Consistency of maximum likelihood estimates

Patrick Breheny

October 5

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

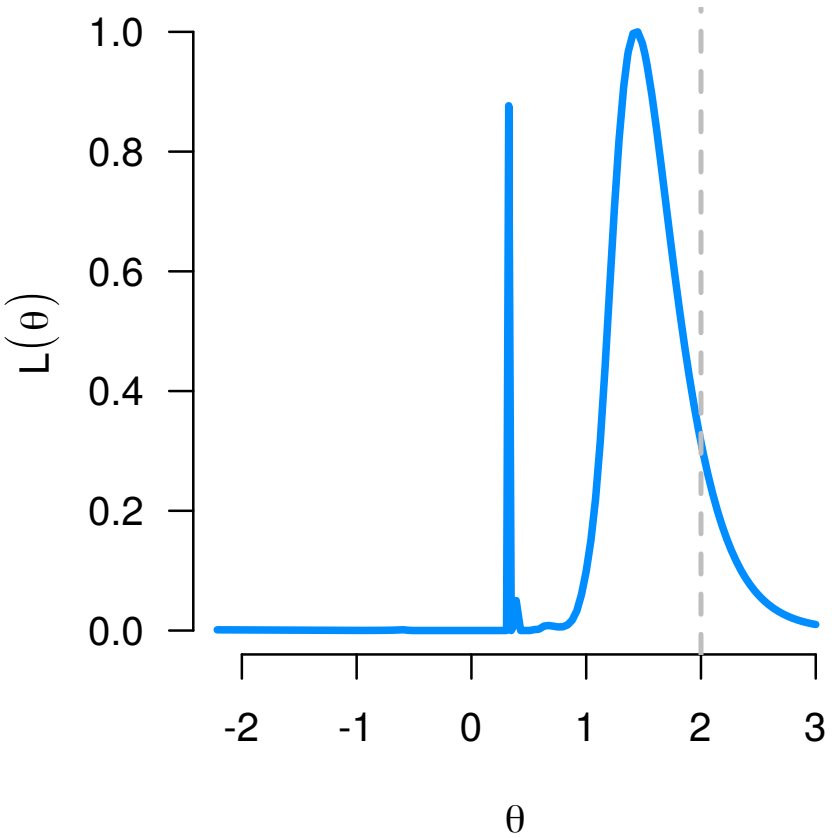
- To get a sense of the problems that arise when the likelihood can have multiple peaks, consider the following model[1]:

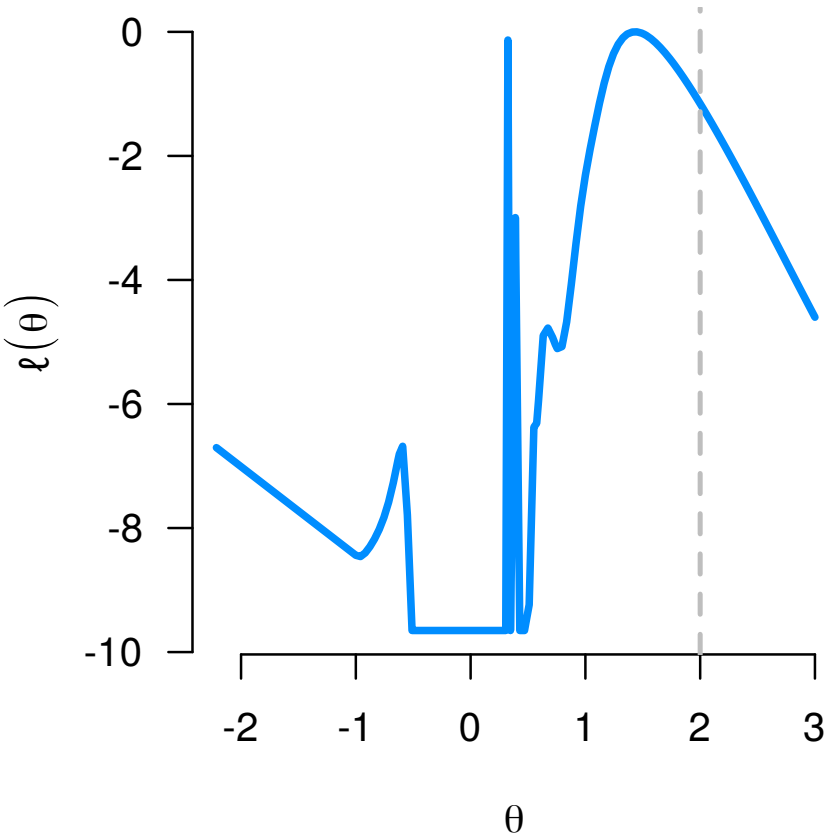
$$X_i \overset{\text{iid}}{\sim} \tfrac{1}{2}\mathrm{N}(0,1) + \tfrac{1}{2}\mathrm{N}(\theta, \exp(-2/\theta^2));$$

  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

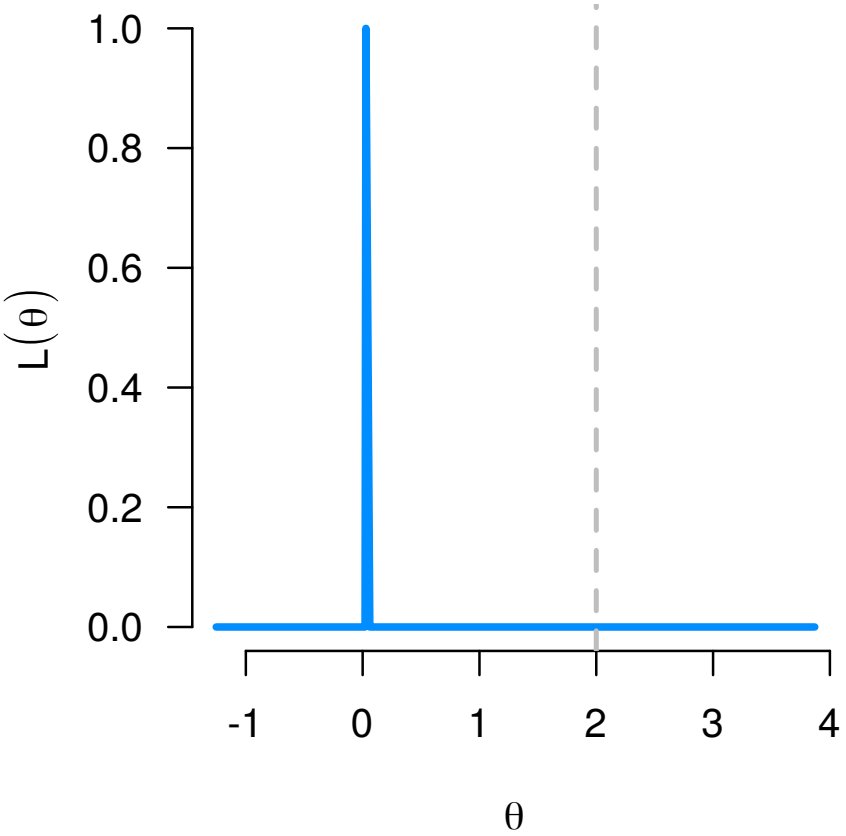
An inconsistent MLE: $n = 10$

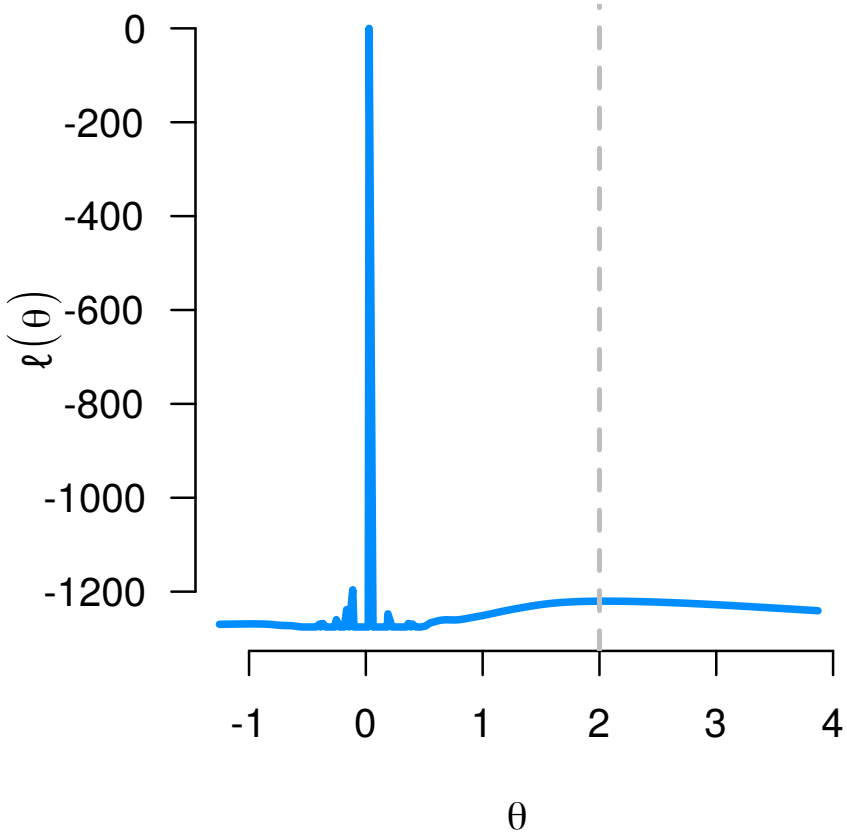




An inconsistent MLE: $n = 40$

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 $H(\mathbf{x})$ 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-Liebler 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

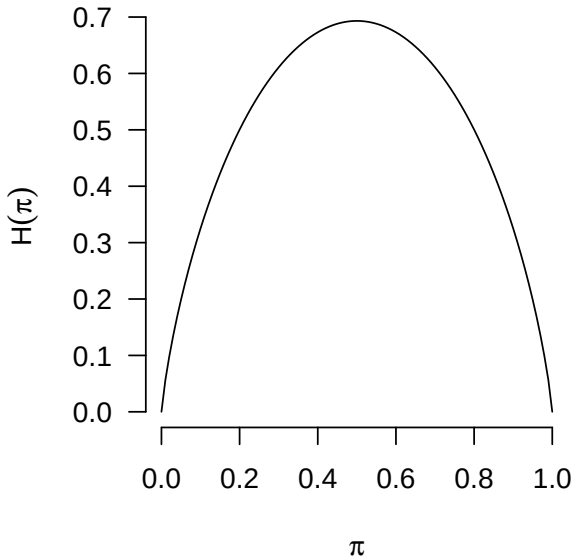
Entropy

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

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

- 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 $\frac{1}{2}KL(p\|q) + \frac{1}{2}KL(q\|p)$ 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

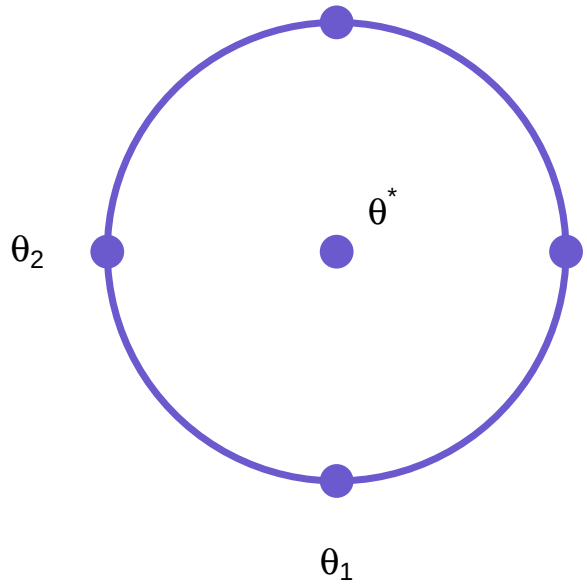
- More quantitatively, the likelihood ratio converges to zero exponentially fast, with a rate given by the KL divergence
- Again, the only condition here is that we do not have $p(x|\boldsymbol{\theta}) = p(x|\boldsymbol{\theta}^*)$ almost everywhere; this is known as *identifiability* and if it is violated, 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

Eigenvalue introduction

- To make progress in the multiparameter case, we will instead use arguments based on taking Taylor series expansions of the log-likelihood about the point $\boldsymbol{\theta}^*$
- This proof is also going to involve eigenvalues, so let's take a moment now to review their meaning and properties (probably should have done this earlier in the course, but oh well)
- This is not going to be a comprehensive treatment of the entire subject, just an overview of the most important properties as they pertain to statistical theory – in particular, we are only concerned with symmetric matrices here

Eigendecompositions

- The most important thing about eigenvalues is that they allow us to "diagonalize" a matrix: if $\mathbf{A}$ is a symmetric $d \times d$ matrix, then it can be factored into:

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top,$$

  where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$ of $\mathbf{A}$ and the columns of $\mathbf{Q}$ are its eigenvectors
- Furthermore, eigenvectors are orthonormal, so we have $\mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$

Eigenvalues and "size"

- This is very helpful from a conceptual standpoint, as it allows us to separate the "size" of a matrix ($\mathbf{\Lambda}$) from its "direction(s)" ($\mathbf{Q}$)
- For example, we have already seen that one measure of the size of a matrix is based on $\lambda_{\max}$ (for a symmetric matrix, its spectral norm is its largest eigenvalue)
- In addition, the trace and determinant, two other ways of quantifying the "size" of a matrix, are simple functions of the eigenvalues:
  - $\text{tr}(\mathbf{A}) = \sum_i \lambda_i$
  - $|\mathbf{A}| = \prod_i \lambda_i$

Eigenvalues and inverses

- Once one has obtained the eigendecomposition of $\mathbf{A}$, calculating its inverse is straightforward
- If $\mathbf{A}$ is not singular, then $\mathbf{A}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^{\top}$; note that since $\mathbf{\Lambda}$ is diagonal, its inverse is trivial to calculate
- Even if $\mathbf{A}$ is singular, we can obtain a generalized inverse: $\mathbf{A}^{-} = \mathbf{Q}\mathbf{\Lambda}^{-}\mathbf{Q}^{\top}$, where $(\mathbf{\Lambda}^{-})_{ii} = \lambda_i^{-1}$ if $\lambda_i \neq 0$ and $(\mathbf{\Lambda}^{-})_{ii} = 0$ otherwise
- Many other important properties of matrices can be deduced entirely from their eigenvalues:
  - $\mathbf{A}$ is positive definite if and only if $\lambda_i > 0$ for all i
  - $\mathbf{A}$ is positive semidefinite if and only if $\lambda_i \geq 0$ for all i
  - If $\mathbf{A}$ has rank r, then $\mathbf{A}$ has r nonzero eigenvalues and the remaining $d - r$ eigenvalues are zero

Extreme values

- Lastly, there is a connection between a matrix's eigenvalues and the extreme values of its quadratic form
- Let the eigenvalues $\lambda_1, \ldots, \lambda_d$ of $\mathbf{A}$ be ordered from largest to smallest. Over the set of all vectors $\mathbf{x}$ such that $\|\mathbf{x}\|_2 = 1$,

$$\max \mathbf{x}^T \mathbf{A} \mathbf{x} = \lambda_1$$

and

$$\min \mathbf{x}^T \mathbf{A} \mathbf{x} = \lambda_d$$

Consistency: Assumptions

OK, back to consistency; 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 can be passed under the integral sign in $\int dP(x|\boldsymbol{\theta})$.
- (ii) The Fisher information $\boldsymbol{\mathcal{I}}_1(\boldsymbol{\theta}^*)$ is positive definite.
- (iii) The third derivatives $\nabla^3 \ell(\boldsymbol{\theta})$ are bounded by M on $\boldsymbol{\Theta}^*$: $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}^*} \left| \nabla^3 \ell(\boldsymbol{\theta})_{jkm} \right| \leq M$ 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 $\mathcal{I}_1(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

- 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)

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:

$$\left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\| \xrightarrow{\mathrm{P}} 0.$$

- 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

Derivative conditions

- It is possible to prove consistency of the MLE under considerably weaker conditions than this; in particular, without any requirements on differentiability
- In particular, Wald (1949) used a compactness argument to show consistency; the gist of it is that if Θ is compact, there exists a finite subcover of Θ, which allows us to use Gibbs' inequality (since we need only use it a finite number of times)
- However, this proof is a bit more abstract and the conditions a bit harder to understand, so we have presented the approach originally published by Cramér (1946), which is also the approach more common in the literature (or at least, the literature I read)

Convergence in non-standard settings

- Keep in mind that we have provided consistency conditions that are sufficient, not necessary
- It is therefore 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)