STRONG LAW OF LARGE NUMBERS AND KINGMAN SUBADDITIVE ERGODIC THEOREM

HUNG T. C. LE

Abstract. In this paper, we use Birkhoff Ergodic Theorem to prove the Strong Law of Large Numbers (SLLN) and the more general Kingman Subadditive Ergodic Theorem. A corollary of this generalization is Furstenberg and Kesten's result on the product of random matrices, which can be seen as a non-commutative analogue of the SLLN.

Contents

1. Introduction

Birkhoff Ergodic Theorem says that if a system mixes things up sufficiently well, the time average of a function eventually approaches its space average. One might get the feeling that this is similar to the context of estimation in statistics, where if a sufficiently representative picture of the underlying distribution can be obtained through sampling, e.g. if the samples do not get "stuck" in some part of the distribution, then one would hope to be able to learn something about the underlying distribution with more and more samples. This is what (a version of) the Strong Law of Large Numbers (SLLN) essentially says. Hence our first goal for this paper is provide a translation from the context of the SLLN to the context of dynamical systems, in particular measure-preserving and ergodic ones, and prove the SLLN through this path, instead of appealing to direct estimates and bounds.

Date: March 11, 2025.

The second goal of our paper is to provide some intuition and the proof to Kingman Subadditive Ergodic Theorem, which can be seen as the generalization of Birkhoff's theorem, as well as the random version of Fekete's Lemma. On a high level, Kingman's result says that as long as the system evolves in a controlled (subadditive) manner, its average growth rate would also stabilize.

To that end, we motivate the following sections with first an overview on measure-theoretic probability and the statement of the SLLN.

2. Probability and Strong Law of Large Numbers

2.1. **Measure-Theoretic Probability.** We assume that the reader is sufficiently familiar with the basics of measure theory. The notion of probability, rigorously defined in terms of a measure space, is as follows.

We imagine that there is an experiment that has a collection of possible outcomes and some notion of probability of things happening. For example, take the experiment of rolling 3 fair, six-sided dices. Let the collection of all outcomes, the **sample space**, be Ω. Then a particular element $\omega \in \Omega$ is an **outcome**.

Perhaps it is most natural to think of probability as assigning a number to each outcome that all add up to 1. However, this approach is problematic, in the sense that it might lead to "paradoxical" things like the Banach-Tarski Paradox. Hence we ought to phrase it in the language of measure theory, where we assign numbers to sets of outcomes, rather than the outcomes themselves. The appropriate collection of sets of outcomes (i.e., subsets of Ω) turns out to be some σ-algebra $\mathcal{F} \subset \mathcal{P}(\Omega)$, and the "probability" is just a measure $\mathbb{P} : \mathcal{F} \to \mathbb{R}_{\geq 0}$ such that $\mathbb{P}(\Omega) = 1$. The measurable sets are appropriately termed **events**, so for each event, $\mathbb{P}$ assigns some probability to that event happening. All in all, we have a probability measure space $(\Omega, \mathcal{F}, \mathbb{P})$.

Example 2.1. $\Omega = \{1, 2, 3, 4, 5, 6\}^3$ is the sample space of outcomes of 3 six-sided dice. We can simply endow Ω with the discrete σ-algebra, i.e., $\mathcal{F} = \mathcal{P}(\Omega)$, and $\mathbb{P}$ the (normalized) counting measure. Then $E = \{(1, a, b) : 2 \leq a, b \leq 6\} \in \mathcal{F}$ is the event that only the first die gets a 1, which "has probability" of happening of $\mathbb{P}(E) = \frac{5^2}{6^3} = \frac{25}{216}$.

One often then wants to go beyond simply "observing" events and their corresponding probability, but also to try to glean some statistics of interest from the probability space; say, the sum of the numbers rolled on the aforementioned dice. Hence the natural notion of a random variable. Through this paper, we endow $\mathbb{R}$ with the Borel σ-algebra $\mathcal{B}_{\mathbb{R}}$.

Definition 2.2 (Random variable)**.** A **random variable** X is a measurable function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

Example 2.3. $X : \Omega \to \mathbb{R}, (\omega_1, \omega_2, \omega_3) \mapsto \omega_1 + \omega_2 + \omega_3$ is a random variable on $\Omega = \{1, 2, 3, 4, 5, 6\}^3, \mathcal{F} = \mathcal{P}(\Omega)$ that extracts the sum of the numbers rolled on 3 six-sided dice.

We also have the corresponding notion of the expected value, or mean, of a random variable X. In the discrete setting it is just the probability-weighted sum of the

values X takes for each outcome; in the general sense of course the sum translates to an integral.

Definition 2.4 (Expected value, Mean)**.** The **expected value**, or **mean** of X is

$$\mathbb{E}X \coloneqq \int_\Omega X \mathrm{d}\mathbb{P}.$$

On a slightly tangential note, one thing that I think is really cool about this formulation of probability is what the notions of sub σ-algebras and σ-algebras generated by random variables mean.

Definition 2.5 (Sub σ-algebra, σ-algebra generated by random variable)**.** $\mathcal{G}$ is a **sub σ-algebra** of $\mathcal{F}$ if $\mathcal{G}$ is also a σ-algebra over Ω and $\mathcal{G} \subset \mathcal{F}$. The **$\sigma$-algebra generated** by a random variable X is the smallest σ-algebra $\sigma(X)$ such that $X : (\Omega, \sigma(X)) \to (\mathbb{R}, \mathcal{B}_\mathbb{R})$ is measurable. In other words it is $\sigma(X) = X^{-1}\mathcal{B}_\mathbb{R}$. Automatically, $\sigma(X)$ is a sub σ-algebra of $\mathcal{F}$.

The notion of (sub) σ-algebras is intimately tied to the notion of the information at hand. In a sense, having a sub σ-algebra means viewing the measure space through an occluded lens, where one loses the fine-grained probabilities assigned to events in the bigger σ-algebra and can only assess events that are coarser in nature.

Example 2.6. In the example above,

$$\sigma(X) = \sigma(\{\{(\omega_1, \omega_2, \omega_3) : \omega_1 + \omega_2 + \omega_3 = k\} : 3 \leq k \leq 18\}) \subsetneq \mathcal{F} = \mathcal{P}(\Omega).$$

This means that through the lens of the random variable X, i.e., as someone who only knows the sum of the 3 rolls, one is unable to differentiate between $(1, 2, 4)$, $(1, 4, 2)$ and $(2, 2, 3)$, and have all these outcomes lumped into some event that contains them all. However, in the discrete σ-algebra $\mathcal{F} = \mathcal{P}(\Omega)$, one has an exact probability for each of those outcomes (since these singletons are measurable). This naturally leads to the notion of the conditional expection of a random variable X with respect to some sub σ-algebra $\mathcal{G} \subset \mathcal{F}$, denoted $\mathbb{E}(X \mid \mathcal{G})$, where it is a $\mathcal{G}$-measurable random variable and is the best estimation of X given the constrained information in $\mathcal{G}$. In the dice example, if $\mathcal{G} = \sigma(\{\{(\omega_1, \omega_2, \omega_3) : \omega_1\omega_2 = l\} : 1 \leq l \leq 36\})$ is the sub σ-algebra of "knowing the product of the first and second dice", then $\mathbb{E}(X \mid \mathcal{G})$ is the best guess of the sum of the 3 rolled numbers only given information about what the product of the first 2 is. The "best" in "best guess" can be quantified when $X \in L^2$, then $\mathbb{E}(X \mid \mathcal{G})$ is the orthogonal projection in L^2 from X to the subspace of $\mathcal{G}$-measurable functions. We will see a remark on how this is related to Birkhoff Ergodic Theorem, but I also think it is just plainly cool.

2.2. **Strong Law of Large Numbers.**

With random variables well-defined, we immediately state the Strong Law of Large Numbers (SLLN).

Theorem 2.7 (SLLN)**.** *Let $Y_0, Y_1, \ldots$ be independent and identically distributed random variables in $L^1(\Omega, \mathcal{F}, \mathbb{P})$. Then*

$$\frac{1}{n}\sum_{i=0}^{n-1} Y_i \xrightarrow{n\to\infty} \mathbb{E}Y_0 \quad a.s.$$

One should immediately imagine the following situation: that there exists an underlying true distribution that we do not know about, but we can get a lot of representative (independent) samples from it (identically distributed). The mean of the samples (the sample mean) is then guaranteed (with probability 1) to converge to the actual mean $\mathbb{E}Y_0$ of the distribution. We make rigorous the notions of independent and identically distributed (i.i.d.) random variables below.

Definition 2.8 (Independence)**.** A sequence of random variables $Y_0, Y_1, \dots$ is said to be **independent** if for any finite subset $\{Y_{n_1}, Y_{n_2}, \dots, Y_{n_k}\}$, the random variables $Y_{n_1}, Y_{n_2}, \dots, Y_{n_k}$ are mutually independent. This means that for any measurable sets $B_1, B_2, \dots, B_k \subset \mathbb{R}$,

$$\mathbb{P}(Y_{n_1} \in B_1, Y_{n_2} \in B_2, \dots, Y_{n_k} \in B_k) = \prod_{i=1}^{k} \mathbb{P}(Y_{n_i} \in B_i).$$

Definition 2.9 (Identically distributed)**.** A sequence of random variables $Y_0, Y_1, \dots$ is said to be **identically distributed** if each Y_n has the same probability distribution. This means that for any measurable set $B \subset \mathbb{R}$ and any $n \in \mathbb{N}_0$,

$$\mathbb{P}(Y_n \in B) = \mathbb{P}(Y_0 \in B).$$

The identically distributed property is obvious for SLLN; there would be no hope of learning something about the underlying distribution if the samples are not from the same distribution in the first place. The independent property suggests that these samples do not affect each other, so overall they can roam around the distribution and are thus representative of the distribution.

With this in mind, we switch gears to the dynamical systems setting to see what happens in general when a system mixes sufficiently well for one to explore everything.

3. Birkhoff Ergodic Theorem for LLN

By default we consider dynamical systems endowed with a probability measure.

3.1. Birkhoff Ergodic Theorem.

Theorem 3.1 (Birkhoff Ergodic Theorem)**.** *If $T \circlearrowright (X, \mathcal{B}, \mu)$ is probability measure-preserving then for every $f \in L^1$, for a.e. $x \in X$, the limit*

$$\frac{1}{n} \sum_{i=0}^{n-1} f \circ T^i(x) \xrightarrow{n\to\infty} \bar{f}(x)$$

exists. And $\bar{f} \in L^1$, is T-invariant and satisfies

$$\int_X \bar{f}(x) \mathrm{d}\mu = \int_X f(x) \mathrm{d}\mu = \mathbb{E}f.$$

We call the above theorem BET for short. The term taken to the limit is the average of f evaluated at the first n points in the orbit of x, hence is the time average. In the case that T is ergodic, which is of particular interest to us, then we get a direct connection between the time average of f and the space average $\mathbb{E}f$ of f through the following corollary.

Corollary 3.2. *If $T \circlearrowright (X, \mathcal{B}, \mu)$ is ergodic, then for every $f \in L^1$, for a.e. $x \in X$,*

$$\frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i(x) \xrightarrow{n\to\infty} \int_X f(x)\mathrm{d}\mu = \mathbb{E}f.$$

The short proof of the corollary uses the following lemma, whose proof can be found in [1].

Lemma 3.3. *If $T \circlearrowright (X, \mathcal{B}, \mu)$ is ergodic then any T-invariant $f \in L^1$ is constant almost everywhere.*

Proof. (of Corollary 3.2) Since $\bar{f} \in L^1$ and is T-invariant, the above lemma implies that $\bar{f} = c \in \mathbb{R}$ almost everywhere, which satisfies:

$$c = \int_X \bar{f}(x)\mathrm{d}\mu = \int_X f(x)\mathrm{d}\mu.$$

It then follows from BET that for a.e. $x \in X$ we get

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i(x) = \bar{f}(x) = c = \int_X f(x)\mathrm{d}\mu = \mathbb{E}f.$$

□

Remark 3.4. Actually, there is a more explicit form of the limit $\bar{f}$ in BET, which is

$$\bar{f} = \mathbb{E}(f \mid \mathcal{I}),$$

where $\mathcal{I}$ is the (sub) σ-algebra of T-invariant sets. What this intuitively means is that the long-term limit of f stabilizes to the part of f that can't be "shaken out" by T, i.e., the part that only involves the T-invariant events. In other words we can see the limit of the time average of f only by looking at the invariant events under T. In this setting it especially helps to imagine the orthogonal projection of f to the subspace of $\mathcal{I}$-measurable functions. This does make sense for the long term limit of f, because through time, the T-associated noise dissipates and only the pattern remains.

Of course, what happens when T is ergodic is that $\mathcal{I}$ only consists of full-measure and null sets, so $\mathbb{E}(f \mid \mathcal{I}) = \mathbb{E}f$ and we recover the expression in Corollary 3.2. We omit the proof of Theorem 3.1, which can also be found in [1], and instead focus to emphasize the embedding of SLLN in the dynamical setting.

3.2. **Dynamics of Sequence of Random Variables.** We first address the more general case of a stationary sequence of random variables, and how it can interact quite nicely with, naturally, a sequence space with the shift map.

Definition 3.5. A sequence of random variables $Y_0, Y_1, \dots$ is **stationary** if for every $m \geq 0$, the joint distribution of the random vector $(Y_0, \dots, Y_m)$ is the same as that of $(Y_k, \dots, Y_{k+m})$.

In the most general case of $Y_0, Y_1, \dots$ being any sequence of random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, one can always define the map

$$\begin{aligned} \mathbf{Y} : \Omega &\to \mathbb{R}^\infty \\ \omega &\mapsto (Y_0(\omega), Y_1(\omega), \dots) \end{aligned}$$

simply encoding (Y_n) in the sequence space $\mathbb{R}^\infty$. Perhaps this is the more natural way to think about an indexed sequence, in that any $(x_i)_{i\in I}$ indexed by I should be thought of as a map on I, hence $(a_k)_{k\in\mathbb{N}_0}$ is a map $a : \mathbb{N}_0 \to \mathbb{R}$, and $\mathbf{Y}$ is just the random version of that.

It is easily checked to be measurable with respect to the product σ-algebra $\mathcal{B}^\infty$, since $Y_0, Y_1, \dots$ were originally $\mathcal{B}$-measurable. As such, it induces a probability measure

$$P = \mathbf{Y}_*\mathbb{P} = \mathbb{P} \circ \mathbf{Y}^{-1}.$$

The upshot of doing this is that instead of working directly with $Y_0, Y_1, \dots$, we work with the coordinates in the sequence space and utilize its structure, namely the sequence of random variables $X_0, X_1, \dots$ from $(\mathbb{R}^\infty, B^\infty, P)$ to $(\mathbb{R}, \mathcal{B})$ that does the projection $X_n : (a_k) \mapsto a_n$. It then follows by construction that the joint distribution of $Y_0, Y_1, \dots$ under $\mathbb{P}$ is the same as the joint distribution of $X_0, X_1, \dots$ under P, i.e. for all m and $B_0, \dots, B_m \in \mathcal{B}_\mathbb{R}$, we have

$$P(X_0 \in B_0, \dots, X_m \in B_m) = P(B_0 \times \cdots \times B_m) = \mathbb{P}(Y_0 \in B_0, \dots, Y_m \in B_m).$$

It would then follow that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} X_i = \int X_0 \mathrm{d}P = \mathbb{E}_P X_0 \quad a.s. \tag{3.6}$$

is equivalent to

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} Y_i = \int Y_0 \mathrm{d}\mathbb{P} = \mathbb{E}_\mathbb{P} Y_0 \quad a.s. \tag{3.7}$$

and we shall use the former to show the latter.

To that end, we lean onto the dynamics on the sequence space. We know that the shift map $T : \mathbb{R}^\infty \to \mathbb{R}^\infty$

$$T(a_0, a_1, \dots) = (a_1, a_2, \dots)$$

is typically nice on the sequence space, depending on the probability measure P. When the sequence (Y_n) is stationary, it is indeed so.

Proposition 3.8. *If (Y_n) is stationary then $T \circlearrowright (\mathbb{R}^\infty, \mathcal{B}^\infty, P)$ is probability measure-preserving.*

Proof. It suffices for us to check for cylinder sets. Indeed,

$$\begin{aligned} T_*P(B_0 \times \cdots \times B_m) &= P(T^{-1}(B_0 \times \cdots \times B_m \times \mathbb{R} \times \dots)) \\ &= P(\mathbb{R} \times B_0 \times \cdots \times B_m \times \mathbb{R} \times \dots) \\ &= \mathbb{P}(Y_1 \in B_0, Y_2 \in B_1, \dots, Y_{m+1} \in B_m) \\ &= \mathbb{P}(Y_0 \in B_0, Y_1 \in B_1, \dots, Y_m \in B_m) \\ &= P(B_0 \times \cdots \times B_m) \end{aligned}$$

so indeed $T_*P = P$. □

We highlight that the second-last equality crucially depends on the fact that (Y_n) is stationary.

Now, in the particular case that $Y_0, Y_1, \dots$ are independent and identically distributed, we get something a lot better. Since they are identically distributed, we can denote the law of all Y_n as $\nu \coloneqq (Y_0)_*\mathbb{P}$. Independence then get us that the induced measure P is simply the product measure ν^∞, because P satisfies

$$\begin{aligned} P(B_0 \times \cdots \times B_m \times \mathbb{R} \times \dots) &= \mathbb{P}(Y_0 \in B_0, Y_1 \in B_1, \dots, Y_m \in B_m) \\ &= \prod_{i=0}^{m} \mathbb{P}(Y_i \in B_i) \\ &= \prod_{i=0}^{m} \nu(B_i). \end{aligned}$$

Such a nice $P = \nu^\infty$ gives us a nice T.

Proposition 3.9. *If $Y_0, Y_1, \dots$ are i.i.d. then $T \circlearrowright (\mathbb{R}^\infty, \mathcal{B}^\infty, \nu^\infty)$ is mixing. Hence it is ergodic.*

Proof. We want to show that for any measurable $U, V \in \mathcal{B}^\infty$ we have

$$\nu^\infty(T^{-n}U \cap V) \xrightarrow{n\to\infty} \nu^\infty(U)\nu^\infty(V).$$

It suffices to show for cylinder U and V. Indeed, if $U = [B_0, \dots, B_m]$ and $V = [B'_0, \dots, B'_{m'}]$, where we denote cylinder sets by their prefix, then for all $n > m'$ we get that

$$T^{-n}(U) = [\underbrace{\mathbb{R}, \dots, \mathbb{R}}_{n \text{ times}}, B_0, \dots, B_m]$$

which intersects V at

$$T^{-n}(U) \cap V = [B'_0, \dots, B'_{m'}, \underbrace{\mathbb{R}, \dots, \mathbb{R}}_{(n-m') \text{ times}}, B_0, \dots, B_m]$$

which has measure

$$\nu^\infty(T^{-n}(U) \cap V) = \prod_{i'=0}^{m'} \nu(B'_{i'}) \prod_{i=0}^{m} \nu(B_i) = \nu^\infty(V)\nu^\infty(U)$$

as desired. □

Of course, now we can apply BET, informed with that we want to have $\mathbb{E}X_0$ on the RHS, to get

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} X_0 \circ T^i = \mathbb{E}_P X_0 \quad a.s.$$

and it is immediate that $X_0 \circ T^i = X_i$ so we are done.

4. Kingman Subadditive Ergodic Theorem

For the second half of this paper, we move away from SLLN and directly prove Kingman's result, which as previously mentioned provides a convergence result similar to BET, but for more general "controlled" processes, namely subadditive ones.

Theorem 4.1 (Kingman Subadditive Ergodic Theorem)**.** *If $T \circlearrowright (\Omega, \mathcal{F}, \mathbb{P})$ is measure-preserving and $\{X(m,n) : 0 \le m < n < \infty\}$ is a family of integrable random variables that satisfies*

(1) $X(m+1, n+1) = X(m, n) \circ T$

(2) $X(0, n) \leq X(0, m) + X(m, n)$

then for a.e. $\omega \in \Omega$*, the limit*

$$\frac{X(0, n)}{n}(\omega) \xrightarrow{n \to \infty} Y(\omega)$$

exists. And Y *is* T*-invariant.*

Remark 4.2. An equivalent formulation is to have a sequence (G_n) of integrable random variables such that $G_{n+m} \leq G_n + G_m \circ T^n$, then $\frac{G_n}{n} \xrightarrow{n \to \infty} Y$ almost surely; the equivalence of the 2 formulations can be seen by setting $X(m, n) = G_{n-m} \circ T^m$. Therefore, in a sense, the "two-dimensional" family $\{X(m, n)\}$ consists of shifted versions of "one-dimensional" G_n, but we proceed with $X(m, n)$ for sake of intuition in the presented proof.

Remark 4.3. To make sense of the assumptions, it is helpful to have in mind some $f \in L^1$, and implicitly

$$X(m, n) = f \circ T^{m+1} + \cdots + f \circ T^n,$$

or in the formalization of G_n above

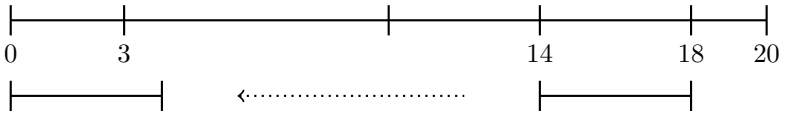
$$G_n = \sum_{i=1}^{n} f \circ T^i,$$

then the shift condition is trivially satisfied, and the subadditive condition actually evaluates to an equality. The theorem then concludes that

$$\frac{1}{n} \sum_{i=1}^{n} f \circ T^i \xrightarrow{n \to \infty} Y \quad a.s.$$

and it is exactly in this sense that Kingman's theorem is a generalization of Birkhoff's. We let this intuition of $X(m, n)$ denoting the sum of things from $m+1$ to n guide us through the proof.

Proof. A general strategy for us throughout this proof is to use both conditions of $X(m, n)$ in the following way: in order to make estimates about $X(0, n)$, we will divide it up into chunks to use the subadditive condition (keep in mind that "the whole is no greater than sum of the parts"), and then use the shift condition to get something that starts from 0.



For example, in the figure above, we divided $[0, 20]$ into several chunks, and realize that $X(14, 18) = X(0, 4) \circ T^{10}$.

Our first step is to perform a kind of normalization to justify that we can assume WLOG that $X(\cdot, \cdot) \leq 0$. Indeed, we can always define

$$\tilde{X}(0, n) = X(0, n) - \sum_{i=0}^{n-1} X(i, i+1)$$

where we've divided $(0, n]$ into chunks of unit length, and then use the shift condition to get

$$\tilde{X}(0,n) = X(0,n) - \sum_{i=0}^{n-1} X(0,1) \circ T^i$$

and the entity on the RHS looks familiar. Indeed,

$$\frac{\tilde{X}(0,n)}{n} = \frac{X(0,n)}{n} - \frac{1}{n}\sum_{i=0}^{n-1} X(0,1) \circ T^i$$

and we know from BET that $\frac{1}{n}\sum_{i=0}^{n-1} X(0,1) \circ T^i$ converges to $\overline{X(0,1)}$ (following notation from BET) that is T-invariant, so if we want to show that $\frac{X(0,n)}{n}$ converges to something that is T-invariant, it suffices to show that $\frac{\tilde{X}(0,n)}{n}$ converges something that is T-invariant too. And if $\tilde{X}(0,n) \leq 0$ then all corresponding $\tilde{X}(m,n) = \tilde{X}(0,n-m) \circ T^m \leq 0$ too. WLOG, assume $X(\cdot,\cdot) \leq 0$.

If the sequence converges, then it better converges to the lim inf. Define

$$Y = \liminf \frac{X(0,n)}{n}.$$

It remains for us to show that $Y = Y \circ T$, and that $\limsup \frac{X(0,n)}{n} \leq Y$. A heuristic of why we want to pursue this route instead of defining $Y' = \limsup \frac{X(0,n)}{n}$, is that it should be easier to show $\limsup \frac{X(0,n)}{n} \leq Y$ than $\liminf \frac{X(0,n)}{n} \geq Y'$, due to that we have a $\leq$ condition on X, not $\geq$.

Indeed, the subadditive structure of X allows us to do

$$\begin{aligned} X(0,n+1) &\leq X(0,1) + X(1,n+1) \\ &= X(0,1) + X(0,n) \circ T \end{aligned}$$

where we chopped off the first unit chunk of $(0, n+1]$, and shift the second chunk by 1. But then that implies

$$\begin{aligned} \frac{X(0,n+1)}{n+1} &\leq \frac{X(0,1)}{n+1} + \frac{n}{n+1}\frac{X(0,n)}{n} \circ T \\ \Rightarrow Y = \liminf \frac{X(0,n+1)}{n+1} &\leq \liminf \left[\frac{X(0,1)}{n+1} + \frac{n}{n+1}\frac{X(0,n)}{n} \circ T\right] \\ &= 0 + Y \circ T = Y \circ T. \end{aligned}$$

This should be an equality: Indeed, for all $t \in \mathbb{R}$, we get $\{Y \geq t\} \subset \{Y \circ T \geq t\} = T^{-1}\{Y \geq t\}$. T is measure-preserving so we also have $\mathbb{P}(\{Y \geq t\}) = \mathbb{P}(T^{-1}\{Y \geq t\})$, which implies $\mathbb{P}(\{Y < t, Y \circ T \geq t\}) = 0$. This holds for all $t \in \mathbb{R}$, so in particular

$$\{Y \neq Y \circ T\} = \{Y < Y \circ T\} = \bigcup_{t \in \mathbb{Q}} \{Y < t, Y \circ T \geq t\}$$

is a countable union of null sets and is therefore also null. This is similar in virtue to the proof of the lemma that if T is measure-preserving and $Y \in L^1$ then $\int Y \mathrm{d}\mathbb{P} = \int (Y \circ T)\mathrm{d}\mathbb{P}$ – indeed it is even easier to see that $Y = Y \circ T$ almost surely in this case, because $Y \leq Y \circ T$ and $\int Y \mathrm{d}\mathbb{P} = \int (Y \circ T)\mathrm{d}\mathbb{P}$ immediately

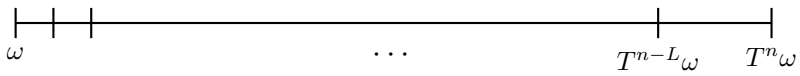
implies that $Y = Y \circ T$ almost surely.

It remains for us to show that $\limsup \frac{X(0,n)}{n} \leq Y$ almost surely. We first truncate Y by defining $Y_M = \max\{-M, Y\}$ for M large, and some tolerance $\varepsilon > 0$. Then, we can define a bad set and a corresponding good set, with the "lookahead window" parameter L

$$B_M(L) = \left\{\omega \in \Omega : \ \forall\, 1 \leq l \leq L, \frac{X(0,l)}{l}(\omega) > Y_M + \varepsilon\right\}, \quad G_M(L) = B_M(L)^C.$$

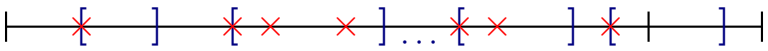
Heuristically, points in the bad set $B_M(L)$ are bad because all their time averages within L steps in the future are away from our limit; but as expected, as $L \to \infty$, it would get increasingly hard for all the averages within the L-window to stay $\varepsilon > 0$ far away from the (truncated) limit Y_M.

Here is the crucial trick: to get an estimate on $X(0,n)(\omega)$, we look at points along the orbit of ω



and leave a window of L at the end. We then find the good points (marked red below) that are not in the excluded L-window at the end. What it means for some $T^k\omega$ to be a good point is, by definition, that there exists some $l \in [L]$ such that $\frac{X(0,l)}{l}(T^k\omega) \leq Y_M + \varepsilon$, i.e., $X(k, l+k)(\omega) \leq l(Y_M + \varepsilon)$, and here we see why we're looking at the orbit of a particular ω, because it is implicitly performing the "cut up and move back" strategy that we mentioned at the start.

After finding all good points, we know that each of them corresponds to some "good window" of size at most L, so we can draw those windows from left to right, such that if a good point is already covered by a good window, we skip that good point and continue to the next one that is not already in a good window. In the figure below, we denote these windows in blue square brackets, and denote the good points that originally correspond to these windows $\{T^{a_i}\omega\}$ with window sizes l_i.



By subadditivity, non-positivity of X and disjointness of the good windows, we then get that

$$\begin{aligned} X(0,n) &\leq \sum X(a_i, a_i + l_i) \\ &\leq \left(\sum l_i\right)(Y_M + \varepsilon) \\ &\leq \left(\sum l_i\right) Y_M + n\varepsilon \\ \Rightarrow \limsup \frac{X(0,n)}{n} &\leq \limsup \left(\frac{\sum l_i}{n} Y_M\right) + \varepsilon \\ &= Y_M \liminf \left(\frac{\sum l_i}{n}\right) + \varepsilon \end{aligned}$$

where we note that the first inequality is only made possible because we originally excluded a window of size L at the end so as to prevent a good window flowing out of $[0, n]$. Of course, it is still possible for a good window to overlap in this excluded window, like in the figure above, but that is fine. And the last equality is because $Y_M \leq 0$.

To further upper bound this, realize that $Y_m \leq 0$ so in fact we need to lower bound $\liminf\left(\frac{\sum l_i}{n}\right)$. We know that these good windows cover all the good points in the first $n - L$ positions. So we get that

$$\begin{aligned}\sum l_i &\geq \sum_{i=0}^{n-L} \mathbb{1}_{G_M(L)}(\omega) \\ &\geq -L + \sum_{i=0}^{n} \mathbb{1}_{G_M(L)}(\omega) \\ \Rightarrow \liminf \frac{\sum l_i}{n} &\geq \lim \frac{1}{n} \sum_{i=0}^{n} \mathbb{1}_{G_M(L)}(\omega) = \lim \frac{1}{n} \sum_{i=0}^{n} (1 - \mathbb{1}_{B_M(L)}) = 1 - \overline{\mathbb{1}_{B_M(L)}}\end{aligned}$$

and it is justified to write lim on the RHS, because it converges almost surely by BET, and we write in terms of the bad sets to make the following step clearer. Plugging this estimate back to above, we get

$$\limsup \frac{X(0,n)}{n} \leq Y_M(1 - \overline{\mathbb{1}_{B_M(L)}}) + \varepsilon \quad a.s.$$

and drive $L \to \infty$ as initially planned, on the almost sure set that all $\overline{\mathbb{1}_{B_M(L)}}$ exist (only excluding a countable union of null sets), and see that it gets increasingly hard to be bad as L increases, so the above evaluates to

$$\limsup \frac{X(0,n)}{n} \leq Y_M(1 - 0) + \varepsilon = Y_M + \varepsilon \quad a.s.$$

when $L \to \infty$; and letting $M \to \infty$ and $\varepsilon \to 0$ completes that $\limsup \frac{X(0,n)}{n} \leq Y$ almost surely.

More rigorously, we know that $B_M(L) \supset B_M(L{+}1)$, so $1 \geq \mathbb{1}_{B_M(L)} \geq \mathbb{1}_{B_M(L+1)} \geq 0$ and $1 \geq \overline{\mathbb{1}_{B_M(L)}} \geq \overline{\mathbb{1}_{B_M(L+1)}} \geq 0$, and from BET we get

$$\int \overline{\mathbb{1}_{B_M(L)}} \mathrm{d}\mu = \int \mathbb{1}_{B_M(L)} \mathrm{d}\mu = \mu(B_M(L)) \xrightarrow{L\to\infty} 0$$

so by Dominated Convergence Theorem,

$$\int \lim_{L\to\infty} \overline{\mathbb{1}_{B_M(L)}} = 0,$$

which suggests that $\lim_{L\to\infty} \overline{\mathbb{1}_{B_M(L)}} = 0$ almost surely. And we are done. □

Remark 4.4. We see that there are multiple applications of BET, using the property that T is measure-preserving. However, in the case that T is ergodic, all BET applications converge to a deterministic limit, which implies that the limit of Kingman is also deterministic.

A direct corollary of Kingman's theorem is the following result by Furstenberg and Kesten [2] on the growth rate of the product of random matrices.

Corollary 4.5 (Furstenberg-Kesten)**.** *Let $A_1, A_2, \dots$ be an i.i.d. sequence of $d \times d$ matrices. Let $\|\cdot\|$ be a submultiplicative matrix norm, such that $\log\|A_1\|$ is integrable. Then it follows that almost surely the limit*

$$\lim_{n\to\infty} \frac{1}{n} \log(\|A_n \dots A_1\|)$$

exists and is a constant.

Proof. Choose $X(m,n) = \log(\|A_n \dots A_{m+1}\|)$ with T being the shift operator on the corresponding sequence space, just like in the proof of SLLN. Then since $\|\cdot\|$ is submultiplicative, we get that $\|A_n \dots A_1\| \leq \|A_n \dots A_{m+1}\|\|A_m \dots A_1\|$, which implies $X(0,n) \leq X(0,m) + X(m,n)$. Applying Kingman yields immediately that

$$\lim_{n\to\infty} \frac{X(0,n)}{n} = \lim_{n\to\infty} \frac{1}{n} \log(\|A_n \dots A_1\|)$$

exists almost surely, and $A_1, A_2, \dots$ are i.i.d. so T is ergodic, and the limit is deterministic. □

Another direct corollary of Kingman's theorem is the Fekete's Subadditive Lemma for sequences.

Corollary 4.6 (Fekete's Lemma)**.** *If the sequence (a_n) is subadditive, then $\frac{a_n}{n} \xrightarrow{n\to\infty} \inf \frac{a_n}{n}$.*

Proof. We use the formulation with $X(m,n) = G_{n-m} \circ T^m$. Then simply let $G_n = a_n$ and we get that $X(0,n) \leq X(0,m) + X(m,n)$ iff $G_n \leq G_m + G_{n-m}$ iff $a_n \leq a_m + a_{n-m}$, which does hold for subadditive (a_n).

Therefore $\frac{a_n}{n} = \frac{X(0,n)}{n}$ converges, and it is not hard to see that the limit is $Y = \liminf \frac{X(0,n)}{n}$ is just $\inf \frac{a_n}{n}$. □

Remark 4.7. An interpretation of Furstenberg-Kesten is a law of large numbers over a non-commutative group, in this case matrices with multiplication. Matrices typically do not commute, so standard techniques in the additive case would have required a lot of extra care, if not unfeasible.

Remark 4.8. The expression in Furstenberg-Kesten is suspiciously familiar to, say, the expression for a system's topological entropy, where we find the long-term average exponential growth rate of the system. In particular, when endowed with matrix operator norms $\|\cdot\|_{op}$ (the spectral norm $\|\cdot\|_{op,2}$ is helpful to aid visualization, as it denotes the maximum stretching along singular vectors), this is more evident as $\frac{1}{n}\log\|A_n \dots A_1\|_{op}$ then measures the long-term average exponential rate that the sequence (A_n) stretches vectors by when consecutively applied (hence the order $A_n \dots A_1$ and not the other way around, since we think of the system's evolution as applying A_1, then A_2, etc.) Of course, there is a formalism that encapsulates this notion, namely **Lyapunov exponents**; however the length of the discussion thereof clearly does not fit within these margins, and nor does the depth of the discussion fit within the author's current capacity.

Acknowledgment

This is a report I wrote for the course MATH 27600: Dynamical Systems at the University of Chicago in Winter 2025, taught by Aaron Calderon. My biggest thanks goes to Aaron, who was an amazing instructor throughout the quarter and infused the classroom with his immense enthusiasm for the subject. One might say that his teaching indeed stirred our intellectual curiosity, perhaps ergodically.

References

[1] Michael Brin and Garrett Stuck. *Introduction to Dynamical Systems*. Cambridge University Press, 2002.

[2] Harry Furstenberg and Harry Kesten. *Products of random matrices*. The Annals of Mathematical Statistics, 31(2):457-469, 1960.