# Asymptotic Normality of Maximum Likelihood Estimators and A Discussion on their Finite Normal Approximations

Naitong Chen

December 8, 2020

## 1 Background

Maximum likelihood estimation, as a dominant player in statistical inference, is a simple and intuitive method of parameter estimation. Given a set of i.i.d. observations, maximum likelihood estimation finds the value, denoted maximum likelihood estimate, that maximizes the likelihood function that describes the observations at hand. Due to the simple and intuitive nature of maximum likelihood estimation, it can be found in many different applications: from simple regression analysis to imputing missing values in a given dataset. Besides the fact that maximum likelihood estimation can be easily and intuitively applied to different problems, there are many other properties that make maximum likelihood estimation applealing.

In this report, we focus on one of these nice properties, namely the *asymptotic normality of maximum likelihood estimators* (MLE). Specifically, as the sample size approaches infinity, the MLE converges in distribution to a normal distribution centred at the true value of the parameter. This property are significant in that it justifies the behaviour of such estimators with theoretical rigour. The fact that, given a sufficiently large sample, some arbitrary MLE approximately follows the well-studied normal distirbution centred at the (unknown) true parameter value unifies how we can interpret this large class of estimators and how we can apply them to solving other problems. One of the immediate applications of this result is that we can quantify uncertainties of the MLEs.

While asymptotic normality is nice to have, a natural question to ask then is how accurate the MLEs are compared to other estimators. It turns out that, for an unbiased estimator, the lowest variance that it can possibly achieve is precisely the asymptotic variance of the MLE. This is formalized as the *Cramer-Rao lowerbound*. This lowerbound on the variance of unbiased estimators tells us that, asymptotically, the MLE is the best we can do in terms of achieving minimal variance.

However, this is still an asymptotic result. In practice, we only work with finitely many observations at a time, and so even with a sufficiently large sample size, the asymptotic normality can only give us approximate results, which is unsatisfying. Besides, there is no clear cut as in what is considered a sufficiently large sample. Fortunately, some work has been done in [AL15], which bounds the distance between an MLE obtained from finitely many observations and its asymptotic normal distribution.

In this report, we develop proofs for both the *asymptotic normality of MLEs* and the *Cramer-Rao lowerbound*, and follow up with a discussion on the work done in [AL15], as well as its potential extensions and applications. Before diving into the proofs, we set up notations and assumptions that will be used throughout the rest of the report.

### 1.1 Notations and Assumptions

For simplicity, we focus on the case where the obsevations are continous and there is one continous parameter to be estimated. Specifically, let $X_1, \cdots, X_n$ be i.i.d. continuous random variables in $\mathbb{R}$ with probability

density function $f(x;\theta)$, where $\theta \in \Theta$ is a unknown parameter. Denote the likelihood function and log-likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta), \qquad l(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

Then the maximum likelihood estimator can be obtained by

$$\hat{\theta} = \arg\max_{\theta\in\Theta} l(\theta) = \arg\max_{\theta\in\Theta} L(\theta).$$

In addition we use $\mathbb{E}_\theta, \mathrm{Var}_\theta$ and $P_\theta$ to denote expectation, variance and probability with respect to $f(\cdot;\theta)$.

Throughout the report, we assume the following regularity conditions.

- (R1) $f(x;\theta)$ is identifiable: $\theta_1 \neq \theta_2 \implies f(x;\theta_1) \neq f(x;\theta_2)$.
- (R2) $f(x;\theta)$ has common support for all $\theta \in \Theta$.
- (R3) $\theta_0$ is an iterior point in $\Theta$.
- (R4) $f(x;\theta)$ is twice differentiable in $\theta$.
- (R5) The integral $\int f(x;\theta)$ can be different twice in $\theta$ under the integral sign.
- (R6) $\hat{\theta}$ is the unique solution to $\frac{\partial l(\theta)}{\theta} = 0$.

# 2 Asymptotic Normality and Efficiency of MLE

The proofs for both the *asymptotic normality of MLE* and the *Cramer-Rao lowerbound* are well-documented in [HMC05], although some of the details and explanations are left out. In this section, we present both of these results following the main ideas in [HMC05], and fill in much of the left-out details. In particular, since the relevance of the regularity conditions that we have assumed in Section 1.1 may not be immediately clear, we point out, in the subsequent sections, how each of them are essential to arriving at both the *asymptotic normality of MLE* and the *Cramer-Rao lowerbound*.

## 2.1 Score Function and Fisher Information

To begin, we define two functions of the derivative of the log-likelihood function. The *score function* $\frac{\partial \log f(x;\theta)}{\partial\theta}$ is defined to be the derivative of the log-likelihood function of a single observation. And the *Fisher information* is defined as

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right)^2\right]. \tag{1}$$

These two quantities repeatedly come up in the proofs of both the *asymptotic normality of MLE* and the *Cramer-Rao lowerbound*. Using (R4) and (R5) from the regularity conditions, we can show the following properties of the score function and Fisher information through some simple algebraic manipulations.

**Lemma 2.1.** *Under regularity conditions and using the notations set up in Section 1.1, we have*

$$\mathbb{E}_\theta\left(\frac{\partial \log f(X_1;\theta)}{\partial\theta}\right) = 0, \text{ and } I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \log f(X_1;\theta)}{\partial\theta^2}\right] = Var_\theta\left(\frac{\partial \log f(X_1;\theta)}{\partial\theta}\right).$$

*Proof.* We begin with differentiating both sides of $1 = \int_{\mathbb{R}} f(x;\theta)dx$.

$$0 = \int_{\mathbb{R}} \frac{\partial}{\partial\theta} f(x;\theta)dx = \int_{\mathbb{R}} \frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)} f(x;\theta)dx = \int_{\mathbb{R}} \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx = \mathbb{E}_\theta\left[\frac{\partial \log f(x;\theta)}{\partial\theta}\right]. \tag{2}$$

Differentiating with respect to $\theta$ again, we get

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}} \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx \\
&= \int_{\mathbb{R}} \frac{\partial^2 \log f(x;\theta)}{\partial\theta^2} f(x;\theta)dx + \int_{\mathbb{R}} \frac{\partial \log f(x;\theta)}{\partial\theta}\frac{\partial f(x;\theta)}{\partial\theta}dx \\
&= \int_{\mathbb{R}} \frac{\partial^2 \log f(x;\theta)}{\partial\theta^2} f(x;\theta)dx + \int_{\mathbb{R}} \frac{\partial \log f(x;\theta)}{\partial\theta}\frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx \qquad (3) \\
&= \mathbb{E}_\theta\left[\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right] + \mathbb{E}_\theta\left[\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2\right],
\end{aligned}$$

where Eq. (3) used the same trick as Eq. (2). By Eq. (1), we have

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right].$$

Finally, since $\mathbb{E}_\theta\left[\frac{\partial \log f(x;\theta)}{\partial\theta}\right] = 0$,

$$\mathrm{Var}\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right) = \mathbb{E}_\theta\left[\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2\right] - \left(\mathbb{E}_\theta\left[\frac{\partial \log f(x;\theta)}{\partial\theta}\right]\right)^2 = I(\theta).$$

Together, we conclude that

$$\mathbb{E}_\theta\left(\frac{\partial \log f(X_1;\theta)}{\partial\theta}\right) = 0, \text{ and } I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \log f(X_1;\theta)}{\partial\theta^2}\right] = \mathrm{Var}_\theta\left(\frac{\partial \log f(X_1;\theta)}{\partial\theta}\right).$$

□

Here we offer one way of interpreting the Fisher information. By (R6) from the regularity conditions, $\hat{\theta}$ is a stationary point of the log-likelihood function. Plugging $\hat{\theta}$ to the Fisher information, we get

$$I(\hat{\theta}) = -\mathbb{E}_{\hat{\theta}}\left[\left.\frac{\partial^2 \log f(X_1;\theta)}{\partial\theta^2}\right|_{\theta=\hat{\theta}}\right] = \mathrm{Var}_{\hat{\theta}}\left(\left.\frac{\partial \log f(X_1;\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}}\right).$$

The first equality says that the Fisher information measures the curvature of the log-likelihood function around the maximum likelihood estimate. Particularly, a large value is associated with a sharp peak around the maximum likelihood estimate, which means that the estimate is very sensitive to changes in the data $\boldsymbol{X}$. Similarly, a small value is associated with a blunt peak around the maximum likelihood estimate, which means that the estimate does not fluctuate much with small changes in the data $\boldsymbol{X}$. This lines up well with the second equality which tells us that the Fisher information at $\hat{\theta}$ measures the variance of the score function at that point.

## 2.2 Asymptotic Normality of MLE

Having established some properties on the score function and the Fisher information, we devote this section to developing the first main result.

**Theorem 2.2.** *Under regularity conditions and following the notations set up in Section 1.1,*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, [I(\theta_0)]^{-1}\right),$$

*where* $I(\theta_0) = \mathbb{E}_{\theta_0}\left[\left(\left.\frac{\partial \log f(X_1;\theta)}{\partial\theta}\right|_{\theta=\theta_0}\right)^2\right]$.

The overall structure of the proof is relatively straightforward. First the mean value theorem is applied to obtain an expression of $\sqrt{n}(\hat{\theta} - \theta_0)$ using a fraction of derivatives of the log-likelihood function $l(\theta)$. The asymptotic properties of the fraction's numerator and denominator are analyzed separately, before being put together to finally get the desired result. However, when studying the asymptotic properties of the fraction, we need to invoke the consistency of MLEs under regularity conditions, which in some sense is more intricate than the proof of Theorem 2.2 itself. We therefore begin with a lemma which will be used to prove the consistency of MLEs. Note that to make clear the dependence of the likelihood functions on the data, we write this dependence explicitly in the lemma below. Let $\boldsymbol{X}$ represent the set of all i.i.d. $X_1, \cdots, X_n$.

**Lemma 2.3.** *Under regularity conditions and following the notations set up in Section 1.1,*

$$\lim_{n\to\infty} P_{\theta_0}\left(L(\theta_0 \mid \boldsymbol{X}) > L(\theta \mid \boldsymbol{X})\right) = 1, \forall \theta \neq \theta_0.$$

*Proof.* We begin by taking the log on both sides of the inequality on the LHS and rearrange.

$$\begin{aligned} L(\theta_0 \mid \boldsymbol{X}) > L(\theta \mid \boldsymbol{X}) &\implies \sum_{i=1}^{n} \log f(X_i;\theta_0) > \sum_{i=1}^{n} \log f(X_i;\theta) \\ &\implies \sum_{i=1}^{n} \left(\log f(X_i;\theta_0) - \log f(X_i;\theta)\right) < 0 \\ &\implies \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) < 0. \end{aligned} \tag{4}$$

Since $X_i$'s are i.i.d. , the summands are independent, and so by the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) \xrightarrow{p} \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right].$$

Note that $-\log(\cdot)$ is strictly convex [1], then by Jensen's inequality, we can establish, on the RHS of the above equation,

$$-\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] = \mathbb{E}_{\theta_0}\left[-\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] > -\log \mathbb{E}_{\theta_0}\left[\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right].$$

Now note

$$\log \mathbb{E}_{\theta_0}\left[\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right] = \log \int_{\mathbb{R}} \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)dx = \log \int_{\mathbb{R}} f(x;\theta)dx = \log 1 = 0. \tag{5}$$

Together, we have

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) \xrightarrow{p} \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] < 0. \tag{6}$$

To show the desired equation, it is equivalent to show, by Eq. (4),

$$\lim_{n\to\infty} P_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) < 0\right) = 1.$$

By Eq. (6), we know that for all $\epsilon > 0$,

$$\lim_{n\to\infty} P_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) - \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right]\right| < \epsilon\right) = 1.$$

[1] Strict convexity implies that there is no more than one minimum, hence the strict inequality sign can be used.

Again by rearranging the inequality inside, we get

$$\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right]-\epsilon<\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right)<\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right]+\epsilon. \tag{7}$$

Note that the probability of event Eq. (7) is less than or equal to that of

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right)<\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right]+\epsilon.$$

Since $\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right]<0$, by fixing $\epsilon=-\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right]>0$, we have

$$\begin{aligned}
&\lim_{n\to\infty}P_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right)-\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right]\right|<\epsilon\right)\\
&\leq\lim_{n\to\infty}P_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right)<\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right]+\epsilon\right)\\
&=\lim_{n\to\infty}P_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right)<0\right)=1.
\end{aligned}$$

Therefore, we conclude that

$$\lim_{n\to\infty}P_{\theta_0}\left(L(\theta_0\mid\boldsymbol{X})>L(\theta\mid\boldsymbol{X})\right)=1,\forall\theta\neq\theta_0.$$

□

The key step in the above proof is obtaining Eq. (6) using Jensen's inequality and the weak law of large numbers. The desired result can then be obtained by manipulating the definition of convergence in probability in Eq. (6). Note Eq. (5) in the above proof requires the use of (R2) from the regularity conditions.

The above lemma says that $\theta_0$ maximizes $L(\theta)$, and therefore $l(\theta)$ as $n$ approaches infinity, under $P_{\theta_0}$. By (R1) of the regularity conditions, it is easy to see that $\theta_0$ is the unique maximizer of $L(\theta)$ and $l(\theta)$ as $n$ approaches infinity, under $P_{\theta_0}$. On the other hand, by (R6) of the regularity conditions, $\hat{\theta}$ uniquely maximizes $l(\theta)$ for all finite $n$. Together with the fact that $\theta_0$ uniquely maximizes $l(\theta)$ in probability, it makes intuitive sense to draw the conclusion that $\lim_{n\to\infty}P_{\theta_0}\left(|\hat{\theta}-\theta_0|<\epsilon\right)=1$, which precisely means that $\hat{\theta}$ is a consistent estimator of $\theta_0$.

The next lemma formalizes this statement. Instead of proving the result directly, using (R3) from the regularity conditions, an arbitrary sequence of solutions $\left(\bar{\theta}_n\right)_{n\geq1}$ to $\frac{\partial l(\theta|\boldsymbol{X})}{\partial\theta}=0$ are first found. Then by (R6) from the regularity conditions, we identify that $\left(\bar{\theta}_n\right)_{n\geq1}$ are indeed the sequence of MLEs. The key observation in the proof below is the existence of a local maximum in the neighbourhood of $\theta_0$.

**Lemma 2.4.** *Under regularity conditions and following the notations set up in Section 1.1, $\hat{\theta}$ is a consistent estimator of $\theta_0$.*

*Proof.* We begin by showing that the equation $\frac{\partial l(\theta|\boldsymbol{X})}{\partial\theta}=0$ has a solution $\bar{\theta}_n$ that converges in probability to $\theta_0$.

Since $\theta_0$ is an iterior point of $\Theta$, we can find $a>0$ such that $\theta_0\in(\theta_0-a,\theta_0+a)\subset\Theta$. Then define the event

$$S_n=\left\{\boldsymbol{X}:l(\theta_0\mid\boldsymbol{X})>l(\theta_0-a\mid\boldsymbol{X})\cap l(\theta_0\mid\boldsymbol{X})>l(\theta_0+a\mid\boldsymbol{X})\right\}.$$

Lemma 2.3 says, under $P_{\theta_0}$, when $n$ approaches infinity, $\theta_0$ is the unique maximizer to $L(\theta \mid \boldsymbol{X})$, which implies that it is also the unique maximizer to $l(\theta \mid \boldsymbol{X})$. Then clearly $\lim_{n\to\infty} P_{\theta_0}(S_n) = 1$.

Note since $l(\theta_0 \mid \boldsymbol{X}) > l(\theta_0 - a \mid \boldsymbol{X})$ and $l(\theta_0 \mid \boldsymbol{X}) > l(\theta_0 + a \mid \boldsymbol{X})$, with $f(\theta; \boldsymbol{X})$ being continous and differentiable, there must exists, for any $\boldsymbol{X} \in S_n$, a local maximum in $(\theta_0 - a, \theta_0 + a)$. Denote this value $\bar{\theta}_n$, then $\frac{\partial l(\theta|\boldsymbol{X})}{\partial \theta}\Big|_{\theta=\bar{\theta}_n} = 0$.

Then for all $a > 0$ small enough, we can find a sequence of $\hat{\theta}$ such that

$$\lim_{n\to\infty} P_{\theta_0}\left(|\bar{\theta}_n - \theta_0| < a\right) = 1.$$

By choosing $\bar{\theta}_n$ to be the one closest to $\theta_0$, denoted $\theta_n^*$, we have identified a sequence $(\theta_n^*)_{n\geq 1}$, independent of $a$, such that

$$\lim_{n\to\infty} P_{\theta_0}\left(|\theta_n^* - \theta_0| < a\right) = 1, \forall a > 0.$$

This precisely means that $\theta *_n \xrightarrow{p} \theta_0$.

Now note that under regularity conditions, $\hat{\theta}$ is the unique solution to $\frac{\partial l(\theta|\boldsymbol{X})}{\partial \theta} = 0$. Therefore, we conclude that $\hat{\theta}$ is a consistent estimator of $\theta_0$. □

Now that we have established the consistency of MLEs, we are equipped with all of the necessary tools to prove Theorem 2.2.

*Proof.* By the mean value theorem, for $f : \mathbb{R} \to \mathbb{R}$ continuous on $[a, b]$ and differentiable on $(a, b)$, for all $c \in (a, b)$, $\frac{f(a)-f(b)}{a-b} = f'(c)$. Let $f(\theta) = l'(\theta) = \frac{\partial l(\theta)}{\partial \theta}, a = \hat{\theta}, b = \theta_0, c = \theta_1 \in (\theta_0, \hat{\theta})$. Then with $l''(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta^2}$, the above equation becomes $\frac{l'(\hat{\theta})-l'(\theta_0)}{\hat{\theta}-\theta_0} = l''(\theta_1)$. We know $l'(\hat{\theta}) = 0$, then the equation above becomes

$$0 = l'(\theta_0) + (\hat{\theta} - \theta_0)l''(\theta_1) \implies \sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\sqrt{n}l'(\theta_0)}{l''(\theta_1)}. \tag{8}$$

We first look at the denominator of Eq. (8). By Lemma 2.4, $\hat{\theta} \xrightarrow{p} \theta_0$. Then since $\theta_1 \in (\theta_0, \hat{\theta})$, we must have $\theta_1 \xrightarrow{p} \theta_0$. Then by Proposition 10.7 from the lecture notes,

$$l''(\theta_1) \xrightarrow{p} l''(\theta_0).$$

Now by the weak law of large numbers and Lemma 2.1,

$$l''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(x_i; \theta)\Big|_{\theta=\theta_0} \xrightarrow{p} \mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial \theta^2} \log f(X_1; \theta)\Big|_{\theta=\theta_0}\right] = -I(\theta_0). \tag{9}$$

Now we look at the numerator of Eq. (8). By Lemma 2.1, $\mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial \theta} \log f(X_1; \theta)\big|_{\theta=\theta_0}\right] = 0$ and $I(\theta_0) = \mathrm{Var}\left(\frac{\partial}{\partial \theta} \log f(X_1; \theta)\big|_{\theta=\theta_0}\right)$. Then by the central limit theorem,

$$\begin{aligned}
\sqrt{n}l'(\theta_0) &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i; \theta)\Big|_{\theta=\theta_0} \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i; \theta)\Big|_{\theta=\theta_0} - \mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial \theta} \log f(X_1; \theta)\Big|_{\theta=\theta_0}\right]\right) \\
&\xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}\left(\frac{\partial}{\partial \theta} \log f(X_1; \theta)\Big|_{\theta=\theta_0}\right)\right) \\
&= \mathcal{N}\left(0, I(\theta_0)\right).
\end{aligned} \tag{10}$$

Together by Eqs. (8) to (10) and Slutsky's Theorem,

$$\sqrt{n}(\hat{\theta}-\theta_0) = -\frac{\sqrt{n}l'(\theta_0)}{l''(\theta_1)} \xrightarrow{d} \frac{\mathcal{N}(0, I(\theta_0))}{I(\theta_0)} = \mathcal{N}\left(0, [I(\theta_0)]^{-1}\right).$$

Therefore we conclude that $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} \mathcal{N}\left(0, [I(\theta_0)]^{-1}\right)$. □

Following the outline of the proof's structure at the beginning of the subsection, the proof above should be relatively easy to follow. Again, the mean value theorem is first applied to obtain a fraction as an expression for $\sqrt{n}(\hat{\theta}-\theta_0)$. Using results from Lemmas 2.1 and 2.4, together with the weak law of large numbers, central limit theorem (both introduced in class) and Slutsky's theorem (proof shown in Appendix B), we obtain the final convergence result. Note that the use of consistency of MLE is to show that $l''(\theta_1)$ converges in probability to $-I(\theta_0)$.

Using our previously developed interpretation on the Fisher information, we get that, when the MLE is sensitive to changes in the data, we get a large Fisher information, which implies by Theorem 2.2 that the estimator has a small asymptotic variance. This again aligns with our intuition. Particularly, a large Fisher information implies that the curvature around the maximum of the log-likelihood function is sharp, which implies that there is less uncertainty around the estimator of $\theta_0$.

Instead of the mean value theorem, some versions of the above proof start with a Taylor expansion of $l'(\theta)$ at $\theta_0$, where consistency of MLE is used to show that the higher order terms are bounded in probability. While they may be theoretically more rigorous, they are less straightforward and harder to follow. Therefore, we have chosen the above version of the proof to be presented.

We now verify Theorem 2.2 through some simulations. From Theorem 2.2, we have that $\hat{\theta} \xrightarrow{d} \mathcal{N}(\theta_0, [nI(\theta_0)]^{-1})$. We generate 1000 trials of $n = 10, 40, 70, 100$ i.i.d. observations from Gamma($\alpha = 2, \theta = 2$), using the shape and scale parameterization. Simple calculation yields $\hat{\theta} = \frac{1}{\bar{x}}$, and $I(\theta_0) = \frac{1}{\theta_0^2}$. The histograms of the MLEs of $\hat{\theta}$ given $\alpha$ are plotted for each sample size, which is compared to their asymptotic distribution, shown as red curves in the histogram (Fig. 1). The code can be found at `src/simulation_normality.R`. `RStudio 1.3.959` is used to generate the plots below. It is worth noting that as the sample size increases, the variance of the MLE decreases, which aligns with Theorem 2.2.

The *asymptotic normality of MLEs* enables us to study the estimators more closely. As mentioned before, one of the immediate applications is that we can quantify the uncertainties around the MLEs. For examle, we can construct approximate confidence intervals of the unknown parameters using the fact that the MLEs are asymptotically normal.

## 2.3 Asymptotic Efficiency of MLE

Having established Theorem 2.2, it is natural to shift our attention to the quality of MLEs. The next result (*Cramer-Rao lowerbound*) gives us some insight to the quality of MLEs compared to other estimators in terms of their variances. Note that by unbiased estimator, we mean estimators whose expectation is the true underlying parameter that is being estimated.

**Theorem 2.5.** *Let $Y = u(X_1, \cdots, X_n)$ be an unbiased estimator of $\theta_0$ such that $\mathbb{E}_{\theta_0}[Y] = \theta_0$. Then under regularity conditions, $Var(Y) \geq \frac{1}{nI(\theta_0)}$.*

*Proof.* We first expand $E_{\theta_0}[Y]$.

$$\theta_0 = E_{\theta_0}[Y] = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} u(x_1, \cdots, x_n) \prod_{i=1}^{n} f(x_i; \theta_0) dx_1 \cdots dx_n.$$

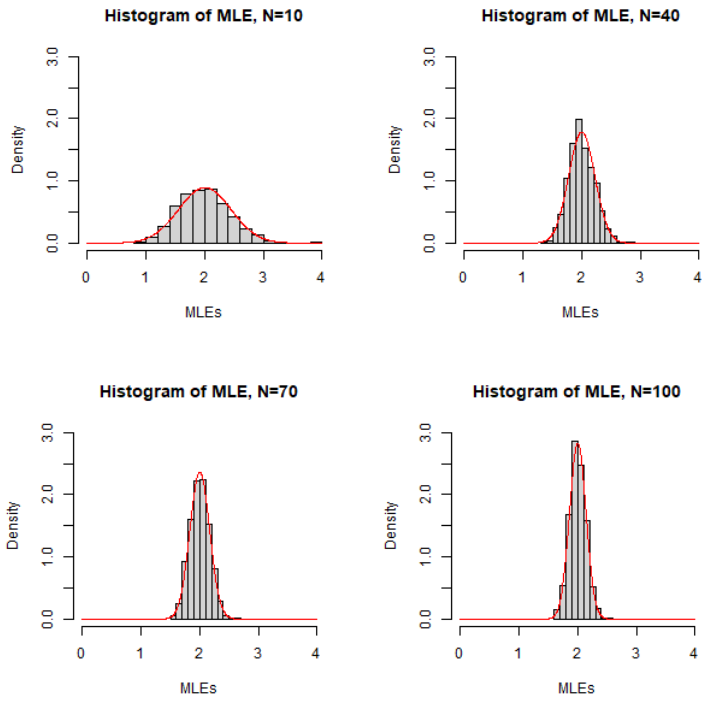


Figure 1: Histograms of MLEs of varying sample sizes

Differentiating both sides with respect to $\theta_0$ gives

$$\begin{aligned} 1 &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} u(x_1, \cdots, x_n) \left( \sum_{i=1}^{n} \frac{\partial f(x_i;\theta_0)}{\partial \theta_0} \frac{1}{f(x_i;\theta_0)} \right) \prod_{i=1}^{n} f(x_i;\theta_0) dx_1 \cdots dx_n \\ &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} u(x_1, \cdots, x_n) \left( \sum_{i=1}^{n} \frac{\partial \log f(x_i;\theta_0)}{\partial \theta_0} \right) \prod_{i=1}^{n} f(x_i;\theta_0) dx_1 \cdots dx_n \end{aligned}$$

By writing $Z = \sum_{i=1}^{n} \frac{\partial \log f(X_i;\theta_0)}{\partial \theta_0}$, we get, with $\rho$ denoting the correlation coefficient between $Y$ and $Z$,

$$1 = \mathbb{E}_{\theta_0}[YZ] = \mathbb{E}_{\theta_0}[Y]\mathbb{E}_{\theta_0}[Z] + \rho\sqrt{\mathrm{Var}(Y)}\sqrt{\mathrm{Var}(Z)} \implies \rho = \frac{1}{\sqrt{\mathrm{Var}(Y)}\sqrt{\mathrm{Var}(Z)}}.$$

We note that since $X_i$'s are i.i.d. ,

$$\mathrm{Var}(Z) = \mathrm{Var}\left( \sum_{i=1}^{n} \frac{\partial \log f(X_i;\theta_0)}{\partial \theta_0} \right) = \sum_{i=1}^{n} \mathrm{Var}\left( \frac{\partial \log f(X_i;\theta_0)}{\partial \theta_0} \right) = n\mathrm{Var}\left( \frac{\partial \log f(X_1;\theta_0)}{\partial \theta_0} \right) = nI(\theta_0).$$

Then $\rho = \frac{1}{\sqrt{nI(\theta_0)}\sqrt{\mathrm{Var}(Y)}}$. By definition, $rho^2 \leq 1$, then

$$\rho^2 = \frac{1}{nI(\theta_0)\mathrm{Var}(Y)} \leq 1 \implies \mathrm{Var}(Y) \geq \frac{1}{nI(\theta_0)}.$$

Therefore for any unbiased estimator $Y$ of $\theta_0$, we have $\mathrm{Var}(Y) \geq \frac{1}{nI(\theta_0)}$. □

The proof itself is relatively straightforward, where we essentially reach the desired conclusion by using

(R5) of the regularity conditions and simple algebraic manipulations. It is, however, the implication of this result that is worth looking at. Note the lower bound on the variance is exactly the asymptotic variance of MLEs. This means that asymptotically, the MLE achieves the smallest possible variance out of all unbiased estimators of $\theta$. Therefore the MLEs are said to be *asymptotically efficient*.

While this is a nice property, it remains rather theoretical. In practice, when we work with finitely many observations, it is not clear how far away we are from the asymptotic normality. And existing heuristics from undergraduate statistics courses such as calling samples larger than 25 or 30 large enough is far from satisfying. Fortunately, [AL15] have done some pioneer work aimed at answering this exact question. We turn in the next section for a brief illustration of their result on the the closeness from the MLE approximated using finitely many observations to the asymptotic normal distribution.

# 3 Quantifying the Distance between MLE to its Asymptotic Distribution

To describe the distance between two distributions, or two random variables following two different distributions, a *Zolotarev-type distance*, as defined in [Zol76], can be used, as it is a natural extension of a metric from non-probabilistic spaces to probabilistic ones. One of the simplist of such distances is

$$D(X,Y) = \sup_{h\in H} \left|\mathbb{E}\left[h(X)\right] - \mathbb{E}\left[h(X)\right]\right|, \qquad H \text{ is the set of continous and bounded real functions,} \tag{11}$$

for it satisfies the following probabilistic versions of properties of a metric:

- $\mathbb{P}(X = Y) \implies D(X,Y) = 0$,
- $D(X,Y) = D(Y,X)$,
- $D(X,Y) \leq D(X,Z) + D(Z,Y)$, for $Z \neq X, Z \neq Y$.

By Theorem 2.2, we have that $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1)$. Then with $Y = \sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ and $Z = \mathcal{N}(0,1)$, $D(Y,Z)$ describes the closeness from the MLE approximated using finitely many observations to the asymptotic normal distribution. [AL15] developed a bound on exactly the quantity $D(Y,Z)$ in the one-dimensional case, for random variables $X$ where there exists a one-to-one twice differentiable mapping $q : \Theta \to \mathbb{R}$ such that $q(\hat{\theta}) = \sum_{i=1}^n g(X_i)$ for some $g : \mathbb{R} \to \mathbb{R}$. The main theorem is presented below.

**Theorem 3.1.** *Under regularity conditions and following the notations set up in Section 1.1 and the beginning of this section, let $q : \Theta \to \mathbb{R}$ be a one-to-one twice differentiable function with $q'(\theta) \neq 0 \forall \theta \in \Theta$ and such that $q\left(\hat{\theta}_n(\boldsymbol{X})\right) = \frac{1}{n}\sum_{i=1}^n g(X_i)$, where the mapping $g : \mathbb{R} \to \mathbb{R}$ is such that $\mathbb{E}\left[|g(X_1) - q(\theta_0)|^3\right] < \infty$ for the true parameter $\theta_0$. Also, there exists a positive constant $0 < \epsilon = \epsilon(\theta_0)$ with $(\theta_0 - \epsilon, \theta_0 + \epsilon) \subset \Theta$. Then for any $h \in H$, we have*

$$\begin{aligned}
&\left|\mathbb{E}\left[h\left(\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)\right)\right] - \mathbb{E}\left[h(Z)\right]\right| \\
&\leq \frac{\|h'\|_\infty}{\sqrt{n}}\left(2 + \frac{[I(\theta_0)]^{3/2}}{|q'(\theta_0)|^3}\mathbb{E}\left[|g(X_1) - q(\theta_0)|^3\right]\right) \\
&+ \mathbb{E}\left[(\hat{\theta}(\boldsymbol{X}) - \theta_0)^2\right]\left(2\frac{\|h\|_\infty}{\epsilon^2}\mathbf{1}\{\exists \theta \in \Theta : q(\theta) \neq \theta\} + \frac{\|h'\|_\infty\sqrt{nI(\theta_0)}}{2|q'(\theta_0)|} sup_{\theta:|\theta-\theta_0|\leq\epsilon}|q''(\theta)|\right).
\end{aligned}$$

Note that for any $h \in H$, the above bound gives a lowerbound of $D(X,Y)$, since *Zolotarev-type distances* take the supremum over all such functions. Compared to Theorems 2.2 and 2.5, the above result provides a quantitative measure on the distance between the unknown distribution of MLE based on finitely many observations and its corresponding asymptotic distribution. The proof of Theorem 3.1 is similar to that of

Theorem 2.2 in that they both start with obtaining an expression of the quantity of interest followed by analysis of each of the terms in the expression. In this particular case, the triangle inequality is first applied on $D(Y, Z)$, and each of the two terms can then be bounded using a combination of the delta method, Stein's method, and Taylor expansions.

To verify the above theorem, we perform another set of simulations, comparing the bound on the MLE of Gamma($\alpha = 2, \theta = 2$) given $\alpha$. In particular, we generate 10000 trials of $n = 10, 100, 1000, 10000, 100000$ i.i.d. observations from Gamma($\alpha = 2, \theta = 2$), using the shape and scale parameterization. With the choice of $h(x) = \frac{1}{x^2+2}$, we have $\|h\|_\infty = 0.5, \|h'\|_\infty = 3\sqrt{1.5}/16, \mathbb{E}[h(Z)] = 0.379$. Then we can compute a Monte-Carlo approximation of $D(Y, Z)$ by averaging over all 10000 trials of $h\left(\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)\right)$. This is then compared to the bound in Theorem 3.1, under the more specific case as laid out in Section 3.2 of [AL15]. (Fig. 2). The code can be found at `src/simulation_bound.R`. `RStudio 1.3.959` is used to generate the plots below.

Note that since there is only one parameter to be estimated, the MLEs are very close to the true value even with a not-so-large sample size, and so the black line on the plot may not always be monotonically decreasing. We do note, however, that the bound developed in Theorem 3.1 becomes tighter as the sample size increases.

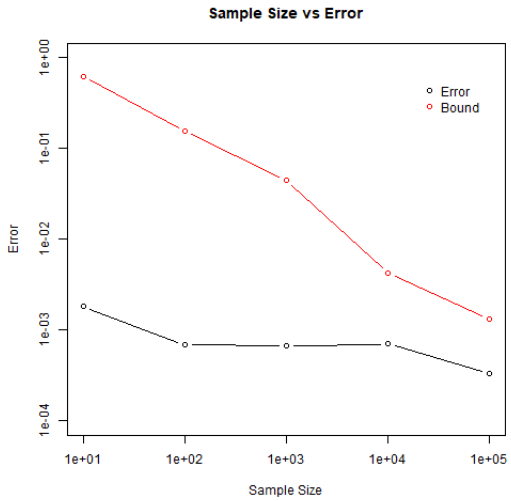


Figure 2: Sample size vs. Zolotarev distance

There are many directions of future research on the topic. As an example, it would be interesting to see if we can transform the bound in Theorem 3.1 to some inequalities on the approximate confidence intervals of $\theta_0$, either deterministically or probabilistically. As another example, we note that [AL15] only discussed the one-dimensional case, and as seen in both of our simulations, the MLE gets quite close to its asymptotic distribution even with a not-so-large sample size. It would be a natural next step to extend the bound in Theorem 3.1 to multi-dimensional settings. We offer some initial ideas on this extension.

Glancing through the proof in [AL15], many of the results applied in the process of achieving the final result has multivariate versions. Particularly, all of the triangle inequality, delta method, and Taylor expansions can be applied directly in the multivariate setting. The bottleneck then becomes developing multivariate versions of two of the lemmas used in the proof. Namely, Lemma 2.1 in [AL15] and Lemma 2.1 in [AR+17].

# A Exercises

Prove Lemma 1 and Theorem 2 using a different method. Prove efficiency for finite parameter space.

# B Proof of Slutsky's Theorem

We begin by noting that the following proofs follow closely the lecture notes from STAT 560 by Professor Ruben Zamar.

**Theorem B.1.** *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $X_n Y_n \xrightarrow{d} cX$.*

*Proof.* We begin by writing $X_n Y_n = X_n(Y_n - c) + cX_n$. Then, in order to achieve the final result, we need to show the following

- If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $X_n + Y_n \xrightarrow{d} X + c$.
- If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} 0$, then $X_n Y_n \xrightarrow{p} 0$.
- If $X_n \xrightarrow{d} X$, then $cX_n \xrightarrow{d} cX$.

If we have the above results, then

$$Y_n - c \xrightarrow{p} 0 \implies X_n(Y_n - c) \xrightarrow{p} 0 \implies \implies X_n(Y_n - c) \xrightarrow{d} 0 \implies X_n Y_n = X_n(Y_n - c) + cX_n \xrightarrow{d} 0 + cX = cX.$$

□

**Lemma B.2.** *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $X_n + Y_n \xrightarrow{d} X + c$.*

*Proof.* Let $t$ be a continuity point of $F_{X+c}$. Then, $t - c$ is a continuity point of $F_X$. Find $\epsilon > 0$ such that $t - c - \epsilon, t - c + \epsilon$ are both continuity points of $F_X$. Then

$$\begin{aligned} F_{X_n+Y_n}(t) &= P(X_n + Y_n \le t) \\ &= P(X_n + Y_n \le t, |Y_n - c| < \epsilon) + P(X_n + Y_n \le t, |Y_n - c| \ge \epsilon) \\ &\le P(X_n \le t - Y_n, c - \epsilon < Y_n < c + \epsilon) + P(|Y_n - c| \ge \epsilon) \\ &\le P(X_n \le t - c + \epsilon) + P(|Y_n - c| \ge \epsilon). \end{aligned}$$

$$\begin{aligned} F_{X_n}(t - c - \epsilon) &= P(X_n \le t - c - \epsilon) \\ &= P(X_n \le t - c - \epsilon, |Y_n - c| < \epsilon) + P(X_n \le t - c - \epsilon, |Y_n - c| \ge \epsilon) \\ &\le P(X_n \le t - c - \epsilon, c - \epsilon < Y_n < c + \epsilon) + P(|Y_n - c| \ge \epsilon) \\ &\le P(X_n + Y_n \le t) + P(|Y_n - c| \ge \epsilon). \end{aligned}$$

Therefore,

$$\begin{aligned} &\limsup_n F_{X_n+Y_n}(t) \le \limsup_n P(X_n \le t - c + \epsilon) + \limsup_n P(|Y_n - c| \ge \epsilon) = F_X(t - c + \epsilon). \\ &\liminf_n F_{X_n}(t - c - \epsilon) \le \liminf_n F_{X_n+Y_n}(t) + \liminf_n P(|Y_n - c| \ge \epsilon). \end{aligned}$$

And so

$$\begin{aligned} \limsup_n F_{X_n+Y_n}(t) &\le F_X(t - c + \epsilon). \\ F_X(t - c - \epsilon) &\le \liminf_n F_{X_n+Y_n}(t). \end{aligned}$$

Together, we have that

$$F_X(t - c - \epsilon) \le \liminf_n F_{X_n+Y_n}(t) \le \limsup_n F_{X_n+Y_n}(t) \le F_X(t - c + \epsilon).$$

Then

$$\lim_{n\to\infty} F_{X_n+Y_n}(t) = F_X(t-c) = F_{X+c}(t).$$

Therefore, if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $X_n + Y_n \xrightarrow{d} X + c$. □

**Lemma B.3.** *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} 0$, then $X_nY_n \xrightarrow{p} 0$.*

*Proof.* By definition, for any $\epsilon > 0$, given $\delta > 0$, we can find $M$ such that $\forall n \geq M$, $X_nY_n \xrightarrow{p} 0 \implies P(|X_nY_n| < \epsilon) > 1 - \delta$.

First, we find $K$ such that $K, -K$ are continuity points of $F_X$, the CDF of $X$, and

$$P(|X| \leq K) = P(-K \leq X \leq K) = F_X(K) - F_X(-K) \geq 1 - \frac{\delta}{4}.$$

Then

$$\begin{aligned} P(|X_n| \leq K) &= P(-K \leq X_n \leq K) \\ &= F_{X_n}(K) - F_{X_n}(-K) \\ &= F_X(K) - F_X(-K) + (F_{X_n}(K) - F_X(K)) - (F_{X_n}(-K) - F_X(-K)) \\ &= F_X(K) - F_X(-K) - (F_X(K) - F_{X_n}(K)) - (F_{X_n}(-K) - F_X(-K)) \end{aligned}$$

Since $\lim_{n\to\infty} F_{X_n}(K) = F_X(K)$, $\lim_{n\to\infty} F_{X_n}(-K) = F_X(-K)$, we can find $N_1, N_2$ such that

$$|F_{X_n}(K) - F_X(K)| < \frac{\delta}{8} \forall n \geq N_1, |F_{X_n}(-K) - F_X(-K)| < \frac{\delta}{8} \forall n \geq N_2.$$

Take $N = \max\{N_1, N_2\}$, then for all $n \geq N$,

$$\begin{aligned} P(|X_n| \leq K) &\geq F_X(K) - F_X(-K) - |F_{X_n}(K) - F_X(K)| - |F_{X_n}(-K) - F_X(-K)| \\ &\geq F_X(K) - F_X(-K) - 2\frac{\delta}{8} \\ &\geq 1 - \frac{\delta}{4} - \frac{\delta}{4} = 1 - \frac{\delta}{2}. \end{aligned}$$

Then $P(|X_n| > K) \leq \frac{\delta}{2}, \forall n \geq N$. Since $Y_n \xrightarrow{p} 0$, there exists $M \geq N$ such that

$$P(|Y_n| < \frac{\epsilon}{K}) \geq 1 - \frac{\delta}{2}, \forall n \geq M.$$

Then for all $n \geq M$,

$$\begin{aligned} P(|X_nY_n| > \epsilon) &= P(|X_nY_n| > \epsilon, |X_n| \leq K) + P(|X_nY_n| > \epsilon, |X_n| > K) \\ &\leq P(|Y_n| \geq \frac{\epsilon}{K}) + P(|X_n| > K) \\ &\leq \frac{\delta}{2} + \frac{\delta}{2} = \delta. \end{aligned}$$

Together, we conclude that if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} 0$, then $X_nY_n \xrightarrow{p} 0$. □

**Lemma B.4.** *If $X_n \xrightarrow{d} X$, then $cX_n \xrightarrow{d} cX$.*

*Proof.* Since $X_n \xrightarrow{d} X$, we have, for any continuity point $t$ of $F_X$ and $c > 0$,

$$\lim_{n\to\infty} P(X_n < t) = P(X < t) \implies \lim_{n\to\infty} P(cX_n < ct) = P(cX < ct).$$

When $c < 0$,

$$\begin{aligned}\lim_{n\to\infty} P(X_n < t) = P(X < t) &\implies \lim_{n\to\infty} P(cX_n > ct) = P(cX > ct) \\ &\implies \lim_{n\to\infty} 1 - P(cX_n > ct) = 1 - P(cX > ct) \\ &\implies \lim_{n\to\infty} P(cX_n < ct) = P(cX < ct).\end{aligned}$$

The case where $c = 0$ is covered in Lemma B.3. Together, we have that if $X_n \xrightarrow{d} X$, then $cX_n \xrightarrow{d} cX$. □

# References

[AL15] A. Anastasiou and C. Ley. "Bounds for the asymptotic normality of the maximum likelihood estimator using the Delta method". In: *arXiv preprint arXiv:1508.04948* (2015).

[AR+17] A. Anastasiou, G. Reinert, et al. "Bounds for the normal approximation of the maximum likelihood estimator". In: *Bernoulli* 23.1 (2017), pp. 191–218.

[HMC05] R. V. Hogg, J. McKean, and A. T. Craig. *Introduction to mathematical statistics*. Pearson Education, 2005. Chap. 6.

[Zol76] V. M. Zolotarev. "Metric distances in spaces of random variables and their distributions". In: *Mathematics of the USSR-Sbornik* 30.3 (1976), p. 373.