Lecture Note: Conditional Expectation, Martingale, and Law of Large Numbers ☆

He ZHANG[1]

Math Department, Penn State, University Park

Abstract

In this note we discuss how to use Doob's convergence theorem to prove Law of Large Numbers under finite variance assumption. As preliminaries, conditional expectation and martingale are briefly introduced. Most of the results are stated without proof. The main reference of this note is the book by Williams [1].

1. Notations

Following notations are important in the note.

- The triple $(\Omega, \mathcal{F}, \mathbb{P})$ denotes the probability space, where $\mathcal{F}$ is a $\sigma-$algebra.
- For a random variable (RV) X over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\sigma(X)$ denotes the $\sigma-$algebra generated by X, that is, $\sigma(X) = \sigma(\{X^{-1}(B); B \in \mathcal{B}\})$, where $\mathcal{B}$ denotes the Borel set over $\mathbb{R}$. Obviously, $\sigma(X)$ is a sub-σ-algebra of $\mathcal{F}$. (e.g. if X is a constant RV, then $\sigma(X) = \{\emptyset, \Omega\}$ the trivial σ-algebra)
- We say a RV X is $\mathcal{L}^k$ $(k \geq 1)$, if it yields finite k^{th} moment, that is, $\mathbb{E}[|X|^k] < +\infty$. Since $(\Omega, \mathcal{F}, \mathbb{P})$ is of finite measure, $\mathcal{L}^p \subset \mathcal{L}^q$ for $p \geq q$.
- For a set $G \in \mathcal{F}$, I_G denotes the characteristic RV over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ given by

$$I_G(\omega) = \begin{cases} 1 & \omega \in G; \\ 0 & \text{otherwise.} \end{cases}$$

2. Conditional Expectation

My favorite motivation for the conditional expectation is the following $\mathcal{L}^2$ approximation problem.

Question. *For a RV $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ and a sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$, find the best approximation for X in the subspace $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ in the $\mathcal{L}^2$ sense.*

☆2017 Winter Reading Group
[1] hqz5159@psu.edu

Since $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ is an inner produce space (the inner product is given by $< X, Y >:= \mathbb{E}[XY]$), this question can be easily answered using orthogonal projection. We can claim that $\exists Y \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$, such that

$$\mathbb{E}[(X-Y)^2] = \inf\{\mathbb{E}[(X-W)^2];\ W \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})\},$$

and

$$X - Y \perp \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P}). \tag{1}$$

As a direct result of (1), we know that $\forall G \in \mathcal{G}$,

$$\mathbb{E}[XI_G] = \mathbb{E}[YI_G] = 0,$$

which is equivalent to

$$\int_G X \mathrm{d}\mathbb{P} = \int_G Y \mathrm{d}\mathbb{P}, \quad \forall G \in \mathcal{G}. \tag{2}$$

In the setting of the probability theory, the σ-algebra represents the available information. (One can think of $\mathcal{G}$ as $\sigma(Z)$ for some RV Z) Thus, (2) implies that X and its approximation Y cannot be separated using the statistics computed from the available information. Further notice that (2), which does not rely on $X \in \mathcal{L}^2$, is also a necessary condition for (1), thus, we can generalize such concept for $X \in \mathcal{L}^1$ based on (2).

Definition 2.1. *(Kolmogorov, 1933) For RV $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and a sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$, there exists a RV Y such that*

1. *Y is $\mathcal{G}$ measurable,*
2. $\mathbb{E}(|Y|) < +\infty$,
3. *For every set $G \in \mathcal{G}$, we have (2).*

Moreover, if $\tilde{Y}$ is another RV with these properties then $Y = \tilde{Y}$ a.s., and we write $Y = \mathbb{E}[X|\mathcal{G}]$.

Comment. *A good definition is the one that looks more like a theorem than a definition.*

We have following natural but not necessarily trivial properties for conditional expectations.

Proposition 2.1. *For RV $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and sub-σ-algebras $\mathcal{G}, \mathcal{H}$ with $\mathcal{H} \subset \mathcal{G}$, the following hold.*

a. $Y = \mathbb{E}[X|\mathcal{G}]$ *then* $\mathbb{E}[Y] = \mathbb{E}[X]$.

b. *If X is $\mathcal{G}$ measurable, then $\mathbb{E}[X|\mathcal{G}] = X$.*

c. $\mathbb{E}[aX_1 + X_2|\mathcal{G}] = a\mathbb{E}[X_1|\mathcal{G}] + \mathbb{E}[X_2|\mathcal{G}]$.

d. *If $X \geq 0$ a.s., then $\mathbb{E}[X|\mathcal{G}] \geq 0$ a.s..*

e. *(cMON) If $\exists\{X_n\} \subset \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ such that $X_n \geq 0$ and $X_n \nearrow X$ a.s., then $\mathbb{E}[X_n|\mathcal{G}] \nearrow \mathbb{E}[X|\mathcal{G}]$ a.s..*

f. *(cFatou) For $\{X_n\} \subset \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathbb{E}[\liminf X_n|\mathcal{G}] \leq \liminf \mathbb{E}[X_n|\mathcal{G}]$.*

g. *(cDOM) For $\{X_n\}$ with $X_n \to X$ a.s., if $\exists V \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ such that $|X_n(\omega)| \leq V(\omega)\ \forall\omega$ for $n = 1, 2, \ldots$, then $\mathbb{E}[X_n|\mathcal{G}] \to \mathbb{E}[X|\mathcal{G}]$ a.s..*

h. *(cJensen) If $c : \mathbb{R} \to \mathbb{R}$ is convex, and $\mathbb{E}[|c(X)|] < +\infty$, then $\mathbb{E}[c(X)|\mathcal{G}] \geq c(\mathbb{E}[X|\mathcal{G}])$. (e.g. $c(x) = |x|^p$ for $p \geq 1$)*

i. *(Tower Property) $\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}\big] = \mathbb{E}[X|\mathcal{H}]$.*

j. *(Taking out what is known) If Z is an a.s. bounded $\mathcal{G}$-measurable RV, then $\mathbb{E}[ZX|\mathcal{G}] = Z\mathbb{E}[X|\mathcal{G}]$.*

k. *If $\mathcal{H}$ is independent of $\sigma(\sigma(X), \mathcal{G})$, then $\mathbb{E}[X|\sigma(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X|\mathcal{G}]$ a.s.. In particular, if X is independent of $\mathcal{H}$, then $\mathbb{E}[X|\mathcal{H}] = \mathbb{E}[X]$.* $\boxed{\text{redundant?}}$

$$\begin{tikzcd} & \sigma(\mathcal{G},\mathcal{H}) \arrow[dl, dash] \arrow[dr, dash] & \\ \mathcal{G} \arrow[dr, dash] & & \mathcal{H} \arrow[dl, dash] \\ & \mathcal{G}\cap\mathcal{H}. & \end{tikzcd}$$

Question. *What is the definition of independent RVs?*

3. Martingales

Following discussion about 'fair game' provides one of the motivations for martingales.

Question. *What makes a series of games fair?*

To answer this question, let's think of $X_n - X_{n-1}$ as your net winnings per unit stake in game n in a series of games, played at times $n = 1, 2, \ldots$. Further assume there is no game at time 0 and $X_0 = 0$.

Intuitively speaking, a series of games is fair if the conditional expectation $\mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}] = 0$ a.s., where $\mathcal{F}_{n-1}$ denotes the available information after the game $n-1$. If this is your first time playing this game, then all the available information $\mathcal{F}_{n-1}$ comes from the outcome of the past games, that is,

$$\mathcal{F}_{n-1} = \sigma(X_0, X_1, \ldots, X_{n-1}). \tag{3}$$

Notice that X_{n-1} is $\mathcal{F}_{n-1}$-measurable, therefore we can conclude that a series of games is fair if the random process $\{X_n\}$ satisfies

$$\mathbb{E}[X_n|\mathcal{F}_{n-1}] = X_{n-1} \; a.s. \quad n = 1, 2, \ldots, \tag{4}$$

where $\mathcal{F}_{n-1}$ is given by (3). This examples provides motivation for the following definitions of adapted process and martingale.

Definition 3.1. *(Adapted process) Consider a filtered space $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathbb{P})$. Here $\{\mathcal{F}_n : n \geq 0\}$ is a filtration, that is, an increasing family of sub-σ-algebra of $\mathcal{F}$. ($\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$) A process $X = \{X_n : n \geq 0\}$ is called adapted (to the filtration $\{\mathcal{F}_n\}$) if for each n, X_n is $\mathcal{F}_n$-measurable.*

Intuitively, if X is adapted, the value $X_n(\omega)$ is known to us at time n. In this note, without specification, $\{\mathcal{F}_n\}$ is always the trivial filtration given by (3), and we define

$$\mathcal{F}_\infty := \sigma\left(\bigcup_n \mathcal{F}_n\right) \subseteq \mathcal{F}.$$

Definition 3.2. *(Martingale, supermartingale, and submartingale) A process $X = \{X_n : n \geq 0\}$ is called a martingale (with respect to $(\{\mathcal{F}_n\}, \mathbb{P})$) if*

1. *X is adapted,*

2. $\mathbb{E}[|X_n|] < +\infty$, $\forall n$,
3. (4).

A supermartingale (with respect to $(\{\mathcal{F}_n\}, \mathbb{P})$) is defined similarly, except that 3. is replaced by

$$\mathbb{E}[X_n|\mathcal{F}_{n-1}] \leq X_{n-1} \ a.s. \quad n = 1, 2, \ldots, \tag{5}$$

and a submartingale is defined with 3. replaced by

$$\mathbb{E}[X_n|\mathcal{F}_{n-1}] \geq X_{n-1} \ a.s. \quad n = 1, 2, \ldots. \tag{6}$$

Remark. *A supermartingale 'decreases on average'; a submartingale 'increases on average'! (Supermartingale corresponds to superharmonic)*

Following examples are crucial in understanding martingales.

Example. *(Sums of independent zero-mean RVs) Let $\{X_n : n \geq 1\}$ be a sequence of independent RVs with $\mathbb{E}[|X_n|] < +\infty$, $\forall n$, and $\mathbb{E}[X_n] = 0$, $\forall n$. Define ($S_0 := 0$ and)*

$$S_n := \sum_{k=1}^{n} X_k, \ n \geq 1.$$

With filtration given by (3), *we have (a.s.)*

$$\mathbb{E}[S_n|\mathcal{F}_{n-1}] = \mathbb{E}[S_{n-1}|\mathcal{F}_{n-1}] + \mathbb{E}[X_n|\mathcal{F}_{n-1}] = S_{n-1} + \mathbb{E}[X_n] = S_{n-1}.$$

Example. *(Accumulating data about a RV) Let $\{\mathcal{F}_n\}$ be our filtration, and let $\xi \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Define $M_n := \mathbb{E}[\xi|\mathcal{F}_n]$ ('some version of'). By the Tower Property, we have (a.s.)*

$$\mathbb{E}[M_n|\mathcal{F}_{n-1}] = \mathbb{E}[\xi|\mathcal{F}_n|\mathcal{F}_{n-1}] = \mathbb{E}[\xi|\mathcal{F}_{n-1}] = M_{n-1}.$$

Thus, $M = \{M_n\}$ is a martingale.

We end this section by stating the famous Doob's convergence theorem without proof.

Theorem 3.1. *(Doob's 'forward' convergence theorem) Let X be a supermartingale bounded in $\mathcal{L}^1$, that is, $\sup_n \mathbb{E}[|X_n|] < +\infty$. Then, almost surely, $X_\infty := \lim X_n$ exists and is finite. For definiteness, we define $X_\infty(\omega) := \limsup X_n(\omega)$, $\forall \omega$, so that X_∞ is $\mathcal{F}_\infty$-measurable and $X_\infty = \lim X_n$ a.s..*

Remark. *It need not be true that $X_n \to X_\infty$ in $\mathcal{L}^1$. For the counter example, see* `https://en.wikipedia.org/wiki/Branching_process`.

4. Martingales Bounded in $\mathcal{L}^2$

This section introduces the basic tools that we are going to use in the proof of the Law of Large Numbers under a finite variance assumption.

Let $M = \{M_n : n \geq 0\}$ be a martingale in $\mathcal{L}^2$, that is, $\mathbb{E}[|M_n|^2] < \infty$, $\forall n$. Recall the orthogonal projection approach we mentioned in Section 2. We know that the formula

$$M_n = M_0 + \sum_{k=1}^{n}(M_k - M_{k-1}), \quad n = 1, 2, \ldots$$

express M_n as the sum of orthogonal terms, and with Pythagoras's theorem we have

$$\mathbb{E}[M_n^2] = \mathbb{E}[M_0^2] + \sum_{k=1}^{n}\mathbb{E}[(M_k - M_{k-1})^2]. \tag{7}$$

Here, the orthogonality is given by the fact that for $s, t, u, v \in \mathbb{Z}^+$, with $s \leq t \leq u \leq v$, we have $M_v - M_u$ is orthogonal to $\mathcal{L}^2(\Omega, \mathcal{F}_u, \mathbb{P})$, and in particular,

$$\mathbb{E}[(M_t - M_s)(M_v - M_u)] = \mathbb{E}\big[\mathbb{E}[(M_t - M_s)(M_v - M_u)|\mathcal{F}_u]\big] = \mathbb{E}\big[(M_t - M_s)\mathbb{E}[(M_v - M_u)|\mathcal{F}_u]\big] = 0.$$

orthogonal (uncorrelated).

As a result of the orthogonality, (7) allows us to derive $\mathcal{L}^2$-convergence from the Doob's convergence theorem.

Theorem 4.1. *Let $M = \{M_n : n \geq 0\}$ be a martingale in $\mathcal{L}^2$. Then M is bounded in $\mathcal{L}^2$ if and only if*

$$\sum_k \mathbb{E}[(M_k - M_{k-1})^2] < +\infty;$$

and when this obtains, $M_n \to M_\infty$ a.s. and in $\mathcal{L}^2$.

L^2 con. to f
$\Rightarrow \exists$ subseq. a.e. con. to f.

Following corollary is important in the proofs of Law of Large Numbers.

Corollary 4.2. *Suppose that $\{X_n : n \geq 1\}$ is a sequence of independent RVs such that*

$$\mathbb{E}[X_k] = 0, \quad \sigma_k^2 := \mathrm{Var}[X_k] \leq +\infty, \quad \forall k.$$

Then

$$\sum_k \sigma_k^2 < +\infty \Rightarrow \sum_k X_k \text{ covergence a.s..}$$

5. Law of Large Numbers

Let's first state the theorem and go through the proof under a finite variance assumption.

Theorem 5.1. *(Kolmogorov's Strong Law of Large Numbers) Let $X_1, X_2, \ldots$ be IID RVs with $\mathbb{E}[|X_k|] < +\infty, \forall k$. Let $\mu = \mathbb{E}[X_k]$, then*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \to \mu \text{ a.s..}$$

Remark. *Since the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is of finite measure, the a.s. convergence leads to the convergence in measure. (results of the Egoroff's theorem)*

To prove this theorem, we need following two lemmas.

Lemma 5.2. *(Cesàro's Lemma) Suppose the $\{b_n\}$ is a sequence of strictly positive real numbers with $b_n \nearrow +\infty$, and $\{v_n\}$ is a convergent sequence of real numbers such that $v_n \to v_\infty \in \mathbb{R}$. Then*

$$\underline{\frac{1}{b_n}\sum_{k=1}^{n}(b_k - b_{k-1})v_k} \to v_\infty \quad (n \to +\infty).$$

generalized Cesaro sum.

Proof. For $\forall \epsilon > 0$, choose N such that $v_k > v_\infty - \epsilon$, $\forall k \geq N$. Then

$$\liminf_{n\to+\infty} \frac{1}{b_n}\sum_{k=1}^{n}(b_k - b_{k-1})v_k \geq \liminf_{n\to+\infty}\left[\frac{1}{b_n}\sum_{k=1}^{N}(b_k - b_{k-1})v_k + \frac{b_n - b_N}{b_n}(v_\infty - \epsilon)\right] = v_\infty - \epsilon.$$

By a similar argument, one can prove that

$$\limsup_{n\to+\infty} \frac{1}{b_n}\sum_{k=1}^{n}(b_k - b_{k-1})v_k \leq v_\infty - \epsilon.$$

□

Lemma 5.3. *(Kronecker's Lemma) Suppose the $\{b_n\}$ is a sequence of strictly positive real numbers with $b_n \nearrow +\infty$. Let $\{x_n\}$ be a sequence of real numbers, and define*

$$s_n := \sum_{k=1}^{n} x_k.$$

Then

$$\sum_n \frac{x_n}{b_n} \text{ converges} \Rightarrow \cancel{\sum_n} \frac{s_n}{b_n} \to 0.$$

Proof. Assume $\sum_n \frac{x_n}{b_n}$ converges. Let $u_n := \sum_{k\leq n}\frac{x_k}{b_k}$, then $u_n \to u_\infty \in \mathbb{R}$. By Abel's summation formula we have

$$s_n = \sum_{k=1}^{n} b_k(u_k - u_{k-1}) = b_n u_n - \sum_{k=1}^{n}(b_k - b_{k-1})u_{k-1},$$

that is,

$$\frac{s_n}{b_n} = u_n - \frac{1}{b_n}\sum_{k=1}^{n}(b_k - b_{k-1})u_{k-1}.$$

By Cesàro's Lemma, we know that $\lim \frac{s_n}{b_n}$ exists and

$$\lim_{n\to+\infty}\frac{s_n}{b_n} = \lim_{n\to+\infty}\left(u_n - \frac{1}{b_n}\sum_{k=1}^{n}(b_k - b_{k-1})u_{k-1}\right) = u_\infty - u_\infty = 0.$$

□

We are ready to prove the following result.

Proposition 5.4. *Let $\{W_n : n \geq 1\}$ be a sequence of independent RVs such that*

$$\mathbb{E}[W_n] = 0, \text{ and } \sum \frac{\mathrm{Var}[W_n]}{n^2} \leq +\infty.$$

Then

$$\frac{1}{n}\sum_{k=1}^{n} W_k \to 0 \ a.s..$$

Proof. By Kronecker's Lemma, it is enough to prove that $\sum \frac{W_n}{n}$ coverage a.s.. But this is immediate from Corollary 4.2. $\square$

Remark. *As a corollary of this proposition, we have the Law of Large Numbers under a finite variance assumption.*

The proof of general Strong Law of Large Numbers (Theorem 5.1) is left to the readers. Following lemma acts as a hint.

Lemma 5.5. *(Kolmogorov's Truncation Lemma) Suppose that $X_1, X_2, \ldots$ are IID RVs each with the same distributions as X, where $\mathbb{E}[|X|] < +\infty$. Set $\mu := \mathbb{E}[X]$, and define*

$$Y_n := \begin{cases} X_n & |X_n| \le n, \\ 0 & \text{otherwise.} \end{cases}$$

Then

1. $\mathbb{E}[Y_n] \to \mu$;
2. $\mathbb{P}[Y_n = X_n \text{ eventually}] = 1$;
3. $\sum n^{-2} \operatorname{Var}[Y_n] < +\infty$.

References

[1] David Williams, *Probability with martingales*, Cambridge University Press, 1991.