university of groningen / faculty of science and engineering / mathematics and applied mathematics

Relation between empirical processes and Z/M-estimation

Bachelor's Project Mathematics

Juli 2021

Student: Y.C. Dijkstra

First supervisor: dr. W.P. Krijnen

Second assessor: dr. M.A. Grzegorczyk

Abstract

The goal of this paper is to investigate the asymptotic properties of Z- and M-estimators in the empirical process theory. The paper will demonstrate that the empirical process theory is a very important tool in asymptotic statistics. Applications of the empirical process theory will be displayed accompanied by some examples to serve as illustration. The applications will be used to approach Z- and M-estimation in a non-parametric way. In practice, non-parametric regression is only called upon when other attempts have failed. It will be seen that this non-parametric approach to regression analysis must be taken into account and should be drawn upon a lot more in the future.

Contents

Introduction

In statistics, there are often assumptions when estimating the relationships between a dependent variable and one or more independent variables. The most common form of this kind of regression analysis is normal regression, but it is also possible to assume that the observations follow from a linear regression. This type of regression analysis, where the predictor takes a predetermined form, is called parametric regression. The opposite, less-used regression analysis is called non-parametric regression, which constructs the predictor according to the information from the observations. It is primarily used when the assumptions of the parametric tests are violated. This paper will give a quick peek in the applications of non-parametric regression and show that this type of regression ought to be used a lot more in the future.

Nowadays, the empirical process theory represents a significant part in non-parametric asymptotic statistics. This theory originated from the study of goodness-of-fit statistics. An example of a goodness-of-fit test, is the Kolmogorov–Smirnov test which quantifies the distance between the null distribution function and the so-called empirical distribution function. With introducing the empirical distribution function, the empirical process theory became more and more important. With applications like the bootstrap, the delta-method and goodness-of-fit testing, the theory plays a key part in asymptotic statistics.

Now that the theory has shown its importance, the question arises what effect this theory has on the asymptotic properties of Z- and M-estimators. Which conditions are needed in order to satisfy uniform consistency for the least squares estimator? And what is the convergence rate of this estimator? This is being investigated by analyzing entropy conditions and researching if these entropy conditions are enough to satisfy the Uniform Law of Large Numbers. To examine the relation between the empirical process theory and Z-estimation, conditions for P-Donsker classes need to be established. What are these conditions and how are they related to asymptotic properties of Z-estimators?

Literature review

It wasn't untill the advent of goodness-of-fit tests that the empirical process theory began taking shape. Before this, there was virtually no literature about empirical processes. At the same time (1930's), the Glivenko-Cantelli theorem was discovered which introduced the empirical distribution function. Later, the consideration of the empirical measure resulted in generalization of the theorem for function classes. These classes are called Glivenko-Cantelli classes. In 1952, the discovery of the Donsker's theorem accounted for the emergence of a whole new branche of the empirical process theory. It gave rise to P-Donsker classes. In the 1970's and 1980's there were many studies into P-Donsker classes and empirical processes and its applications by David Pollard, Evarist Giné, Joel Zinn and many others. These studies have laid the foundation for further studies like van der Vaart and Wellner (1996) [2] and van de Geer (2000) [9].

The book from van de Geer [9] presents the empirical process theory and its applications with displaying the effectiveness of the theory in non-parametric models. The main goal of the book is to demonstrate the relation between empirical processes and the asymptotic properties of M-estimation. The book starts with explaining the entropy of a function class and proofs the property of finite entropy, is a necessary condition of a Glivenko-Cantelli class. The book differs here from van der Vaart and Wellner [2]. Both proofs apply Hoeffding's inequality and the symmetrization method, however, van der Vaart and Wellner use the Hoeffding's inequality with respect to an other norm and the symmetrization method was used with respect to the expectation instead of the probability. Another difference is that van de Geer covers P-Donsker classes but doesn't address Z-estimation, as van der Vaart and Wellner do. Van der Vaart and Wellner demonstrate the applications of P-Donsker classes in Z-estimation. Nonetheless, consistency and the convergence rates of M-estimators are covered by both books.

Bodhisattva Sen [15] has made a document about the empirical processes theory that is easier understood than other books about this topic. The proofs for consistency of M-estimators and ensuring Glivenko-Cantelli are approached from another point of view.

Empirical process

This chapter will introduce the empirical process, which is the main topic of this paper. We start with considering i.i.d random variables $X_1, X_2, ..., X_n$ defined on $\mathcal{X}$ with a common cumulative distribution function F. Before we can obtain the empirical process, the empirical distribution function has to be defined. The empirical distribution function is given by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(X_i)$$

where $\mathbb{1}$ is the indicator function. In other words, the empirical distribution function at a given point is equal to the proportion of observations that are less than or equal to that point. If we fix x, $\mathbb{1}_(x)X$ can be seen as a random variable with sample mean $F_n(X)$ and expectation $F(x)$. Recalling the Strong Law of Large Numbers, we have

$$P\left(\lim_{n\to\infty} F_n(x) = F(x)\right) = 1$$

for each $x \in \mathcal{X}$. Hence, $F_n(x)$ converges almost sure to F. In chapter 'Glivenko-Cantelli classes', it will be seen that this convergence is uniform (see theorem (1.1)). Now we know the definition of the empirical distribution function, we can construct the corresponding empirical process. An empirical process indexed by a particular function class $\mathcal{F}$, is given by

$$(\mathbb{G}_n(f))_{f\in\mathcal{F}} = \left\{ \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(x) \right) \ : \ f \in \mathcal{F} \right\}$$

If we take $\mathcal{F} = \{\mathbb{1}_{(-\infty,x]} : x \in \mathbb{R}\}$, then

$$(\mathbb{G}_n(f))_{f\in\mathcal{F}} = (\sqrt{n}(F_n(x) - F(x)))_{x\in\mathbb{R}}$$

By the Central Limit theorem, all elements of this empirical process converge in distribution to a normal random variable.

$$\mathbb{G}_n(x) \xrightarrow{d} N(0, F(x)(1 - F(x))$$

We will discover later that this empirical process converges in distribution to a standard Brownian motion (see Donsker's theorem (3.1)).

Empirical measure

Let $A \subset \mathcal{X}$, then the empirical measure is given as

$$\mathbb{P}_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(X_i)$$

which is just $1/n$ times the number of variables X_i in A. Note that all indicator functions can be written as Dirac measures, so

$$\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(A)$$

Therefore if we let $\mathcal{F}$ be a class of functions defined on $\mathcal{X}$, then for $f \in \mathcal{F}$ we can write

$$\int f \; d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

Recalling the definition of the expectation $\mathbb{E}$, we write

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f = \int f \; d(\mathbb{P}_n - \mathbb{P}) \tag{0.1}$$

With this equation we can write the Uniform Law of Large Numbers in an other way. This trick comes in handy for proofing theorems for certain classes. We say a function class $\mathcal{F}$ satisfies the Uniform Law of Large Numbers if

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f \right| \xrightarrow{a.s.} 0$$

But now by equation (0.1), we can write

$$\sup_{f\in\mathcal{F}} \left| \int f \; d(\mathbb{P}_n - \mathbb{P}) \right| \xrightarrow{a.s.} 0$$

A class satisfying this condition is called a Glivenko-Cantelli class (see definition (1.1)). From earlier understandings, we know that the function class $\{\mathbb{1}_{(-\infty,x]} : x \in \mathcal{X}\}$ can be called Glivenko-Cantelli. But what about other function classes? That will be figured out in the next chapter.

Glivenko-Cantelli classes

This chapter will be about identifying Glivenko-Cantelli classes. Particular function classes will be proven to be Glivenko-Cantelli, but first we need to know what a Glivenko-Cantelli class is.

Definition 1.1 *(Glivenko-Cantelli class) A function class $\mathcal{F}$ is called Glivenko-Cantelli if*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} \left| \int f \; d(\mathbb{P}_n - \mathbb{P}) \right| \xrightarrow{a.s.} 0 \tag{1.2}$$

i.e. it satisfies the Uniform Law of Large Numbers.

Later on, we will use lemma 2.4.5 from [2] to show that condition (1.2) is equivalent to $\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \to 0$ (convergence in mean). This lemma will help us to construct one of the two theorems which proof Glivenko-Cantelli for function classes with certain entropy conditions. Before we will demonstrate these theorems, we will proof that the class $\{\mathbb{1}_{(-\infty,x]} : x \in \mathcal{X}\}$ is Glivenko-Cantelli. To proof this, we will make use of the following theorem.

Theorem 1.1 *(Glivenko-Cantelli theorem)*

$$\|F_n - F\|_\infty = \sup_{x\in\mathcal{X}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

Proof See theorem (A1.1) from Appendix

To demonstrate that the function class $\{\mathbb{1}_{(-\infty,x]} : x \in \mathcal{X}\}$ is Glivenko-Cantelli, note that

$$F_n(x) = \int \mathbb{1}_{(-\infty,x]} \; d\mathbb{P}_n$$

and

$$F(x) = \int \mathbb{1}_{(-\infty,x]} \; d\mathbb{P}$$

Therefore,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\{\mathbb{1}_{(-\infty,x]} : x\in\mathcal{X}\}} = \sup_{x\in\mathcal{X}} |F_n(x) - F(x)|$$

Using theorem (1.1), we find that this class is Glivenko-Cantelli.

Entropy

It is already mentioned that two theorems will be demonstrated based on entropy conditions. To obtain the entropy of a function class $\mathcal{F}$ we first need to construct a ϵ-net. A ϵ-net is defined on a metric space, so we consider for $1 \le p \le \infty$ the lebesgue space $L_p(\mu) = \{f : \mathcal{X} \to \mathbb{R} : \int |f|^p d\mu < \infty\}$ with respect to the measure

μ. The p-norm on $L_p(\mu)$ will be referred to as $\|.\|_p$. Let $\mathcal{F} \subset L_p(\mu)$, then for $f \in \mathcal{F}$ we have

$$\|f\|_p = \left(\int |f|^p d\mu \right)^{1/p}$$

if $1 \leq p < \infty$ and

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$$

Now that everything is established, we can construct the ϵ-net.

Definition 1.2 *(ϵ-net) Suppose $\epsilon > 0$, then $\mathcal{G}$ is a ϵ-net for a function class $\mathcal{F}$ with respect to the p-norm if there exist a $g \in G$ for each $f \in \mathcal{F}$ such that*

$$\|f - g\|_p \leq \epsilon$$

Definition 1.3 *(Covering number) The ϵ-covering number of $\mathcal{F}$ is*

$$N_p(\epsilon, \mathcal{F}, \mu) = min\{N \in \mathbb{N} : \exists \text{ a } \epsilon\text{-net } g_1, g_2, ..., g_N \text{ of } \mathcal{F}$$

$$\text{with respect to the p-norm}\}$$

We see that a function class $\mathcal{G}$ can be called a ϵ-net of $\mathcal{F}$ if all functions of $\mathcal{F}$ can be contained in the union of the closed balls of radius ϵ around the functions of $\mathcal{G}$. $N_p(\epsilon, \mathcal{F}, \mu)$ is the cardinality of the smallest ϵ-net. By taking the logarithm of $N_p(\epsilon, \mathcal{F}, \mu)$ we obtain the entropy $H_p(\epsilon, \mathcal{F}, \mu)$.

Definition 1.4 *(Entropy) The entropy of $\mathcal{F}$ is defined as*

$$H_p(\epsilon, \mathcal{F}, \mathbb{P}) = logN_p(\epsilon, \mathcal{F}, \mu)$$

and for the supremum norm

$$H_\infty(\epsilon, \mathcal{F}) = logN_\infty(\epsilon, \mathcal{F})$$

The entropy for the supremum norm can be written this way, because the norm doesn't depend on the measure. $\mathcal{F}$ is totally bounded if the entropy $H_\infty(\epsilon, \mathcal{F})$ is finite for all $\epsilon > 0$.

Bracket entropy

Besides the normal entropy we also need the bracket entropy to construct the two theorems. The computation of the bracket entropy is based on function brackets that encapsulate the functions of the function class. Such a function bracket is called a ϵ-bracket.

Definition 1.5 *(ϵ-bracket) Let l, u denote functions such that $l \leq f \leq u$ then the bracket $[l, u]$ is called a ϵ-bracket with respect to the p-norm if*

$$\|u - l\|_p \leq \epsilon$$

Definition 1.6 *(Bracket number) The bracket number $N_{p,B}(\epsilon, \mathcal{F}, \mu)$ of $\mathcal{F}$ is the minimal number of ϵ-brackets with respect to the p-norm needed to cover $\mathcal{F}$.*

Again is $N_{p,B}(\epsilon, \mathcal{F}, \mu)$ defined as the smallest number of ϵ-brackets such that the ϵ-brackets cover all of $\mathcal{F}$. Taking the logarithm of again $N_{p,B}(\epsilon, \mathcal{F}, \mu)$ will result in the bracket entropy $H_{p,B}(\epsilon, \mathcal{F}, \mu)$.

Definition 1.7 *(Bracket entropy) The bracket entropy of $\mathcal{F}$ is defined as*

$$H_{p,B}(\epsilon, \mathcal{F}, \mu) = log N_{p,B}(\epsilon, \mathcal{F}, \mu)$$

You probably figured out by now that the treated entropy's are very alike. The only difference is that the normal entropy depends on ϵ-nets and the bracket entropy on ϵ-brackets, but the construction is almost the same. Logically, there must be a relation between the two. The following lemma describes this relation.

Lemma 1.2 *For all $\epsilon > 0$ and $1 \leq p < \infty$,*

$$H_p(\epsilon, \mathcal{F}, \mu) \leq H_{p,B}(\epsilon, \mathcal{F}, \mu)$$

and if μ is a probability measure, we have

$$H_{p,B}(\epsilon, \mathcal{F}, \mu) \leq H_\infty \left(\frac{\epsilon}{2}, \mathcal{F}\right)$$

Proof See lemma (A1.2) from Appendix

From this lemma we can conclude that if the bracket entropy is finite, the normal entropy has to be finite. Also, if the entropy for the supremum norm is finite, the bracket entropy with respect to a probability measure has to be finite.

Symmetrization

Symmetrization is a very powerfull technique and plays an important role in the empirical process theory. It involves i.i.d. random variables $X_1, X_2, ..., X_n$ called the test set and independent copies $X'_1, X'_2, ..., X'_n$ called the training set. The training set is used to calculate expectation and the test set to check the performance, but the real trick of symmetrization relies on the fact that $f(X_i) - f(X'_i)$ has the same distribution as $f(X'_i) - f(X)$ for all possible functions f and all $1 \leq i \leq n$. Here is where the Rademacher sequence comes in. A Rademacher sequence $W_1, W_2, ..., W_n$ is a sequence of independent variables satisfying

$$\mathbb{P}(W_i = 1) = \mathbb{P}(W_i = -1) = \frac{1}{2}, \qquad 1 \leq i \leq n$$

By the independence of W_i, one can tell that $f(X_i) - f(X'_i)$ has the same distribution as $W_i(f(X_i) - f(X'_i))$ for all i. This transformation is such a powerfull technique because the symmetrized version is much easier to control than the original version.

Theorem 1.3 *(Symmetrization) Let $\mathcal{F}$ be a class of functions, then*

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} W_i f(X_i)\right\|_{\mathcal{F}}$$

where $(W_1, W_2, ..., W_n)$ is a Rademacher sequence.

Proof See theorem (A1.8) from Appendix

ULLN conditions

Now we will construct two theorems that ensure the Uniform Law of Large Numbers for particular function classes. The entropy with respect to the empirical measure will be used as a condition for these theorems.

Theorem 1.4 *(Bracketing) Let $\mathcal{F}$ be a function class such that*

$$H_{1,B}(\epsilon, \mathcal{F}, \mathbb{P}_n) < \infty, \qquad \forall \epsilon > 0$$

then $\mathcal{F}$ satisfies the ULLN.

Proof From the supposition it follows that $\mathcal{F}$ can be covered by finitely many pairs of functions $\{[g_i^U, g_i^L]\}_{i=1}^N$, so for every $f \in \mathcal{F}$ then there exist a function pair $\left[g_i^L, g_i^U\right]$ such that

$$\left\|g_i^L - g_i^U\right\|_1 \leq \epsilon \text{ and } g_i^L \leq f \leq g_i^U$$

Therefore

$$\begin{aligned}\int f \; d(\mathbb{P}_n - P) &\geq \int g_i^L d(\mathbb{P}_n - P) + \int (g_i^L - f) \; d(\mathbb{P}_n - P) \\ &\geq \int g_i^L d(\mathbb{P}_n - \mathbb{P}) - \epsilon\end{aligned}$$

and

$$\begin{aligned}\int f \; d(\mathbb{P}_n - \mathbb{P}) &\leq \int g_i^U d(\mathbb{P}_n - \mathbb{P}) + \int (g_i^U - f) \; d(\mathbb{P}_n - P) \\ &\leq \int g_i^U d(\mathbb{P}_n - \mathbb{P}) + \epsilon\end{aligned}$$

Consequently for every $f \in \mathcal{F}$,

$$\min_{0\leq i\leq n} \int g_i^L \; d(\mathbb{P}_n - \mathbb{P}) - \epsilon \leq \int f \; d(\mathbb{P}_n - \mathbb{P}) \leq \max_{0\leq i\leq n} \int g_i^U \; d(\mathbb{P}_n - \mathbb{P}) + \epsilon$$

By the Strong Law of Large Numbers we have that the bounds converge to $-\epsilon$ and ϵ. In other words,

$$\sup_{f\in\mathcal{F}} \left| \int f \; d(\mathbb{P}_n - \mathbb{P}) \right| \leq \epsilon, \qquad a.s.$$

Because this holds for all $\epsilon > 0$, we conclude that

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{a.s.} 0$$

From this theorem we can conclude that the bracket entropy is a necessary and sufficient property of a Glivenko-Cantelli class. The next theorem will demonstrate that the entropy combined with the envelope condition ensure the Uniform Law of Large Numbers for a function class. The supremum of a function class is called the envelope. The envelope condition is satisfied if the supremum is finite.

Theorem 1.5 *Let $\mathcal{F}$ be a class of functions such that the envelope condition $\sup_{f\in\mathcal{F}}\|f\|_2 < \infty$ is satisfied. Assume for all $\epsilon > 0$ that*

$$\frac{1}{n} H_1(\epsilon, \mathcal{F}, \mathbb{P}_n) \xrightarrow{n\to\infty} 0 \tag{1.3}$$

then $\mathcal{F}$ satisfies the ULLN.

Proof Let $X_1, X_2, ..., X_n$ be i.i.d. random variables and assume $\mathcal{G}$ is a ϵ-net of $\mathcal{F}$ such that the cardinality of $\mathcal{G}$ is $N_1(\epsilon, \mathcal{F}, \mathbb{P}_n)$. We can write

$$\left|\frac{1}{n}\sum_{i=1}^{n} W_i f(X_i)\right| \leq \left|\frac{1}{n}\sum_{i=1}^{n} W_i g(X_i)\right| + \left|\frac{1}{n}\sum_{i=1}^{n} W_i (f(X_i) - g(X_i))\right|$$

$$\leq \left|\frac{1}{n}\sum_{i=1}^{n} W_i g(X_i)\right| + \epsilon$$

Therefore also

$$\left\|\frac{1}{n}\sum_{i=1}^{n} W_i f(X_i)\right\|_{\mathcal{F}} \leq \left\|\frac{1}{n}\sum_{i=1}^{n} W_i g(X_i)\right\|_{\mathcal{G}} + \epsilon$$

Take the expectation with respect to $W_1, W_2, ..., W_n$ on both sides.

$$\mathbb{E}_W\left\|\frac{1}{n}\sum_{i=1}^{n} W_i f(X_i)\right\|_{\mathcal{F}} \leq \mathbb{E}_W\left\|\frac{1}{n}\sum_{i=1}^{n} W_i g(X_i)\right\|_{\mathcal{G}} + \epsilon \tag{1.4}$$

To construct a bound for the right side of the equation, we are using the Orlicz norm defined as

$$\|.\|_{\psi} := \inf\{k \in (0,\infty) : \mathbb{E}[\psi(|.|/k)] \leq 1\}$$

for a monotone non-decreasing, convex function $\psi : \mathcal{X} \to \mathbb{R}$. For convenience, define

$$\langle W, g\rangle_{\mathbb{P}_n} := \frac{1}{n}\sum_{i=1}^{n} W_i g(X_i)$$

where $W = [W_1, W_2, ..., W_n]$. Then note that

$$\mathbb{E}_W \left[\psi \left(\frac{|\langle W, g \rangle_{\mathbb{P}_n}|}{\|\langle W, g \rangle_{\mathbb{P}_n}\|_\psi)} \right) \right] \leq 1$$

for all $g \in \mathcal{G}$, which means that we can apply lemma (A1.7) with $\psi(x) = e^{x^2} - 1$ to bound the term with

$$\sqrt{1 + logN_1(\epsilon, \mathcal{F}, \mathbb{P}_n)} \sup_{g \in G} \left\| \frac{1}{n} \sum_{i=1}^{n} W_i g(X_i) \right\|_\psi \tag{1.5}$$

By Hoeffding's inequality for the Orlicz norm (see lemma (A1.5) from Appendix), we have

$$\sup_{g \in G} \left\| \frac{1}{n} \sum_{i=1}^{n} W_i g(X_i) \right\|_\psi \leq \sqrt{\frac{6}{n}} \sup_{g \in \mathcal{G}} \|g\|_2$$

Hence we can bound term (1.5) with

$$\sqrt{\frac{1 + logN_1(\epsilon, \mathcal{F}, \mathbb{P}_n)}{n}} \sqrt{6} \sup_{g \in \mathcal{G}} \|g\|_2 \tag{1.6}$$

From assumption (1.3) and because $\sup_{g \in \mathcal{G}} \|g\|_2 < \infty$, we can conclude that term (1.6) converges to zero. We chose ϵ arbitrary, so we can choose ϵ such that the left side of (1.4) converges to zero. If we take

$$h_n = \mathbb{E}_W \left\| \frac{1}{n} \sum_{i=1}^{n} W_i f \right\|_{\mathcal{F}}$$

we have that h_n is bounded by M for all $n \in \mathbb{N}$ and converges pointwise to zero. Using the Dominated Convergence Theorem, deduce that

$$\mathbb{E}_x \mathbb{E}_W \left\| \frac{1}{n} \sum_{i=1}^{n} W_i f(X_i) \right\|_{\mathcal{F}} = \int_{\mathcal{X}} h_n \, d\mathbb{P} \xrightarrow{n \to \infty} 0$$

Then by symmetrization (see theorem (1.3)), we have

$$\mathbb{E} \|P_n - P\|_{\mathcal{F}} \leq 2\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} W_i f(X_i) \right\|_{\mathcal{F}} \xrightarrow{n \to \infty} 0$$

Conclude that $\|P_n - P\|_{\mathcal{F}}$ converges in mean to zero. Applying lemma 2.4.5 from [2] to the reverse-martingale $\|P_n - P\|_{\mathcal{F}}$ with respect to a suitable filtration ensures convergence in probability to zero.

Examples

Here there will be function classes shown that are or aren't Glivenko-Cantelli. The first two examples are demonstrated and proven by checking if the Uniform Law of Large Numbers is satisfied. The other examples will make use of the theorem that is discussed. We will construct upper bounds for the entropy such that condition (1.3) is met.

Theorem 1.6 *Let $\mathcal{D}$ be a Glivenko-Cantelli class. The class $\{\mathbb{1}_D : D \in \mathcal{D}\}$ is also Glivenko-Cantelli.*

Proof For convenience, denote $\mathcal{F} = \{\mathbb{1}_D : D \in \mathcal{D}\}$. Then

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\left|\int f \; d(\mathbb{P}_n - \mathbb{P})\right| = \sup_{D\in\mathcal{D}}\left|\int \mathbb{1}_D \; d(\mathbb{P}_n - \mathbb{P})\right|$$

$$= \sup_{D\in\mathcal{D}}\left|\int_D d(\mathbb{P}_n - \mathbb{P})\right| = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{D}}$$

We know $\mathcal{D}$ is Glivenko-Cantelli, so the last obtained term converges in probability. Therefore,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{D}} \xrightarrow{n\to\infty} 0$$

i.e. $\mathcal{F}$ is Glivenko-Cantelli.

Theorem 1.7 *There exist a function class that satisfies the law of large numbers, but isn't Glivenko-Cantelli.*

Proof Take $\mathcal{F} = \{\mathbb{1}_D : D \subset \mathbb{R},\ |D| < \infty\}$, hence $\mathcal{F}$ consists of all indicator functions of sets with finite cardinality. Let $\mathbb{P}$ be a continuous probability measure, then for a $f \in \mathcal{F}$ we have

$$\int f \; d\mathbb{P} = \int \mathbb{1}_D \; d\mathbb{P} = \int_D d\mathbb{P} = 0$$

The last equality follows from the fact that the D has finite cardinality, so the integral exist only of points. These points have zero probability for a continuous probability measure. Also, $\sup_{f\in\mathcal{F}} \int f \; d\mathbb{P}_n = 1$ for all $n \in \mathbb{N}$, hence $\mathcal{F}$ is not Glivenko-Cantelli. We only need to proof that

$$\max_{f\in\mathcal{F}}\left|\int f \; d(\mathbb{P}_n - \mathbb{P})\right| \xrightarrow{a.s.} 0$$

Note that

$$\max_{f\in\mathcal{F}} \int f \; d\mathbb{P}_n = \frac{1}{n}\max_{f\in\mathcal{F}} \sum_{i=1}^{n} f(X_i) = \frac{1}{n}\max_{D\in\mathcal{D}} \sum_{i=1}^{n} \mathbb{1}_D(X_i)$$

We have $|D| < \infty$, hence

$$\frac{1}{n} \max_{D \in \mathcal{D}} \sum_{i=1}^{n} \mathbb{1}_D(X_i) \xrightarrow{a.s.} 0$$

Therefore, $\mathcal{F}$ satisfies the strong law of large numbers.
Now there will be examples shown which have finite entropy. We proof that by obtaining an upper bound. To understand how such a bound is derived, we start with an easy example.

Theorem 1.8 *Let $\mathcal{F}$ be a class of all increasing bounded functions $f : \mathcal{X} \to [0, 1]$, such that $\mathcal{X}$ has n elements. Then*

$$H_\infty(\epsilon, \mathcal{F}) \leq \frac{1}{\epsilon} log\left(n + \frac{1}{\epsilon}\right), \qquad 0 \leq \epsilon \leq \frac{1}{n}$$

Proof Let $x_1 \leq x_2, ... \leq x_n$ denote all elements of $\mathcal{X}$, then define

$$g(x_i) = \epsilon * \left\lfloor \frac{f(x_i)}{\epsilon} \right\rfloor, \qquad i = 1, 2, ..., n$$

Note that

$$|g(x_i) - f(x_i)| = \left| \epsilon * \left\lfloor \frac{f(x_i)}{\epsilon} \right\rfloor - f(x_i) \right| \leq \epsilon$$

for all $i = 1, 2, ..., n$. The number of possibilities g can be chosen is

$$\binom{n + \lfloor 1/\epsilon \rfloor}{\lfloor 1/\epsilon \rfloor}$$

because we need to pick $\lfloor 1/\epsilon \rfloor$ elements out of $n + \lfloor 1/\epsilon \rfloor$ elements. Expanding Stirling approximation for a binomial coefficient gives

$$log\binom{n + \lfloor 1/\epsilon \rfloor}{\lfloor 1/\epsilon \rfloor} \approx (n + \lfloor 1/\epsilon \rfloor) log(n + \lfloor 1/\epsilon \rfloor) - n log(n) - \lfloor 1/\epsilon \rfloor log(\lfloor 1/\epsilon \rfloor)$$

$$\leq \frac{1}{\epsilon} log\left(n + \frac{1}{\epsilon}\right)$$

where the last inequality follows from the fact that $\frac{1}{\epsilon} \geq n$.

Because this class consists of bounded functions, they must have bounded norms. Hence applying lemma (1.2) gives that this class satisfies both conditions of theorem (1.5). Applying this theorem, proofs that this function class is a Glivenko-Cantelli class. The same holds for the following example.

Theorem 1.9 *Assume $f : [a, b] \to [0, M]$ is a Lipschitz continuous bounded function. Let $\mathcal{F}$ be the class generated by f, i.e. $\mathcal{F}$ consists of all Lipschitz continuous bounded functions. Then for some constant A,*

$$H_\infty(\epsilon, \mathcal{F}) \leq A \frac{1}{\epsilon}, \qquad \forall \epsilon \geq 0$$

Proof We consider the class of all Lipschitz continuous bounded functions. Let $\mathcal{F} := \{f : [a,b] \to [0,M] \ : |f(x) - f(y)| \leq K|x-y|, \ \forall x, y \in [a,b]\}$ denote this class. Before we can find the entropy number, we first need to construct a ϵ-net which covers the class.

To construct a ϵ-net, we start with a sequence $(x_i)_{i\geq 0}$ such that

$$x_{i+1} = x_i + \frac{\epsilon}{K}$$

with initial values $x_0 = a$ and $x_n = b$. From the definition of x_i it follows that

$$n = \left\lceil \frac{K(b-a)}{\epsilon} \right\rceil$$

For a $f \in \mathcal{F}$, we can define

$$g(x) = \sum_{i=0}^{n-1} f(x_i) * \mathbb{1}_{I_i} \tag{1.7}$$

where $I_i = [x_i, x_{i+1}]$. This just means that $g(x) = f(x_i)$ with i the number satisfying $x \in I_i$. We know that

$$|f(x_i) - f(x)| \leq K|x_i - x| \leq K\left|\frac{\epsilon}{K}\right| = \epsilon, \qquad \text{for } x \in I_i$$

hence from (1.7) it follows that

$$|g(x) - f(x)| = |f(x_i) - f(x)| \leq \epsilon, \qquad \text{for } x \in I_i$$

This confirms that the family generated by (1.7) is a ϵ-net for $\mathcal{F}$. To establish an upper bound for the entropy number we need to count the number of ways g can be chosen. It is easy to see that there are $\lceil M/\epsilon \rceil$ possible choices for $g(x_0)$. We also have

$$|g(x_{i+1}) - g(x_i)| = |f(x_{i+1}) - f(x_i)| \leq \epsilon$$

This means that there are at most 3 choices for $g(x_{i+1})$, if $g(x_i)$ is known. Therefore

$$N_\infty(\epsilon, \mathcal{F}) \leq \left\lceil \frac{M}{\epsilon} \right\rceil * 3^{\lfloor K(b-a)/\epsilon \rfloor}$$

Least Squares Estimation

Problem description

In regression analysis is least square estimation a very valuable tool to approximate the solution of a regression model. This regression model exist of observed response variables $Y_1, Y_2, ..., Y_n$ of random variables $z_1, z_2, ..., z_n$ (covariates) out of a space $\mathcal{Z}$, independent errors $W_1, W_2, ..., W_n$ with expectation zero and finite variance, and finally the unknown regression function f_0 which will be approximated.The model is given by

$$Y_i = f_0(z_i) + W_i, \qquad i = 1, 2, ..., n \tag{2.8}$$

The aim of least square estimation is to find the correct f_0, which will be done by minimizing the errors W_i. If $\mathcal{F}$ is a function class and we assume f_0 lies in $\mathcal{F}$, then the least squares estimator $\hat{f}_n$ is given by

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(z_i))^2 \tag{2.9}$$

If we assumed the errors were normally distributed the least square estimator was equal to the maximum likelihood estimator, but we only assume zero expectation and finite variance. Recall the definition of the empirical measure $\mathbb{P}_n$, then

$$\mathbb{P}_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(z_i)$$

denotes the empirical measure of the covariates $z_1, z_2, ..., z_n$ for $A \subset \mathcal{Z}$. The norm with respect to the empirical measure of a function $f : \mathcal{Z} \to \mathbb{R}$ is written as

$$\|f\|^2_{\mathbb{P}_n} = \frac{1}{n} \sum_{i=1}^{n} f^2(z_i)$$

For convenience, we write

$$\|Y - f\|^2_{\mathbb{P}_n} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(z_i))^2$$

and for $W = [W_1, W_2, ..., W_n]$,

$$\langle W, f \rangle_{\mathbb{P}_n} = \frac{1}{n} \sum_{i=1}^{n} W_i f(z_i)$$

The least squares problem aims to minimize $\left\| \hat{f}_n - f_0 \right\|_{\mathbb{P}_n}$. This means we need to figure out the conditions of $\mathcal{F}$ for which the norm converges in probability to zero.

Consistency

We say the estimator $\hat{f}_n$ is a consistent estimator of f_0, if

$$\lim_{n\to\infty} \mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} > \delta\right) = 0, \qquad \forall \delta > 0$$

i.e. if it converges in probability to the unknown regression function. To establish certain conditions for consistency of the least square estimator, we will be using this following inequality.

Lemma 2.1 *Let $\hat{f}_n$ be the estimator (see (2.9)) for the function f_0. Suppose $W = [W_1, W_2, ..., W_n]$ is the random error sequence of the observed variable sequence Y, then*

$$\left\|\hat{f}_n - f_0\right\|^2_{\mathbb{P}_n} \leq 2\langle W, \hat{f}_n - f_0\rangle_{\mathbb{P}_n}$$

Proof See lemma (A2.1) from Appendix

Using this inequality we can establish conditions for which consistency of the least square estimator is guaranteed. These conditions are not dependent on the full class $\mathcal{F}$, but only the subclass $\mathcal{F}_n(R)$ defined as

$$\mathcal{F}_n(R) = \{f \in \mathcal{F} : \|f - f_0\|_{\mathbb{P}_n} \leq R\} \tag{2.10}$$

In other words, $\mathcal{F}_n(R)$ denotes all functions of $\mathcal{F}$ that are in the ball around f_0 with radius R. This class will be used to proof the following theorem. This proof relies on condition (2.11) which is met, because we assumed that all errors have finite variance and zero expectation.

Theorem 2.2 *Let $\mathcal{F}_n(R)$ defined as (2.10) for a function class $\mathcal{F}$. For all errors of the observed variables, assume that*

$$\lim_{K\to\infty} \lim_{n\to\infty} \sup \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(W_i^2 * \mathbb{1}_{\{|W_i|>K\}}) = 0 \tag{2.11}$$

and for $\mathcal{F}_n(R)$, we have

$$\frac{1}{n} H_1(\delta, \mathcal{F}_n(R), \mathbb{P}_n) \to 0, \qquad \forall \delta, R > 0$$

Then the the least square estimator $\hat{f}_n$ is consistent, i.e.

$$\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} \xrightarrow{\mathbb{P}} 0$$

.

Proof To proof consistency we need to show that $\mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} > \delta\right)$ goes to zero when n goes to infinity. We have that $R > 0$, hence

$$\mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} > \delta\right) \leq \mathbb{P}\left(\delta < \left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} \leq R\right) + \mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} > R\right)$$

We start with first term of the right equation. From lemma (2.1), it follows that

$$\mathbb{P}\left(\delta < \left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} \leq R\right) \leq \mathbb{P}\left(2\langle r, \hat{f}_n - f_0\rangle > \delta\right)$$

We have that $\hat{f}_n \in \mathcal{F}_n(R)$, so we can write

$$\begin{aligned}
\mathbb{P}\left(2\langle r, \hat{f}_n - f_0\rangle > \delta\right) &\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}_n(R)} \langle r, f - f_0\rangle_{\mathbb{P}_n} \geq \frac{\delta^2}{4}\right) \\
&\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}_n(R)} \langle r * \mathbb{1}_{\{|r|>K\}}, f - f_0\rangle_{\mathbb{P}_n} \geq \frac{\delta^2}{4}\right) \\
&+ \mathbb{P}\left(\sup_{f\in\mathcal{F}_n(R)} \langle r * \mathbb{1}_{\{|r|\leq K\}}, f - f_0\rangle_{\mathbb{P}_n} \geq \frac{\delta^2}{4}\right) \qquad (2.12)
\end{aligned}$$

Using Cauchy-Schwarz on the first term, we find

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in\mathcal{F}_n(R)} \langle r * \mathbb{1}_{\{|r|>K\}}, f - f_0\rangle_{\mathbb{P}_n} \geq \frac{\delta^2}{4}\right) &\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}_n(R)} \|f - f_0\|_{\mathbb{P}_n} \left\|r * \mathbb{1}_{\{|r|>K\}}\right\|_{\mathbb{P}_n} \geq \frac{\delta^2}{4}\right) \\
&\leq \mathbb{P}\left(\left\|r * \mathbb{1}_{\{|r|>K\}}\right\|_{\mathbb{P}_n} \geq \frac{\delta^2}{4R}\right)
\end{aligned}$$

where the last inequality follows from the fact that $\sup_{f\in\mathcal{F}_n(R)} \|f - f_0\|_{\mathbb{P}_n} \leq R$. Applying Markov's inequality bounds the last term,

$$\mathbb{P}\left(\left\|r * \mathbb{1}_{\{|r|>K\}}\right\|_{\mathbb{P}_n} \geq \frac{\delta^2}{4R}\right) \leq \left(\frac{4R}{\delta^2}\right)^2 \mathbb{E}\left\|r * \mathbb{1}_{\{|r|>K\}}\right\|_{\mathbb{P}_n}$$

so we can write

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_n(R)} \langle r * \mathbb{1}_{\{|r|>K\}}, f - f_0\rangle_{\mathbb{P}_n} \geq \frac{\delta^2}{4}\right) \leq \left(\frac{4R}{\delta^2}\right)^2 \mathbb{E}\left\|r * \mathbb{1}_{\{|r|>K\}}\right\|_{\mathbb{P}_n} = \eta$$

The term from (2.12) can also be bounded by Markov's inequality,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_n(R)} \langle r * \mathbb{1}_{\{|r|\leq K\}}, f - f_0\rangle_{\mathbb{P}_n} \geq \frac{\delta^2}{4}\right) \leq \frac{4}{\delta^2}\mathbb{E}\left\|\langle r * \mathbb{1}_{\{|r|\leq K\}}, f - f_0\rangle_{\mathbb{P}_n}\right\|_{\mathcal{F}_n(R)}$$

Let $\mathcal{G}$ be a ϵ-net of $\mathcal{F}_n(R)$ with cardinality $N_1(\delta, \mathcal{F}_n(R), \mathbb{P}_n)$, then

$$\mathbb{E}\left\|\langle r * \mathbb{1}_{\{|r|\leq K\}}, f - f_0\rangle_{\mathbb{P}_n}\right\|_{\mathcal{F}_n(R)} \leq \mathbb{E}\left\|\langle r * \mathbb{1}_{\{|r|\leq K\}}, g - g_0\rangle_{\mathbb{P}_n}\right\|_{\mathcal{G}} + K\epsilon$$

In almost the same way as before, we can construct a bound for the right side of the equation using the Orlicz norm. For convenience we assume that all $W_i * \mathbb{1}_{\{|r|\leq K\}}$ are symmetric, so all expectations remain zero. It can also be proven without this assumption, but we will not get into that. Note that the inner product is bounded,

$$\langle r * \mathbb{1}_{\{|r|\leq K\}}, g - g_0\rangle_{\mathbb{P}_n} = \frac{1}{n}\sum_{i=1}^{n} r * \mathbb{1}_{\{|r|\leq K\}}(g(z_i) - g_0(z_i))$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} K(g(z_i) - g_0(z_i)) \leq K\|g - g_0\|_{\mathbb{P}_n}$$

for all $g \in \mathcal{G}$. Now we can apply lemma (A1.3) (from Appendix), such that

$$e^{s\langle r*\mathbb{1}_{\{|r|\leq K\}}, g-g_0\rangle_{\mathbb{P}_n}} \leq e^{s^2\sigma^2/2}$$

where $\sigma = \frac{1}{\sqrt{2n}}K\|g - g_0\|_{\mathbb{P}_n}$. The condition of lemma (A1.9) (from Appendix) is satisfied, hence applying this lemma gives

$$\mathbb{E}\left\|\langle r * \mathbb{1}_{\{|r|\leq K\}}, g - g_0\rangle_{\mathbb{P}_n}\right\|_{\mathcal{G}} \leq \sqrt{log(N_1(\delta, \mathcal{F}_n(R), \mathbb{P}_n))}\frac{1}{\sqrt{n}}K \sup_{g\in\mathcal{G}}\|g - g_0\|_{\mathbb{P}_n}$$

Bringing it all together we can bound $\mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} > \delta\right)$ with

$$\frac{\delta^2}{4}K\left(\sqrt{log(N_1(\delta, \mathcal{F}_n(R), \mathbb{P}_n))}\frac{1}{\sqrt{n}}R + \epsilon\right) + \eta$$

From the second condition we know that the square root of the entropy divided by n converges to zero. Besides this condition, ϵ is chosen arbitrary, so we can make the left side smaller than η such that

$$\mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} > \delta\right) \leq 2\eta$$

By the first condition we can make K so large that η converges to zero, so we have

$$\mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} > \delta\right) \xrightarrow{n\to\infty} 0$$

Convergence rate

Our first approach to the least squares problem presumed the general case where the errors W_i were randomly chosen with zero expectation and finite variance. We will now assume that the errors are uniformly sub-Gaussian, but not necessary independent.

Definition 2.1 *A sequence of random variables $W_1, W_2, ..., W_n$ is called uniformly sub-Gaussian if there exist a $\sigma_0 > 0$ such that*

$$\max_{1\leq i\leq n} m^2 \left(\mathbb{E}e^{|W_i|^2/m^2} - 1\right) \leq \sigma_0^2$$

for a constant $m > 0$.

Lemma 2.3 *If a sequence of random variables $W_1, W_2, ..., W_n$ is uniformly sub-Gaussian, then*

$$\mathbb{E}\|W\|_{\mathbb{P}_n}^2 \leq \sigma_0^2$$

where

$$W = [W_1, W_2, ..., W_n]$$

Proof Note that

$$\mathbb{E}\|W\|_{\mathbb{P}_n}^2 = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} W_i^2\right) \leq \max_{1\leq i\leq n} \mathbb{E}|W_i|^2$$

From the fact that $ln(y) \leq y - 1$ for all $y \in \mathbb{R}$, it follows that

$$\max_{1\leq i\leq n} \mathbb{E}|W_i|^2 = \max_{1\leq i\leq n} m^2 \left(\mathbb{E}\left(\frac{|W_i|^2}{m^2}\right)\right) \leq m^2 \left(\max_{1\leq i\leq n} e^{\mathbb{E}\left(|W_i|^2/m^2\right)}\right)$$

Applying Jensen's inequality (see theorem (A1.6)) gives

$$m^2 \left(\max_{1\leq i\leq n} e^{\mathbb{E}\left(|W_i|^2/m^2\right)}\right) \leq m^2 \left(\max_{1\leq i\leq n} \mathbb{E}e^{|W_i|^2/m^2}\right) \leq \sigma_0^2$$

We will use the assumption that all errors are uniformly sub-Gaussian to calculate the convergence rate of $\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n}$, along with the function $J(\delta, \mathcal{F}_n(\delta), \mathbb{P}_n)$ defined for a fixed $\sigma > 0$, as

$$J(\delta, \mathcal{F}_n(\delta), \mathbb{P}_n) = \int_{\delta^2/(2^6\sigma)}^{\delta} H^{1/2}(u, \mathcal{F}_n(\delta), \mathbb{P}_n)\ du$$

where $0 < \delta < 2^6\sigma$. Besides that, we will also make use of the so called peeling device.

Lemma 2.4 *If a sequence of random variables $W_1, W_2, ..., W_n$ with zero expectations is uniformly sub-Gaussian, then for all $1 \leq i \leq n$*

$$\mathbb{E}|W_i|^k \leq \frac{k!}{2} m^{k-2} \sigma_0^2, \qquad k = 3, 4, ...$$

Proof See lemma (A2.2) from Appendix

Lemma 2.5 *(Peeling device) Let τ be a function which takes its input from a function class $\mathcal{F}$. Suppose $\{m_s\}_{s=0}^S$ is a strictly increasing sequence such that*

$$\mathcal{F} \subseteq \bigcup_{s=0}^{S} \mathcal{F}_s$$

where

$$\mathcal{F}_s = \{f \in \mathcal{F} : m_{s-1} \leq \tau(f) < m_s\}$$

then for a stochastic process Z_n indexed by $\mathcal{F}$, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{|Z_n(f)|}{\tau(f)} > a\right) \leq \sum_{s=0}^{S} \mathbb{P}\left(\sup_{f\in\mathcal{F},\ \tau(f)<m_s} |Z_n(f)| > a m_{s-1}\right)$$

Proof See lemma (A2.3) from Appendix

With this lemma, we can split the function class $\mathcal{F}_n(R)$ in subclasses so it becomes easier to calculate the convergence rate.

Theorem 2.6 *Suppose all errors are uniformly sub-Gaussian. Take $\psi(\delta) \geq J(\delta, \mathcal{F}_n(\delta), \mathbb{P}_n)$ such that $\psi(\delta)/\delta^2$ is a non-decreasing function. Then for a $\delta_n > 0$ such that*

$$\sqrt{n}\delta_n^2 \geq c\psi(\delta_n)$$

there exist constants $c_1, c_2 > 0$ depending on m and σ_0 such that for all $\delta \geq \delta_n$,

$$\mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} > \delta\right) \leq c_1 e^{-n\delta^2/c_2}$$

for the least square estimator $\hat{f}_n$ of f_0.

Proof Fix $\sigma > 0$, then we have

$$\mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} > \delta\right) \leq \mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} * \mathbb{1}_{\{\|W\|\leq\sigma\}} > \delta\right) + \mathbb{P}\left(\|W\|_{\mathbb{P}_n} > \sigma\right)$$

By applying the Cauchy-Schwarz inequality to lemma (2.1), we get

$$\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n}^2 \leq 2\langle W, \hat{f} - f_0\rangle \leq 2\|W\|_{\mathbb{P}_n}\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n}$$

Hence on $\{\|W\|_{\mathbb{P}_n} \leq \sigma\}$, we have

$$\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} \leq 2\|W\|_{\mathbb{P}_n} \leq 2\sigma$$

Which means that on $\{\|W\|_{\mathbb{P}_n} \leq \sigma\}$ we only need to consider $\delta \leq 2\sigma$. Therefore

$$\mathbb{P}\left(\left\|\hat{f} - f_0\right\|_{\mathbb{P}_n} * \mathbb{1}_{\{\|W\|\leq\sigma\}} > \delta\right) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}_n(2\sigma)} \|f - f_0\|_{\mathbb{P}_n} * \mathbb{1}_{\{\|W\|\leq\sigma\}} > \delta\right)$$

$$\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}_n(2\sigma)} \frac{\langle W, f-f_0\rangle_{\mathbb{P}_n}}{\|f-f_0\|_{\mathbb{P}_n}} * \mathbb{1}_{\{\|W\|\leq\sigma\}} > 2^{-1}\delta\right)$$

where the last inequality follows from using lemma (2.1) again. Now we divide the function class $\mathcal{F}_n(2\sigma)$ such that we can apply the peeling device. Define

$$\mathcal{F}_s = \{f \in \mathcal{F} : 2^s\delta \leq \|f-f_0\|_{\mathbb{P}_n} < 2^{s+1}\delta\}$$

then

$$\mathcal{F} \subseteq \bigcup_{s=0}^{S} \mathcal{F}_s$$

where

$$S = \min\{s \in \mathbb{N} : 2^s\delta > 2\sigma\}$$

Applying the peeling device (see lemma (2.5)) gives

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_n(2\sigma)} \frac{\langle W, f-f_0\rangle_{\mathbb{P}_n}}{\|f-f_0\|_{\mathbb{P}_n}} * \mathbb{1}_{\{\|W\|\leq\sigma\}} > 2^{-1}\delta\right) \leq \sum_{s=0}^{S} \mathbb{P}\left(\sup_{f\in\mathcal{F}_n(2^{s+1}\delta)} \langle W, f-f_0\rangle_{\mathbb{P}_n} * \mathbb{1}_{\{\|W\|\leq\sigma\}} \geq 2^{2s-1}\delta^2\right)$$

Take $C = \frac{1}{16}c$, then

$$\frac{1}{16}\sqrt{n}\delta_n^2 \geq C\psi(\delta_n)$$

which means for $0 \leq s \leq S$,

$$\sqrt{n}2^{2s-1}\delta^2 \geq C\psi(2^{s+1}\delta)$$

If we define

$$\mathbb{P}_s = \mathbb{P}\left(\sup_{f\in\mathcal{F}_n(2^{s+1}\delta)} \langle W, f-f_0\rangle_{\mathbb{P}_n} * \mathbb{1}_{\{\|W\|\leq\sigma\}} \geq 2^{2s-1}\delta^2\right)$$

we can apply Corollary 8.3 from [9] to all $\mathbb{P}_s$, to obtain

$$\sum_{s=0}^{S} \mathbb{P}_s \leq \sum_{s=0}^{S} Ce^{-(n2^{4s-2}\delta^4)/(4C^2 2^{2s+2}\delta^2)} = \sum_{s=0}^{S} Ce^{-n(2^{2s-5}\delta^2)/C^2}$$

So we can write,

$$\mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} > \delta\right) \leq \sum_{s=0}^{S} Ce^{-n(2^{2s-5}\delta^2)/C^2} + \mathbb{P}\left(\|W\|_{\mathbb{P}_n} > \sigma\right)$$

We assumed the errors were sub-Gaussian, so applying lemma (2.3) and (2.4) gives

$$\mathbb{E}\|W\|_{\mathbb{P}_n}^2 \leq \sigma_0^2 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|W_i|^m \leq \frac{k!}{2}m^{k-2}\sigma_0^2$$

Using Bernstein's inequality (see theorem (A2.4) from Appendix), we get

$$\mathbb{P}(\|W\|_{\mathbb{P}_n}^2 > 2\sigma_0^2) \leq 2e^{-(n(2\sigma_0^2)^2)/(2(2\sigma_0^2 m+\sigma_0^2)} \leq e^{-(n\sigma_0^2)/m}$$

$$\leq e^{-(n2^{S-1}\delta^2)/m} \leq e^{-(n2^{S-1}\delta^2)/m}$$

Taking $\sigma = 2\sigma_0$ shows that there exist constants $c_1, c_2 > 0$ depending on m and σ_0 such that

$$\mathbb{P}\left(\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n} > \delta\right) \leq e^{-n(2^{S-1}\delta^2)/m} + \sum_{s=0}^{S} Ce^{-n(2^{2s-5}\delta^2)/C^2} \leq c_1 e^{-n\delta^2/c_2}$$

P-Donsker classes

The study of empirical processes has led to the discovery of P-Donsker classes which represent a whole new branch in the empirical process theory. The main cause was Donsker's theorem, which can be seen as a functional extension of the Central Limit theorem. This theorem states that the empirical process indexed by $\mathcal{F} = \{\mathbb{1}_{(-\infty,x]} : x \in \mathbb{R}\}$ converges in distribution to a standard Brownian motion.

Definition 3.1 *(Brownian motion) A Brownian process is a random process* $(B(t))_{t\geq 0}$ *satisfying these conditions*

(i) The process has stationary increments. That is, for every $0 \leq s < t$ the distribution of $B(t) - B(s)$ is the same as the distribution of $B(t-s)$.
(ii) The process has independent increments. That is, for all $0 < t_1 < ... < t_n$ the random variables $B(t_1), B(t_2) - B(t_1), ..., B(t_n) - B(t_{n-1})$ are all independent.
(iii) For all $t \in (0, \infty)$, $B(t)$ is normally distributed with zero mean and variance t.
(iv) With probability 1, the function $t \mapsto B(t)$ is continuous.

We speak of a standard Brownian motion if $B(0) = 0$. The function classes whose empirical process converges in distribution to a tight Brownian motion are called P-Donsker. Portmanteau lemma (see (A3.1) from Appendix) states that for all empirical processes $\mathbb{G}_n$, we have

$$\mathbb{G}_n \xrightarrow{d} \mathbb{G} \quad \longleftrightarrow \quad \mathbb{E}f(\mathbb{G}_n) \to \mathbb{E}f(\mathbb{G})$$

for all continuous bounded functions f.

Definition 3.2 *(P-Donsker class) A function class* $\mathcal{F}$ *is called P-Donsker if the empirical process* $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ *converges in distribution to a tight Brownian motion* $\mathbb{G}$ *in the space* $L_\infty(\mathcal{F})$ *of uniformly bounded functions on* $\mathcal{F}$*. In other words,*

$$\mathbb{E}f(\mathbb{G}_n) \to \mathbb{E}f(\mathbb{G})$$

for every continuous bounded function $f : L_\infty(\mathcal{F}) \to \mathbb{R}$ *and a tight Brownian process* $\mathbb{G}$*.*

An example of a P-Donsker class is the function class $\{\mathbb{1}_{(-\infty,x]} : x \in \mathbb{R}\}$. Donsker's theorem proofs that the empirical process of this function class converges in distribution to a tight Brownian motion in the space of uniformly bounded functions.

Theorem 3.1 *(Donsker's theorem) The function class* $\mathcal{F} = \{\mathbb{1}_{(-\infty,x]} : x \in \mathbb{R}\}$ *is P-Donsker.*

Proof The empirical process of $\mathcal{F}$ is given by

$$(\mathbb{G}_n(f))_{f\in\mathcal{F}} = (\sqrt{n}(F_n(x) - F(x)))_{x\in\mathbb{R}}$$

where $F_n(x) = \mathbb{P}_n(X_i \leq x)$ denotes the empirical distribution function and $F(x) = \mathbb{P}(X_i \leq x)$ the true distribution function. By the Central Limit theorem, $\mathbb{G}_n$ indexed by $x \in \mathbb{R}$ converges in distribution to a Gaussian process G with zero mean and variance $F(x)(1-F(x))$. Therefore, $\mathcal{F}$ can be called P-Donsker.

P-Donsker conditions

One of the necessary conditions of P-Donsker classes is asymptotic equicontinuity. Asymptotic equicontinuity is defined for functions with elements from any normed metric space. We consider a metric space with norm $\|\cdot\|$.

Definition 3.3 *(asymptotic equicontinuity) An empirical process $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ can be called asymptotic equicontinuous at f_0 if for each $\eta, \delta > 0$ there exist a $\epsilon > 0$ such that*

$$\lim_{n\to\infty} \sup_n \mathbb{P}\left(\sup_{f\in\mathcal{F}, \|f-f_0\|\leq\epsilon} |\mathbb{G}_n(f) - \mathbb{G}_n(f_0)| > \eta \right) < \delta$$

Asymptotic equicontinuity and totally boundedness are sufficient and necessary conditions of P-Donsker classes. The following theorem will only proof the necessity of the properties.

Theorem 3.2 *Suppose that $\mathcal{F}$ is totally bounded and that the empirical process G_n of $\mathcal{F}$ is asymptotic equicontinuous at all $f_0 \in \mathcal{F}$, then $\mathcal{F}$ can be called P-Donsker.*

Proof We know that there exist a process $\mathbb{G}$ such that $\mathbb{G}_n \xrightarrow{d} \mathbb{G}$. Let $\mathcal{F}_k$ with $k \in \mathbb{N}$ be finite sets increasing to a finite subset $\mathcal{F}_0$ of $\mathcal{F}$.

$$\mathbb{P}\left(\max_{f_1,f_2\in\mathcal{F}_k,\ \|f_1-f_2\|\leq\epsilon} |\mathbb{G}(f_1) - \mathbb{G}(f_2)| > \eta \right) \leq \lim_{n\to\infty} \inf_n \mathbb{P}\left(\max_{f_1,f_2\in\mathcal{F}_k, \|f_1-f_2\|\leq\epsilon} |\mathbb{G}_n(f_1) - \mathbb{G}_n(f_2)| > \eta \right)$$

$$\leq \lim_{n\to\infty} \inf_n \mathbb{P}\left(\sup_{f_1,f_2\in\mathcal{F}_0,\ \|f_1-f_2\|\leq\epsilon} |\mathbb{G}_n(f_1) - \mathbb{G}_n(f_2)| > \eta \right)$$

Taking $k \to \infty$ for the left side of the equation gives

$$\mathbb{P}\left(\sup_{f_1,f_2\in\mathcal{F}_0,\ \|f_1-f_2\|\leq\epsilon} |\mathbb{G}(f_1) - \mathbb{G}(f_2)| > \eta \right) \leq \lim_{n\to\infty} \inf_n \mathbb{P}\left(\sup_{f_1,f_2\in\mathcal{F}_0,\ \|f_1-f_2\|\leq\epsilon} |\mathbb{G}_n(f_1) - \mathbb{G}_n(f_2)| > \eta \right)$$

From the assumption of asymptotic equicontinuity there exist a sequence $\epsilon_r > 0$ with $\epsilon_r \xrightarrow{r\to\infty} 0$ such that

$$\mathbb{P}_r = \mathbb{P}\left(\sup_{f_1,f_2\in\mathcal{F}_0,\ \|f_1-f_2\|\leq\epsilon_r} |\mathbb{G}(f_1) - \mathbb{G}(f_2)| > 2^{-r} \right) \leq 2^{-r}$$

Note that,

$$\sum_{r=0}^{\infty} \mathbb{P}_r \leq \sum_{r=0}^{\infty} 2^{-r} < \infty \tag{1.13}$$

Hence from applying Borel-Cantelli lemma, conclude that there exist a $r(w) < \infty$ almost surely, such that for all w

$$\sup_{f_1, f_2 \in \mathcal{F}_0,\ \|f_1 - f_2\| \leq \epsilon_r} |\mathbb{G}(f_1; w) - \mathbb{G}(f_2; w)| \leq 2^{-r}, \qquad \forall r > r(w)$$

Therefore, $\mathbb{G}(f; w)$ is uniformly continuous for the norm $\|\cdot\|$ for almost all w. $\mathcal{F}$ is totally bounded, so $\mathbb{G}(f; w)$ is also bounded. (1.13) can be extended to all of $\mathcal{F}$ on the w set where $\mathbb{G}$ is uniformly continuous. This extension is a version of $\mathbb{G}$ whose trajectories are all uniformly continuous in $\mathcal{F}$ which proofs through applying proposition 2.1.7 from [8] that $\mathcal{F}$ is P-Donsker.

Z-estimation

One of the fields where P-Donsker classes can be applicable lies in Z-estimation. It will be used to ensure asymptotical normality for Z-estimators.

Definition 3.4 *(Z-estimator) $\hat{\theta}_n$ is called a Z-estimator of θ_0 for an estimating function $\psi_\theta(\cdot)$ if*

$$\mathbb{P}(\psi_{\theta_0}) = 0 \quad \textit{and} \quad \mathbb{P}_n(\psi_{\hat{\theta}_n}) \xrightarrow{n \to \infty} 0$$

Z-estimators look a lot like M-estimators, but not every M-estimator can be written as a Z-estimator. One of the asymptotic properties of a Z-estimator is asymptotic normality. It shows the relation between the estimator and the estimated parameter.

Definition 3.5 *A Z-estimator $\hat{\theta}_n$ of θ_0 is called asymptotically normal if*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma_0^2)$$

for a constant $\sigma_0^2 > 0$.

The following lemma states a property for P-Donsker classes. We will apply this lemma in the proof that ensures asymptotic normality for Z-estimators.

Lemma 3.3 *Suppose the function class $\mathcal{F}$ is P-Donsker. Let $\hat{f}_n \in \mathcal{F}$ depending on $X_1, X_2, ..., X_n$ be such that*

$$\left\| \hat{f}_n - f_0 \right\| \xrightarrow{\mathbb{P}} 0 \tag{1.14}$$

for a fixed function f_0, then also

$$\left| G_n(\hat{f}_n) - G_n(f_0) \right| \xrightarrow{\mathbb{P}} 0$$

Proof $\mathcal{F}$ is P-Donsker, hence for all $\eta, \delta > 0$, there exist a $\epsilon > 0$ such that

$$\lim_{n\to\infty} \sup_n \mathbb{P}\left(\sup_{f\in\mathcal{F}, \|f-f_0\|\leq\epsilon} |G_n(f) - G_n(f_0)| > \eta \right) < \delta$$

Take

$$\Omega_n = \left\{ \left\| \hat{f}_n - f_0 \right\| \leq \epsilon \right\}$$

and

$$\tilde{\Omega}_n = \left\{ \sup_{\|\hat{f}_n - f_0\|\leq\epsilon} \left| G_n(\hat{f}_n) - G_n(f_0) \right| \leq \eta \right\}$$

then by the property of asymptotic equicontinuity and assumption (1.14),

$$\mathbb{P}(\Omega_n) \xrightarrow{n\to\infty} 1 \quad \text{and} \quad \mathbb{P}(\tilde{\Omega}_n) \xrightarrow{n\to\infty} 1$$

Note that $\left\{ \left| G_n(\hat{f}_n) - G_n(f_0) \right| \leq \eta \right\} \subset \Omega_n \cap \tilde{\Omega}_n$, thus

$$\mathbb{P}\left(\left| G_n(\hat{f}_n) - G_n(f_0) \right| \leq \eta \right) \geq \mathbb{P}\left(\Omega_n \cap \tilde{\Omega}_n \right) \xrightarrow{n\to\infty} 1$$

which concludes the proof.

Theorem 3.4 *Suppose $\hat{\theta}_n \in \Theta$ is a Z-estimator of θ_0 for a function $\psi_\theta \in \{\psi_\theta : \Theta\}$, where $\{\psi_\theta : \theta \in \Theta\}$ is P-Donsker. If $\mathbb{P}(\psi_\theta)$ is differentiable at θ_0 and*

$$\left| \hat{\theta}_n - \theta_0 \right| \xrightarrow{\mathbb{P}} 0$$

then $\hat{\theta}_n$ is asymptotically normal.

Proof From supposition it follows that we can apply lemma (3.3), so

$$\left| \mathbb{G}_n(\psi_{\hat{\theta}_n}) - \mathbb{G}_n(\psi_{\theta_0}) \right| \xrightarrow{\mathbb{P}} 0$$

Which is equivalent to

$$\left| \sqrt{n} \left(\mathbb{P}_n \left(\psi_{\hat{\theta}_n} \right) - \mathbb{P}(\psi_{\hat{\theta}_n}) \right) - \sqrt{n} \left(\mathbb{P}_n \left(\psi_{\theta_0} \right) - \mathbb{P}(\psi_{\theta_0}) \right) \right| \xrightarrow{\mathbb{P}} 0$$

Rearrange the terms, to obtain

$$\left| \sqrt{n} \left(\mathbb{P} \left(\psi_{\hat{\theta}_n} \right) - \mathbb{P}(\psi_{\theta_0}) \right) - \sqrt{n} \left(\mathbb{P}_n \left(\psi_{\hat{\theta}_n} \right) - \mathbb{P}_n(\psi_{\theta_0}) \right) \right| \xrightarrow{\mathbb{P}} 0$$

By definition of a Z-estimator, note that

$$\sqrt{n} \left(\mathbb{P}_n \left(\psi_{\hat{\theta}_n} \right) - \mathbb{P}_n(\psi_{\theta_0}) \right) \xrightarrow{P} -\sqrt{n} \left(\mathbb{P}_n(\psi_{\theta_0}) \right)$$

$$= -\sqrt{n}\left(\mathbb{P}_n(\psi_{\theta_0}) - \mathbb{P}(\psi_{\theta_0})\right) = -\mathbb{G}_n(\psi_{\theta_0})$$

Substituting this, yields

$$\left|\sqrt{n}\left(\mathbb{P}\left(\psi_{\hat{\theta}_n}\right) - \mathbb{P}(\psi_{\theta_0})\right) + \mathbb{G}_n(\psi_{\theta_0})\right| \xrightarrow{\mathbb{P}} 0$$

Take $m(\theta) = \mathbb{P}(\psi_\theta)$, then

$$\sqrt{n}\left(\mathbb{P}\left(\psi_{\hat{\theta}_n}\right) - \mathbb{P}(\psi_{\theta_0})\right) = \sqrt{n}(m(\hat{\theta}_n) - m(\theta_0))$$

$$= \sqrt{n}(\hat{\theta}_n - \theta_0)\frac{m(\hat{\theta}_n) - m(\theta_0)}{\hat{\theta}_n - \theta_0}$$

We know that $\left|\hat{\theta}_n - \theta_0\right| \xrightarrow{\mathbb{P}} 0$, hence

$$\frac{m(\hat{\theta}_n) - m(\theta_0)}{\hat{\theta}_n - \theta_0} \xrightarrow{\mathbb{P}} -m'(\theta_0)$$

where $m'(\theta)$ denotes the derivative of $m(\theta_0)$. Conclude that

$$\left|m'(\theta_0)\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) - \mathbb{G}_n(\psi_{\theta_0})\right| \xrightarrow{\mathbb{P}} 0$$

$\{\psi_\theta : \theta \in \Theta\}$ is P-Donsker, so $\mathbb{G}_n(\psi_{\theta_0})$ converges in distribution to a tight Brownian motion. Therefore,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N(0, m'(\theta_0)^{-1})$$

Note that $m'(\theta_0)$ is the Fisher information matrix if ψ is set to be the derivative of the likelihood. This means that this theorem proofs that the maximum likelihood estimator is asymptotically normal.

Conclusion

The goal of this paper was to investigate the asymptotic properties of Z- and M-estimators in the empirical process theory. Properties like asymptotic equicontinuity, asymptotic normallity, consistency, convergence rate have been treated according to the empirical process theory. It has been demonstrated that asymptotic equicontinuity is a necessary conditions of P-Donsker classes and has been used to proof asymptotic normality for Z-estimators. This has been applied to the maximum likelihood estimator. Furthermore we have established consistency for the least square estimator and computed the convergence rate of the estimator. With these applications of the empirical process theory in Z- and M-estimation, it has been shown how influenceable this theory can be on non-parametric models.

Appendix

Theorem A1.1 *(Glivenko-Cantelli theorem)*

$$\|F_n - F\|_\infty = \sup_{x\in\mathcal{X}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

Proof By the Strong law of large numbers, we know

$$\max_{x\in\mathcal{X}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

Therefore we only need to proof that there exist a uniform bound for $\sup_{x\in\mathcal{X}} |F_n(x) - F(x)|$ that converges to zero. Before we can construct the bound, we define the sequence $(x_i)_{i\leq m}$ such that,

$$F(x_i) - F(x_{i-1}) = \frac{1}{m}, \qquad 0 \leq i \leq m$$

with $-\infty = x_0 < x_1 < ... < x_m = \infty$. Now for $x \in [x_{i-1}, x_i]$ we have

$$F_n(x) - F(x) \geq F_n(x_{i-1}) - F(x_i) = F_n(x_{i-1}) - F(x_{i-1}) - \frac{1}{m}$$

and

$$F_n(x) - F(x) \leq F_n(x_i) - F(x_{i-1}) = F_n(x_i) - F(x_i) + \frac{1}{m}$$

hence

$$\sup_{x\in\mathcal{X}} |F_n(x) - F(x)| \leq \max_{x\in\mathcal{X}} |F_n(x) - F(x)| + \frac{1}{m} \xrightarrow{a.s.} \frac{1}{m}$$

The result follows from the fact that m is chosen arbitrary, so $1/m$ can be made as small as possible.

Lemma A1.2 *For all $\epsilon > 0$ and $1 \leq p < \infty$,*

$$H_p(\epsilon, \mathcal{F}, \mu) \leq H_{p,B}(\epsilon, \mathcal{F}, \mu)$$

and if μ is a probability measure, we have

$$H_{p,B}(\epsilon, \mathcal{F}, \mu) \leq H_\infty\left(\frac{\epsilon}{2}, \mathcal{F}\right)$$

Proof Let $\mathcal{F}$ be a function class with bracket number $N_{p,B}(\epsilon, \mathcal{F}, \mu)$ with respect to the p-norm and measure μ for a $\epsilon > 0$. Then $\mathcal{F}$ can be covered by $N_{p,B}$ pairs of functions $\{[g_i^U, g_i^L]\}_{i=1}^{N_{p,B}}$. Which is equal to saying that there exist a pair of functions $[g_i^U, g_i^L]$ for every $f \in \mathcal{F}$ such that

$$\left\|g_i^L - g_i^U\right\|_p \leq \epsilon \text{ and } g_i^L \leq f \leq g_i^U$$

Therefore,

$$\left\|g_i^L - f\right\|_p \leq \epsilon$$

Which means that $\{g_i^L\}_{i=1}^{N_{p,B}}$ is an ϵ-net of $\mathcal{F}$, so we can conclude that $H_p(\epsilon, \mathcal{F}, \mu) \leq H_{p,B}(\epsilon, \mathcal{F}, \mu)$.

Lemma A1.3 *Let X be a random variable with $\mathbb{E}X = 0$ and $X \in [a, b]$ then*

$$\mathbb{E}\left[e^{sX}\right] \leq e^{\frac{s^2(b-a)^2}{8}}$$

Proof X can be written as a convex combination of a and b, because X is bounded by a and b. Therefore for $\alpha = \frac{X-a}{b-a}$, we have

$$X = \alpha b + (1-\alpha)a$$

The function $x \to e^{sx}$ is also convex, hence

$$e^{sX} = \alpha e^{sb} + (1-\alpha)e^{sa} = \left(\frac{X-a}{b-a}\right)e^{sb} + \left(1 - \frac{X-a}{b-a}\right)e^{sa}$$

$$= \left(\frac{X-a}{b-a}\right)e^{sb} + \left(\frac{b-X}{b-a}\right)e^{sa}$$

Taking the expectation on both sides gives

$$\mathbb{E}\left(e^{sX}\right) \leq \left(\frac{\mathbb{E}(X)-a}{b-a}\right)e^{sb} + \left(\frac{b-\mathbb{E}(X)}{b-a}\right)e^{sa}$$

By the supposition,

$$\mathbb{E}\left(e^{sX}\right) = \left(\frac{-a}{b-a}\right)e^{sb} + \left(\frac{b}{b-a}\right)e^{sa} = e^{f(y)}$$

with $y = t(b-a)$ and

$$f = \left(\frac{a}{b-a}\right)y + log\left(1 + \frac{a}{b-a} - \left(\frac{a}{b-a}\right)e^y\right)$$

Because we have $f(0) = f'(0) = 0$ and $f''(y) \leq 1/4$ for all $y > 0$, we can apply Taylor's theorem. Therefore, there exist a $m \in (0, y)$ such that

$$f(y) = f(0) + yf'(0) + \frac{y^2}{2}f''(m) = \frac{y^2}{2}f''(m) \leq \frac{y^2}{8} = \frac{s^2(b-a)^2}{8}$$

Conclude that

$$\mathbb{E}(e^{sX}) \leq e^{\frac{s^2(b-a)^2}{8}}$$

Lemma A1.4 *(Hoeffding's inequality for the Rademacher sequence) Let* $a = (a_1, a_2, ..., a_n)$ *be a n-dimensional vector where* $a_1, a_2, ..., a_n$ *are constants. If* $W_1, W_2, ..., W_n$ *is a Rademacher sequence, then*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} a_i W_i\right| \geq t\right) \leq 2e^{-t^2/(2\|a\|_2^2)}$$

Proof For any $s \geq 0$ and Rademacher value W_i, we have

$$\mathbb{E}e^{sW_i} = \frac{e^s + e^{-s}}{2}$$

Writing this in power series, we obtain

$$\mathbb{E}e^{sW_i} = \frac{1}{2}\left(\sum_{n=0}^{\infty}\frac{s^n}{n!} + \sum_{n=0}^{\infty}(-1)^n\frac{s^n}{n!}\right) =$$

$$\sum_{n\ even}\frac{s^n}{n!} = \sum_{n=0}^{\infty}\frac{s^{2n}}{(2n)!} \leq \sum_{n=0}^{\infty}\frac{(\frac{s}{2})^{2n}}{n!} = e^{s^2/2}$$

For any $s \geq 0$ we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} a_i W_i\right| \geq t\right) = \mathbb{P}\left(e^{s|\sum_{i=1}^{n} a_i W_i|} \geq e^{st}\right)$$

By the Markov inequality and the obtained upper bound, we get

$$\mathbb{P}\left(e^{s|\sum_{i=1}^{n} a_i W_i)|} \geq e^{st}\right) \leq e^{-st}\mathbb{E}\left[e^{s\left(\sum_{i=1}^{n} a_i W_i\right)}\right] \leq e^{\left(s^2/2\right)\|a\|_2^2 - st}$$

where the last inequality follows from the Cauchy-Schwarz inequality. The proof is complete if we substitute $s = \frac{t}{\|a\|_2^2}$ in the already obtained upper bound.

Lemma A1.5 *(Hoeffding's inequality for the Orlicz norm) Let* $W_1, W_2, ..., W_n$ *be Rademacher variables and* $a = (a_1, a_2, ..., a_n)$ *be a constant n-dimensional vector. Then*

$$\left\|\sum_{i=1}^{n} a_i W_i\right\|_{\psi} \leq \sqrt{6}\|a\|_2$$

where $\|.\|_{\psi}$ *denotes the Orliz norm with* $\psi(x) = e^{x^2} - 1$.

Proof Define $X = |\sum_{i=1}^{n} a_i W_i|$ then the Orlicz norm is defined as

$$\|X\|_{\psi} = \inf\{p > 0 : \mathbb{E}\left[\psi(X/p)\right] \leq 1\}$$

Therefore $\|X\|_{\psi} \leq p$ means $\mathbb{E}\left[\psi(X/p)\right] \leq 1$.

$$\mathbb{E}\left[\psi(X/p)\right] = \mathbb{E}\left[e^{X^2/p^2}\right] - 1 = \int_0^{\infty} P(e^{X^2/p^2} > x)\, dx - 1 = \int_1^{\infty} P(X > p\sqrt{log(x)})\, dx$$

From the Hoeffding's inequality (see lemma (A1.4)) we know that

$$P(X \geq t) \leq 2e^{-t^2/(2\|a\|_2^2)}$$

So if we pick $t = p\sqrt{log(x)}$ and $p = \sqrt{6}\|a\|_2$ then we get

$$\mathbb{P}(X \geq p\sqrt{log(x)}) \leq 2e^{-3log(x)}$$

If we substitute that we have

$$\mathbb{E}\left[\psi(X/p)\right] \leq 2\int_1^\infty e^{-3log(x)}dx < 2\int_1^\infty x^{-3}dx = 1$$

This means that

$$\|X\|_\psi \leq \sqrt{6}\|a\|_2$$

Theorem A1.6 *(Jensen's inequality) Suppose f is a convex function and X is a random variable, then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Proof f is a convex function, hence for all $\lambda \in [0,1]$ we have

$$f(\lambda y + (1-\lambda)x) \geq \lambda f(y) + (1-\lambda)f(x), \qquad x, y \in \mathbb{R}$$

Rewriting gives,

$$f(x + \lambda(y-x) \geq f(x) + \lambda(f(y) - f(x))$$

$$\implies f(y) - f(x) \geq \frac{f(x + \lambda(y-x)) - f(x))}{\lambda}$$

Letting λ go to zero gives that there exist an $a \in \mathbb{R}$ such that

$$f(y) - f(x) \geq a(y-x)$$

Therefore, we have that

$$f(x) - f\left(\mathbb{E}\left[X\right]\right) \geq a\left(x - \mathbb{E}\left[X\right]\right)$$

Define $b := a\mathbb{E}\left[X\right] + f\left(\mathbb{E}\left[X\right]\right)$, then

$$f(x) \geq ax + b$$

and

$$f\left(\mathbb{E}\left[X\right]\right) = a\mathbb{E}\left[X\right] + b$$

for all $x \in \mathbb{R}$. This also means that

$$f\left(X(x)\right) \geq aX(x) + b$$

Using this inequality will get us the result.

$$\mathbb{E}\left[f\left(X\right)\right] = \int f(X(x))\ d\mathbb{P} \geq \int aX(x) + b\ d\mathbb{P}$$

$$= a\int X(x)\ d\mathbb{P} + b\int d\mathbb{P} = a\mathbb{E}\left[X\right] + b = f\left(\mathbb{E}\left[X\right]\right)$$

Lemma A1.7 *Let $X_1, X_2, ..., X_n$ be random variables and f a strictly increasing, convex, non-negative function such that for all $1 \leq i \leq n$,*

$$\mathbb{E}\left[f(X_i/c_i)\right] \leq L$$

where $c_1, c_2, ..., c_n$ and L are positive constants, then

$$E\max_{1\leq i\leq n}|X_i| \leq f^{-1}(Ln)\max_{1\leq i\leq n}|c_i|$$

Proof From the fact that

$$\frac{\mathbb{E}\max|X_i|}{\max c_i} \leq \mathbb{E}\max\frac{|X_i|}{c_i}$$

and Jensen's inequality (see theorem (A1.6)) it follows that

$$f\left(\frac{\mathbb{E}\max|X_i|}{\max c_i}\right) \leq \mathbb{E}\left[f\left(\max\frac{|X_i|}{c_i}\right)\right]$$

Because f is strictly increasing, we have that

$$f\left(\frac{\mathbb{E}\max|X_i|}{\max c_i}\right) \leq \sum_{i=1}^{n}\mathbb{E}\left[f\left(\frac{|X_i|}{c_i}\right)\right] \leq Ln$$

Taking f^{-1} on both sides and multiplying with $\max c_i$ gives the result.

Theorem A1.8 *(Symmetrization) Let $\mathcal{F}$ be a class of functions, then*

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}W_i f(X_i)\right\|_{\mathcal{F}}$$

where $(W_1, W_2, ..., W_n)$ is a Rademacher sequence.

Proof Let $X_1', X_2', ..., X_n'$ be independent copies of *i.i.d.* random variables $X_1, X_2, ..., X_n$

defined on the same probability space. If we apply Jensen's inequality (see (A1.6)) on the norm we get

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i')\right| \leq \mathbb{E}_{X'} \sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n} f(X_i) - f(X_i')\right|$$

Because $f(X_i') - f(X_i)$ has the same distribution as $f(X_i) - f(X_i')$, we can also write

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \mathbb{E}_W \mathbb{E}_{X'} \sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n} W_i(f(X_i) - f(X_i'))\right|$$

where $(W_1, W_2, ..., W_n)$ is a *Rademacher* sequence. If we take the expectation $\mathbb{E}$ with respect to $X_1, X_2, ..., X_n$ on both sides, we get

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \mathbb{E} \sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n} W_i(f(X_i) - f(X_i'))\right| = \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} W_i(f(X_i) - f(X_i'))\right\|_{\mathcal{F}}$$

Using the triangle inequality we derive the final expression

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} W_i f(X_i)\right\|_{\mathcal{F}}$$

Lemma A1.9 *Let $X_1, X_2, ..., X_n$ be i.i.d. random variables such that for all $\lambda > 0$,*

$$\mathbb{E}\left[e^{\lambda X_i}\right] \leq e^{\lambda^2\sigma^2/2}, \qquad \text{for all } 1 \leq i \leq n$$

then

$$\mathbb{E} \max_{1\leq i\leq n} X_i \leq \sigma\sqrt{2logN}$$

Proof Taking Jensen's inequality (see theorem (A1.6)) for the convex function $e^{\lambda x}$, it follows that

$$e^{\lambda \mathbb{E}\max_i X_i} \leq \mathbb{E}e^{\lambda \max_i X_i} \leq \sum_{i=1}^{n} \mathbb{E}e^{\lambda X_i} \leq ne^{\lambda^2\sigma^2/2}$$

Taking the logarithm on both sides, we conclude that

$$\mathbb{E} \max_{1\leq i\leq n} X_i \leq \frac{logN}{\lambda} + \frac{\lambda\sigma^2}{2}$$

Now define

$$g(x) = \frac{logN}{x} + \frac{x\sigma^2}{2}$$

then its derivative is given by

$$g'(x) = -\frac{logN}{x^2} + \frac{\sigma^2}{2}$$

and its second derivative by

$$g''(x) = \frac{2logN}{x^3}$$

Note that $g''(x) > 0$ which means that $g'(\lambda) = 0$ gives the maximum value. Hence substituting $\lambda = \sqrt{2logN}/\sigma$ in the upper bound yields the result.

Lemma A2.1 *Let $\hat{f}_n$ be the estimator (see (2.9)) for the function f_0. Suppose $W = [W_1, W_2, ..., W_n]$ is the random error sequence of the observed variable sequence Y, then*

$$\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n}^2 \leq 2\langle W, \hat{f}_n - f_0\rangle_{\mathbb{P}_n}$$

Proof From (2.8) we know that

$$2\langle W, \hat{f}_n - f_0\rangle_{\mathbb{P}_n} = 2\langle Y - f_0, \hat{f}_n - f_0\rangle_{\mathbb{P}_n}$$

Expanding this, we get

$$2\langle W, \hat{f}_n - f_0\rangle_{\mathbb{P}_n} = \|Y - f_0\|_{\mathbb{P}_n}^2 + \left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n}^2 - \left\|Y - \hat{f}_n\right\|_{\mathbb{P}_n}^2$$

$\hat{f}_n$ minimizes the norm of the difference with Y, hence

$$\left\|Y - \hat{f}_n\right\|_{\mathbb{P}_n}^2 \leq \|Y - f_0\|_{\mathbb{P}_n}^2$$

Therefore

$$\left\|\hat{f}_n - f_0\right\|_{\mathbb{P}_n}^2 \leq 2\langle W, \hat{f}_n - f_0\rangle_{\mathbb{P}_n}$$

Lemma A2.2 *If a sequence of random variables $W_1, W_2, ..., W_n$ with zero expectations is uniformly sub-Gaussian, then for all $1 \leq i \leq n$*

$$\mathbb{E}|W_i|^k \leq \frac{k!}{2} m^{k-2} \sigma_0^2, \qquad k = 3, 4, ...$$

Proof From supposition, we know

$$\max_{1 \leq i \leq n} m^2 \left(\mathbb{E}e^{|W_i|^2/m^2} - 1\right) \leq \sigma_0^2$$

Expanding the exponential function gives

$$\max_{1 \leq i \leq n} \mathbb{E}\left[\sum_{k=1}^{\infty} \frac{|W_i|^{2k}}{k! m^{2k-2}}\right] \leq \sigma_0^2$$

Hence,

$$\max_{1\le i\le n}\mathbb{E}\left[\sum_{k\ even,\ k\ge 2}\frac{|W_i|^k}{k!m^{k-2}}\right]\le \max_{1\le i\le n}\mathbb{E}\left[\sum_{k=1}^{\infty}\frac{|W_i|^{2k}}{(2k)!m^{2k-2}}\right]\le\frac{1}{2}\sigma_0^2$$

Therefore, we must have

$$\max_{1\le i\le n}\mathbb{E}|W_i|^k\le\frac{k!}{2}m^{k-2}\sigma_0^2,\qquad k=2,4,...$$

Besides that, note that

$$\max_{1\le i\le n}\mathbb{E}\left[\frac{|W_i|^k}{k!m^{k-2}}\right]\le \max_{1\le i\le n}\mathbb{E}\left[\sum_{k\ even,\ k\ge 2}\frac{|W_i|^k}{k!m^{k-2}}\right]\le\frac{1}{2}\sigma_0^2,\qquad k=3,5,...$$

Again, we see that

$$\max_{1\le i\le n}\mathbb{E}|W_i|^k\le\frac{k!}{2}m^{k-2}\sigma_0^2,\qquad k=3,5,...$$

Finally, conclude that

$$\max_{1\le i\le n}\mathbb{E}|W_i|^k\le\frac{k!}{2}m^{k-2}\sigma_0^2,\qquad k=3,4,...$$

Lemma A2.3 *(Peeling device) Let τ be a strictly increasing function which takes its input from a function class $\mathcal{F}$. Suppose $\{m_s\}_{s=0}^S$ is a strictly increasing sequence such that*

$$\mathcal{F}\subseteq\bigcup_{s=0}^{S}\mathcal{F}_s$$

where

$$\mathcal{F}_s=\{f\in\mathcal{F}:m_{s-1}\le\tau(f)<m_s\}$$

then for a stochastic process Z_n indexed by $\mathcal{F}$, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{|Z_n(f)|}{\tau(f)}>a\right)\le\sum_{s=0}^{S}\mathbb{P}\left(\sup_{f\in\mathcal{F},\ \tau(f)<m_s}|Z_n(f)|>am_{s-1}\right)$$

Proof We have

$$\mathcal{F}\subseteq\bigcup_{s=1}^{S}\mathcal{F}_s$$

hence

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{|Z_n(f)|}{\tau(f)}>a\right)\le\sum_{s=0}^{S}\mathbb{P}\left(\sup_{f\in\mathcal{F}_s}\frac{|Z_n(f)|}{\tau(f)}>a\right)$$

Furthermore, from the definition of $\mathcal{F}_s$ it follows

$$\sum_{s=0}^{S} \mathbb{P}\left(\sup_{f \in \mathcal{F}_s} \frac{|Z_n(f)|}{\tau(f)} > a \right) \leq \sum_{s=0}^{S} \mathbb{P}\left(\sup_{f \in \mathcal{F}_s} \frac{|Z_n(f)|}{m_{s-1}} > a \right)$$

$$= \sum_{s=0}^{S} \mathbb{P}\left(\sup_{f \in \mathcal{F}_s} |Z_n(f)| > a m_{s-1} \right)$$

Theorem A2.4 *(Bernstein inequality) Let $Z_1, Z_2, ..., Z_n$ be random variables with zero expectation such that*

$$\mathbb{E}\,|Z_i|^k \leq \frac{k!}{2} m^{k-2} \sigma_0^2, \qquad k = 2, 3, ... \tag{2.15}$$

then for all $t > 0$,

$$\mathbb{P}\left(\frac{1}{n} \left| \sum_{i=1}^{n} Z_i \right| \geq t \right) \leq 2e^{-(nt^2)/(2(\sigma^2 + mt))}$$

Proof Using Markov's inequality, we get

$$\mathbb{P}\left(\frac{1}{n} \left| \sum_{i=1}^{n} Z_i \right| \geq t \right) = \mathbb{P}\left(\left| \sum_{i=1}^{n} Z_i \right| \geq nt \right)$$

$$\leq \mathbb{P}\left(e^{\lambda \left| \sum_{i=1}^{n} Z_i \right|} \geq e^{\lambda n t} \right) \leq e^{-\lambda n t} \mathbb{E}\left(e^{\lambda \left| \sum_{i=1}^{n} Z_i \right|} \right) \leq e^{-\lambda n t} \mathbb{E}\left(e^{n \lambda \max_{1 \leq i \leq n} |Z_i|} \right)$$

From the power series expansion of the exponential function, we can write for $Z = \max_{1 \leq i \leq n} Z_i$,

$$\mathbb{E}\left(e^{\lambda Z} \right) = 1 + \lambda \mathbb{E} Z + \sum_{k \geq 2} \frac{\lambda^k \mathbb{E} Z^k}{k!}$$

From the supposition, it follows that

$$\mathbb{E}(Z) = \mathbb{E}\left(\sum_{i=1}^{n} Z_i \right) = \sum_{i=1}^{n} \mathbb{E}(Z_i) = 0$$

which means that

$$\mathbb{E}\left(e^{\lambda Z} \right) \leq 1 + \sum_{k \geq 2} \frac{\lambda^k \mathbb{E} Z^k}{k!} \leq 1 + \left(\frac{\lambda^2 \sigma_0^2}{2} \right) \sum_{k \geq 2} (\lambda m)^{k-2}$$

where the last inequality follows from (2.15). Now if we take λ such that $0 < \lambda < 1/m$, then

$$\sum_{k \geq 3} (\lambda m)^{k-2} \leq \frac{\lambda m}{1 - \lambda m}$$

Hence,

$$\mathbb{E}\left(e^{\lambda Z}\right) \leq 1 + \left(\frac{\lambda^2\sigma_0^2}{2}\right)\left(1 + \frac{\lambda m}{1-\lambda m}\right)$$

$$= 1 + \frac{\lambda^2\sigma_0^2}{2(1-\lambda m)} \leq e^{(\lambda^2\sigma_0^2)/(2(1-\lambda m))}$$

where the last inequality follows from the fact that $ln(y) \leq y - 1$ for all $y \in \mathbb{R}$. Take $\lambda = t/(\sigma_0^2 + mt)$, then $\lambda < 1/m$, because

$$\lambda = \frac{t}{\sigma_0^2 + mt} = \frac{1}{\frac{\sigma^2}{t} + m} < \frac{1}{m}$$

Conclude that,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \geq t\right) \leq e^{-(nt^2)/(2(\sigma_0^2+mt))}$$

Do the same with $-Z_1, -Z_2, ..., -Z_n$ and conclude that

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} Z_i\right| \geq t\right) \leq 2e^{-(nt^2)/(2(\sigma_0^2+mt))}$$

Lemma A3.1 *(Portmanteau's lemma) Let $(X_i)_{1\leq i\leq n}$ be a sequence of random variables, then $X_n \xrightarrow{d} X$ for a random variable X if and only if $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all continuous bounded functions f.*

Proof Assume $\lim_{n\to\infty} \mathbb{E}f(\mathbb{X}_n) = \mathbb{E}f(\mathbb{X}_n)$, then define

$$f_{x,\epsilon}(y) = \begin{cases} 1, & y \leq x \\ 0, & y \geq x + \epsilon \\ \frac{y-x}{\epsilon} & x \leq y \leq x + \epsilon \end{cases}$$

$f_{x,\epsilon}$ is continuous, so

$$\lim_{n\to\infty} \sup_n \mathbb{P}(X_n \leq x) \leq \lim_{n\to\infty} \sup_n \mathbb{E}f(\mathbb{X}_n) = \mathbb{E}f(\mathbb{X}_n)$$

$$\leq \mathbb{P}(X \leq x) + \mathbb{P}(x \leq X \leq x + \epsilon)\left(\frac{x + \epsilon - x}{\epsilon}\right) = \mathbb{P}(X \leq x + \epsilon)$$

Let $\epsilon \to 0$, then

$$\lim_{n\to\infty} \sup_n \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x)$$

In the same way, we can proof that

$$\lim_{n\to\infty} \inf_n \mathbb{P}(X_n \leq x) \geq \mathbb{P}(X \leq x)$$

Therefore,

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$$

which means

$$X_n \xrightarrow{d} X$$

We only need to proof the other way, so assume now that $X_n \xrightarrow{d} X$. Let Y_n be a sequence of variables with the same distribution as all X_n, but converges almost surely to Y. f is continuous, so we have $f(Y_n) \to f(Y)$. Applying the bounded convergence theorem gives

$$\mathbb{E}f(X_n) = \mathbb{E}f(Y_n) \to \mathbb{E}f(Y) = \mathbb{E}f(X)$$

This concludes the proof.

Lemma A3.2 *(Portmanteau's second lemma) Let $(X_i)_{1\leq i\leq n}$ be a sequence of random variables, such that*

$$X_n \xrightarrow{d} X \tag{3.16}$$

then

$$\lim_{n\to\infty} \inf_n \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G), \qquad \text{for all open } G \subset \mathbb{R}$$

Proof Condition (3.16) implies that for all uniform bounded functions f, we have

$$\lim_{n\to\infty} \mathbb{P}(f(X_n) \leq x) \geq \mathbb{P}(f(X) \leq x), \qquad \forall x \in \mathbb{R}$$

For a $\epsilon > 0$ and a closed set S, take

$$f(x) = (1 - \frac{\min_{s\in S}|x-s|}{\epsilon})^+$$

where the plus sign means that $f(x)$ can't be less than zero. Note that,

$$|f(x) - f(x)| \leq \frac{\min_{s\in S}|x-s|}{\epsilon}, \qquad \forall x, y \in \mathbb{R}$$

which means f is uniform continuous. f is also bounded,

$$\mathbb{1}_S(x) \leq f(x) \leq \mathbb{1}_{S^\epsilon}(x)$$

where

$$S^\epsilon = \{x : \min_{s\in S}|x-s| \leq \epsilon\}$$

Therefore,

$$\lim_{n\to\infty} \sup_n \mathbb{P}(X_n \in S) = \lim_{n\to\infty} \sup_n \mathbb{P}(\mathbb{1}_S(X_n) = 1)$$
$$\leq \lim_{n\to\infty} \sup_n \mathbb{P}(f(X_n) = 1) = \mathbb{P}(f(X) = 1) \leq \mathbb{P}(X \in S^\epsilon)$$

Letting $\epsilon \to 0$ gives

$$\lim_{n\to\infty} \sup_n \mathbb{P}(X_n \in S) \leq \mathbb{P}(X \in S)$$

Taking the open set G as the complement of S concludes the proof.

References

[1] "A new derivation of Stirling's approximation of n!" In: (1990).

[2] Jon A. Wellner Aad W. van der Vaart. *Weak Convergence and Empirical Processes: With Applications to Statistics*. 1996.

[3] Patrick Billingsley. *Convergence of Probability Measures*. 1999.

[4] Rick Durret. *Probability: Theory and Examples*. 2019. URL: https://services.math.duke.edu/~rtd/PTE/PTE5_011119.pdf.

[5] Joel Zinn Evarist Gine. *Empirical Processes Indexed by Lipschitz Functions*. 1986.

[6] Joel Zinn Evarist Gine. *Gaussian Characterization of Uniform Donsker Classes of Functions*. 1991.

[7] Joel Zinn Evarist Giné. "The Central Limit Theorem for Empirical Processes Under Local Conditions: The Case of Radon Infinitely Divisible Limits without Gaussian Component". In: (1988).

[8] Richard Nickl Evarist Giné. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. 2016.

[9] Sara van de Geer. *Empirical Processes in M-estimation*. 2000.

[10] Wassily Hoeffding. *Probability inequalities for sums of bounded random variables*. 1963.

[11] Oliver C. Ibe. *Markov Processes for Stochastic Modeling*. Second Edition. 2013.

[12] Jon A. Wellner J. Dehardt. *Generalizations of the Glivenko-Cantelli Theorem*. 1971.

[13] David Pollard. *Convergence of stochastic processes*. 1984.

[14] Jon Wellner Qiyang Han. "Convergence rates of least squares regression estimators with heavy-tailed errors". In: (2019). URL: https://sites.stat.washington.edu/jaw/JAW-papers/jaw-han.aos.2019.pdf.

[15] Bodhisattva Sen. *A Gentle Introduction to Empirical Process Theory and Applications*. 2021. URL: http://www.stat.columbia.edu/~bodhi/Talks/Emp-Proc-Lecture-Notes.pdf.

[16] Howard G. Tucker. *A Generalization of the Glivenko-Cantelli Theorem*. 1959.

[17] Aad W. van der Vaart. *New Donsker classes*. 1996.

[18] Pierre Youssef. *Bennett-Bernstein inequality and the spectral gap of random regular graphs*. URL: https://www.lpsm.paris/pageperso/youssef/recherche_files/Mini-course-fanar.pdf.