Lecture 4: Convergence and Consistency

Tianpei Xie

Jul. 26th., 2015

Contents

1 Recall: Modes of Convergence

1.1 Definitions

- **Remark** (***Two Basic Modes of Convergence***) [Royden and Fitzpatrick, 1988, Tao, 2011]

  1. **Definition** (***Pointwise Convergence***)
     We say that f_n converges to f ***pointwise*** if, for any $x \in X$ and $\epsilon > 0$, there exists $N > 0$ (*that **depends** on ϵ and x*) such that for all $n \geq N$, $|f_n(x) - f(x)| \leq \epsilon$. Denoted as $f_n(x) \to f(x)$.

  2. **Definition** (***Uniform Convergence***)
     We say that f_n converges to f ***uniformly*** if, for any $\epsilon > 0$, there exists $N > 0$ (*that **depends** on ϵ only*) such that for all $n \geq N$, $|f_n(x) - f(x)| \leq \epsilon$ for every $x \in X$. Denoted as $f_n \to f$, *uniformly*.

     Unlike pointwise convergence, the time N at which $f_n(x)$ must be permanently ϵ-close to $f(x)$ is not permitted to depend on x, but must instead be chosen *uniformly* in x.

- **Remark** (***Modes of Convergence of Measurable Functions***)
  When the domain X is equipped with the structure of a measure space $(X, \mathscr{B}, \mu)$, and the functions f_n (and their limit f) are measurable with respect to this space. In this context, we have some *additional modes of convergence*:

  1. **Definition** (***Pointwise Almost Everywhere Convergence***)
     We say that f_n converges to f ***pointwise almost everywhere*** if, for μ***-almost everywhere*** $x \in X$, $f_n(x)$ converges to $f(x)$. It is denoted as $f_n \overset{a.e.}{\to} f$. In probabilty, it is called ***almost sure convergence*** or ***convergence with probability*** 1. It is denoted as $f_n \overset{a.s.}{\to} f$.

     In other words, there exists ***a null set*** E, $(\mu(E) = 0)$ such that for *any* $x \in X \setminus E$ and any $\epsilon > 0$, there exists $N > 0$ (*that **depends** on ϵ and x*) such that for all $n \geq N$, $|f_n(x) - f(x)| \leq \epsilon$.

  2. **Definition** (***Uniformly Almost Everywhere Convergence***) [Tao, 2011]
     We say f_n converges to f ***uniformly almost everywhere***, ***essentially uniformly***, or ***in L^∞ norm*** if, for every $\epsilon > 0$, there exists N such that for every $n \geq N$, $|f_n(x) - f(x)| \leq \epsilon$, ***for*** μ***-almost every*** $x \in X$.

     That is, $f_n \to f$ *uniformly* in $x \in X \setminus E$, for some E with $\mu(E) = 0$.

     We can also formulate in terms of L^∞ ***norm*** as

     $$\|f_n(x) - f(x)\|_{L^\infty(X)} \overset{n\to\infty}{\longrightarrow} 0,$$

     where $\|f\|_{L^\infty(X)} = \operatorname{ess\,sup}_x |f(x)| \equiv \inf_{\{E:\mu(E)=0\}} \sup_{x \in X \setminus E} |f(x)|$ is the ***essential bound***. It is denoted as $f_n \overset{L^\infty}{\to} f$.

  3. **Definition** (***Almost Uniform Convergence***) [Tao, 2011]
     We say that f_n converges to f ***almost uniformly*** if, for every $\epsilon > 0$, there exists an ***exceptional set*** $E \in \mathscr{B}$ of *measure* $\mu(E) \leq \epsilon$ such that f_n converges ***uniformly*** to f on the ***complement*** of E.

That is, for arbitrary δ there exists some E with $\mu(E) \le \delta$ such that $f_n \to f$ *uniformly* in $x \in X \setminus E$.

4. **Definition** (***Convergence in L^1 Norm***)
We say that f_n converges to f ***in L^1 norm*** if the quantity

$$\|f_n - f\|_{L^1(X)} = \int_X |f_n(x) - f(x)|\, d\mu \overset{n\to\infty}{\longrightarrow} 0.$$

In probability theory, it is called the ***convergence in mean***. Denoted as $f_n \overset{L^1}{\to} f$.

5. **Definition** (***Convergence in Measure***)
We say that f_n converges to f ***in measure*** if, for every $\epsilon > 0$, the measures

$$\mu\left(\{x \in X : |f_n(x) - f(x)| \ge \epsilon\}\right) \overset{n\to\infty}{\longrightarrow} 0.$$

Denoted as $f_n \overset{\mu}{\to} f$.

In probability theory, it is called ***convergence in probability*** and is denoted as $f_n \overset{p}{\to} f$.

1.2 Modes of Convergence via Tail Support and Width

- **Remark** (***Tail Support and Width***)

Definition Let $E_{n,m} := \{x \in X : |f_n(x) - f(x)| \ge 1/m\}$. Define ***the N-th tail support set***

$$T_{N,m} := \{x \in X : |f_n(x) - f(x)| \ge 1/m,\ \ \exists n \ge N\} = \bigcup_{n \ge N} E_{n,m}.$$

Also let $\mu(E_{n,m})$ be the ***width*** of n-th event $\mathbb{1}\{E_{n,m}\}$. Note that $T_{N,m} \supseteq T_{N+1,m}$ is ***monotone nonincreasing*** and $T_{N,m} \subseteq T_{N,m+1}$ is ***monotone nondecreasing***.

1. The ***pointwise convergence*** of f_n to f indicates that for every x, every $m \ge 1$, there exists some $N \equiv N(m,x) \ge 1$ such that $T^c_{N,m} \ni x$ or $T_{N,m} \not\ni x$. Equivalently, ***the tail support shrinks to emptyset***:

$$\bigcap_{N \in \mathbb{N}} T_{N,m} = \lim_{N\to\infty} T_{N,m} = \limsup_{n\to\infty} E_{n,m} = \emptyset, \quad \text{for all } m.$$

2. The ***pointwise almost everywhere convergence*** indicates that there exists ***a null set*** F with $\mu(F) = 0$ such that for every $x \in X \setminus F$ and any $m \ge 1$, there exists some $N \equiv N(m,x) \ge 1$ such that $(T_{N,m} \setminus F) \not\ni x$. Equivalently, ***the tail support shrinks to a nulll set***. Note that it makes no assumption on $(T_{N,m} \cap F)$.

$$\begin{aligned} & \lim_{N\to\infty} T_{N,m} \setminus F = \limsup_{n\to\infty} E_{n,m} \setminus F = \emptyset, \quad \text{for all } m. \\ \Leftrightarrow\ & \bigcap_{N \in \mathbb{N}} T_{N,m} = \lim_{N\to\infty} T_{N,m} = F \\ \Leftrightarrow\ & \mu\left(\lim_{N\to\infty} T_{N,m}\right) = \mu\left(\bigcap_{N \in \mathbb{N}} T_{N,m}\right) = 0 \end{aligned}$$

3. The ***uniform convergence*** indicates that for each $m \geq 1$, there exists some $N(m) \geq 1$ (not depending on x) such that $T_{N,m} = \emptyset$. (i.e. $T_{N,m} \not\ni x$ for all $x \in X$.) So ***the tail support is an empty set***

4. The ***uniformly almost everywhere convergence*** indicates that there exists some null set F with $\mu(F) = 0$ such that for each $m \geq 1$, there exists some $N(m) \geq 1$ (not depending on x) such that $(T_{N,m} \setminus F) = \emptyset$. (i.e. $T_{N,m} \not\ni x$ for all $x \in X \setminus F$.) Equivalently, ***the tail support is a null set***:

$$
\begin{aligned}
& T_{N,m} = F \\
\Leftrightarrow \quad & \mu\left(T_{N,m}\right) = 0
\end{aligned}
$$

5. The ***almost uniform convergence*** indicates that for every δ, there exists *some measurable set* F_δ with $\mu(F_\delta) < \delta$ such that for each $m \geq 1$ there exists some $N(m) \geq 1$ (not depending on x) such that $(T_{N,m} \setminus F_\delta) = \emptyset$. (i.e. $T_{N,m} \not\ni x$ for all $x \in X \setminus F_\delta$.) Equivalently, ***the measure of tail support shrinks to zero***:

$$
\begin{aligned}
\mu\left(T_{N,m}\right) \leq \delta \quad &\Leftrightarrow \quad T_{N,m} = F_\delta \\
\lim_{N \to \infty} \mu\left(T_{N,m}\right) = 0 &
\end{aligned}
$$

6. The ***convergence in measure*** indicates that for any $m \geq 1$ and any $\delta > 0$, there exists $N \equiv N(m, \delta) \geq 1$ such that for all $n \geq N$, *the* ***width*** *of n-th event* ***shrinks to zero***:

$$
\begin{aligned}
& \mu(E_{n,m}) \leq \delta \\
\lim_{n \to \infty} & \mu(E_{n,m}) := \lim_{n \to \infty} \mu\left(\{x \in X : |f_n(x) - f(x)| \geq \epsilon\}\right) = 0
\end{aligned}
$$

1.3 Relationships between Different Modes of Convergence

- **Remark** This diagram shows the *relative strength* of different *modes of convergence*. The direction arrows $A \to B$ means "if A holds, then B holds".

$$
\begin{tikzcd}
\textit{uniform} \arrow[r] \arrow[dd] \arrow[rr, bend left] & \textit{uniformly a.e.} \arrow[r, leftrightarrow] \arrow[d] & \textit{in } L^\infty \textit{ norm} \arrow[dl] \\
& \textit{almost uniform} \arrow[d] \arrow[dr] & \textit{in } L^1 \textit{ norm} \arrow[d] \\
\textit{pointwise} \arrow[r] & \textit{pointwise a.e.} & \textit{in measure}
\end{tikzcd}
$$

Moreover, here are some counter statements:

- $L^\infty \not\to L^1$: see the "*Escape to Width Infinity*" example below.
- ***uniform*** $\not\to L^1$: see the "*Escape to Width Infinity*" example below.
- $L^1 \not\to$ ***uniform*** : see the "*Typewriter Sequence*" example below.

- ***pointwise*** $\not\to L^1$: see the "*Escape to Horizontal Infinity*" example below.
- ***pointwise*** $\not\to$ ***uniform***: see the "$f_n = x/n$" example above.
- For finite measure space, ***pointwise a.e.*** $\to$ ***almost uniform***: see the Egorov's theorem.
- ***almost uniform*** $\not\to L^1$: see the "*Escape to Vertical Infinity*" example below.
- ***almost uniform*** $\not\to L^\infty$: see the "*Escape to Vertical Infinity*" example below. The *converse* is true, however.
- For bounded $f_n \le G, a.e.\ \forall n$, then ***pointwise a.e.*** $\to L^1$: see *Dominated Convergence Theorem*.
- $L^1 \not\to$ ***pointwise a.e.*** : see the "*Typewriter Sequence*" example below.
- ***in measure*** $\not\to$ ***pointwise a.e.*** : see the "*Typewriter Sequence*" example below.
- $L^1 \to$ ***convergence in integral***: by *triangle inequality*. Note that *the other modes of convergence does* ***not directly*** *lead to convergence in integral*.

- **Remark** For finite measure space such as *the probability space*,

$$
\begin{tikzcd}
\textit{uniform} \arrow[r] \arrow[rr, bend left] \arrow[dd] & \textit{uniformly a.e.} \arrow[r, leftrightarrow] \arrow[d] & \textit{in } L^\infty \textit{ norm} \arrow[dl] \arrow[d, red] \\
& \textit{almost uniform} \arrow[d] \arrow[dr] & \textit{in } L^1 \textit{ norm} \arrow[d] \\
\textit{pointwise} \arrow[r] & \textit{pointwise a.e.} \arrow[u, red, bend left] & \textit{in measure}
\end{tikzcd}
$$

1.4 Comparison

Table 1: Comparison of Modes of Convergence

	tail support	***width***	***maximum variation***	***subgraph***
definition	$T_{N,\epsilon} = \bigcup_{n \geq N} E_{n,\epsilon}$	$\mu(E_{n,\epsilon})$	$\sup_{x \in X}\{\lvert f_n(x) - f(x)\rvert\}$	$\Gamma(f_n) = \{(x,t) : 0 \leq t \leq f_n(x)\}$
pointwise	$\bigcap_{N=1}^{\infty} T_{N,\epsilon} = \emptyset$		*or*, $\to 0$ on X	
point-wise a.e.	$\mu\left(\bigcap_{N=1}^{\infty} T_{N,\epsilon}\right) = 0$		*or*, $\to 0$ on $X \setminus E$	
uniform	$T_{N,\epsilon} = \emptyset$		equivalently, $\to 0$ on X	
uniform a.e.* / L^∞ *norm	$\mu\left(T_{N,\epsilon}\right) = 0$		equivalently, $\to 0$ on $X \setminus E$	
almost uniform	$\lim_{N\to\infty} \mu\left(T_{N,\epsilon}\right) = 0$		or, $\to 0$ on $X \setminus E$	
in measure		$\lim_{n\to\infty} \mu\left(E_{n,\epsilon}\right) = 0$	or, $\to 0$ on $X \setminus E$	
L^1 ***norm***			$\to 0$ and support fixed or non-increasing	area of $\Gamma(f_n) = \mathcal{A}(\Gamma(f_n))$ $\lim_{n\to\infty} \mathcal{A}(\Gamma(\lvert f_n - f\rvert)) = 0$

2 Consistency

2.1 Weak and Strong Consistency

- **Definition** (***Weak Consistency***) [Lehmann and Casella, 1998, Resnick, 2013]
  Suppose $X_1, \cdots, X_n, \cdots$ are *i.i.d. random variables* on $(\Omega, \mathscr{F}, \mathcal{P}_\theta)$ with $\theta \in \Theta$ being parameter of distribution $\mathcal{P}_\theta$. Let the ***estimand*** be $g(\theta)$ and the estimator be $\delta_n \equiv \delta_n (X_1, \ldots, X_n)$, which is also a random variable.

  A sequence of estimator δ_n of $g(\theta)$ is ***(weak) consistent*** if for every $\theta \in \Theta$,

  $$\delta_n \xrightarrow{p} g(\theta).$$

  i.e. δ_n ***converges to*** $g(\theta)$ ***in probability*** for every parameter θ.

- **Remark** In other word, the statistic

  $$\widehat{g}_n := \widehat{g}_n(X_1, \ldots, X_n)$$

  is ***consistent***, if for every $\theta \in \Theta$, any $\epsilon, \delta > 0$, $\exists N = N(\epsilon, \delta) \in \mathbb{N}$, such that for all $n \geq N$

  $$\mathcal{P}\left(\{\omega \in \Omega : |\widehat{g}_n(\omega, \theta) - g(\theta)| \geq \epsilon\}\right) < \delta.$$

- In constrasts, we can define the strong consistency based on *pointwise almost everywhere (almost sure) convergence.*

  Definition (***Strong Consistency***) [Lehmann and Casella, 1998, Resnick, 2013]
  A sequence of estimator δ_n of $g(\theta)$ is ***strong consistent*** if for every $\theta \in \Theta$,

  $$\delta_n \xrightarrow{a.s.} g(\theta).$$

  i.e. δ_n ***converges to*** $g(\theta)$ ***almost surely*** for every parameter θ.

- **Remark** In other word, the statistic

  $$\widehat{g}_n := \widehat{g}_n(X_1, \ldots, X_n)$$

  is ***strong consistent***, if for every $\theta \in \Theta$, any $\epsilon > 0$,

  $$\mathcal{P}\left(\bigcap_{N \geq 1} \{\omega \in \Omega :\ \exists n \geq N \text{ such that } |\widehat{g}_n(\omega, \theta) - g(\theta)| \geq \epsilon\}\right) = 0.$$
  $$\Rightarrow \mathcal{P}\left(\limsup_{n \to \infty} \{\omega \in \Omega :\ |\widehat{g}_n(\omega, \theta) - g(\theta)| \geq \epsilon\}\right) = 0.$$

- **Remark** (***Consistency*** $=$ ***Asymptotic Analysis***)
  The ***consistency*** property of a statistic is based on ***the asymptotic analysis*** of the $\mathscr{F}$-measurable functions $(X_1, X_2, \ldots)$. It states that ***the statistical estimator*** will ***converge*** to ***the true value*** of the ***estimand***, given ***infinite amount*** *of data.* So the consistency statement is to say that *the estimator will* ***reveal the ground truth*** *given enough data.*

  The ***asymptotic random variable*** such as $(\limsup_n X_n)$, $(\liminf_n X_n)$, $\lim\limits_{n \to \infty} \frac{S_n}{n}$ are all ***tail random variables***, i.e. they are measurable with respect to *tail σ-algebra* $\mathscr{T}$. They tends to behave regularly given that all samples are i.i.d.

- We distinguish the consistency with the unbiasedness

  Definition (***Unbiasedness***)
  An estimator $\delta(X)$ of $g(\theta)$ is ***unbiased*** if

$$\mathbb{E}_{\mathcal{P}_\theta}\left[\delta(X)\right] = g(\theta), \quad \forall\, \theta \in \Theta$$

- **Remark** The statistic is unbiased if it fits a ***linear functional equation***

$$\int_{\Omega} \delta(X(\omega))\, d\mathcal{P}_\theta(\omega) = g(\theta), \quad \forall \theta \in \Theta$$

- **Remark** (***Consistency vs. Unbiasedness***)
  In general, there is ***no direct relationship*** between *consistency* and *unbiaseness*:

  - An *estimator* is ***unbiased*** if it is ***centered*** around *the true value*. It does not guarantee that when the sample size increases, the estimator ***itself*** will ***converge*** to its ***mean value*** $g(\theta)$ for any choice of $\theta \in \Theta$.

    For instance, the samples (X_n) are *independent uniformly distributed* in *a unit circle centered at* 0, each sample is an *estimator* of constant $\theta_0 = 0$. The expected value of X_n is 0 so each X is ***unbiased***. But X_n *does not converge to* 0 in any sense.

  - An *estimator* is ***consistent*** if it ***converges to*** *the true value*. It does not guarantee that (X_n) are *distributed around* the true value for each n.

    For instance, the sample $X_n(\omega) = \frac{1}{n}\omega$. We see that $X_n \to 0$ almost surely, but

$$\int_{\Omega} \left[X_n(\omega) - g(\theta)\right] d\mathcal{P}_\theta(\omega) = \frac{1}{n}$$

    is ***nonzero*** for each n. So (X_n) is ***consistent*** but ***biased***.

  In other word, the ***unbiasedness*** is about *the* ***distribution*** *of estimators* $\{\widehat{g}_n\}$, while the ***consistency*** is about *the* ***trends*** *of estimators* $\widehat{g}_n$.

- **Remark** (***Consistency is just Approximation***)
  The notion of consistency *describes the* ***ideal behavior*** *of estimator*. But a *consistency performance* just provides ***an approximation in theory***, since most asymptotic result only works when $n \to \infty$, which is ***impractical***.

- **Theorem 2.1** ***(Gilvenko-Cantelli Theorem)****[Devroye et al., 2013, Resnick, 2013]*
  Let $Z_1, \ldots, Z_n$ be i.i.d. real valued random variables with ***distribution functional*** $F(\lambda) = \mathcal{P}\left[Z \le \lambda\right]$. *Denote* ***the standard empirical distribution functional*** *by*

$$\widehat{F}_n(\lambda) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{[Z_i \le \lambda]\right\}.$$

  Then for any $\lambda \in \mathbb{R}$,

$$\widehat{F}_n(\lambda) \overset{a.s.}{\to} F(\lambda),$$

  that is, $\widehat{F}_n(\lambda)$ is ***strongly consistent***.

- **Remark** It is shown that

$$\mathcal{P}\left\{\sup_{\lambda\in\mathbb{R}}\left|F(\lambda)-\widehat{F}_n(\lambda)\right|\geq\epsilon\right\}\leq 8(n+1)\exp\left(-n\epsilon^2/32\right),$$

and, in particular, by *the Borel-Cantelli lemma*,

$$\mathcal{P}\left(\limsup_{n\to\infty}\left\{\omega:\sup_{\lambda\in\mathbb{R}}\left|F(\lambda)-\widehat{F}_n(\lambda,\omega)\right|\geq\epsilon\right\}\right)=0$$

Note that we may define $\nu(A)=\mathcal{P}\circ Z^{-1}(A)$ as the induced measure on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$, and let $\mathcal{A}=\{(-\infty,\lambda],\lambda\in\mathbb{R}\}$, then it is equivalent to

$$\mathcal{P}\left\{\sup_{A\in\mathcal{A}\subset\mathscr{B}}|\nu(A)-\nu_n(A)|\geq\epsilon\right\}\leq 8(n+1)\exp\left(-n\epsilon^2/32\right)$$

2.2 Consistency in Statistical Learning Theory

- Finally, we introduce similar notion of consistency used in *statistical learning*.

  Definition (***Bayes Error***) [Devroye et al., 2013]
  Given a sequence of i.i.d. training variables $\mathcal{D}_n\equiv\{(X_i,Y_i),1\leq i\leq n\}$ on $(\Omega,\mathscr{F},\mathcal{P})$ and a sequence of ***classification rules*** $g_n\equiv g_n(X;X_1,\ldots,X_n)$ such that the error probability is defined as

$$L_n\equiv\mathcal{P}\left\{g_n(X,\mathcal{D}_n)\neq Y|\mathcal{D}_n\right\}.$$

  The optimal error probability is given by the ***Bayes error***

$$L^*=\inf_g\mathcal{P}\left\{g(X)\neq Y\right\}.$$

- **Definition** (***Consistent Classification Rules***)
  A classification rule is ***consistent (asymoptotically Bayes-risk efficient) for a certain distribution*** $\mathcal{P}(X,Y)$ if

$$\mathbb{E}_{\mathcal{P}}\left[L_n(g_n)\right]\equiv\mathcal{P}\left\{g_n(X,\mathcal{D}_n)\neq Y\right\}\to L^*,\quad\text{as } n\to\infty$$

  Since $1\geq L_n\geq L^*$, the above is equivalent to ***convergence in probability***

$$\lim_{n\to\infty}\mathcal{P}\left\{L_n(g_n)-L^*\geq\epsilon\right\}=0.$$

  Also the classification rule is the ***strongly consistent*** if

$$L_n\to L^*\;\;a.s.$$

- **Remark** Note that for bounded continuous L_n, $\mathbb{E}_{\mathcal{P}}\left[L_n(g_n)\right]\to L^*$, as $n\to\infty$ means that $L_n\rightsquigarrow L^*$ in distribution. It implies convergence in probability since the limiting random variable L^* is a constant.

- A stronger version of consistency when the underlying distribution $\mathcal{P}$ is unknown

Definition (***Universal Consistency***)
A sequence of *classification rules* is called ***universally consistent (strongly) consistent*** if it is ***(strongly) consistent*** for ***any distribution*** $\mathcal{P}(X,Y)$, i.e.

$$\lim_{n\to\infty} \mathcal{P}\left\{L_n(g_n) - L^* \geq \epsilon\right\} = 0, \quad \forall\, \mathcal{P} \text{ on } (\Omega, \mathscr{F})$$

and

$$\mathcal{P}\left\{\limsup_{n\to\infty}\left\{L_n - L^* \geq \epsilon\right\}\right\} = 0, \quad \forall\, \mathcal{P} \text{ on } (\Omega, \mathscr{F}).$$

3 Laws of Large Number

3.1 Weak Laws of Large Number

•

3.2 Strong Laws of Large Number

•

3.3 Fisher Consistency

- A different type of consistency is *Fisher consistency*:

  Definition ***Fisher Consistency***
  Suppose $X_1, \cdots, X_n, \cdots$ are i.i.d. random variables on $(\Omega, \mathscr{F}, \mathcal{P}_\theta)$ with $\theta \in \Theta$ being parameter of distribution $\mathcal{P}_\theta$. If an estimator of $g(\theta)$, $\delta_n \equiv \delta_n(X_1, \ldots, X_n)$ can be represented as ***functional of empirical distributions***

  $$\widehat{\mathcal{P}}_n X(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{X_i \leq \lambda\right\}$$

  such as $\delta'_n \equiv \delta'(\widehat{\mathcal{P}}_n X)$. Then the estimator δ'_n of $g(\theta)$ is ***Fisher consistent*** if

  $$\delta'\left(\mathcal{P}_\theta\right) = g(\theta)$$

  or equivalently under the strong law of large numbers, $\widehat{\mathcal{P}}_n X(\lambda) \to \mathcal{P}_\theta(\lambda)$ ***almost surely***, so

  $$\delta'\left(\lim_{n\to\infty} \widehat{\mathcal{P}}_n X\right) = g(\theta)$$

- **Remark** As long as the X_i are exchangeable, an estimator δ defined in terms of the X_i can be converted into an estimator δ' that can be defined in terms of $\widehat{\mathcal{P}}_n$ by averaging δ over all permutations of the data. The resulting estimator will have the same expected value as δ and its variance will be no larger than that of δ.

4 Weak Convergence

4.1 Definitions

- **Remark** (***Weak*** * ***Convergence***)
  Convergence in distribution is also called ***weak convergence*** in probability theory [Folland, 2013]. In general, we can see that it is actually ***not*** *a mode of* ***convergence of random variables*** X_n ***itself*** but instead is *the* ***convergence*** *of* ***their distributions*** $\int f d\mu_n$. Equivalently, it is *the* ***convergence of probability measures*** $\mathcal{P}_{X_n} = \mathcal{P} \circ X_n^{-1}$ *on* $\mathcal{B}(\mathbb{R})$.

  Note that in functional analysis, however, ***weak convergence*** is actually for a different mode of convergence (i.e. $\int f_n d\mu \to \int f d\mu$ for all $\mu \in \mathcal{M}(X)$), while ***the convergence in distribution*** is ***the weak*** * ***convergence***.

  Definition (***Weak*** * ***Topology on Banach Space***)
  Let X be a *normed vector space* and X^* be its dual space. The ***weak*** * ***topology*** on X^* is *the weakest topology* on X^* so that $f(x)$ *is* ***continuous for all*** $x \in X$.

  The weak* *topology* on space of regular Borel measures $\mathcal{M}(X) \simeq (\mathcal{C}_0(X))^*$ on a ***compact Hausdorff*** space X, is often called ***the vague topology***. Note that $\mu_n \xrightarrow{w^*} \mu$ if and only if $\int f d\mu_n \to \int f d\mu$ for all $f \in \mathcal{C}_0(X)$.

- **Definition** (***Cumulative Distribution Function***) [Van der Vaart, 2000]
  Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a probability space. Given any real-valued measurable function $\xi : \Omega \to \mathbb{R}$, we define the ***cumulative distribution function*** $F : \mathbb{R} \to [0, \infty]$ of ξ to be the function

$$F_\xi(\lambda) := \mathcal{P}\left(\{\omega \in \Omega : \xi(\omega) \leq \lambda\}\right) = \int_X \mathbb{1}\left\{\xi(\omega) \leq \lambda\right\} d\mathcal{P}(\omega).$$

- **Definition** (***Converge in Distribution***) [Van der Vaart, 2000]
  Let $\xi_n : \Omega \to \mathbb{R}$ be a sequence of real-valued *measurable functions*, and $\xi : \Omega \to \mathbb{R}$ be another measurable function. We say that ξ_n ***converges in distribution*** to ξ if *the cumulative distribution function* $F_n(\lambda)$ of ξ_n ***converges pointwise*** to the cumulative distribution function $F(\lambda)$ of ξ at all $\lambda \in \mathbb{R}$ for which F is continuous. Denoted as $\xi_n \xrightarrow{F} \xi$ or $\xi_n \xrightarrow{d} \xi$ or $\xi_n \rightsquigarrow \xi$.

$$\xi_n \xrightarrow{d} \xi \;\Leftrightarrow\; F_n(\lambda) \to F(\lambda), \text{ for all } \lambda \in \mathbb{R}$$

- **Theorem 4.1** ***(Portmanteau Theorem)****. [Van der Vaart, 2000]*
  For any random vectors X_n *and* X *the followings are* ***equivalent***

  1. $\mathcal{P}\{X_n \leq \lambda\} \to \mathcal{P}\{X \leq \lambda\}$ *for all* ***continuity point*** $\lambda \mapsto \mathcal{P}\{X \leq \lambda\}$*;*
  2. $\mathbb{E}_{\mathcal{P}}[f(X_n)] \to \mathbb{E}_{\mathcal{P}}[f(X)]$ *for all* ***bounded, continuous*** *function* f*;*
  3. $\mathbb{E}_{\mathcal{P}}[f(X_n)] \to \mathbb{E}_{\mathcal{P}}[f(X)]$ *for all bounded, Lipschitz continuous function* f*;*
  4. $\liminf_{n\to\infty} \mathbb{E}_{\mathcal{P}}[f(X_n)] \geq \mathbb{E}_{\mathcal{P}}[f(X)]$ *for all nonnegative continuous function* f*;*
  5. $\liminf_{n\to\infty} \mathcal{P}\{X_n \in G\} \geq \mathcal{P}\{X \in G\}$ *for every open set* G*;*
  6. $\limsup_{n\to\infty} \mathcal{P}\{X_n \in F\} \leq \mathcal{P}\{X \in F\}$ *for every closed set* F*;*

7. *$\mathcal{P}\{X_n \in B\} \to \mathcal{P}\{X \in B\}$ for all Borel sets B with $\mathcal{P}\{X \in \delta B\} = 0$, where $\delta B = \overline{B} - int(B)$ is the boundary of B.*

Proof: 1. 1) $\Rightarrow$ 2) Assume that the distribution function F_X of X is continuous. Then condition 1) implies that $\mathcal{P}\{X_n \in I\} \to \mathcal{P}\{X \in I\}$ for any box $I \in \mathbb{R}^d$. Choose I be sufficiently large and compact, so that $\mathcal{P}(X \notin I) < \epsilon$. A continuous function f is uniformly continuous on compact I and $I = \bigcup_{k=1}^n I_k$ has partition into finitely many boxes I_k such that f varies at most ϵ in I_k.

   Define a simple function $f_\epsilon(x) = \sum_{k=1}^n f(x_k)\mathbb{1}\{I_k\}$ where $x_k \in I_k$ is arbitrary chosen. Then $|f - f_\epsilon| < \epsilon$ for $x \in I$, given that f is bounded e.g. within $[-1, 1]$.

$$
\begin{aligned}
|\mathbb{E}_{\mathcal{P}}[f(X_n)] - \mathbb{E}_{\mathcal{P}}[f_\epsilon(X_n)]| &\le \epsilon + \mathcal{P}\{X_n \notin I\} \\
|\mathbb{E}_{\mathcal{P}}[f(X)] - \mathbb{E}_{\mathcal{P}}[f_\epsilon(X)]| &\le \epsilon + \mathcal{P}\{X \notin I\} < 2\epsilon
\end{aligned}
$$

   For sufficiently large n, the right side of the first equation is smaller than 2ϵ as well (convergence in distribution). We combine this with

$$
\begin{aligned}
|\mathbb{E}_{\mathcal{P}}[f_\epsilon(X_n)] - \mathbb{E}_{\mathcal{P}}[f_\epsilon(X)]| &\le \sum_{k=1}^n |f(x_k)|\,|\mathcal{P}\{X_n \in I_k\} - \mathcal{P}\{X \in I_k\}| \\
&\to 0
\end{aligned}
$$

   together with the triangle inequality we can get $|\mathbb{E}_{\mathcal{P}}[f(X_n)] - \mathbb{E}_{\mathcal{P}}[f(X)]|$ is bounded by 5ϵ eventually for any $\epsilon > 0$, so the result hold. ∎

2. 1) to 3) is similar to 1) to 2).

3. 3) to 5) For every open set G there exists a sequence of Lipschitz functions with $0 \le f_m \uparrow \mathbb{1}\{G\}$. For instance $f_m = \min\{1, m\, d(x, G^c)\}$. For every fixed m, by assumption on convergence in expectation,

$$
\liminf_{n\to\infty} \mathcal{P}\{X_n \in G\} \ge \liminf_{n\to\infty} \mathbb{E}_{\mathcal{P}}[f_m(X_n)] = \mathbb{E}_{\mathcal{P}}[f_m(X)].
$$

   As $m \to \infty$, the RHS increases to $\mathcal{P}\{X \in G\}$ by monotone convergence theorem. ∎

4. 5) to 6) take the complements.

5. 5) + 6) to 7). Let $\mathrm{int}(B)$ and $\overline{B}$ be the interior and closure of B, respectively. By 5) and 6)

$$
\begin{aligned}
\mathcal{P}\{X \in \mathrm{int}(B)\} &\le \liminf_{n\to\infty} \mathcal{P}\{X_n \in \mathrm{int}(B)\} \\
&\le \limsup_{n\to\infty} \mathcal{P}\left\{X_n \in \overline{B}\right\} \\
&\le \mathcal{P}\left\{X \in \overline{B}\right\}.
\end{aligned}
$$

   If $\mathcal{P}\{X \in \delta B\} = 0$ then the LHS and RHS will be equal. Note that by remark below, we can almost find such B in practice. The probability $\mathcal{P}\{X \in B\} = \lim_{n\to\infty} \mathcal{P}\{X_n \in B\}$, since they lies in between these inequalities.

6. 7) to 1) Each cell $(-\infty, x]$ such that x is a continuity point of $x \mapsto \mathcal{P}\{X \le x\}$ is a continuity set. Then the convergence results follows as a specification $B \equiv (-\infty, \lambda]$.

7. 4) to 2). Given any f is bounded, continuous, we need to prove that $\mathbb{E}_{\mathcal{P}}[f(X)] \geq \limsup_{n\to\infty} \mathbb{E}_{\mathcal{P}}[f(X_n)]$ and $\mathbb{E}_{\mathcal{P}}[f(X)] \leq \liminf_{n\to\infty} \mathbb{E}_{\mathcal{P}}[f(X_n)]$.

   Note that $(\sup\{f(x)\} - f)$ and $(f - \inf\{f(x)\})$ are nonnegative, bounded continuous. Then

$$
\begin{aligned}
\mathbb{E}_{\mathcal{P}}[(\sup\{f(x)\} - f(X))] &\leq \liminf_{n\to\infty} \mathbb{E}_{\mathcal{P}}[(\sup\{f(x)\} - f(X_n))] \\
\Rightarrow \sup\{f(x)\} - \mathbb{E}_{\mathcal{P}}[f(X))] &\leq \sup\{f(x)\} + \liminf_{n\to\infty} \mathbb{E}_{\mathcal{P}}[-f(X_n)] \\
\mathbb{E}_{\mathcal{P}}[f(X))] &\geq \limsup_{n\to\infty} \mathbb{E}_{\mathcal{P}}[f(X_n)]
\end{aligned}
$$

   similarly

$$
\begin{aligned}
\mathbb{E}_{\mathcal{P}}[(f(X) - \inf\{f(x)\})] &\leq \liminf_{n\to\infty} \mathbb{E}_{\mathcal{P}}[(f(X_n) - \inf\{f(x)\}))] \\
\Rightarrow -\inf\{f(x)\} + \mathbb{E}_{\mathcal{P}}[f(X))] &\leq -\inf\{f(x)\} + \liminf_{n\to\infty} \mathbb{E}_{\mathcal{P}}[f(X_n)] \\
\mathbb{E}_{\mathcal{P}}[f(X))] &\leq \liminf_{n\to\infty} \mathbb{E}_{\mathcal{P}}[f(X_n)]
\end{aligned}
$$

   which completes the proof. ■

8. 2) to 4) Define $f_M = \min\{f, M\}$ for nonegative f and any real $M \geq 0$, so f_M is bounded continuous, as

$$
\begin{aligned}
\mathbb{E}_{\mathcal{P}}[f_M(X))] &= \lim_{n\to\infty} \mathbb{E}_{\mathcal{P}}[f_M(X_n)] \\
&\leq \liminf_{n\to\infty} \mathbb{E}_{\mathcal{P}}[f(X_n)]
\end{aligned}
$$

   Take $M \to \infty$, we have RHS $\mathbb{E}_{\mathcal{P}}[f_M(X))] \to \mathbb{E}_{\mathcal{P}}[f(X))]$ by monotone convergence theorem, so completes the proof. ■

- **Remark** A *continuity* set B has boundary of measure zero $\mathcal{P}\{X \in \delta B\} = 0$. Since for any collection of pairwise disjoint measureable sets, at most countable many sets can have positive measures, o.w. the total measure will be infinite. Thus given $\{B_\alpha\}_{\alpha\in A}$ all except at most countable many sets are continuity sets. For each k, at most countably sets of form $\{x : x_k \leq \alpha\}$ are not continuity sets. As a conclusion, there exists a dense subsets $Q_1, \cdots, Q_j$ so that each box with corner in $Q_1 \times Q_j$ is a continuity set. We can then choose I side this box.

- **Remark** The c.d.f. $F(\lambda) := \mathcal{P}_f((-\infty, \lambda]) = \mathcal{P}(\{x \in X : f(x) \leq \lambda\})$ where $\mathcal{P}_f = \mathcal{P} \circ f^{-1}$ is a ***measure*** on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ induced by function f. Thus $f_n \xrightarrow{d} f$ if and only if

$$
\mathcal{P}_{f_n}(A) \to \mathcal{P}_f(A), \quad \forall A \in \mathscr{B}(\mathbb{R}).
$$

- We can reformulate the definition of *convergence in distribution* as below:

  Definition [Wellner et al., 2013]
  Let $(\mathcal{X}, d)$ be a *metric space*, and $(\mathcal{X}, \mathscr{B})$ be *a measurable space*, where $\mathscr{B}$ is ***the Borel σ-field on*** $\mathcal{X}$, the smallest σ-field containing *all the open balls* (as the basis of *metric topology* on $\mathcal{X}$). Let $\{\mathcal{P}_n\}$ and $\mathcal{P}$ be ***Borel probability measures*** on $(\mathcal{X}, \mathscr{B})$.

Then the sequence $\mathcal{P}_n$ ***converges in distribution*** to $\mathcal{P}$, which we write as $\mathcal{P}_n \rightsquigarrow \mathcal{P}$, if and only if

$$\int_{\Omega} f d\mathcal{P}_n \to \int_{\Omega} f d\mathcal{P}, \quad \text{for all } f \in \mathcal{C}_b(\mathcal{X}).$$

Here $\mathcal{C}_b(\mathcal{X})$ denotes the set of *all **bounded**, **continuous**, real functions on* $\mathcal{X}$.

We can see that ***the convergence in distribution*** is actually ***a weak* convergence***. That is, it is ***the weak convergence*** of ***bounded linear functionals*** $I_{\mathcal{P}_n} \overset{w^*}{\to} I_{\mathcal{P}}$ on *the space of all probability measures* $\mathcal{P}(\mathcal{X}) \simeq (\mathcal{C}_b(\mathcal{X}))^*$ on $(\mathcal{X}, \mathscr{B})$ where

$$I_{\mathcal{P}} : f \mapsto \int_{\Omega} f d\mathcal{P}.$$

Note that the $I_{\mathcal{P}_n} \overset{w^*}{\to} I_{\mathcal{P}}$ is equivalent to $I_{\mathcal{P}_n}(f) \to I_{\mathcal{P}}(f)$ *for all* $f \in \mathcal{C}_b(\mathcal{X})$.

4.2 Properties

- **Theorem 4.2** ***(Continuous Mapping Theorem)*** *[Van der Vaart, 2000]*
  Let $g : \mathbb{R}^k \to \mathbb{R}^m$ *be* ***continuous*** *at every point of a set* $C \subset \mathbb{R}^k$ *such that* $\mathcal{P}(X \in C) = 1$. *Then*

  1. *If* $X_n \overset{a.s.}{\to} X$, *then* $g(X_n) \overset{a.s.}{\to} g(X)$;
  2. *If* $X_n \overset{\mathcal{P}}{\to} X$, *then* $g(X_n) \overset{\mathcal{P}}{\to} g(X)$;
  3. *If* $X_n \rightsquigarrow f$, *then* $g(X_n) \rightsquigarrow g(X)$.

  Proof: 1. Directly by the property of continuous map, since $g(\lim_{n\to\infty} y_{n,\omega}) = \lim_{n\to\infty} g(y_{n,\omega})$, where $y_{n,\omega} = X_n(\omega)$ for $\omega \in \Omega/E$, $\mathcal{P}(E) = 0$.

  2. For any $\epsilon > 0$, there exists $\delta > 0$ such that the set

  $$B_\delta \equiv \left\{ z \in \mathbb{R}^k \mid \exists y, \; \|z - y\| \le \delta, \; \|g(z) - g(y)\| > \epsilon \right\}.$$

  Clearly, if $X \notin B_\delta$ and $\|g(X_n) - g(X)\| > \epsilon$, then $\|X_n - X\| > \delta$. So

  $$\mathcal{P}\left\{\|g(X_n) - g(X)\| > \epsilon\right\} \le \mathcal{P}\left\{\|X_n - X\| > \delta\right\} + \mathcal{P}\left\{X \in B_\delta\right\}$$

  The first term on RHS converges to 0 as $n \to \infty$ for every fixed $\delta > 0$ due to the convergence in measure. Since $B_\delta \cap C \downarrow 0$, by continuity of g, the second term converges to 0 as $\delta \to 0$.

  3. The event $\{g(X_n) \in F\} \equiv \left\{X_n \in g^{-1}(F)\right\}$ for any closed/open set F. Note that

  $$g^{-1}(F) \subseteq \overline{g^{-1}(F)} \subset g^{-1}(F) \cup C^c$$

  Thus there exists a sequence of $y_m \to y$ and $g(y_m) \in F$ for every closed F. If $y \in C$, then $g(y_m) \to g(y)$, which is in F, since F is closed. Otherwise, $y \in C^c$. By the portmanteau lemma, since X_n converges to X in distribution,

  $$\begin{aligned} \limsup_{n\to\infty} \mathcal{P}\left\{g(X_n) \in F\right\} &\le \limsup_{n\to\infty} \mathcal{P}\left\{X_n \in \overline{g^{-1}(F)}\right\} \\ &\le \mathcal{P}\left\{X \in \overline{g^{-1}(F)}\right\} \end{aligned}$$

Since $\mathcal{P}(C^c) = 0$, the RHS

$$\begin{aligned}\mathcal{P}\left\{X \in \overline{g^{-1}(F)}\right\} &= \mathcal{P}\left\{X \in g^{-1}(F)\right\} \\ &= \mathcal{P}\left\{g(X) \in F\right\}.\end{aligned}$$

Again by applying the portmanteau lemma, $g(X_n)$ converges to $g(X)$ in distribution.
$\blacksquare$

References

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Gerald B Folland. *Real analysis: modern techniques and their applications.* John Wiley & Sons, 2013.

Erich Leo Lehmann and George Casella. *Theory of point estimation*, volume 31. Springer Science & Business Media, 1998.

Sidney I Resnick. *A probability path.* Springer Science & Business Media, 2013.

Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*, volume 198. Prentice Hall, Macmillan New York, 1988.

Terence Tao. *An introduction to measure theory*, volume 126. American Mathematical Soc., 2011.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Jon Wellner et al. *Weak convergence and empirical processes: with applications to statistics.* Springer Science & Business Media, 2013.