the likelihood can have multiple peaks, consider the following model[1]:

$$X_i \overset{\text{iid}}{\sim} \tfrac{1}{2}\mathrm{N}(0,1) + \tfrac{1}{2}\mathrm{N}(\theta, \exp(-2/\theta^2));$$

  in words, an equal mixture of a standard normal and a normal distribution whose variance goes to zero (fast!) as the mean goes to zero
- Let's generate some samples from this model with $\theta = 2$ and take at a look at its likelihood and what happens to it as $n \to \infty$

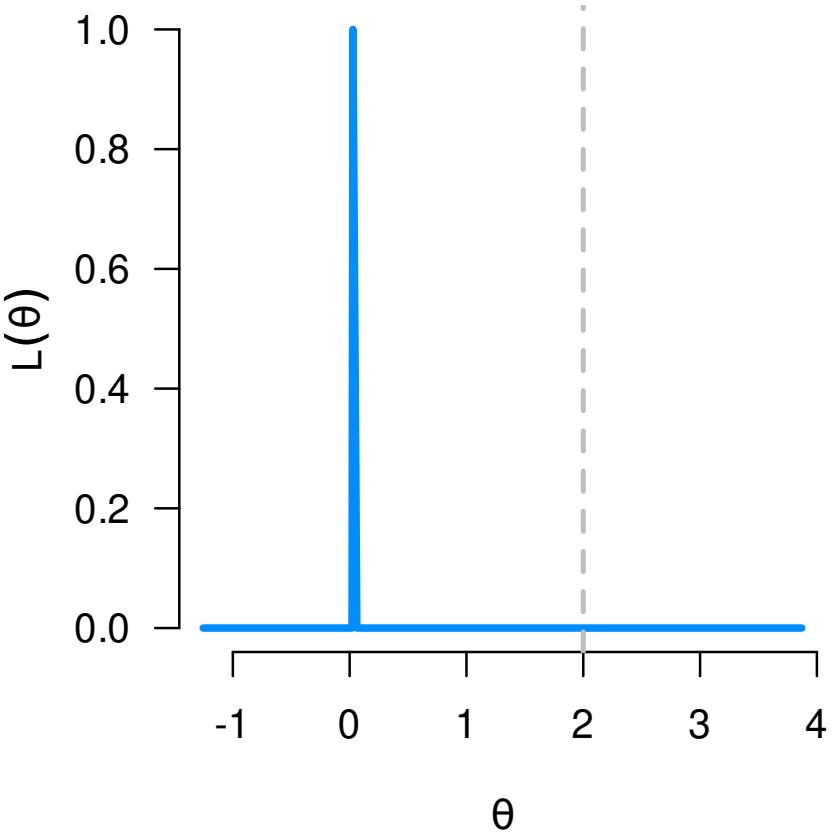
[1] This example comes from Radford Neal

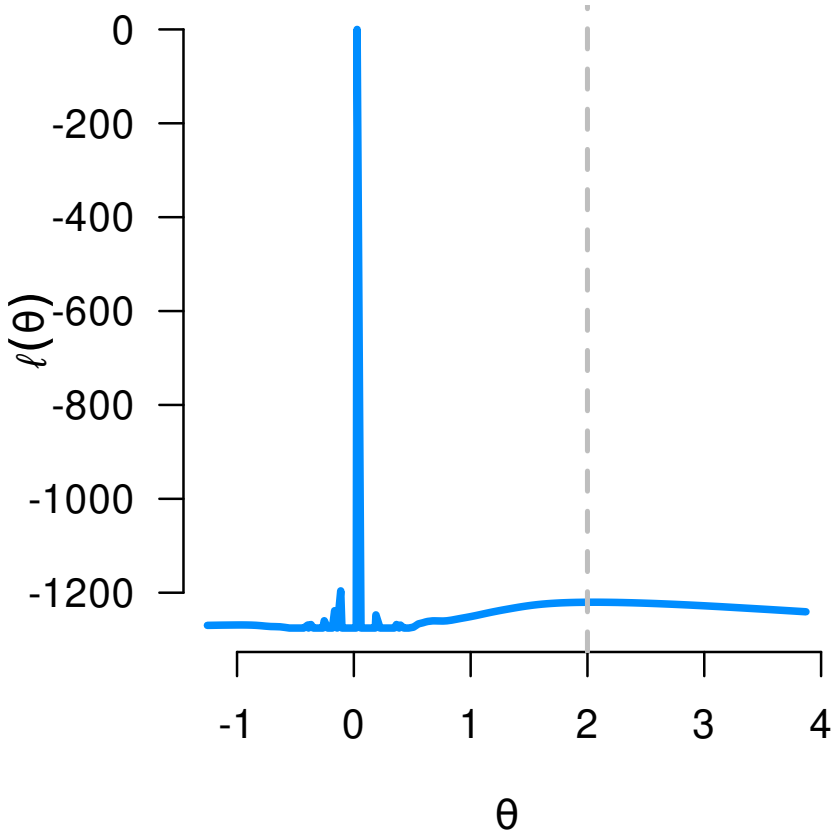
An inconsistent MLE: $n = 10$



An inconsistent MLE: $n = 40$

As $n \to \infty$, it is increasingly certain that a giant spike will occur near zero: $\hat{\theta} \xrightarrow{\mathrm{P}} 0 \neq 2$





Unimodal functions

- To rule out such situations, let's restrict attention to unimodal likelihoods, starting with a definition of "unimodal"
- In one dimension, a function f is unimodal if there exists a point m such that f is monotonically increasing for $x \leq m$ and monotonically decreasing for $x \geq m$
- Extending to multiple dimensions, a function $f : \mathbb{R}^d \to \mathbb{R}$ is unimodal if there exists a point $\mathbf{m}$ such that for all $\|\mathbf{u}\| = 1$, $f(\mathbf{m} + x\mathbf{u})$ is a monotone decreasing function of x
- A point $\mathbf{m} \in \mathbb{R}^d$ is a *strict local maximum* of a function $f : \mathbb{R}^d \to \mathbb{R}$ if there exists a neighborhood $N_r(\mathbf{m})$ such that $f(\mathbf{m}) > f(\mathbf{x})$ for all $\mathbf{x} \in N_r(\mathbf{m})$ with $\mathbf{x} \neq \mathbf{m}$
- A unimodal function has exactly one such point, and that point is the global maximum

Sufficient conditions for unimodality

- Proving that a function is unimodal is typically challenging unless we can resort to derivatives
- For any function that is twice differentiable, a sufficient (but not necessary) condition for unimodality is that its Hessian matrix $H(\mathbf{x}) = \nabla^2 f(\mathbf{x})$ is negative definite for all $\mathbf{x}$
- In the likelihood context, this means that the information matrix is positive definite for all $\boldsymbol{\theta}$

Log concavity

- Furthermore, if its Hessian is negative definite at all points, the function is concave
- In the likelihood context, then, if the information matrix is positive definite for all $\boldsymbol{\theta}$, then its log-likelihood is a concave function
- Such probability models are said to be *log-concave*
- Many common parametric models, including everything in the exponential family, are log-concave

Kullback-Leibler divergence

- Next, we need something like a "norm" that measures the distance between two probability distributions
- **Definition:** For two distributions p and q, the *Kullback-Leibler divergence* (commonly abbreviated KL divergence, also known as KL information) is defined as

$$\mathrm{KL}(p\|q) = \mathbb{E}_p \log \frac{p}{q} = \int \log \frac{p(x)}{q(x)} dP(x),$$

  where the integrand is defined to be $+\infty$ if $q(x) = 0, p(x) > 0$ and 0 if $p(x) = 0$
- Essentially, the KL divergence is measuring the ability of the likelihood ratio to distinguish between two distributions

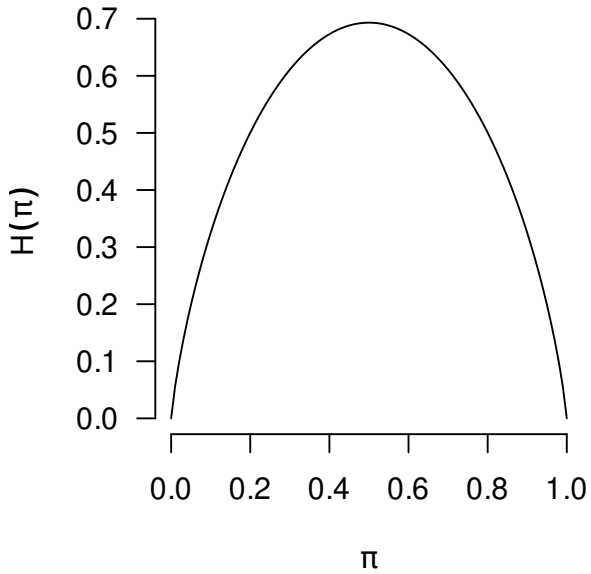
Entropy

- The KL divergence is related to a concept in physics and information theory called *entropy*, which is defined as

$$H(p) = -\mathbb{E} \log p(X)$$

- Entropy measures the degree of uncertainty in a distribution, with the uniform and constant distributions representing the extremes
- Note that $H(p) = -\mathrm{KL}(p||u) + \mathrm{Const}$, where u is a uniform distribution

For example, in the Bernoulli distribution:



Gibbs' inequality

- Note that the KL divergence is not symmetric: it is measuring the distance from distribution p to distribution q, not the other way around[2]
- Furthermore, the KL divergence does not satisfy the triangle inequality, so is not a norm; hence the term "divergence" as opposed to "distance"
- However, it does satisfy positivity
- **Theorem (Gibbs' inequality):** For any two distributions p and q, $\mathrm{KL}(p\|q) \geq 0$. Furthermore, $\mathrm{KL}(p\|q) = 0$ if and only if $p = q$ almost everywhere.
- This theorem is also known as the Shannon-Kolmogorov information inequality

[2] the symmetric version $\frac{1}{2}\mathrm{KL}(p\|q) + \frac{1}{2}\mathrm{KL}(q\|p)$ is known as the Jensen-Shannon divergence

Consistency

- So, what does this have to do with consistency?
- By the WLLN, we have

$$\frac{1}{n}\log\frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}^*)} = \frac{1}{n}\sum_i \log\frac{L_i(\boldsymbol{\theta})}{L_i(\boldsymbol{\theta}^*)} \xrightarrow{\mathrm{P}} -\mathrm{KL}(\boldsymbol{\theta}^* \| \boldsymbol{\theta}),$$

  which is less than 0 unless $p(x|\boldsymbol{\theta}) = p(x|\boldsymbol{\theta}^*)$ almost everywhere
- In other words, $\mathbb{P}\{L(\boldsymbol{\theta}) < L(\boldsymbol{\theta}^*)\} \to 1$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$

Identifiability

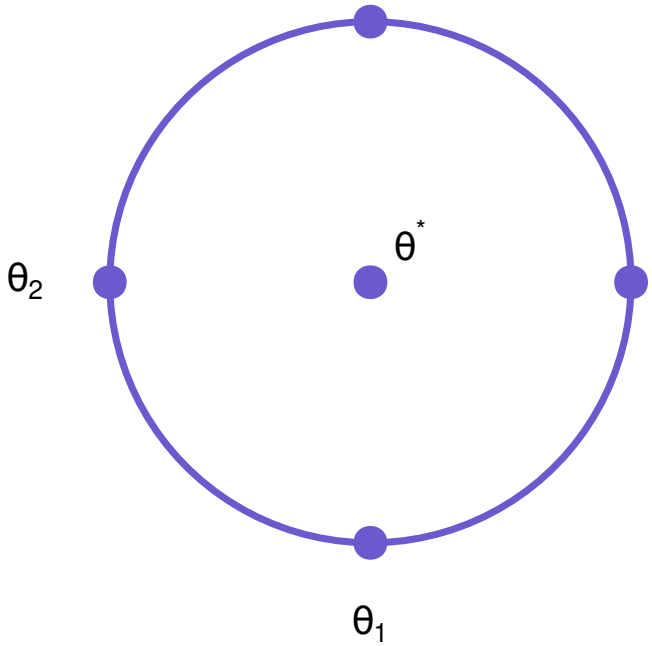
- More quantitatively, the likelihood ratio converges to zero exponentially fast, with a rate given by the KL divergence
- Again, the only condition here is that we do not have $p(x|\boldsymbol{\theta}) = p(x|\boldsymbol{\theta}^*)$ almost everywhere; this is known as *identifiability* and if it is violated, the models $p(x|\boldsymbol{\theta})$ and $p(x|\boldsymbol{\theta}^*)$ are said to be not identifiable
- For example, suppose $\mathbf{x}_{1i} \overset{\text{iid}}{\sim} \mathrm{N}(\mu + \alpha, 1)$ and $\mathbf{x}_{2i} \overset{\text{iid}}{\sim} \mathrm{N}(\mu + \beta, 1)$; this is not identifiable because $\{\mu, \alpha, \beta\} = \{0, 2, 4\}$ specifies the same distribution as $\{\mu, \alpha, \beta\} = \{3, -1, 1\}$ (along with infinitely many other combinations)

Consistency?

- Are we done? Have we established consistency?
- In one dimension, yes!
- **Theorem:** Let $\{p(x|\theta) : \theta \in \Theta \subset \mathbb{R}\}$ be a probability model that is unimodal (with respect to θ) and identifiable, and suppose $X_i \overset{\text{iid}}{\sim} p(x|\theta^*)$. Then $\hat{\theta} \xrightarrow{\text{P}} \theta^*$.
- The argument also works if the parameter space $\boldsymbol{\Theta}$ is finite

Multiple dimensions

- Unfortunately, this argument breaks down even with $d = 2$:



- To apply our earlier argument, we need to show that $\mathbb{P}\{L(\boldsymbol{\theta}^*) > L(\boldsymbol{\theta})\} \to 1$ for the entire ring; use Gibbs' inequality all we like, but it's no help – the ring contains an infinite number of points

Consistency: Assumptions

What assumptions do we need?

(A) IID: $X_1, \ldots, X_n$ are iid with density $p(x|\boldsymbol{\theta}^*)$.

(B) Interior point: There exists an open set $\boldsymbol{\Theta}^* \subset \boldsymbol{\Theta} \subset \mathbb{R}^d$ that contains $\boldsymbol{\theta}^*$.

(C) Smoothness: For all x, $p(x|\boldsymbol{\theta})$ is continuously differentiable with respect to $\boldsymbol{\theta}$ up to third order on $\boldsymbol{\Theta}^*$, and satisfies the following conditions:

- (i) Derivatives up to second order exist and can be passed under the integral sign in $\int dP(x|\boldsymbol{\theta})$.
- (ii) The Fisher information $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)$ is positive definite.
- (iii) The third derivatives are bounded: there exists $M(x)$ satisfying $\mathbb{E}M(X) < \infty$ such that $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}^*} \left| \nabla^3 \ell(\boldsymbol{\theta}|x)_{jkm} \right| \leq M(x)$ for all j, k, m.

Consistency: Assumptions (cont'd)

- To avoid the possibility of multiple local maxima, I'll also add the following assumption:

(D) Log-concavity: The Fisher information $\mathcal{I}(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and $\boldsymbol{\Theta}$ is a convex set

- Obviously, Assumption (D) implies much of assumption (C); I give them as separate assumptions here since assumptions (A)-(C) are standard, while assumption (D) is "extra"
- Next time, we will consider what happens when we remove it, retaining only (A)-(C)

Remarks: IID

- Keep in mind that Condition (C) describes what happens for a *single observation*, whereas Condition (A) describes how these observations are related to each other (iid)
- We are covering the IID case because it is the obvious place to start, but keep in mind that IID is not at all a necessary condition: the theoretical properties we will prove apply to many non-IID settings (likelihood would not be terribly useful if it only worked in IID settings)
- However, additional conditions may be required in non-IID cases, as we saw in the lecture on the Lindeberg-Feller CLT

Remarks: C(i)

- This condition is necessary in order to ensure $\mathbb{E}\mathbf{u}(\theta^*) = \mathbf{0}$ and $\mathbb{V}\mathbf{u}(\theta^*) = \mathbb{E}(\boldsymbol{\mathcal{I}})$; some authors assume this directly instead
- Whether we can pass derivatives under the integration sign is governed by the DCT; in this case, it requires that

$$\frac{|p(x|\boldsymbol{\theta}_1) - p(x|\boldsymbol{\theta}_2)|}{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|} \leq g(x, \boldsymbol{\theta}^*)$$

  for every x and for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}^*$ and that $g(x, \boldsymbol{\theta}^*)$ is integrable (this is for ∇; the condition for ∇^2 is similar)
- This condition (known as a Lipschitz condition) limits how much the derivative can change within $\boldsymbol{\Theta}^*$ (alternatively, we could require $\nabla_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta})$ to be continuous over $\boldsymbol{\Theta}^*$)

Remarks: C(ii) and C(iii)

Note that C(ii) applies to the Fisher information, while C(iii) applies to the derivative of the observed information – this is important!

- The observed information could randomly fail to be positive definite; this does not cause problems (well, not asymptotically)
- Meanwhile, we need a bound on the observed derivatives, which can include x; this means that our bound $M(x)$ must be allowed to be random

Remarks: Continuity of information

- Although not explicitly stated, the above conditions also ensure that both the observed information and Fisher information are continuous functions of $\boldsymbol{\theta}$
- All differentiable functions are continuous; thus, by requiring the third derivative to exist, we require that the second derivative (the observed information) is continuous (by the same reasoning, the score must be continuous)
- Also, if the third derivative is bounded, then the first and second derivatives are bounded; this allows us to use the dominated convergence theorem and

$$\lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)$$

Remarks: Uniform continuity

- In fact, these conditions ensure that the observed information is uniformly continuous over $\boldsymbol{\Theta}^*$
- In other words, we can find a single δ such that $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ is close to $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)$ for any $\boldsymbol{\theta}, \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}^*$ whenever $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta$
- Uniform continuity is important because it provides uniform convergence of the observed information:

$$\tfrac{1}{n}\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) \xrightarrow{\mathrm{P}} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)$$

as $\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{P}} \boldsymbol{\theta}^*$; note that we can't simply use the law of large numbers or the continuous mapping theorem here because both the information *and* the point at which the information is being evaluated are changing simultaneously

Consistency of the MLE

- OK, let's now prove the following important theorem
- **Theorem (Consistency of the MLE):** Suppose assumptions (A)-(D) are met. Then the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is consistent:

$$\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{P}} \boldsymbol{\theta}^*.$$

- Connecting this to our earlier remarks on uniform convergence towards the beginning of the course, note that pointwise convergence of the likelihood ratio around the boundary of $\boldsymbol{\Theta}^*$ was not enough; we needed uniform convergence over the entire boundary

Alternative proof

- It is possible to prove consistency of the MLE under considerably weaker conditions than this; in particular, without any requirements on differentiability
- This was the approach taken by Wald (1949), who used a compactness argument, which involves the existence of finite subcovers of open sets and is considerably more abstract than the approach we have taken here
- Our approach follows that of Cramér (1946), and is more common in the literature (or at least, the literature I read)

Convergence in non-standard settings

- Because of this, however, it is possible for the MLE to be consistent even in situations that do not meet our regularity conditions
- For example:
  - $X_i \overset{\text{iid}}{\sim} \text{Bern}(\theta)$; $\hat{\theta} \xrightarrow{\text{P}} \theta^*$ even if $\theta^* = 1$ (on the boundary)
  - $X_i \overset{\text{iid}}{\sim} \text{Laplace}(\theta)$; $\hat{\theta} \xrightarrow{\text{P}} \theta^*$ even though likelihood not differentiable at θ^*
  - $X_i \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$; $\hat{\theta} \xrightarrow{\text{P}} \theta^*$ even though likelihood isn't even continuous at θ^* (let alone differentiable)