Three different proofs of the Shannon–McMillan–Breiman Theorem

by

Jiacheng Guo

Department of Mathematics and Statistics

McGill University, Montréal

June 2024

A thesis submitted to McGill University
in partial fulfillment of the requirements of the degree of Master of Science.

Abstract

This thesis presents three different proofs of the Shannon–McMillan–Breiman theorem, a cornerstone in information theory. The three proofs exploit distinct mathematical approaches and shed light on different facets of the theorem. We begin by establishing some classical results from ergodic theory and probability theory which will be required for the first two proofs. We then give the statement of the theorem, followed by some important mathematical constructions as technical preparations for the proofs. Finally, we give three different proofs to the theorem. For the first and second proofs, which adapt Kingman's subadditive ergodic theorem and the martingale convergence theorem respectively, the theorem is justified for general shift-invariant measures. The third proof, which is due to D. Ornstein and B. Weiss, takes a combinatorial approach and only shows the ergodic case of the theorem.

Résumé

Ce mémoire présente trois preuves différentes du théorème de Shannon–McMillan–Breiman, une pierre angulaire de la théorie de l'information. Les trois démonstrations exploitent des approches mathématiques distinctes et jettent un éclairage sur différentes facettes du théorème. On commence par établir quelques résultats classiques de la théorie ergodique et de la théorie des probabilités qui sont nécessaires aux deux premières preuves. Nous donnons l'énoncé du théorème, suivi de constructions mathématiques qui interviennent dans les démonstrations du théorème. Finalement, les trois démonstrations sont exposées. Pour les première et seconde preuves, qui adaptent respectivement le théorème ergodique sous-additif de Kingman et le théorème de convergence des martingales, le théorème est justifié pour les mesures générales invariantes sous translation. La troisième preuve, due à D. Ornstein et B. Weiss, prend une approche combinatoire qui est restreinte au cas ergodique du théorème.

Acknowledgements

I would like to extend my heartfelt gratitude to my supervisor, Professor Vojkan Jakšić, and my co-supervisor, Professor Dmitry Jakobson, for their generous support and invaluable guidance throughout my Master's degree in mathematics here at McGill University. Both Professor Jakšić and Professor Jakobson have continuously engaged in my mathematical training since my junior and senior undergraduate years, and I truly appreciate them for imparting profound insights into mathematical knowledge and research to me all along and offering me this precious opportunity, both financially and academically, of doing Master's research in entropic information theory.

Moreover, I am grateful to the Department of Mathematics and Statistics at McGill University, for their financial support through departmental stipend and graduate fellowship, notably a Wolfe Fellowship in Science and Technology Literacy for the first academic year of my Master's.

Sincere thanks are also due to my friends, Arnab Chakraborty and Kenz Zaghib, for their companionship and help. In particular, I want to thank Arnab for all the insightful discussions, and I want to thank Kenz for the energizing conversations we have had and for his immense help with proofreading the document and translating the abstract into French.

Finally, I would like to convey deep appreciation to my parents and my sister. Over the years throughout my life journey, they have always stood by me with limitless love and unwavering support, infusing me with courage and strength. Without them, I would never reach the current stage as a person. I consider myself incredibly fortunate to have them by my side, and I am forever thankful for this.

Contents

Chapter 1

Introduction

A foundational pillar in information theory is the Shannon–McMillan–Breiman (SMB) theorem, which sometimes is also referred as the asymptotic equipartition property in certain contexts. Initially proven for Markov chains by Claude E. Shannon in his well-known 1948 paper [Sha48], the theorem was later generalized to discrete stochastic processes by Brockway McMillan in 1953 [Mc53] and finally shown for almost-sure convergence by Leo Breiman in 1957 [Bre57]. The SMB theorem illuminates fundamental yet profound connections between information-theoretic notion of entropy and randomness carried in classical discrete dynamical systems.

In this thesis, we are concerned with the SMB theorem in the setting of one-sided shift over finite alphabet and we summarize three different proofs of the SMB theorem in this setting. These proofs employ distinct approaches involving different mathematical theorems and ideas, which serve as key ingredients in the proofs and shed light on different facets of the SMB theorem.

The primary mathematical languages throughout the context of this thesis will be measure theory, probability theory, and dynamical systems. For basic and classical measure-theoretic techniques such as the monotone convergence theorem, the dominated convergence theorem, and Fatou's lemma, one may refer to [Fo99] and [Ru87]. For sufficient probability background, one may consult [Bill95] and [Wil91], where preliminary probability concepts and important elementary theorems such as the Borel–Cantelli lemma and Chebyshev's inequality are covered to full details. One may also find the lecture notes [AB19] very good, useful, and well-organized. If one wants to grasp materials of ergodic theory and dynamical systems, they may consult [Wa82] and [KH95], both of which are decent for this purpose.

Moreover, there are important concepts, notably entropy, and mathematical ideas such as covering and packing that will be discussed and used in this thesis, and close references for reviewing these are [Jak19] and [Sh96]. Some of the basic aspects of entropy will also be introduced in Appendix D of the thesis.

In Chapter 2, some important classical theorems from ergodic theory and probability theory, along with their detailed proofs, are presented. These theorems are Birkhoff's ergodic theorem, Kingman's subadditive ergodic theorem, and Doob's martingale convergence theorem respectively, and they will also be introduced in this order. The versions of these theorems to be presented will be specified in the chapter and they will all be applied as crucial technical ingredients in the first two proofs of the SMB theorem.

In Chapter 3, we quickly introduce the basic mathematical setting of one-sided shift over finite alphabet, and then give the statement of the SMB theorem. After that, we perform two important mathematical constructions, namely extension from one-sided shift setting to two-sided shift and the establishment of some functions Z_n for $n \geq 2$ which will be used to derive the so-called entropic function $\log Z$. These mathematical constructions are essential to the first two proofs of the SMB theorem, where it will be convenient for us to work on the two-sided shift setting for the first proof and the entropic function $\log Z$ plays quite an important role in the second proof.

In Chapter 4, we give the three proofs of the SMB theorem. They are referred as the subaddtive proof, the martingale proof, and the Ornstein–Weiss proof respectively. The three proofs will also be introduced in this order. It can be told from their names that the first two proofs will employ Kingman's subadditive ergodic theorem and the martingale convergence theorem respectively. For the third proof, it is quite self-contained in a way that it adapts information-theoretic ideas of packing, which are never mentioned in the preceding chapters. For this reason, the Ornstein–Weiss proof is presented at the very end along with basic introductions to covering and packing. Moreover, the subadditive and martingale proofs will justify the SMB theorem for general shift-invariant probability measures, while the Ornstein–Weiss proof will only show the ergodic case of the theorem.

Chapter 2

Classical theorems for the proofs

We begin with introduction and proofs of three well-known classical results from ergodic theory and probability theory, which will play as key ingredients in the first two proofs of the SMB theorem. These results are: a general version of Birkhoff's ergodic theorem for measure-preserving dynamical systems, a general version of Kingman's subadditive ergodic theorem that allows an additional error term in the subadditive property, and a version of Doob's martingale convergence theorem for non-negative supermartingales.

2.1 Birkhoff's ergodic theorem

As one of the most classical ergodic theorems, Birkhoff's ergodic theorem was initially proven by George D. Birkhoff in 1931 [Bir31], who was a pioneer of ergodic theory. Its statement and proof are given below.

Theorem 2.1.1 (Birkhoff). *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $T : \Omega \longrightarrow \Omega$ be a measure-preserving transformation. Suppose $X \in L^1(\Omega, \mathrm{d}\mathbb{P})$, then the limit*

$$\overline{X}(x) := \lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} X \circ T^m(x)$$

exists for $\mathbb{P}$-almost all $x \in \Omega$. $\overline{X}$ is T-invariant and the convergence also holds in $L^1(\Omega, \mathrm{d}\mathbb{P})$. Moreover,

$$\int_\Omega \overline{X}\, \mathrm{d}\mathbb{P} = \int_\Omega X\, \mathrm{d}\mathbb{P}.$$

Proof. Denote $\mathcal{F}_T$ as the T-invariant σ-algebra. Namely, $\mathcal{F}_T := \{E \in \mathcal{F} : T^{-1}(E) = E\}$. It is trivial to check that $\mathcal{F}_T$ is a sub-σ-algebra of $\mathcal{F}$.

For the given $X \in L^1(\Omega, \mathrm{d}\mathbb{P})$, we shall justify the above claimed pointwise convergence by showing that the limit superior of its Birkhoff average, namely $\frac{1}{n}\sum_{m=0}^{n-1} X \circ T^m$, is essentially bounded above by the limit function $\overline{X}$ $\mathbb{P}$-a.s. and the limit inferior is bounded below by $\overline{X}$ $\mathbb{P}$-a.s.

Let us first consider some arbitrary $Y \in L^1(\Omega, \mathrm{d}\mathbb{P})$. For each $n \in \mathbb{N}$, define the following function on Ω:

$$G_n(x) := \max_{1\le j\le n} \sum_{m=0}^{j-1} Y \circ T^m(x).$$

If for each $x \in \Omega$, we treat $\{G_n(x)\}_{n\in\mathbb{N}}$ as a sequence in $\mathbb{R}$, then clearly this is an increasing sequence.

Now that the functions G_n's are defined in terms of Y, we associate Y with a set of points which admit divergent behavior under G_n's asymptotically:

$$A_Y := \left\{x \in \Omega : \lim_{n\to\infty} G_n(x) = \infty\right\}.$$

Under this definition, we can write A_Y alternatively as follows:

$$A_Y = \left\{x \in \Omega : \lim_{n\to\infty} G_n(x) = \infty\right\} = \bigcap_{N=1}^{\infty}\bigcup_{\ell=1}^{\infty}\bigcap_{n=\ell}^{\infty}\{G_n > N\}.$$

By the definition of our function G_n, it is $\mathcal{F}$-measurable. Hence, A_Y is $\mathcal{F}$-measurable.

Then on $\Omega\backslash A_Y$, $\lim_{n\to\infty} G_n(x) < \infty$ and

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1} Y \circ T^m(x) \le \lim_{n\to\infty} \frac{1}{n} G_n(x) = 0. \tag{2.1}$$

We are interested in the measure of A_Y. If for some Y, A_Y has measure 0, then (2.1) will hold $\mathbb{P}$-a.s.

To proceed, consider the following manipulation:

$$\begin{aligned} G_{n+1}(x) - G_n(T(x)) &= \max_{1\le j\le n+1} \sum_{m=0}^{j-1} Y \circ T^m(x) - \max_{1\le j\le n} \sum_{m=1}^{j} Y \circ T^m(x) \\ &= \max_{1\le j\le n+1} \sum_{m=0}^{j-1} Y \circ T^m(x) - \max_{2\le j\le n+1} \sum_{m=1}^{j-1} Y \circ T^m(x) \end{aligned}$$

$$
\begin{aligned}
&= \max\left\{Y(x), Y(x) + \max_{2\leq j\leq n+1} \sum_{m=1}^{j-1} Y \circ T^m(x)\right\} \\
&\quad - \max_{2\leq j\leq n+1} \sum_{m=1}^{j-1} Y \circ T^m(x) \\
&= \max\{Y(x), Y(x) + G_n(T(x))\} - G_n(T(x)) \\
&= Y(x) - \min\{0, G_n(T(x))\},
\end{aligned}
$$

and we have derived a useful relation

$$
G_{n+1}(x) = G_n(T(x)) + Y(x) - \min\{0, G_n(T(x))\}, \tag{2.2}
$$

which holds for all $x \in \Omega$.

For any $x \in A_Y$, $\lim_{n\to\infty} G_{n+1}(x) = \infty$. Then by the relation (2.2), if taking limit as $n \to \infty$, we must also have $\lim_{n\to\infty} G_n(T(x)) = \infty$. Thus, $T(x) \in A_Y$ and $A_Y \subseteq T^{-1}(A_Y)$.

On the other hand, if $T(x) \in A_Y$, then for $n \in \mathbb{N}$ large enough, $G_n(T(x)) > 0$ and $G_{n+1}(x) = G_n(T(x)) + Y(x)$. Then again if we take limit as $n \to \infty$ on both sides, we must have

$$
\lim_{n\to\infty} G_{n+1}(x) = \infty.
$$

This gives us $x \in A_Y$ and therefore, $T^{-1}(A_Y) \subseteq A_Y$.

Hence, $T^{-1}(A_Y) = A_Y$ and $A_Y \in \mathcal{F}_T$.

For any fixed $x \in \Omega$, $G_n(x)$ is monotonically increasing in n by its definition, then $G_{n+1}(x) - G_n(x) \geq 0$ for all $x \in \Omega$, and we naturally have

$$
\begin{aligned}
0 &\leq \int_{A_Y} [G_{n+1}(x) - G_n(x)] \,\mathrm{d}\mathbb{P} \\
&= \int_{A_Y} G_{n+1}(x) \,\mathrm{d}\mathbb{P} - \int_{A_Y} G_n(x) \,\mathrm{d}\mathbb{P} \\
&= \int_{A_Y} G_{n+1}(x) \,\mathrm{d}\mathbb{P} - \int_{A_Y} G_n(T(x)) \,\mathrm{d}\mathbb{P} \qquad (2.3) \\
&= \int_{A_Y} [G_{n+1}(x) - G_n(T(x))] \,\mathrm{d}\mathbb{P},
\end{aligned}
$$

where in the step (2.3), we are using the fact that the pushforward measure of $\mathbb{P}$ given by T, denoted as $\mathbb{P}_T$, is identical to $\mathbb{P}$, because of T-invariance of $\mathbb{P}$:

$\forall\, F \in \mathcal{F}, \mathbb{P}_T(F) = \mathbb{P} \circ T^{-1}(F) = \mathbb{P}(F)$. Then for any $W \in L^1(\Omega, \mathrm{d}\mathbb{P})$, we would have

$$\int_{A_Y} W \,\mathrm{d}\mathbb{P} = \int_{A_Y} W \,\mathrm{d}\mathbb{P}_T = \int_{T^{-1}(A_Y)} W \circ T \,\mathrm{d}\mathbb{P} = \int_{A_Y} W \circ T \,\mathrm{d}\mathbb{P}.$$

For $x \in A_Y$, $T(x) \in A_Y$ and the sequence $\{G_n(T(x))\}_{n\in\mathbb{N}}$ diverges to infinity, so by (2.2), $G_{n+1}(x) - G_n(T(x))$ decreases to $Y(x)$ as $n \to \infty$. Besides, for every $x \in A_Y$, $G_n(T(x)) \geq Y(T(x))$ by definition, so

$$\begin{aligned} -G_n(T(x)) \leq -Y(T(x)) \;&\Rightarrow\; \max\{0, -G_n(T(x))\} \leq \max\{0, -Y(T(x))\} \\ &\Rightarrow\; -\min\{0, G_n(T(x))\} \leq -\min\{0, Y(T(x))\}. \end{aligned}$$

Thus,

$$\begin{aligned} \forall\, x \in A_Y : |G_{n+1}(x) - G_n(T(x))| &\leq |Y(x)| - \min\{0, G_n(T(x))\} \\ &\leq |Y(x)| - \min\{0, Y(T(x))\} \\ &\leq |Y(x)| + |Y(T(x))|. \end{aligned}$$

Then $|G_{n+1} - G_n \circ T| \leq |Y| + |Y \circ T|$ on A_Y. By the dominated convergence theorem,

$$0 \leq \int_{A_Y} [G_{n+1}(x) - G_n(T(x))]\,\mathrm{d}\mathbb{P} \xrightarrow{n\to\infty} \int_{A_Y} Y(x)\mathrm{d}\mathbb{P}.$$

Next, we want to have (2.1) realized $\mathbb{P}$-a.s. and that requires us to come up with some Y such that A_Y has measure 0. On the other hand, as we showed before, A_Y is always a T-invariant set, which means that A_Y is a set in the sub-σ-algebra $\mathcal{F}_T$. If for a given Y, we want to find a function whose integral on A_Y agrees with that of Y, and at the same time, it could be "the best T-invariant approximation" of Y, then the conditional expectation[1] of Y conditioning on $\mathcal{F}_T$ is a good candidate to be used here and we denote it as $Y_T := \mathbb{E}[Y|\mathcal{F}_T]$.

Then we shall have

$$0 \leq \int_{A_Y} Y \,\mathrm{d}\mathbb{P} = \int_{A_Y} Y_T \,\mathrm{d}\mathbb{P},$$

and if we can find some Y such that $Y_T < 0$ on A_Y, then we must have $\mathbb{P}(A_Y) = 0$. This is where we start to construct such Y's by using our given L^1 function X.

[1] We refer the readers to Appendix A for a heuristic derivation of conditional expectation and some of its useful properties.

Define $Y^{(k)} := X - X_T - \frac{1}{k}$ for $k \in \mathbb{N}$, where $X_T = \mathbb{E}[X|\mathcal{F}_T]$. For the corresponding set $A_{Y^{(k)}}$ for each $Y^{(k)}$, we simply write it as A_k.

By properties of conditional expectation, $Y_T^{(k)} = X_T - X_T - \frac{1}{k} = -\frac{1}{k}$. Thus, $Y_T^{(k)}$ is a negative constant on Ω and we shall have $\mathbb{P}(A_k) = 0$.

Therefore, (2.1) holds $\mathbb{P}$-a.s. for $Y^{(k)}$. Plug in its expression and by T-invariance of X_T, we get

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} X \circ T^m(x) \leq X_T(x) + \frac{1}{k} \tag{2.4}$$

on $\Omega \backslash A_k$.

Since (2.4) holds for each $k \in \mathbb{N}$ $\mathbb{P}$-a.s., then

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} X \circ T^m(x) \leq X_T(x)$$

on $\cap_{k=1}^{\infty}(\Omega\backslash A_k) = \Omega\backslash(\cup_{n=1}^{\infty} A_k)$, and

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} \mathbb{P}(A_k) = 0 \;\Rightarrow\; \mathbb{P}\left(\bigcap_{k=1}^{\infty}(\Omega\backslash A_k)\right) = 1.$$

Hence, $\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} X \circ T^m(x) \leq X_T(x)$ $\mathbb{P}$-a.s.

By replacing X with $-X$, we are getting the symmetric relation

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} X \circ T^m(x) \geq X_T(x) \;\; \mathbb{P}\text{-a.s.}$$

Therefore, we have the limit

$$\overline{X}(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} X \circ T^m(x)$$

exists for $\mathbb{P}$-almost all $x \in \Omega$, and the limit function is the conditional expectation of X conditioning on $\mathcal{F}_T$: $\overline{X} = X_T = \mathbb{E}[X|\mathcal{F}_T]$, where X_T is T-invariant.

It remains to show the convergence also holds in $L^1(\Omega, \mathrm{d}\mathbb{P})$. For the case that $X \in L^1(\Omega, \mathrm{d}\mathbb{P})$ is bounded, the convergence in L^1 follows from $\mathbb{P}$-a.s. pointwise convergence and the dominated convergence theorem.

Now consider the general case. Define $U : L^1(\Omega, \mathrm{d}\mathbb{P}) \longrightarrow L^1(\Omega, \mathrm{d}\mathbb{P})$ by sending $X \mapsto X \circ T$.

Since $\mathbb{P}$ is T-invariant, U preserves the L^1-norm:

$$\int_\Omega |X| \, \mathrm{d}\mathbb{P} = \int_\Omega |X| \circ T \, \mathrm{d}\mathbb{P}.$$

Set $U_n := \frac{1}{n} \sum_{j=0}^{n-1} U^j$ and it is easy to check that $\|U_n X\|_1 \leq \|X\|_1$.

We want to show that the sequence $\{U_n X\}_{n\in\mathbb{N}}$ is Cauchy in $L^1(\Omega, \mathrm{d}\mathbb{P})$. Let $\varepsilon > 0$ be arbitrary. Since X is L^1, there exists some $C > 0$ such that

$$\int_{\{|X|>C\}} |X| \, \mathrm{d}\mathbb{P} = \int_\Omega \left| X - X\mathbb{1}_{\{|X|\leq C\}} \right| \, \mathrm{d}\mathbb{P} < \frac{\varepsilon}{3}.$$

Let $Y := X\mathbb{1}_{\{|X|\leq C\}}$. Then Y is a bounded integrable function and $\{U_n Y\}_{n\in\mathbb{N}}$ converges to Y_T in L^1 and $\{U_n Y\}_{n\in\mathbb{N}}$ is Cauchy in $L^1(\Omega, \mathrm{d}\mathbb{P})$, so there exists some $N \in \mathbb{N}$ such that for all $n, m \geq N$, $\|U_n Y - U_m Y\|_1 < \frac{\varepsilon}{3}$.

Then for $n, m \geq N$, we have the following

$$\begin{aligned} \|U_n X - U_m X\|_1 &= \|U_n X - U_n Y + U_m Y - U_m X + U_n Y - U_m Y\|_1 \\ &\leq \|U_n X - U_n Y\|_1 + \|U_m Y - U_m X\|_1 + \|U_n Y - U_m Y\|_1 \\ &\leq 2\|X - Y\|_1 + \|U_n Y - U_m Y\|_1 \\ &< 2 \cdot \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

Hence, $\{U_n X\}_{n\in\mathbb{N}}$ is Cauchy in $L^1(\Omega, \mathrm{d}\mathbb{P})$. Since $L^1(\Omega, \mathrm{d}\mathbb{P})$ is a complete metric space, we have that $\{U_n X\}_{n\in\mathbb{N}}$ converges to its limit $\lim_{n\to\infty} U_n X = \overline{X}$ in L^1.

Furthermore, if we set $X_n := U_n X$, we have

$$0 \leq \left| \int_\Omega \left(X_n - \overline{X} \right) \mathrm{d}\mathbb{P} \right| \leq \int_\Omega \left| X_n - \overline{X} \right| \mathrm{d}\mathbb{P}, \ \ \forall\, n \in \mathbb{N}.$$

On the other hand, by T-invariance,

$$\int_\Omega X_n \, \mathrm{d}\mathbb{P} = \int_\Omega X \, \mathrm{d}\mathbb{P}, \ \ \forall\, n \in \mathbb{N},$$

so

$$0 \leq \left| \int_\Omega X \, \mathrm{d}\mathbb{P} - \int_\Omega \overline{X} \, \mathrm{d}\mathbb{P} \right| \leq \int_\Omega \left| X_n - \overline{X} \right| \mathrm{d}\mathbb{P}, \ \ \forall\, n \in \mathbb{N}.$$

Finally, as $\{X_n\}_{n\in\mathbb{N}}$ converges to $\overline{X}$ in L^1, $\left\|X_n - \overline{X}\right\|_1 \to 0$ as $n \to \infty$, and it gives that

$$\left|\int_\Omega X \,\mathrm{d}\mathbb{P} - \int_\Omega \overline{X} \,\mathrm{d}\mathbb{P}\right| = 0 \;\Rightarrow\; \int_\Omega \overline{X} \,\mathrm{d}\mathbb{P} = \int_\Omega X \,\mathrm{d}\mathbb{P}.$$

This finishes the proof. □

In the ergodic case, we have $\mathcal{F}_T$ consists of sets with either zero measure or full measure. Hence, the limit function, which is the conditional expectation of X conditioning on $\mathcal{F}_T$ and is $\mathcal{F}_T$-measurable, must be constant for $\mathbb{P}$-almost all $x \in \Omega$. The constant value is given by what we have shown in the end of the proof:

$$\int_\Omega \overline{X} \,\mathrm{d}\mathbb{P} = \int_\Omega X \,\mathrm{d}\mathbb{P} \;\Rightarrow\; \overline{X} = \int_\Omega X \,\mathrm{d}\mathbb{P}.$$

Thus, in the case when $\mathbb{P}$ is ergodic for T, we have

$$\frac{1}{n}\sum_{m=0}^{n-1} X \circ T^m \xrightarrow{n\to\infty} \int_\Omega X \,\mathrm{d}\mathbb{P}$$

$\mathbb{P}$-a.s. and the convergence also holds in L^1. This is exactly what the famous saying "the time average agrees with the space average almost everywhere" refers to.

To conclude, given an integrable function X, the asymptotic time average is given by the best T-invariant approximation, which is the conditional expectation of X conditioning on $\mathcal{F}_T$ for general T-invariant $\mathbb{P}$ and is the space average for ergodic $\mathbb{P}$.

For this proof, the general routine we followed is inspired from the one presented in [KH95, Theorem 4.1.2].

Remark. In probability theory, an important result regarding the asymptotic behavior in the large n-limit of the basic averaging of a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is the law of large numbers. In the most classical version of the result, namely Kolmogorov's strong law of large numbers (SLLN), we assume the i.i.d. (independent and identically distributed) condition and integrability for $\{X_n\}_{n\in\mathbb{N}}$. We can see in this section that, bringing in the same conclusion, Birkhoff's ergodic theorem relaxes the independent assumption in the setting, which provides a generalized version of SLLN.

2.2 Kingman's subadditive ergodic theorem

It appears that in ergodic theory, possessing a property known as subadditivity can lead to highly useful ergodic results. The subadditive property turns out to be important and useful in many contexts and one fundamental result about real numbers is Fekete's lemma[2]. Analogous to the case of real numbers, an ergodic theorem was obtained as a consequence of assuming pointwise subadditive property to a sequence of random variables, and it was originally due to Sir John Kingman in 1968 [King68]. After that, further generalizations of the theorem followed subsequently[3], and one notable result of our interest relaxes the original strict subadditivity to allow an additional additive error term. We refer this generalized subadditivity as weak subadditivity and we shall present this generalized version of the subadditive ergodic theorem in this section.

The version of the theorem we are about to show was originally due to Yves Derriennic [Der83]. However, in his paper, a much more general result is presented and it encompasses the theorem to be shown in this section as one particular assertion. The readers may consult [Der83] for his full result.

Theorem 2.2.1 (Kingman–Derriennic). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and T be a measure-preserving transformation on it. Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of functions on Ω with $X_1^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$ and*

$$X_{n+m} \leq X_n + X_m \circ T^n + Y_m \circ T^n, \quad \forall\, n, m \in \mathbb{N}, \tag{2.5}$$

where $\{Y_n\}_{n\in\mathbb{N}}$ is a sequence of non-negative functions with the following two properties:

$$\sup_{n\geq 1} \|Y_n\|_1 < \infty \qquad \textit{and} \qquad \lim_{n\to\infty} \frac{1}{n} Y_n = 0 \ \ \mathbb{P}\text{-}a.s.$$

Then the limit

$$X(x) := \lim_{n\to\infty} \frac{1}{n} X_n(x)$$

exists for $\mathbb{P}$-almost all $x \in \Omega$. X is T-invariant, $X^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$, and

$$\int_\Omega X \,\mathrm{d}\mathbb{P} = \lim_{n\to\infty} \frac{1}{n} \int_\Omega X_n \,\mathrm{d}\mathbb{P}.$$

[2] See Lemma B.1 in Appendix B.

[3] For readers who are interested, the most general version (up-to-date) of Kingman's subadditive ergodic theorem can be found in the paper by Renaud Raquépas [Raq23].

In addition, if $\{X_n\}_{n\in\mathbb{N}}$ is non-negative, then $X \in L^1(\Omega, \mathrm{d}\mathbb{P})$, X is non-negative, and $\frac{1}{n}X_n$ converges to X in L^1 as well.

Proof. In the proof, we will consider the functions $\{\frac{1}{n}X_n\}_{n\in\mathbb{N}}$ in most of the cases. Denote the functions given by the pointwise limit inferiors and limit superiors of the sequence as below:

$$\underline{X} := \liminf_{n\to\infty} \frac{1}{n}X_n, \quad \overline{X} := \limsup_{n\to\infty} \frac{1}{n}X_n.$$

To show the limit function exists $\mathbb{P}$-a.s., the goal is then to show $\overline{X} = \underline{X}$ $\mathbb{P}$-a.s.

For the given sequence of functions $\{X_n\}_{n\in\mathbb{N}}$, although no explicit information is given regarding the integrability, we can show, using the given almost-subadditive property (2.5) of the sequence, that the positive parts $\{X_n^+\}_{n\in\mathbb{N}}$ are L^1.

First, iterating (2.5) for n times and using the fact that Y_1 is non-negative, we have

$$\begin{aligned} X_n &\le X_{n-1} + X_1 \circ T^{n-1} + Y_1 \circ T^{n-1} \\ &\le X_{n-2} + X_1 \circ T^{n-2} + Y_1 \circ T^{n-2} + X_1 \circ T^{n-1} + Y_1 \circ T^{n-1} \\ &\le \cdots \le X_1 + \sum_{j=1}^{n-1}(X_1+Y_1)\circ T^j \le \sum_{j=0}^{n-1}(X_1+Y_1)\circ T^j \\ \Rightarrow \quad X_n &\le \sum_{j=0}^{n-1}(X_1+Y_1)\circ T^j. \end{aligned} \tag{2.6}$$

This holds for all $n \in \mathbb{N}$.

It is trivial to see the following two facts: for any two $\mathcal{F}$-measurable functions W_1 and W_2, $(W_1 + W_2)^+ \le W_1^+ + W_2^+$; for an $\mathcal{F}$-measurable function W, $(W \circ T)^+ = W^+ \circ T$.

Thus, (2.6) implies that

$$X_n^+ \le \sum_{j=0}^{n-1}(X_1^+ + Y_1^+)\circ T^j.$$

Integrating both sides and dividing by n give

$$\frac{1}{n}\int_\Omega X_n^+ \,\mathrm{d}\mathbb{P} \le \frac{1}{n}\sum_{j=0}^{n-1}\int_\Omega (X_1^+ + Y_1^+)\circ T^j \,\mathrm{d}\mathbb{P} = \int_\Omega (X_1^+ + Y_1^+)\,\mathrm{d}\mathbb{P},$$

where we used T-invariance of $\mathbb{P}$.

Since Y_1 is non-negative, $\sup_{n\geq 1} \|Y_n\|_1 < \infty$, and $X_1^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$, we have

$$\frac{1}{n}\int_\Omega X_n^+ \,\mathrm{d}\mathbb{P} \leq \int_\Omega (X_1^+ + Y_1^+)\,\mathrm{d}\mathbb{P} = \int_\Omega X_1^+ \,\mathrm{d}\mathbb{P} + \|Y_1\|_1 \leq \int_\Omega X_1^+ \,\mathrm{d}\mathbb{P} + \sup_{n\geq 1} \|Y_n\|_1 < \infty.$$

Hence, $X_n^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$ for all $n \in \mathbb{N}$ and $a_n := \int_\Omega X_n \,\mathrm{d}\mathbb{P} \in [-\infty, \infty)$. We next show very quickly that the limit of the integral values of the functions $\left\{\frac{1}{n}X_n\right\}_{n\in\mathbb{N}}$ exists and yields in $[-\infty, \infty)$.

Integrate both sides of (2.5), we have

$$\begin{aligned}\int_\Omega X_{n+m}\,\mathrm{d}\mathbb{P} &\leq \int_\Omega X_n\,\mathrm{d}\mathbb{P} + \int_\Omega X_m \circ T^n\,\mathrm{d}\mathbb{P} + \int_\Omega Y_m \circ T^n\,\mathrm{d}\mathbb{P} \\ (\text{by } T\text{-invariance of } \mathbb{P}) \quad &= \int_\Omega X_n\,\mathrm{d}\mathbb{P} + \int_\Omega X_m\,\mathrm{d}\mathbb{P} + \int_\Omega Y_m\,\mathrm{d}\mathbb{P} \\ (Y_m \text{ is non-negative}) \quad &= \int_\Omega X_n\,\mathrm{d}\mathbb{P} + \int_\Omega X_m\,\mathrm{d}\mathbb{P} + \|Y_m\|_1 \\ &\leq \int_\Omega X_n\,\mathrm{d}\mathbb{P} + \int_\Omega X_m\,\mathrm{d}\mathbb{P} + \sup_{n\geq 1}\|Y_n\|_1 .\end{aligned}$$

If replacing the above integral terms by a_n's and letting $M := \sup_{n\geq 1} \|Y_n\|_1 < \infty$, then we have that the real sequence $\{a_n\}_{n\in\mathbb{N}}$ satisfies

$$a_{n+m} \leq a_n + a_m + M,$$

for all $n, m \in \mathbb{N}$. By Lemma B.1, we have the limit of the sequence $\left\{\frac{1}{n}a_n\right\}_{n\in\mathbb{N}}$ exists in $[-\infty, \infty)$:

$$\lim_{n\to\infty} \frac{a_n}{n} = \inf_{n\geq 1} \frac{a_n}{n}.$$

Denote this limit as L:

$$L := \lim_{n\to\infty} \frac{a_n}{n} = \lim_{n\to\infty} \frac{1}{n}\int_\Omega X_n\,\mathrm{d}\mathbb{P}.$$

Next, let us show that $\underline{X}^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$, which will help to show $X^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$ in the end and will also act implicitly as one of the conditions for applying the monotone convergence theorem in some context below.

Note that

$$\underline{X}^+ = \max(\underline{X}, 0) = \liminf_{n\to\infty} \frac{1}{n} X_n^+ =: X_\sharp. \tag{2.7}$$

The reason is as follows:

If for some $x \in \Omega$, $\underline{X}^+(x) = \underline{X}(x) \geq 0$, then there exists some $N \in \mathbb{N}$ such that for all $n \geq N$, $\frac{1}{n}X_n(x) \geq 0$. By the definitions of X_n^+ and $X_\sharp$, for such x, we have $X_\sharp(x) = \underline{X}(x)$. If for some $x \in \Omega$, $\underline{X}(x) < 0 \Rightarrow \underline{X}^+(x) = 0$, then that means the terms $\frac{1}{n}X_n(x)$ fall below 0 infinitely often. In other words, for all $n \in \mathbb{N}$, there exists some $N \geq n$ such that $\frac{1}{N}X_N(x) < 0$. Then again by the definitions of X_n^+ and $X_\sharp$, for such x, $X_\sharp(x) = 0$.

Then we can apply Fatou's lemma:

$$\int_\Omega \underline{X}^+ \, \mathrm{d}\mathbb{P} \leq \liminf_{n\to\infty} \frac{1}{n} \int_\Omega X_n^+ \, \mathrm{d}\mathbb{P} \leq \int_\Omega X_1^+ \, \mathrm{d}\mathbb{P} + \sup_{n\geq 1} \|Y_n\|_1 < \infty,$$

so $\underline{X}^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$.

To proceed, we shall first show the desired relation $\overline{X} = \underline{X}$ $\mathbb{P}$-a.s. for the special case when the sequence of functions $\left\{\frac{1}{n}X_n\right\}_{n\in\mathbb{N}}$ is bounded from below. After that, this result will, under simple arguments, naturally extend to the general case when assumption on boundedness from below is erased.

For the special case, our strategy to achieve the desired relation $\overline{X} = \underline{X}$ $\mathbb{P}$-a.s. is to show: **(i).** $L = \int_\Omega \underline{X} \, \mathrm{d}\mathbb{P}$ and **(ii).** $L \geq \int_\Omega \overline{X} \, \mathrm{d}\mathbb{P}$. Since $\overline{X} \geq \underline{X}$ always holds true on Ω and once **(i)** and **(ii)** are justified, we get

$$0 \leq \int_\Omega \left(\overline{X} - \underline{X}\right) \mathrm{d}\mathbb{P} \leq L - L = 0 \;\Rightarrow\; \overline{X} = \underline{X} \;\; \mathbb{P}\text{-a.s.}$$

In addition, we will also show the relation **(i)** holds for the general case when the assumption on boundedness from below is removed, for the purpose of showing the integral of the limit function equals the limit of the integrals of $\frac{1}{n}X_n$ in the end, namely (2.15).

Moreover, in the context of the special case of **(i)**, we will show T-invariance of $\underline{X}$ and $\overline{X}$, as T-invariance of $\underline{X}$ will play an important role in the proof of a claimed relation (2.12) that we will give in support of the proof of **(i)**. T-invariance of $\underline{X}$ and $\overline{X}$ in the general case will follow immediately in the end once the property is justified for the special case.

Special case - (i). Suppose there exists some $C \in \mathbb{N}$ such that $\frac{1}{n}X_n \geq -C$ for every $n \in \mathbb{N}$. Then

Fatou's lemma can be applied and it gives

$$\int_\Omega \underline{X}\,\mathrm{d}\mathbb{P} \le \liminf_{n\to\infty} \int_\Omega \frac{1}{n} X_n\,\mathrm{d}\mathbb{P} = \lim_{n\to\infty} \frac{1}{n} \int_\Omega X_n\,\mathrm{d}\mathbb{P} = L.$$

Before moving on to prove the other direction, let us first show T-invariance of $\underline{X}$ and $\overline{X}$ for this special case.

A simple case of the almost-subadditive property (2.5) is

$$X_{n+1} \le X_1 + X_n \circ T + Y_n \circ T,$$

so if we divide both sides by $n+1$ and take limit inferior, we get

$$\underline{X} = \liminf_{n\to\infty} \frac{1}{n+1} X_{n+1} \le \liminf_{n\to\infty} \frac{1}{n+1}\left(X_1 + X_n \circ T + Y_n \circ T\right). \tag{2.8}$$

However, for any $x \in \Omega$, $X_1(x)$ is just a constant and for one of the conditions regarding $\{Y_n\}_{n\in\mathbb{N}}$, we have

$$\lim_{n\to\infty} \frac{1}{n} Y_n = 0 \;\; \mathbb{P}\text{-a.s.} \tag{2.9}$$

Thus,

$$\lim_{n\to\infty} \frac{1}{n+1}(X_1 + Y_n \circ T) = 0 \;\; \mathbb{P}\text{-a.s.}$$

and (2.8) would become

$$\begin{aligned}
\underline{X} &\le \liminf_{n\to\infty} \frac{1}{n+1}\left(X_1 + X_n \circ T + Y_n \circ T\right) \\
&= \liminf_{n\to\infty} \left(\frac{n}{n+1}\frac{1}{n} X_n \circ T\right) + \lim_{n\to\infty} \frac{1}{n+1}(X_1 + Y_n \circ T) \\
&= \liminf_{n\to\infty} \frac{1}{n} X_n \circ T + 0 = \underline{X} \circ T \quad (\mathbb{P}\text{-a.s.})\,.
\end{aligned}$$

Therefore, we have shown that $\underline{X} \le \underline{X} \circ T$ $\mathbb{P}$-a.s. Similarly, in fact, with fewer arguments, it also follows that $\overline{X} \le \overline{X} \circ T$ $\mathbb{P}$-a.s.

However, T-invariance of $\mathbb{P}$ tells us that

$$\int_\Omega \underline{X}\,\mathrm{d}\mathbb{P} = \int_\Omega \underline{X} \circ T\,\mathrm{d}\mathbb{P},$$

then

$$0 \le \int_\Omega (\underline{X} \circ T - \underline{X}) \, \mathrm{d}\mathbb{P} = 0 \quad \Rightarrow \quad \underline{X} = \underline{X} \circ T \;\; \mathbb{P}\text{-a.s.}$$

Similarly, we also have $\overline{X} = \overline{X} \circ T$ $\mathbb{P}$-a.s. This shows $\mathbb{P}$-almost sure T-invariance of both $\underline{X}$ and $\overline{X}$ for the special case, which will be useful in the later proof as well as for showing the conclusion of T-invariance of the limit function.

We continue with showing the other direction of **(i)**. Consider the following construction of a sequence of sets. Let $\varepsilon > 0$ be fixed. For each $k \in \mathbb{N}$, we define

$$E_k^\varepsilon := \left\{ x \in \Omega : \exists\, j \in \{1, ..., k\} \text{ such that } \frac{1}{j} X_j(x) + \frac{1}{j} Y_j(x) \le \underline{X}(x) + \varepsilon \right\}.$$

We have $\{E_k^\varepsilon\}_{k \in \mathbb{N}}$ is a sequence of monotonically increasing nested sets: If $x \in E_k^\varepsilon$ for some k, then that means there exists some $j \in \{1, ..., k\}$ such that $\frac{1}{j} X_j(x) + \frac{1}{j} Y_j(x) \le \underline{X}(x) + \varepsilon$. For this j, it is also in $\{1, ..., k, k+1\}$, so $x \in E_{k+1}^\varepsilon$ as well and $E_k^\varepsilon \subseteq E_{k+1}^\varepsilon$.

We also have the union of all sets in $\{E_k^\varepsilon\}_{k \in \mathbb{N}}$ has full measure. To put it explicitly, suppose the set where (2.9) holds is Ω'. Then $\Omega' \subseteq \Omega$ and $\mathbb{P}(\Omega') = 1$. We claim the following:

$$\Omega' \subseteq \bigcup_{k=1}^{\infty} E_k^\varepsilon.$$

Let $x \in \Omega'$. (2.9) tells us that there exists some $N_x \in \mathbb{N}$ such that for all $n \ge N_x$,

$$\frac{1}{n} Y_n(x) \le \frac{\varepsilon}{2}. \tag{2.10}$$

On the other hand, by the definition of limit inferior, the terms in the sequence $\left\{\frac{1}{n} X_n(x)\right\}_{n \in \mathbb{N}}$ will be less than or equal to $\underline{X}(x) + \frac{\varepsilon}{2}$ infinitely often. That is, for all $N \in \mathbb{N}$, there exists some $K \ge N$ such that $\frac{1}{K} X_K(x) \le \underline{X}(x) + \frac{\varepsilon}{2}$. Hence, for $N = N_x$, we denote the corresponding K as K_x, and we would have

$$\frac{1}{K_x} X_{K_x}(x) \le \underline{X}(x) + \frac{\varepsilon}{2}. \tag{2.11}$$

Combining (2.10) and (2.11), we are having

$$\frac{1}{K_x} X_{K_x}(x) + \frac{1}{K_x} Y_{K_x}(x) \leq \underline{X}(x) + \varepsilon,$$

so for any $k \geq K_x$, $x \in E_k^\varepsilon \subseteq \cup_{k=1}^{\infty} E_k^\varepsilon$. This justifies the claimed inclusion.

As $\{E_k^\varepsilon\}_{k\in\mathbb{N}}$ is a sequence of monotonically increasing nested sets with the union achieves full measure, by continuity from below,

$$\lim_{k\to\infty} \mathbb{P}\left(E_k^\varepsilon\right) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} E_k^\varepsilon\right) = 1,$$

so there would exist some $k_0 \in \mathbb{N}$, such that for all $k \geq k_0$, $\mathbb{P}(E_k^\varepsilon) > 0$.

We next define the following functions for $k \geq k_0$:

$$H_k^\varepsilon := \left(\underline{X} + \varepsilon\right) \mathbb{1}_{E_k^\varepsilon} + \left(X_1 + Y_1\right)\left(1 - \mathbb{1}_{E_k^\varepsilon}\right);$$

$$R_k^\varepsilon := \max\left(H_k^\varepsilon, X_1 + Y_1\right).$$

It might be difficult to see why we define these two functions this way at first, but it turns out that they give a good upper bound (2.12) for X_n (as will be presented below), which will lead to the desired other direction of **(i)**. In the proof of (2.12), one will see the reasoning behind the definitions of H_k^ε and R_k^ε.

Let $k \geq k_0$ be fixed. For any $n > k$, the two functions H_k^ε and R_k^ε, in the form of ergodic sum, give an upper bound for X_n:

$$X_n \leq \sum_{j=0}^{n-k-1} H_k^\varepsilon \circ T^j + \sum_{j=n-k}^{n-1} R_k^\varepsilon \circ T^j \quad \mathbb{P}\text{-a.s.} \tag{2.12}$$

The proof of (2.12) takes quite an amount of space, so to avoid affecting the continuity of the main proof of the theorem, we shall accept the claimed relation (2.12) for now and give its own proof after the main proof.

Integrating (2.12) and dividing n on both sides give that

$$\frac{1}{n}\int_\Omega X_n \,\mathrm{d}\mathbb{P} \leq \frac{n-k}{n}\int_\Omega H_k^\varepsilon \,\mathrm{d}\mathbb{P} + \frac{k}{n}\int_\Omega R_k^\varepsilon \,\mathrm{d}\mathbb{P},$$

where $\frac{1}{n}\int_\Omega X_n \,\mathrm{d}\mathbb{P}$ converges to L and both $\int_\Omega H_k^\varepsilon \,\mathrm{d}\mathbb{P}$ and $\int_\Omega R_k^\varepsilon \,\mathrm{d}\mathbb{P}$ are constants (we assume the two constants are both finite, since if one of them reaches $-\infty$, then $L = -\infty$ and the result follows trivially). Let $n \to \infty$, we get

$$L \leq \int_\Omega H_k^\varepsilon \,\mathrm{d}\mathbb{P} = \int_{E_k^\varepsilon} (\underline{X} + \varepsilon) \,\mathrm{d}\mathbb{P} + \int_{\Omega \backslash E_k^\varepsilon} (X_1 + Y_1) \,\mathrm{d}\mathbb{P}.$$

As discussed earlier, $\{E_k^\varepsilon\}_{k\in\mathbb{N}}$ is nested and will eventually achieve full measure, so if we let $k \to \infty$, we obtain

$$L \leq \int_\Omega (\underline{X} + \varepsilon) \,\mathrm{d}\mathbb{P} + 0 = \int_\Omega \underline{X} \,\mathrm{d}\mathbb{P} + \varepsilon.$$

This relation holds for all $\varepsilon > 0$, so letting $\varepsilon \downarrow 0$ and we finally get

$$L \leq \int_\Omega \underline{X} \,\mathrm{d}\mathbb{P}.$$

Thus, we have shown **(i).** $L = \int_\Omega \underline{X} \,\mathrm{d}\mathbb{P}$ in this special case.

Next, we remove the boundedness from below assumption temporarily and show that the relation **(i)** also holds for general $\{X_n\}_{n\in\mathbb{N}}$.

Continuation of (i) - general case. For our sequence of functions $\{X_n\}_{n\in\mathbb{N}}$ that is not necessarily bounded from below, consider a corresponding sequence defined as follows:

$$X_n^C := \max(X_n, -nC),$$

where $C \in \mathbb{N}$.

If in addition, we denote

$$\underline{X}^C := \liminf_{n\to\infty} \frac{1}{n} X_n^C,$$

then by what we just showed in the special case,

$$\int_\Omega \underline{X}^C \,\mathrm{d}\mathbb{P} = \lim_{n\to\infty} \frac{1}{n} \int_\Omega X_n^C \,\mathrm{d}\mathbb{P} = \inf_{n\geq 1} \frac{1}{n} \int_\Omega X_n^C \,\mathrm{d}\mathbb{P}. \tag{2.13}$$

With the same reasoning as showing (2.7), we essentially have $\underline{X}^C(x) = \max(\underline{X}, -C)$.

As it is clear that $\underline{X}^C \downarrow \underline{X}$, then the monotone convergence theorem tells us

$$\int_\Omega \underline{X}\,\mathrm{d}\mathbb{P} = \lim_{C\to\infty}\int_\Omega \underline{X}^C\,\mathrm{d}\mathbb{P} = \inf_{C\geq 1}\int_\Omega \underline{X}^C\,\mathrm{d}\mathbb{P}.$$

Plugging in (2.13) gives

$$\int_\Omega \underline{X}\,\mathrm{d}\mathbb{P} = \inf_{C\geq 1}\int_\Omega \underline{X}^C\,\mathrm{d}\mathbb{P} = \inf_{C\geq 1}\inf_{n\geq 1}\frac{1}{n}\int_\Omega X_n^C\,\mathrm{d}\mathbb{P} = \inf_{n\geq 1}\frac{1}{n}\inf_{C\geq 1}\int_\Omega X_n^C\,\mathrm{d}\mathbb{P},$$

and by the monotone convergence theorem again (since $X_n^C \downarrow X_n$), we get

$$\int_\Omega \underline{X}\,\mathrm{d}\mathbb{P} = \inf_{n\geq 1}\frac{1}{n}\int_\Omega X_n\,\mathrm{d}\mathbb{P} = L.$$

Therefore, we have justified $L = \int_\Omega \underline{X}\,\mathrm{d}\mathbb{P}$ for the general case as well.

We proceed in the special case that $\left\{\frac{1}{n}X_n\right\}_{n\in\mathbb{N}}$ is bounded below by $-C$ for some $C \in \mathbb{N}$, and move on to the justification of **(ii).** $L \geq \int_\Omega \overline{X}\,\mathrm{d}\mathbb{P}$.

Special case - (ii). For this part, we shall look along a subsequence of $\{X_n\}_{n\in\mathbb{N}}$ which will be an intermediate step to justify the desired relation. Let $k \in \mathbb{N}$ be fixed and we refer k to be "the size of the steps" in this auxiliary subsequence we are looking at.

For $n \geq k$, we can always write it as $n = km_n + q_n$, where m_n is the quotient and q_n is the remainder, so $q_n \in [0, k-1]$. We would then like to show the following inequality:

$$X_n = X_{km_n+q_n} \leq X_{km_n} + \sum_{j=1}^{k}\left(X_j^+ + Y_j^+\right)\circ T^{km_n}. \tag{2.14}$$

If $q_n = 0$, then this is obvious. Otherwise, by (2.5),

$$X_n = X_{km_n+q_n} \leq X_{km_n} + X_{q_n}\circ T^{km_n} + Y_{q_n}\circ T^{km_n}.$$

Since $0 < q_n < k$ in this case, we have

$$(X_{q_n} + Y_{q_n})\circ T^{km_n} \leq \left(X_{q_n}^+ + Y_{q_n}^+\right)\circ T^{km_n} \leq \sum_{j=1}^{k}\left(X_j^+ + Y_j^+\right)\circ T^{km_n},$$

as the sum contains the q_n-th term.

Hence,

$$X_n = X_{km_n+q_n} \le X_{km_n} + (X_{q_n} + Y_{q_n}) \circ T^{km_n} \le X_{km_n} + \sum_{j=1}^{k} \left(X_j^+ + Y_j^+ \right) \circ T^{km_n},$$

so (2.14) follows.

Now denote

$$F := \sum_{j=1}^{k} \left(X_j^+ + Y_j^+ \right).$$

As we showed before, $X_n^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$ for all $n \in \mathbb{N}$. Together with the integrability of every Y_n, we have $F \in L^1(\Omega, \mathrm{d}\mathbb{P})$. Thus, by Theorem 2.1.1 (Birkhoff's ergodic theorem),

$$\lim_{n\to\infty} \frac{1}{km_n} \sum_{i=1}^{km_n} F \circ T^i < \infty \quad \mathbb{P}\text{-a.s.}$$

and this gives

$$\lim_{n\to\infty} \frac{1}{km_n} F \circ T^{km_n} = 0 \quad \mathbb{P}\text{-a.s.}$$

Return to (2.14), divide both sides by km_n, and take limit superior as $n \to \infty$, we get

$$\limsup_{n\to\infty} \frac{1}{km_n} X_n \le \limsup_{n\to\infty} \frac{1}{km_n} X_{km_n} + \limsup_{n\to\infty} \frac{1}{km_n} F \circ T^{km_n} = \limsup_{n\to\infty} \frac{1}{km_n} X_{km_n}$$

$\mathbb{P}$-a.s.

Since m_n is the quotient of n divided by k or $n = km_n + q_n$ for $q_n \in [0, k-1]$, then asymptotically, $\frac{n}{km_n} \to 1$. Hence,

$$\limsup_{n\to\infty} \frac{1}{km_n} X_n = \limsup_{n\to\infty} \frac{n}{km_n} \frac{1}{n} X_n = \limsup_{n\to\infty} \frac{1}{n} X_n.$$

On the other hand, $m_{n_1} = m_{n_2}$ if $n_1, n_2 \in [kN, k(N+1))$ for some $N \in \mathbb{N}$. Then essentially,

$$\limsup_{n\to\infty} \frac{1}{km_n} X_{km_n} = \limsup_{n\to\infty} \frac{1}{kn} X_{kn}.$$

Therefore, we obtain that

$$\limsup_{n\to\infty} \frac{1}{n} X_n \le \limsup_{n\to\infty} \frac{1}{kn} X_{kn} \quad \mathbb{P}\text{-a.s.}$$

As $\left\{\frac{1}{kn}X_{kn}\right\}_{n\in\mathbb{N}}$ is a subsequence of $\left\{\frac{1}{n}X_n\right\}_{n\in\mathbb{N}}$, then

$$\limsup_{n\to\infty}\frac{1}{kn}X_{kn} \le \limsup_{n\to\infty}\frac{1}{n}X_n.$$

We finally get the $\mathbb{P}$-almost everywhere equality between $\overline{X}$ and the limit superior taken along the subsequence with gaps of size k:

$$\overline{X} = \limsup_{n\to\infty}\frac{1}{n}X_n = \limsup_{n\to\infty}\frac{1}{kn}X_{kn} \ \ \mathbb{P}\text{-a.s.} \ \Rightarrow \ \limsup_{n\to\infty}\frac{1}{n}X_{kn} = k\overline{X} \ \ \mathbb{P}\text{-a.s.}$$

Next, define the following functions:

$$S_n^{(k)} := -\sum_{j=0}^{n-1}(X_k+Y_k)\circ T^{jk}, \quad \underline{S^{(k)}} := \liminf_{n\to\infty}\frac{1}{n}S_n^{(k)}.$$

Technically, we are summing over n values of X_k+Y_k taken at every k step and the minus sign in the front will help with turning limit inferior to limit superior later.

For $S_1^{(k)}$, as we have supposed that $\frac{1}{n}X_n \ge -C$ for some $C>0$, then

$$\left[S_1^{(k)}\right]^+ = (-X_k-Y_k)^+ \le (-X_k)^+ + (-Y_k)^+ \le kC,$$

so $\left[S_1^{(k)}\right]^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$.

On the other hand, $\left\{S_n^{(k)}\right\}_{n\in\mathbb{N}}$ is a (sub)additive sequence:

$$\begin{aligned}
S_{n+m}^{(k)} &= -\sum_{j=0}^{n+m-1}(X_k+Y_k)\circ T^{jk} \\
&= -\sum_{j=0}^{n-1}(X_k+Y_k)\circ T^{jk} - \sum_{j=n}^{n+m-1}(X_k+Y_k)\circ T^{jk} \\
&= S_n^{(k)} - \sum_{j=0}^{m-1}(X_k+Y_k)\circ T^{jk}\circ T^{nk} \\
&= S_n^{(k)} + S_m^{(k)}\circ\left(T^k\right)^n.
\end{aligned}$$

Then we can apply the result of **(i)** for the general case to $S_n^{(k)}$'s and get

$$\int_\Omega \underline{S^{(k)}}\,\mathrm{d}\mathbb{P} = \lim_{n\to\infty} \frac{1}{n}\int_\Omega S_n^{(k)}\,\mathrm{d}\mathbb{P}.$$

Substitute the expression for $S_n^{(k)}$ and apply T-invariance of $\mathbb{P}$, it follows

$$\int_\Omega \underline{S^{(k)}}\,\mathrm{d}\mathbb{P} = \lim_{n\to\infty}\left(-\int_\Omega (X_k+Y_k)\,\mathrm{d}\mathbb{P}\right) = -\int_\Omega (X_k+Y_k)\,\mathrm{d}\mathbb{P}.$$

Let us now look at $-\underline{S^{(k)}}$. By applying subadditive property (backwards) for n times, we are getting

$$\begin{aligned}
-\underline{S^{(k)}} &= -\liminf_{n\to\infty}\frac{1}{n}S_n^{(k)} = \limsup_{n\to\infty} -\frac{1}{n}S_n^{(k)} \\
&= \limsup_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}(X_k+Y_k)\circ T^{jk} \\
&\geq \limsup_{n\to\infty}\frac{1}{n}(X_{kn}+Y_k) \\
&\geq \limsup_{n\to\infty}\frac{1}{n}X_{kn} = k\overline{X} \quad (\mathbb{P}\text{-a.s.})\,.
\end{aligned}$$

Taking integration on both sides implies

$$\int_\Omega -\underline{S^{(k)}}\,\mathrm{d}\mathbb{P} \geq k\int_\Omega \overline{X}\,\mathrm{d}\mathbb{P} \;\Rightarrow\; \frac{1}{k}\int_\Omega (X_k+Y_k)\,\mathrm{d}\mathbb{P} \geq \int_\Omega \overline{X}\,\mathrm{d}\mathbb{P}.$$

This holds for all $k \in \mathbb{N}$. Note that Y_k is non-negative for all $k\in\mathbb{N}$, so

$$0 \leq \frac{1}{k}\int_\Omega Y_k\,\mathrm{d}\mathbb{P} = \frac{1}{k}\left\|Y_k\right\|_1 \leq \frac{\sup_{n\geq 1}\left\|Y_n\right\|_1}{k} = \frac{M}{k},$$

where $M = \sup_{n\geq 1}\left\|Y_n\right\|_1$ as given before and is finite. Then by taking $k\to\infty$, we are getting

$$\lim_{k\to\infty}\frac{1}{k}\int_\Omega Y_k\,\mathrm{d}\mathbb{P} = 0.$$

We also have

$$\lim_{k\to\infty}\frac{1}{k}\int_\Omega X_k\,\mathrm{d}\mathbb{P} = L,$$

so

$$\int_\Omega \overline{X}\,\mathrm{d}\mathbb{P} \le \lim_{k\to\infty} \frac{1}{k} \int_\Omega (X_k + Y_k)\,\mathrm{d}\mathbb{P} = L.$$

This justifies the relation **(ii).** $L \ge \int_\Omega \overline{X}\,\mathrm{d}\mathbb{P}$ for the special case.

Therefore, for the special case, $\overline{X} = \underline{X}$ $\mathbb{P}$-a.s. We next move to the general case, where $\frac{1}{n}X_n$'s are not necessarily bounded from below.

General case. We adapt the defined bounded version of the functions X_n^C and $\underline{X}^C$ for $C \in \mathbb{N}$ as introduced in the above context of continuation of **(i)** for the general case. We define $\overline{X}^C$ analogously and with the same reasoning, $\overline{X}^C = \max(\overline{X}, -C)$.

By **(i)** and **(ii)** for the special case, we have

$$\int_\Omega \overline{X}^C\,\mathrm{d}\mathbb{P} \le \lim_{n\to\infty} \frac{1}{n} \int_\Omega X_n^C\,\mathrm{d}\mathbb{P} = \int_\Omega \underline{X}^C\,\mathrm{d}\mathbb{P},$$

so

$$0 \le \int_\Omega \left(\overline{X}^C - \underline{X}^C\right)\mathrm{d}\mathbb{P} \le 0 \quad\Rightarrow\quad \overline{X}^C = \underline{X}^C \;\; \mathbb{P}\text{-a.s.}$$

This holds for all $C \in \mathbb{N}$. For each $C \in \mathbb{N}$, if we denote the set where $\overline{X}^C = \underline{X}^C$ holds as Ω_C, then $\Omega_C \subseteq \Omega$ and $\mathbb{P}(\Omega_C) = 1$.

Then on $\cap_{C=1}^\infty \Omega_C$ which is the set where $\overline{X}^C = \underline{X}^C$ holds for all $C \in \mathbb{N}$, we have $\overline{X} = \underline{X}$ and

$$\mathbb{P}\left(\Omega\backslash\left(\bigcap_{C=1}^\infty \Omega_C\right)\right) = \mathbb{P}\left(\bigcup_{C=1}^\infty (\Omega\backslash\Omega_C)\right) \le \sum_{C=1}^\infty \mathbb{P}\left(\Omega\backslash\Omega_C\right) = 0$$

implies $\cap_{C=1}^\infty \Omega_C$ has full measure.

Therefore, we eventually get $\overline{X} = \underline{X}$ $\mathbb{P}$-a.s. and this tells us that the limit $X := \lim_{n\to\infty} \frac{1}{n}X_n$ exists $\mathbb{P}$-a.s.

Then it directly follows, from **(i)** for the general case, that

$$\int_\Omega X\,\mathrm{d}\mathbb{P} = \int_\Omega \underline{X}\,\mathrm{d}\mathbb{P} = \lim_{n\to\infty} \frac{1}{n} \int_\Omega X_n\,\mathrm{d}\mathbb{P}. \tag{2.15}$$

Since $\underline{X}^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$ and $X = \underline{X}$ $\mathbb{P}$-a.s., then $X^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$.

We have also shown $\mathbb{P}$-almost sure T-invariance of both $\underline{X}$ and $\overline{X}$ for the special case, then here we would have $\underline{X}^C$ and $\overline{X}^C$ are T-invariant $\mathbb{P}$-a.s. for each $C \in \mathbb{N}$. By letting $C \to \infty$ and by $X = \underline{X} = \overline{X}$ $\mathbb{P}$-a.s., we have X is T-invariant $\mathbb{P}$-a.s.

If we further assume that X_n is non-negative for all $n \in \mathbb{N}$, then clearly X is also non-negative by $\mathbb{P}$-almost sure pointwise convergence. We also have $X \in L^1$ since $X = X^+ \in L^1(\Omega, \mathrm{d}\mathbb{P})$. In addition, for the corresponding sequence $\left\{\frac{1}{n}X_n\right\}_{n\in\mathbb{N}}$, which is a sequence of L^1 functions now, we have $\frac{1}{n}X_n \to X$ $\mathbb{P}$-a.s. as just concluded above, then by Scheffé's lemma[4], (2.15) implies $\frac{1}{n}X_n \to X$ in L^1. □

We now turn to the proof of the claimed relation (2.12):

Proof of (2.12). Recall that there exists some $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$, $\mathbb{P}(E_k^\varepsilon) > 0$. Next, for each $k \geq k_0$ fixed, we defined the two functions

$$H_k^\varepsilon = (\underline{X} + \varepsilon)\, \mathbb{1}_{E_k^\varepsilon} + (X_1 + Y_1)\left(1 - \mathbb{1}_{E_k^\varepsilon}\right);$$

$$R_k^\varepsilon = \max\left(H_k^\varepsilon, X_1 + Y_1\right).$$

Then for every $n > k$, we have the inequality:

$$X_n \leq \sum_{j=0}^{n-k-1} H_k^\varepsilon \circ T^j + \sum_{j=n-k}^{n-1} R_k^\varepsilon \circ T^j \quad \mathbb{P}\text{-a.s.}$$

To prove this claimed inequality, we shall fix our $k \geq k_0$ and also fix $x \in \Omega$. The reason why we need a positive measure for E_k^ε is to use Poincaré's recurrence theorem[5]. That is, given $\mathbb{P}(E_k^\varepsilon) > 0$, we have that for $\mathbb{P}$-almost every $x \in E_k^\varepsilon$, the sequence $\{T^j(x)\}_{j\in\mathbb{N}}$ revisits E_k^ε infinitely often.

Then it suffices to consider only two cases: **(1).** $\{T^j(x)\}_{j\in\mathbb{N}}$ never meets E_k^ε, in other words, $T^j(x) \notin E_k^\varepsilon$ for all $j \in \mathbb{N}_0 := \{0\} \cup \mathbb{N}$; **(2).** $T^j(x) \in E_k^\varepsilon$ for some $j \in \mathbb{N}_0$, and thus, there are infinitely many $i \in \mathbb{N}$ such that $i \geq j$ and $T^i(x) \in E_k^\varepsilon$, by Theorem B.2.

For case **(1)**, (2.12) would simply be given by the almost-subadditive property (2.5): We already have (2.6) obtained from (2.5), then we further split the sum into two parts:

$$X_n \leq \sum_{j=0}^{n-1}(X_1 + Y_1) \circ T^j = \sum_{j=0}^{n-k-1} (X_1 + Y_1) \circ T^j + \sum_{j=n-k}^{n-1} (X_1 + Y_1) \circ T^j.$$

[4] See Lemma B.3 in Appendix B.

[5] See Theorem B.2 in Appendix B.

For $x \in \Omega$ such that $T^j(x) \notin E_k^\varepsilon$, or equivalently, $T^j(x) \in \Omega \backslash E_k^\varepsilon$, for all $j \in \mathbb{N}_0$, we have $(X_1 + Y_1) \circ T^j(x) = H_k^\varepsilon \circ T^j(x)$ for all $j \in \mathbb{N}_0$ by the definition of H_k^ε. With $R_k^\varepsilon \geq H_k^\varepsilon$, (2.12) directly follows:

$$X_n = \sum_{j=0}^{n-k-1} H_k^\varepsilon \circ T^j + \sum_{j=n-k}^{n-1} H_k^\varepsilon \circ T^j \leq \sum_{j=0}^{n-k-1} H_k^\varepsilon \circ T^j + \sum_{j=n-k}^{n-1} R_k^\varepsilon \circ T^j.$$

We move on to case **(2)**. For our fixed $x \in \Omega$ that satisfies case **(2)**, we inductively construct infinite sequences $\{m_j\}_{j\in\mathbb{N}_0}$ and $\{n_j\}_{j\in\mathbb{N}}$ satisfying

$$m_j \leq n_{j+1} < m_{j+1} \leq n_{j+2}, \quad \forall\, j \in \mathbb{N}_0.$$

The definitions of the two sequences are as follows: Set $m_0 = 0$, then

- Given m_{j-1}, define

$$n_j := \inf\{i \geq m_{j-1} : T^i(x) \in E_k^\varepsilon\}.$$

  Since $\{T^i(x)\}_{i\in\mathbb{N}}$ revisits E_k^ε infinitely often, n_j is always well-defined.

- With n_j defined, $T^{n_j}(x) \in E_k^\varepsilon$, so by the definition of E_k^ε, there exists some $\ell \in \{1, ..., k\}$ such that $\frac{1}{\ell}X_\ell(T^{n_j}(x)) + \frac{1}{\ell}Y_\ell(T^{n_j}(x)) \leq \underline{X}(T^{n_j}(x)) + \varepsilon = \underline{X}(x) + \varepsilon$, where we used T-invariance of $\underline{X}$. Take the first such ℓ:

$$\ell_j := \min\left\{\ell \in \{1, ..., k\} : \frac{1}{\ell}X_\ell(T^{n_j}(x)) + \frac{1}{\ell}Y_\ell(T^{n_j}(x)) \leq \underline{X}(T^{n_j}(x)) + \varepsilon\right\},$$

  and set $m_j := n_j + \ell_j$.

Based on this definition, the intervals on natural numbers $\{[m_j, m_{j+1}) \cap \mathbb{N}_0\}_{j\in\mathbb{N}_0}$ form a partition of $\mathbb{N}_0$. Hence, for our given $n > k$, we can always find some $J \in \mathbb{N}_0$ such that $m_J \leq n < m_{J+1}$. Then by the almost-subadditive property (2.5), we get that

$$X_n \leq X_{m_J} + \sum_{i=m_J}^{n-1} (X_1 + Y_1) \circ T^i.$$

If $J = 0$, then we simply have (2.6) here. If $m_J \neq 0$, then we can continue with the almost-subadditive property (2.5) with $m_J = n_J + \ell_J$ we defined above:

$$X_{m_J} \leq X_{n_J} + (X_{\ell_J} + Y_{\ell_J}) \circ T^{n_J},$$

and this gives, by substituting back, that

$$X_n \leq X_{n_J} + (X_{\ell_J} + Y_{\ell_J}) \circ T^{n_J} + \sum_{i=m_J}^{n-1} (X_1 + Y_1) \circ T^i.$$

Then for the X_{n_J} term, we again use (2.5) to "downgrade from n_J to m_{J-1}":

$$X_{n_J} \leq X_{m_{J-1}} + \sum_{i=m_{J-1}}^{n_J-1} (X_1 + Y_1) \circ T^i,$$

and again by (2.5), "to downgrade from m_{J-1} to n_{J-1}":

$$X_{m_{J-1}} \leq X_{n_{J-1}} + \left(X_{\ell_{J-1}} + Y_{\ell_{J-1}}\right) \circ T^{n_{J-1}}$$

Substitute all these back, we obtain

$$X_n \leq X_{n_{J-1}} + \sum_{i=m_{J-1}}^{n_J-1} (X_1 + Y_1) \circ T^i + \sum_{i=m_J}^{n-1} (X_1 + Y_1) \circ T^i + \sum_{j=J-1}^{J} \left(X_{\ell_j} + Y_{\ell_j}\right) \circ T^{n_j}.$$

We iterate like this by applying (2.5) repeatedly and this finally leads to the result:

$$X_n \leq X_{n_1} + \sum_{j=1}^{J-1} \sum_{i=m_j}^{n_{j+1}-1} (X_1 + Y_1) \circ T^i + \sum_{i=m_J}^{n-1} (X_1 + Y_1) \circ T^i + \sum_{j=1}^{J} \left(X_{\ell_j} + Y_{\ell_j}\right) \circ T^{n_j},$$

but "the next stop after downgrading from n_1" is $m_0 = 0$, so for the term X_{n_1}, the upper bound is simply given by (2.6):

$$X_{n_1} \leq \sum_{i=0}^{n_1-1} (X_1 + Y_1) \circ T^i.$$

Therefore, we obtain the following upper bound for X_n in terms of a bunch of sums:

$$X_n \leq \underbrace{\sum_{j=1}^{J} \sum_{i=m_{j-1}}^{n_j-1} (X_1 + Y_1) \circ T^i}_{\textbf{(I)}} + \underbrace{\sum_{i=m_J}^{n-1} (X_1 + Y_1) \circ T^i}_{\textbf{(II)}} + \underbrace{\sum_{j=1}^{J} \left(X_{\ell_j} + Y_{\ell_j}\right) \circ T^{n_j}}_{\textbf{(III)}}, \tag{2.16}$$

where we label the three summation terms in (2.16) as **(I)**, **(II)**, and **(III)** respectively (as shown above). We will show that, in the end, it turns out that **(I)** and **(III)** will be bounded by sums in terms of H_k^ε; **(II)** will be bounded by sums that partially involve H_k^ε and partially involve R_k^ε.

Let us look at **(III)** first. By how we defined each ℓ_j, we have

$$\frac{1}{\ell_j}\left(X_{\ell_j} + Y_{\ell_j}\right) \circ T^{n_j} \leq \underline{X} \circ T^{n_j} + \varepsilon, \quad \forall\, j \in \{1, ..., J\}.$$

Besides, by the definition of H_k^ε and E_k^ε, if $x \in E_k^\varepsilon$, $H_k^\varepsilon(x) = \underline{X}(x) + \varepsilon$; if $x \in \Omega \backslash E_k^\varepsilon$, $H_k^\varepsilon(x) = X_1(x) + Y_1(x) > \underline{X}(x) + \varepsilon$. Hence, we always have $\underline{X}(x) + \varepsilon \leq H_k^\varepsilon(x)$.

With the above information, we proceed with the following manipulation:

$$\begin{aligned} \sum_{j=1}^{J} \left(X_{\ell_j} + Y_{\ell_j}\right) \circ T^{n_j} &\leq \sum_{j=1}^{J} \ell_j \left(\underline{X} \circ T^{n_j} + \varepsilon\right) \\ &= \sum_{j=1}^{J} \sum_{i=n_j}^{m_j - 1} \left(\underline{X} \circ T^i + \varepsilon\right) \qquad (2.17) \\ &\leq \sum_{j=1}^{J} \sum_{i=n_j}^{m_j - 1} H_k^\varepsilon \circ T^i, \end{aligned}$$

where for the step (2.17), we are technically rewriting our ℓ_j copies of $\underline{X} \circ T^{n_j} + \varepsilon$, using T-invariance of $\underline{X}$, in terms of the sum of $\underline{X} \circ T^i + \varepsilon$ ranging from $i = n_j$ to $i = m_j - 1$ (there are exactly ℓ_j of them).

Hence, we have obtained an upper bound in terms of sums involving H_k^ε for **(III)**.

Next, we look at the double sum **(I)**. By how n_j's are defined, we have that for $i = m_j, ..., n_{j+1} - 1$, $T^i(x) \notin E_k^\varepsilon$, which is equivalent to $T^i(x) \in \Omega \backslash E_k^\varepsilon$.

This implies that $(X_1 + Y_1) \circ T^i(x) > \underline{X} \circ T^i(x) + \varepsilon$ and $(X_1 + Y_1) \circ T^i(x) = H_k^\varepsilon \circ T^i(x)$. Therefore, **(I)** becomes

$$\sum_{j=1}^{J} \sum_{i=m_{j-1}}^{n_j - 1} (X_1 + Y_1) \circ T^i = \sum_{j=1}^{J} \sum_{i=m_{j-1}}^{n_j - 1} H_k^\varepsilon \circ T^i.$$

Note that if $T^{m_j}(x) \in E_k^\varepsilon$ for some m_j, then by definition, $n_{j+1} = m_j$ and the sum over $(X_1 + Y_1) \circ T^i$ from $i = m_j$ to $n_{j+1} - 1$ would be an empty sum, and thus, skipped.

Then the bound for **(I)** + **(III)** would be

$$\textbf{(I)} + \textbf{(III)} \leq \sum_{j=1}^{J} \sum_{i=n_j}^{m_j-1} H_k^\varepsilon \circ T^i + \sum_{j=1}^{J} \sum_{i=m_{j-1}}^{n_j-1} H_k^\varepsilon \circ T^i = \sum_{i=0}^{m_J-1} H_k^\varepsilon \circ T^i,$$

so (2.16) becomes

$$X_n \leq \sum_{i=0}^{m_J-1} H_k^\varepsilon \circ T^i + \sum_{i=m_J}^{n-1} (X_1 + Y_1) \circ T^i. \tag{2.18}$$

We shall finally deal with the remaining sum **(II)**. The problem with **(II)** is that this sum involves the given n, which is between m_J and m_{J+1}, but we do not know the relative position of n_{J+1} with respect to n. It is possible that $m_J \leq n \leq n_{J+1}$ while it is also possible that $n_{J+1} < n < m_{J+1}$.

If we have $m_J \leq n \leq n_{J+1}$, then everything is fine and it would be the same situation as discussed in **(I)**, that $(X_1 + Y_1) \circ T^i = H_k^\varepsilon \circ T^i$ for $i \in \{m_J, ..., n-1\}$. However, if $n_{J+1} < n < m_{J+1}$, then for i such that $n_{J+1} \leq i \leq n-1$, there is at least one term, which is n_{J+1}, such that $T^i(x) \in E_k^\varepsilon$. For $T^i(x) \in E_k^\varepsilon$, we have $H_k^\varepsilon \circ T^i(x) = \underline{X} \circ T^i(x) + \varepsilon$, but it is possible that $(X_1 + Y_1) \circ T^i(x) > \underline{X} \circ T^i(x) + \varepsilon$ for this $T^i(x) \in E_k^\varepsilon$. This is where the function R_k^ε comes into the place.

To guarantee the bound will hold, we shall bound the terms with indices greater than or equal to n_{J+1} by R_k^ε, as $X_1 + Y_1 \leq R_k^\varepsilon$ always holds. Then the question comes to how many such terms do we have with the given n. As discussed earlier, it is also possible that $m_J \leq n \leq n_{J+1}$ and in that case, it would be safe to have the bound involves H_k^ε only. Nevertheless, we always have $H_k^\varepsilon \leq R_k^\varepsilon$, so it never hurts to preserve the last several terms and have them be bounded by R_k^ε.

Notice that since $m_J \leq n < m_{J+1}$ and $m_{J+1} - n_{J+1} = \ell_{J+1} \leq k$, then if n would ever exceed n_{J+1}, n cannot exceed n_{J+1} by k. Hence, we shall preserve the last k terms and have them be bounded above by R_k^ε.

Then there again involve two cases for consideration: One is that there are at least k terms in the sum **(II)**; the other is the number of terms in **(II)** is less than k.

For the first case, we can split **(II)** into two parts:

$$\sum_{i=m_J}^{n-1} (X_1 + Y_1) \circ T^i = \sum_{i=m_J}^{n-k-1} (X_1 + Y_1) \circ T^i + \sum_{i=n-k}^{n-1} (X_1 + Y_1) \circ T^i$$

$$\leq \sum_{i=m_J}^{n-k-1} H_k^\varepsilon \circ T^i + \sum_{i=n-k}^{n-1} R_k^\varepsilon \circ T^i.$$

Substitute this bound back into (2.18), we get

$$\begin{aligned} X_n &\leq \sum_{i=0}^{m_J-1} H_k^\varepsilon \circ T^i + \sum_{i=m_J}^{n-k-1} H_k^\varepsilon \circ T^i + \sum_{i=n-k}^{n-1} R_k^\varepsilon \circ T^i \\ &= \sum_{i=0}^{n-k-1} H_k^\varepsilon \circ T^i + \sum_{i=n-k}^{n-1} R_k^\varepsilon \circ T^i, \end{aligned}$$

which exactly gives us (2.12).

For the second case when there are no enough terms in **(II)**, we could borrow the rest of the terms from $\sum_{i=0}^{m_J-1} H_k^\varepsilon \circ T^i$ and have them also be bounded by R_k^ε. Suppose $n - m_J < k$. Then we split the last $(k - n + m_J)$ terms from $\sum_{i=0}^{m_J-1} H_k^\varepsilon \circ T^i$ and (2.18) becomes

$$\begin{aligned} X_n &\leq \sum_{i=0}^{n-k-1} H_k^\varepsilon \circ T^i + \sum_{i=n-k}^{m_J-1} H_k^\varepsilon \circ T^i + \sum_{i=m_J}^{n-1} (X_1 + Y_1) \circ T^i \\ &\leq \sum_{i=0}^{n-k-1} H_k^\varepsilon \circ T^i + \sum_{i=n-k}^{m_J-1} R_k^\varepsilon \circ T^i + \sum_{i=m_J}^{n-1} R_k^\varepsilon \circ T^i \\ &= \sum_{i=0}^{n-k-1} H_k^\varepsilon \circ T^i + \sum_{i=n-k}^{n-1} R_k^\varepsilon \circ T^i. \end{aligned}$$

Thus, we have also derived (2.12) under this second case.

This completes our proof to the claimed inequality (2.12). □

Given how nicely Fekete's lemma shows the existence of limit of a real sequence satisfying subadditivity, Kingman's subadditive ergodic theorem indicates that when given a sequence of random variables such that they satisfy subadditivity in a pointwise sense, there would almost-surely exist a limit function coming out of this property, so one can regard Kingman's subadditive ergodic theorem as a random variable version of Fekete's lemma.

Remark. Our presented proof of Kingman's subadditive ergodic theorem is an adaptation of the unpublished work of Artur Avila and Jairo Bochi [AB]. Although we used Birkhoff's ergodic theorem as an intermediate step in our proof, Avila and Bochi provided an alternative approach to prove the result without using Birkhoff's ergodic theorem, namely [AB, Lemma 2], where they employed very elegant arguments involving an application of the Borel–Cantelli lemma.

As one may have already observed, Kingman's subadditive ergodic theorem is a generalization of Birkhoff's ergodic theorem. In the same setting, when function $X_0 \in L^1(\Omega, \mathrm{d}\mathbb{P})$ is given, one can construct a sequence of functions $\{X_n\}_{n\in\mathbb{N}}$ by defining

$$X_n := \sum_{j=0}^{n-1} X_0 \circ T^j, \quad \forall\, n \in \mathbb{N},$$

and $\{X_n\}_{n\in\mathbb{N}}$ is clearly (sub)additive with the auxiliary sequence $\{Y_n\}_{n\in\mathbb{N}}$ simply a sequence of zero functions.

From here, we can see that Birkhoff's ergodic theorem can be a direct consequence of Kingman's subadditive ergodic theorem.

2.3 The martingale convergence theorem

The theory of martingales was once dramatically developed by Joseph L. Doob and there are many related classical results named after him, such as Doob's martingale inequality, Doob's decomposition theorem, and Doob's martingale convergence theorem, all of which can be traced back to his celebrated 1953 treatise [Doob]. In this section, we introduce a weaker version of Doob's (super)martingale convergence theorem, which is adapted for non-negative supermartingales only. This is because in the martingale proof of the SMB theorem, our constructed functions $\{Z_n\}_{n\geq 2}$, as will be introduced in Subsection 3.2.2, turn out to be non-negative supermartingales, and it suffices to use this weaker version of the theorem.

Throughout its proof, we will also use an important result from probability theory, which is the optional stopping theorem, and we refer it as Theorem B.4 in Appendix B so that the readers can consult for more details.

Theorem 2.3.1 (Doob). *Let $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n\in\mathbb{N}}, \mathbb{P})$ be a filtered probability space and $\{X_n\}_{n\in\mathbb{N}}$ a non-negative supermartingale with respect to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$. Then there exists a non-negative random variable X such that*

$$\lim_{n\to\infty} X_n = X$$

for $\mathbb{P}$-almost all $x \in \Omega$ and $X \in L^1(\Omega, \mathrm{d}\mathbb{P})$.

Proof. For each point $x \in \Omega$ where the limit does not exist, the strict inequality

$$\liminf_{n\to\infty} X_n(x) < \limsup_{n\to\infty} X_n(x)$$

holds. Then by the density of $\mathbb{Q}$ in $\mathbb{R}$, there exist $a_x, b_x \in \mathbb{Q}$ such that

$$\liminf_{n\to\infty} X_n(x) < a_x < b_x < \limsup_{n\to\infty} X_n(x). \tag{2.19}$$

In other words, by definitions of limit inferior and limit superior, $X_n(x)$ upcrosses the interval $[a_x, b_x]$ infinitely often: There are infinitely many $i \in \mathbb{N}$ such that $X_i(x) < a_x$ and infinitely many $j \in \mathbb{N}$ such that $X_j(x) > b_x$. If we take some $i \in \mathbb{N}$ where $X_i(x) < a_x$ and choose $j = \inf\{k \geq i : X_k(x) > b_x\}$, then from i to j, our process completes one upcrossing from a_x to b_x. Hence, by characterization using upcrossings, we can see that (2.19) is equivalent to the situation where $\{X_n\}_{n\in\mathbb{N}}$ evaluated at x has infinite numbers of upcrossings from a_x to b_x.

Thus, in contrast, we want to show that for every pair of $a, b \in \mathbb{Q}$ such that $a < b$, the number of upcrossings from a to b is finite for $\mathbb{P}$-almost all $x \in \Omega$. As our supermartingale is non-negative, it suffices to just consider non-negative rationals. First, let us formulate everything mathematically and this will use stopping times.

Fix arbitrary $a, b \in \mathbb{Q}$ such that $0 \leq a < b$ and fix $x \in \Omega$. Set $S_0(x) = T_0(x) = 0$ and then define the sequences $\{S_k(x)\}_{k\in\mathbb{N}}$ and $\{T_k(x)\}_{k\in\mathbb{N}}$ inductively as below:

- Given $T_{k-1}(x)$, define

$$S_k(x) := \inf\{m \geq T_{k-1}(x) : X_m(x) < a\};$$

- Given $S_k(x)$, define

$$T_k(x) := \inf\{m \geq S_k(x) : X_m(x) > b\}.$$

From this definition, we know that the sequences $\{S_k(x)\}_{k\in\mathbb{N}}$ and $\{T_k(x)\}_{k\in\mathbb{N}}$ take values in $\mathbb{N}\cup\{\infty\}$, and both could be defined for all $x \in \Omega$, so for each $k \in \mathbb{N}$, we have $S_k : \Omega \longrightarrow \mathbb{N}\cup\{\infty\}$ and $T_k : \Omega \longrightarrow \mathbb{N}\cup\{\infty\}$. It is trivial to check, by showing inductively using elementary measure-theoretic arguments, that for every $k \in \mathbb{N}$, S_k and T_k are stopping times.

We next define, for any $N \in \mathbb{N}$,

$$U_N^{[a,b]}(x) := \max\{k \in \mathbb{N} : T_k(x) \leq N\},$$

which gives the number of times that our process evaluated at x upcrosses $[a,b]$ by time N.

Apparently, $U_N^{[a,b]}(x)$ is non-decreasing in N. We then take

$$U^{[a,b]}(x) := \lim_{N\to\infty} U_N^{[a,b]}(x),$$

which gives the total number of times that $\{X_n(x)\}_{n\in\mathbb{N}}$ upcrosses $[a,b]$.

Our goal then is to show that $U^{[a,b]}(x) < \infty$ for $\mathbb{P}$-almost all $x \in \Omega$ for our fixed $a, b \in \mathbb{Q}$. To approach this, we shall show that the expectation of $U^{[a,b]}$ is finite, since $U^{[a,b]}$ is clearly non-negative.

Given arbitrary $N \in \mathbb{N}$, the constant N is also a stopping time. Then for any $k \in \mathbb{N}$, $S_k \wedge N$ and $T_k \wedge N$ are both bounded stopping times. Fix $N \in \mathbb{N}$ and incorporate with stopping times S_k's and T_k's, we shall be looking at our process at finite times $S_k \wedge N$ and $T_k \wedge N$: $X_{S_k\wedge N}$ and $X_{T_k\wedge N}$. The reasoning why we want to avoid just looking at the process at S_k and T_k is either quantity could be infinity and in that case, we might run into trouble seeing what X_{S_k} or X_{T_k} could be.

We now look at the gap $X_{T_k\wedge N} - X_{S_k\wedge N}$ for each $k \in \mathbb{N}$ and claim that for each $x \in \Omega$,

$$\sum_{k=1}^{\infty} \left(X_{T_k\wedge N} - X_{S_k\wedge N}\right)(x) \geq (b-a)U_N^{[a,b]}(x) - a. \tag{2.20}$$

This claim, together with the use of the optional stopping theorem, will later show the desired finite expectation for each $U_N^{[a,b]}$.

We look at each gap in the above sum: $\left(X_{T_k\wedge N} - X_{S_k\wedge N}\right)(x)$. Depending on the comparison between k and $U_N^{[a,b]}(x)$, there are three cases for us to discuss:

Case 1. When $k \leq U_N^{[a,b]}(x)$. To have simpler notation, let us denote $U_N^{[a,b]}(x) = M$. Then $k \leq M$ implies both $S_k(x) \leq N$ and $T_k(x) \leq N$. In this case, we shall have

$$S_k(x) \wedge N = S_k(x) \ \Rightarrow\ X_{S_k(x)\wedge N}(x) = X_{S_k(x)}(x) < a;$$
$$T_k(x) \wedge N = T_k(x) \ \Rightarrow\ X_{T_k(x)\wedge N}(x) = X_{T_k(x)}(x) > b.$$

Thus, when $k \leq U_N^{[a,b]}(x)$, we have

$$\left(X_{T_k\wedge N} - X_{S_k\wedge N}\right)(x) > b - a.$$

Case 2. When $k > U_N^{[a,b]}(x) + 1$. $k > M + 1$ implies both $S_k(x) > N$ and $T_k(x) > N$. It is clear to see that we must have $T_k(x) > N$, since otherwise, $M = U_N^{[a,b]}(x) \geq k > M+1$, which is a contradiction. We must also have $S_k(x) > N$, because if this were not true, then we would have

$$S_k(x) \leq N \;\Rightarrow\; T_{k-1}(x) \leq S_k(x) \leq N,$$

meaning that (together with $T_k(x) > N$), $M = U_N^{[a,b]}(x) = k-1 \Rightarrow k = M+1$, contradicting $k > M+1$.

Hence, in this case, we shall have $S_k(x) \wedge N = N$ and $T_k(x) \wedge N = N$ and this gives

$$\left(X_{T_k \wedge N} - X_{S_k \wedge N}\right)(x) = X_N(x) - X_N(x) = 0,$$

for all $k > U_N^{[a,b]}(x) + 1$.

Case 3. When $k = U_N^{[a,b]}(x) + 1$. With the same reasoning as used in the previous case, we must have $T_k(x) > N$ here, so $T_k(x) \wedge N = N$ and $X_{T_k(x)\wedge N}(x) = X_N(x)$.

However, as for $S_k(x)$, it is possible that either $S_k(x) > N$ or $S_k(x) \leq N$ happens. If we have $T_{k-1}(x) < N$ and there exists some $m \in \mathbb{N} \cap [T_{k-1}(x)+1, N]$ such that $X_m(x) < a$, then we shall get $S_k(x) \leq N$. If there is no existence of such m or $T_{k-1}(x) = N$, then $S_k(x) > N$ would happen.

Let us further split this third case into two subcases: **(i).** $S_k(x) \leq N$; **(ii).** $S_k(x) > N$. Under **(i)**, $S_k(x) \wedge N = S_k(x)$, so

$$X_{S_k(x)\wedge N}(x) = X_{S_k(x)}(x) < a.$$

On the other hand, our supermartingale is non-negative, so $X_N(x) \geq 0$. These give us

$$\left(X_{T_k \wedge N} - X_{S_k \wedge N}\right)(x) = X_N(x) - X_{S_k(x)}(x) \geq 0 - a = -a.$$

Under **(ii)**, $S_k(x) \wedge N = N$, then we would simply get

$$\left(X_{T_k \wedge N} - X_{S_k \wedge N}\right)(x) = X_N(x) - X_N(x) = 0.$$

As we have assumed that $a \geq 0$, then $-a \leq 0$ and under both subcases, we would have

$$\left(X_{T_k \wedge N} - X_{S_k \wedge N}\right)(x) \geq -a.$$

Hence, when $k = U_N^{[a,b]}(x) + 1$, we get $\left(X_{T_k\wedge N} - X_{S_k\wedge N}\right)(x) \geq -a$.

With the discussion of the above three cases, it is now clear to see why the claim (2.20) holds:

$$\begin{aligned}\sum_{k=1}^{\infty}\left(X_{T_k\wedge N} - X_{S_k\wedge N}\right)(x) &= \sum_{k=1}^{M}\left(X_{T_k\wedge N} - X_{S_k\wedge N}\right)(x) + \left(X_{T_{M+1}\wedge N} - X_{S_{M+1}\wedge N}\right)(x)\\ &\quad + \sum_{k=M+2}^{\infty}\left(X_{T_k\wedge N} - X_{S_k\wedge N}\right)(x)\\ &\geq M(b-a) - a + 0\\ &= (b-a)U_N^{[a,b]}(x) - a.\end{aligned}$$

Note that (2.20) holds for all $x \in \Omega$, so

$$\sum_{k=1}^{\infty}\left(X_{T_k\wedge N} - X_{S_k\wedge N}\right) \geq (b-a)U_N^{[a,b]} - a,$$

and by taking expectations of both sides, we get

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{\infty}\left(X_{T_k\wedge N} - X_{S_k\wedge N}\right)\right] \geq \mathbb{E}\left[(b-a)U_N^{[a,b]} - a\right]\\ \Rightarrow\ &\sum_{k=1}^{\infty}\mathbb{E}\left[X_{T_k\wedge N} - X_{S_k\wedge N}\right] \geq (b-a)\mathbb{E}\left[U_N^{[a,b]}\right] - a. \qquad (2.21)\end{aligned}$$

By the fact that $\{X_n\}_{n\in\mathbb{N}}$ is a supermartingale and the optional stopping theorem, for each $k \in \mathbb{N}$, we have

$$S_k \wedge N \leq T_k \wedge N \leq N \ \Rightarrow\ \mathbb{E}\left[X_{T_k\wedge N}\right] \leq \mathbb{E}\left[X_{S_k\wedge N}\right] \ \Rightarrow\ \mathbb{E}\left[X_{T_k\wedge N} - X_{S_k\wedge N}\right] \leq 0.$$

Hence, (2.21) further becomes

$$0 \geq \sum_{k=1}^{\infty}\mathbb{E}\left[X_{T_k\wedge N} - X_{S_k\wedge N}\right] \geq (b-a)\mathbb{E}\left[U_N^{[a,b]}\right] - a \ \Rightarrow\ \mathbb{E}\left[U_N^{[a,b]}\right] \leq \frac{a}{b-a}.$$

This shows the expectation of $U_N^{[a,b]}$ is finite for all $N \in \mathbb{N}$.

On the other hand,

$$U^{[a,b]} = \lim_{N\to\infty} U_N^{[a,b]} = \sup_{N\geq 1} U_N^{[a,b]}$$

on Ω, then by the monotone convergence theorem,

$$\mathbb{E}\left[U^{[a,b]}\right] = \lim_{N\to\infty} \mathbb{E}\left[U_N^{[a,b]}\right] \leq \frac{a}{b-a},$$

which shows that the expectation of $U^{[a,b]}$ is finite.

Since $U^{[a,b]}$ is clearly a non-negative random variable, having a finite expectation implies that $U^{[a,b]}$ is finite $\mathbb{P}$-a.s.

Therefore, if we denote

$$\Omega_{[a,b]} := \left\{x \in \Omega : U^{[a,b]}(x) < \infty\right\},$$

then $\mathbb{P}(\Omega_{[a,b]}) = 1$. Since our $a, b \in \mathbb{Q}$ are arbitrarily chosen, we then have $\mathbb{P}(\Omega_{[a,b]}) = 1$ for any pair of $a, b \in \mathbb{Q}$ such that $0 \leq a < b$.

Taking

$$\tilde{\Omega} := \bigcap_{\substack{a,b\in\mathbb{Q}\\ 0\leq a<b}} \Omega_{[a,b]}$$

and we shall have $\mathbb{P}(\tilde{\Omega}) = 1$, as the intersection is countable.

This in turn completes the proof that $\liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n$ $\mathbb{P}$-a.s., so the limit

$$X := \lim_{n\to\infty} X_n$$

exists for $\mathbb{P}$-almost all $x \in \Omega$, and is clearly non-negative since our supermartingale is non-negative.

To see X is L^1, note that

$$\mathbb{E}\left[|X|\right] = \mathbb{E}[X] \leq \liminf_{n\to\infty} \mathbb{E}[X_n] \leq \mathbb{E}[X_1] < \infty,$$

where we use Fatou's lemma and the fact that $\{X_n\}_{n\in\mathbb{N}}$ is a supermartingale. □

Chapter 3

The Shannon–McMillan–Breiman Theorem

In this chapter, we introduce our basic information-theoretic setting and present the statement of the SMB theorem, followed by some important mathematical constructions in preparation for its first two proofs.

3.1 The setup and the statement

We use $\mathcal{A}$ to denote a finite alphabet, namely a finite set with its elements called letters. If $\mathcal{A} = \{a_1, a_2, ..., a_\ell\}$ for some $\ell \in \mathbb{N}$, then $|\mathcal{A}| = \ell$. Occasionally, $\mathcal{A}$ is taken to be specific and in the context of Markov chains, $\mathcal{A}$ is simply taken to be $\mathcal{A} = \{0, 1, 2, ..., \ell - 1\}$ for some $\ell \in \mathbb{N}$. In telecommunication, one often takes $\mathcal{A} = \{0, 1\}$. To make it a metric space, we take the discrete metric $d_{\mathcal{A}}$ on $\mathcal{A}$: $d_{\mathcal{A}}(a, a') = 0$ if $a = a'$ and $d_{\mathcal{A}}(a, a') = 1$ if $a \neq a'$. This makes $(\mathcal{A}, d_{\mathcal{A}})$ a compact metric space.

In this thesis, especially for our information-theoretic setting for the SMB theorem, we have our space Ω is taken to be the set of all sequences indexed by natural numbers and taking values in $\mathcal{A}$, namely $\Omega = \mathcal{A}^{\mathbb{N}}$. For every $x \in \Omega$, we refer its n-th entry via simple indexing: $x_n \in \mathcal{A}$. In particular, we refer a finite string (or a finite part of $x \in \Omega$) by the following notation:

$$x_m^n = (x_m, x_{m+1}, ..., x_n) \in \mathcal{A}^{n-m+1} = \underbrace{\mathcal{A} \times \cdots \times \mathcal{A}}_{(n-m+1)\text{-fold}},$$

for any natural numbers $m \leq n$.

We usually equip Ω with topology of pointwise convergence, which is generated by the metric

$$d(x,y) := \sum_{n=1}^{\infty} 2^{-n} d_{\mathcal{A}}(x_n, y_n),\ \forall\, x, y \in \Omega.$$

This makes (Ω, d) a compact metric space, which could be shown using diagonal arguments.

When given some $x_m^n \in \mathcal{A}^{n-m+1}$, a cylinder set, which is a subset of Ω, determined by x_m^n, is then defined to be

$$[x_m^n] := \{y \in \Omega : y_m^n = x_m^n\}.$$

If $m = 1$, we say the cylinder set is standard and x_1^n is the prefix word of the cylinder. Standard cylinder sets are fundamentally important in our one-sided shift context, since they coincide with open balls in Ω. We then take the Borel σ-algebra $\mathcal{F}$ for Ω to be the one generated by standard cylinder sets.

The shift map $T : \Omega \longrightarrow \Omega$ is defined by $(T(x))_n := x_{n+1}$ for all $n \in \mathbb{N}$, for all $x \in \Omega$. It gives the desired one-sided shifting dynamic on our space $(\Omega, \mathcal{F})$. A T-invariant probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F}, T)$ is then called shift-invariant, and we denote the set of all shift-invariant probability measures on Ω by $\mathcal{P}_{\text{inv}}(\Omega)$. Recall that if $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$ satisfies that $\mathbb{P}(E) \in \{0, 1\}$ for all $E \in \mathcal{F}_T$, then $\mathbb{P}$ is said to be ergodic. We denote the set of all ergodic elements in $\mathcal{P}_{\text{inv}}(\Omega)$ by $\mathcal{P}_{\text{erg}}(\Omega)$.

There are some fundamental aspects of shift-invariant measures defined on Ω worth mentioning, and we include this part in Appendix C. Another notion that is absolutely important in our information-theoretic context is entropy. It is usually denoted by S and we refer the readers to Appendix D for a brief introduction of entropy and some of its basic facets.

Now with the basic setup and related notions reviewed, here we present the SMB theorem:

Theorem 3.1.1 (Shannon–McMillan–Breiman). *Let $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$. Then the limit*

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}([x_1^n]) = h_{\mathbb{P}}(x)$$

exists for $\mathbb{P}$-almost all $x \in \Omega$ and the convergence also holds in $L^1(\Omega, \mathrm{d}\mathbb{P})$.

The limit function $h_{\mathbb{P}}$ satisfies $h_{\mathbb{P}} \geq 0$ $\mathbb{P}$-a.s. and $h_{\mathbb{P}} \circ T = h_{\mathbb{P}}$ (shift-invariant) and

$$\int_{\Omega} h_{\mathbb{P}}(x)\, \mathrm{d}\mathbb{P} = S(\mathbb{P}).$$

In particular, if $\mathbb{P} \in \mathcal{P}_{\mathrm{erg}}(\Omega)$, then

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}([x_1^n]) = S(\mathbb{P})$$

for $\mathbb{P}$-almost all $x \in \Omega$ and in $L^1(\Omega, \mathrm{d}\mathbb{P})$.

The proofs of the SMB theorem will be given in Chapter 4. For the moment, we go through some essential mathematical constructions which will be used in the first two proofs. They are a bit long to discuss and take up much space, so instead of being introduced in Chapter 4, they are included in this chapter right after the statement of the SMB theorem.

3.2 Preparations for the proofs

3.2.1 Extension to two-sided shift

Analogous to one-sided shift, there is also two-sided shift (also known as full shift), where potential links between the two setups can be drawn such as extension from one to another. Here we shall briefly discuss about two-sided shift over finite alphabet, which we still use $\mathcal{A}$ to denote, and suppose $|\mathcal{A}| = \ell$.

The setup for two-sided shift over $\mathcal{A}$ is then $\widehat{\Omega} = \mathcal{A}^{\mathbb{Z}}$, namely $\widehat{\Omega}$ is the set of all maps $x : \mathbb{Z} \longrightarrow \mathcal{A}$ or sequences of the form $(x_k)_{k\in\mathbb{Z}}$ with $x_k \in \mathcal{A}$ for each $k \in \mathbb{Z}$. We also equip $\widehat{\Omega}$ with topology of pointwise convergence. There are many metrics that generate this topology and a canonical choice is $\hat{d}(x, y) := \lambda^{n(x,y)}$ for any $x, y \in \widehat{\Omega}$, where $\lambda \in (0, 1)$ is given and $n(x, y) := \min\{|k| : x_k \neq y_k\}$. One can check that $(\widehat{\Omega}, \hat{d})$ is a compact metric space.

Finite strings and cylinder sets in the two-sided shift context are denoted and defined analogously. For two integers $m \leq n$, we denote $x_m^n = (x_m, ..., x_n) \in \mathcal{A}^{n-m+1}$. A cylinder set determined by x_m^n is

$$[x_m^n] := \left\{ y \in \widehat{\Omega} : y_k = x_k, \ \ \forall\, m \leq k \leq n \right\}.$$

Cylinder sets are again both open and closed in $\widehat{\Omega}$ and the family of all cylinder sets generates the Borel σ-algebra in $\widehat{\Omega}$.

The dynamic given in the two-sided shift setting is the left shift $\widehat{T}$, which is a map $\widehat{T} : \widehat{\Omega} \longrightarrow \widehat{\Omega}$ defined by $(\widehat{T}(x))_n = x_{n+1}$ for all $n \in \mathbb{Z}$. $\widehat{T}$ is a continuous bijection and its inverse is the right shift on $\widehat{\Omega}$. In particular, $\widehat{T}$ is a homeomorphism on $(\widehat{\Omega}, \hat{d})$. Note that a big difference between one-sided and two-sided shifts is that in one-sided shift, T is onto, but not bijective.

For any $\widehat{\mathbb{P}} \in \mathcal{P}(\widehat{\Omega})$ and integers $m \leq n$, we define its (m, n)-marginal $\widehat{\mathbb{P}}_m^n$ on $\mathcal{A}^{n-m+1}$ by

$$\widehat{\mathbb{P}}_m^n(x_m^n) = \widehat{\mathbb{P}}([x_m^n]), \quad \forall\, x_m^n \in \mathcal{A}^{n-m+1}.$$

The marginals in this case also satisfy a consistency condition and a version of Kolmogorov's consistency theorem in this case follows. We refer the readers to the end of Appendix C. Shift-invariance is defined in the same way as one-sided shift case: $\widehat{\mathbb{P}} \in \mathcal{P}_{\text{inv}}(\widehat{\Omega})$ if $\widehat{\mathbb{P}} \circ \widehat{T}^{-1} = \widehat{\mathbb{P}}$.

As mentioned in the beginning, one-sided shift space Ω and two-sided shift space $\widehat{\Omega}$ possess potential yet deep links, and there is one close relation between $\mathcal{P}_{\text{inv}}(\Omega)$ and $\mathcal{P}_{\text{inv}}(\widehat{\Omega})$:

Any element $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$ uniquely extends to an element $\widehat{\mathbb{P}} \in \mathcal{P}_{\text{inv}}(\widehat{\Omega})$ by setting that, for any integers $m \leq n$,

$$\widehat{\mathbb{P}}([x_m^n]) := \mathbb{P}([x_{m+k}^{n+k}]), \quad \text{where } m + k \geq 1.$$

Note that by shift-invariance, it does not matter what value k is, as long as $m + k \geq 1$.

On the other hand, every $\widehat{\mathbb{P}} \in \mathcal{P}_{\text{inv}}(\widehat{\Omega})$ determines a unique element $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$ by setting

$$\mathbb{P}([x_1^n]) := \widehat{\mathbb{P}}([x_1^n]), \quad \forall\, n \in \mathbb{N}.$$

Therefore, we have bijective correspondence between $\mathcal{P}_{\text{inv}}(\Omega)$ and $\mathcal{P}_{\text{inv}}(\widehat{\Omega})$, and this offers convenience to construct the other shift setting, either by extension or restriction, from the given setting.

3.2.2 Functions Z_n and $Z_{\max}$

In this subsection, we would like to construct a series of functions, denoted as Z_n for $n \geq 2$, defined on the one-sided shift space $\Omega = \mathcal{A}^{\mathbb{N}}$. The sequence of the functions Z_n possess good convergence behavior and it extends some consequences and subsequent constructions which contribute a lot to the first and second proofs of the SMB theorem. We shall first give the definition.

Given some $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$, let

$$Z_n(x) := \frac{\mathbb{P}([x_2^n])}{\mathbb{P}([x_1^n])} = \frac{\mathbb{P}_{n-1}(x_2^n)}{\mathbb{P}_n(x_1^n)}$$

for any $n \geq 2$.

Note that Z_n is well-defined $\mathbb{P}$-a.s. If we set

$$F_n := \{x \in \Omega : \mathbb{P}([x_1^n]) = 0\} \text{ and } E_n := \{x_1^n \in \mathcal{A}^n : [x_1^n] \subseteq F_n\}$$

for any $n \geq 1$, then trivially,

$$F_n \subseteq \bigsqcup_{x_1^n \in E_n} [x_1^n] \quad \Rightarrow \quad \mathbb{P}(F_n) \leq \sum_{x_1^n \in E_n} \mathbb{P}([x_1^n]) = 0 \quad \Rightarrow \quad \mathbb{P}(F_n) = 0.$$

Hence, $\mathbb{P}([x_1^n]) > 0$ for $\mathbb{P}$-almost all $x \in \Omega$, and this shows $\mathbb{P}$-almost sure well-definedness of Z_n for any $n \geq 2$. Note that, on the domain with full measure where Z_n is well-defined, $Z_n \geq 1$, due to compatibility implied by shift-invariance (c.2).

There is a crucial property of Z_n which relates closely to the entropy associated to $\mathbb{P}$ and its marginals, and in turn motivates its definition:

$$S(\mathbb{P}_{n+1}) - S(\mathbb{P}_n) = \int_\Omega \log Z_{n+1} \,\mathrm{d}\mathbb{P}. \tag{3.1}$$

We approach this crucial relation from some nice properties of entropy of shift-invariant measures on Ω. However, before delving into this discussion, we quote two useful results: one is an inequality observed and proven in [Jak19, Proposition 4.7], named the log-sum inequality, and the other is about distribution functions from [Ru87, Chapter 8]. We refer their proofs to these references.

Lemma 3.2.1. *Let $N \in \mathbb{N}$ and suppose a_j and b_j are non-negative numbers for $j \in \{1, ..., N\}$. Then*

$$\sum_{j=1}^N a_j \log \frac{a_j}{b_j} \geq \sum_{j=1}^N a_j \log \left(\frac{\sum_{k=1}^N a_k}{\sum_{k=1}^N b_k} \right),$$

with the usual convention that "$0 \cdot \log 0 = 0$", "$0/0 = 0$", and so on.

Theorem 3.2.2. *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable function $f : \Omega \longrightarrow [0, \infty]$,*

$$\int_\Omega f \,\mathrm{d}\mathbb{P} = \int_0^\infty \mathbb{P}(\{x \in \Omega : f(x) > t\}) \,\mathrm{d}t.$$

Now let us introduce the important properties of entropy of shift-invariant measures on Ω as mentioned above.

Theorem 3.2.3. *Let* $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$ *and set*

$$A_n = S(\mathbb{P}_{n+1}) - S(\mathbb{P}_n), \quad \forall\, n \in \mathbb{N}.$$

Then:

(1). $0 \leq A_n \leq \log \ell$ *for all* $n \in \mathbb{N}$*, where* $|\mathcal{A}| = \ell$*;*

(2). $A_n \geq A_{n+1}$ *for all* $n \in \mathbb{N}$*;*

(3). $\lim_{n\to\infty} A_n = S(\mathbb{P})$.

Proof. Let $n \in \mathbb{N}$ be fixed. For any $x_1^n \in \operatorname{supp} \mathbb{P}_n$, we define a probability measure on $\mathcal{A}$ as follows:

$$\mathbb{P}_{n+1}^{x_1^n}(a) := \frac{\mathbb{P}_{n+1}(x_1, ..., x_n, a)}{\mathbb{P}_n(x_1, ..., x_n)}, \quad \forall\, a \in \mathcal{A}.$$

One can trivially check it is a probability measure. We then claim the following relation:

$$S(\mathbb{P}_{n+1}) = S(\mathbb{P}_n) + \sum_{x_1^n \in \operatorname{supp} \mathbb{P}_n} \mathbb{P}_n(x_1^n) S\left(\mathbb{P}_{n+1}^{x_1^n}\right). \tag{3.2}$$

We show (3.2) by explicitly expanding and computing the right hand side. For each $x_1^n \in \operatorname{supp} \mathbb{P}_n$,

$$\begin{aligned} S\left(\mathbb{P}_{n+1}^{x_1^n}\right) &= -\sum_{a\in\mathcal{A}} \mathbb{P}_{n+1}^{x_1^n}(a) \log \mathbb{P}_{n+1}^{x_1^n}(a) \\ &= -\sum_{a\in\mathcal{A}} \frac{\mathbb{P}_{n+1}(x_1, ..., x_n, a)}{\mathbb{P}_n(x_1, ..., x_n)} \log\left(\frac{\mathbb{P}_{n+1}(x_1, ..., x_n, a)}{\mathbb{P}_n(x_1, ..., x_n)}\right) \\ &= -\sum_{a\in\mathcal{A}} \frac{\mathbb{P}_{n+1}(x_1, ..., x_n, a)}{\mathbb{P}_n(x_1, ..., x_n)} \left(\log \mathbb{P}_{n+1}(x_1, ..., x_n, a) - \log \mathbb{P}_n(x_1, ..., x_n)\right). \end{aligned}$$

Then,

$$\begin{aligned} &\sum_{x_1^n \in \operatorname{supp} \mathbb{P}_n} \mathbb{P}_n(x_1^n) S\left(\mathbb{P}_{n+1}^{x_1^n}\right) \\ &= -\sum_{x_1^n \in \operatorname{supp} \mathbb{P}_n} \mathbb{P}_n(x_1^n) \sum_{a\in\mathcal{A}} \frac{\mathbb{P}_{n+1}(x_1, ..., x_n, a)}{\mathbb{P}_n(x_1, ..., x_n)} \left(\log \mathbb{P}_{n+1}(x_1, ..., x_n, a) - \log \mathbb{P}_n(x_1, ..., x_n)\right) \\ &= -\sum_{x_1^n \in \operatorname{supp} \mathbb{P}_n} \sum_{a\in\mathcal{A}} \mathbb{P}_{n+1}(x_1, ..., x_n, a) \left(\log \mathbb{P}_{n+1}(x_1, ..., x_n, a) - \log \mathbb{P}_n(x_1, ..., x_n)\right) \end{aligned}$$

$$=-\sum_{x_1^n\in\operatorname{supp}\mathbb{P}_n}\sum_{a\in\mathcal{A}}(\mathbb{P}_{n+1}(x_1,...,x_n,a)\log\mathbb{P}_{n+1}(x_1,...,x_n,a)-$$

$$\mathbb{P}_{n+1}(x_1,...,x_n,a)\log\mathbb{P}_n(x_1,...,x_n))$$

$$=-\sum_{x_1^n\in\operatorname{supp}\mathbb{P}_n}\sum_{a\in\mathcal{A}}\mathbb{P}_{n+1}(x_1,...,x_n,a)\log\mathbb{P}_{n+1}(x_1,...,x_n,a)+$$

$$\sum_{x_1^n\in\operatorname{supp}\mathbb{P}_n}\log\mathbb{P}_n(x_1,...,x_n)\sum_{a\in\mathcal{A}}\mathbb{P}_{n+1}(x_1,...,x_n,a)$$

$$=-\sum_{(x_1,...,x_n,a)\in\mathcal{A}^{n+1}}\mathbb{P}_{n+1}(x_1,...,x_n,a)\log\mathbb{P}_{n+1}(x_1,...,x_n,a)+\sum_{x_1^n\in\mathcal{A}^n}\mathbb{P}_n(x_1^n)\log\mathbb{P}_n(x_1^n)$$

$$=S(\mathbb{P}_{n+1})-S(\mathbb{P}_n).$$

This proves the claim (3.2). By Proposition D.1, entropy is always non-negative, so

$$\sum_{x_1^n\in\operatorname{supp}\mathbb{P}_n}\mathbb{P}_n(x_1^n)S\left(\mathbb{P}_{n+1}^{x_1^n}\right)\geq 0.$$

On the other hand, Proposition D.1 also suggests

$$S\left(\mathbb{P}_{n+1}^{x_1^n}\right)\leq\log\ell$$

for all $x_1^n\in\operatorname{supp}\mathbb{P}_n$.

Then (3.2) implies that

$$S(\mathbb{P}_{n+1})\geq S(\mathbb{P}_n)\ \Rightarrow\ A_n=S(\mathbb{P}_{n+1})-S(\mathbb{P}_n)\geq 0,$$

and

$$S(\mathbb{P}_{n+1})\leq S(\mathbb{P}_n)+\log\ell\sum_{x_1^n\in\operatorname{supp}\mathbb{P}_n}\mathbb{P}_n(x_1^n)=S(\mathbb{P}_n)+\log\ell$$

which gives $A_n=S(\mathbb{P}_{n+1})-S(\mathbb{P}_n)\leq\log\ell$.

This proves Property (1).

To prove (2), we take use of Lemma 3.2.1. Note that, by adopting the reverse thinking of the steps in the derivation of claim (3.2), we have

$$\begin{aligned}
S(\mathbb{P}_{n+1})-S(\mathbb{P}_n) &= -\sum_{x_1^{n+1}\in\mathcal{A}^{n+1}} \mathbb{P}_{n+1}(x_1^{n+1})\log\mathbb{P}_{n+1}(x_1^{n+1}) + \sum_{x_2^{n+1}\in\mathcal{A}^{n}} \mathbb{P}_{n}(x_2^{n+1})\log\mathbb{P}_{n}(x_2^{n+1}) \\
\text{(By (c.2))} \quad &= -\sum_{x_1^{n+1}\in\mathcal{A}^{n+1}} \mathbb{P}_{n+1}(x_1^{n+1})\log\mathbb{P}_{n+1}(x_1^{n+1}) + \sum_{x_1^{n+1}\in\mathcal{A}^{n+1}} \mathbb{P}_{n+1}(x_1^{n+1})\log\mathbb{P}_{n}(x_2^{n+1}) \\
&= -\sum_{x_1^{n+1}\in\mathcal{A}^{n+1}} \mathbb{P}_{n+1}(x_1^{n+1})\log\left(\frac{\mathbb{P}_{n+1}(x_1^{n+1})}{\mathbb{P}_{n}(x_2^{n+1})}\right) \\
\text{(By (c.1))} \quad &= -\sum_{x_1^{n+1}\in\mathcal{A}^{n+1}} \left(\sum_{a\in\mathcal{A}} \mathbb{P}_{n+2}(x_1,...,x_{n+1},a)\log\left(\frac{\sum_{a\in\mathcal{A}}\mathbb{P}_{n+2}(x_1,...,x_{n+1},a)}{\sum_{a\in\mathcal{A}}\mathbb{P}_{n+1}(x_2,...,x_{n+1},a)}\right)\right)
\end{aligned}$$

and we can see from here, that the terms inside the big bracket have the form that matches what appears in the log-sum inequality. Thus, by Lemma 3.2.1,

$$\begin{aligned}
S(\mathbb{P}_{n+1})-S(\mathbb{P}_n) &= -\sum_{x_1^{n+1}\in\mathcal{A}^{n+1}} \left(\sum_{a\in\mathcal{A}} \mathbb{P}_{n+2}(x_1,...,x_{n+1},a)\log\left(\frac{\sum_{a\in\mathcal{A}}\mathbb{P}_{n+2}(x_1,...,x_{n+1},a)}{\sum_{a\in\mathcal{A}}\mathbb{P}_{n+1}(x_2,...,x_{n+1},a)}\right)\right) \\
&\geq -\sum_{x_1^{n+1}\in\mathcal{A}^{n+1}} \left(\sum_{a\in\mathcal{A}} \mathbb{P}_{n+2}(x_1,...,x_{n+1},a)\log\left(\frac{\mathbb{P}_{n+2}(x_1,...,x_{n+1},a)}{\mathbb{P}_{n+1}(x_2,...,x_{n+1},a)}\right)\right) \\
&= -\sum_{x_1^{n+2}\in\mathcal{A}^{n+2}} \mathbb{P}_{n+2}(x_1^{n+2})\log\left(\frac{\mathbb{P}_{n+2}(x_1^{n+2})}{\mathbb{P}_{n+1}(x_2^{n+2})}\right) \\
&= S(\mathbb{P}_{n+2})-S(\mathbb{P}_{n+1}).
\end{aligned}$$

This shows $A_n \geq A_{n+1}$, which is Property (2).

Property (3) then simply follows from (1) and (2). Indeed, (1) and (2) suggest that $\{A_n\}_{n\in\mathbb{N}}$ is a decreasing real sequence bounded below by 0, so $\lim_{n\to\infty} A_n$ exists by the monotone convergence theorem. Its value, on the other hand, can be derived by the so-called Cesàro mean.

We claim that

$$\lim_{n\to\infty} A_n = \lim_{n\to\infty} \frac{A_1+\cdots+A_n}{n}, \tag{3.3}$$

where the right hand side is known as the Cesàro mean. We can quickly justify this relation. Suppose that

$$\lim_{n\to\infty} A_n = L$$

for some $L \in [0, \log \ell]$. Let $\varepsilon > 0$ be arbitrary. Then there exists some $N \in \mathbb{N}$ such that for all $n \geq N$, $|A_n - L| < \frac{\varepsilon}{3}$. Besides, let $d = \left(\frac{3}{\varepsilon}L - 1\right) N$. Then for any $n \geq \max(N, d)$,

$$\begin{aligned}
\left| \frac{1}{N+n} \sum_{k=N+1}^{N+n} A_k - L \right| &= \left| \frac{\sum_{k=N+1}^{N+n} A_k - nL - NL}{N+n} \right| \\
&\leq \frac{\sum_{k=N+1}^{N+n} |A_k - L| + NL}{N+n} \\
&\leq \frac{\sum_{k=N+1}^{N+n} |A_k - L|}{n} + \frac{NL}{N+n} \\
&< \frac{\varepsilon}{3} + \frac{NL}{N+d} = \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \frac{2\varepsilon}{3}.
\end{aligned}$$

Now pick $M \in \mathbb{N}$ large enough so that $\frac{|A_1 + \cdots + A_N|}{M} < \frac{\varepsilon}{3}$ and $M \geq N + \max(N, d)$. Denote $R = M - N \geq \max(N, d)$ and observe the following:

$$\begin{aligned}
\left| \frac{1}{M} \sum_{k=1}^{M} A_k - L \right| &= \left| \frac{1}{M} \sum_{k=1}^{N} A_k + \frac{1}{M} \sum_{k=N+1}^{M} A_k - L \right| \\
&\leq \left| \frac{1}{M} \sum_{k=1}^{N} A_k \right| + \left| \frac{1}{N+R} \sum_{k=N+1}^{N+R} A_k - L \right| \\
&< \frac{\varepsilon}{3} + \frac{2\varepsilon}{3} = \varepsilon.
\end{aligned}$$

This proves the claim (3.3). Plugging in the expression for each A_n, we then get a telescoping sum in the expression of the Cesàro mean. This gives

$$\lim_{n\to\infty} A_n = \lim_{n\to\infty} \frac{A_1 + \cdots + A_n}{n} = \lim_{n\to\infty} \frac{S(\mathbb{P}_{n+1}) - S(\mathbb{P}_1)}{n} = \lim_{n\to\infty} \frac{S(\mathbb{P}_n)}{n} = S(\mathbb{P}).$$

This finishes the proof of Property (3). □

Repeating the same manipulation as done in some steps in the proof of Theorem 3.2.3, together with the definition of Z_n, we bring forward the relation (3.1):

$$\begin{aligned}
A_n = S(\mathbb{P}_{n+1}) - S(\mathbb{P}_n) &= \sum_{x_1^{n+1} \in \mathcal{A}^{n+1}} \mathbb{P}_{n+1}(x_1^{n+1}) \log \left(\frac{\mathbb{P}_n(x_2^{n+1})}{\mathbb{P}_{n+1}(x_1^{n+1})} \right) \\
&= \sum_{x_1^{n+1} \in \mathcal{A}^{n+1}} \mathbb{P}([x_1^{n+1}]) \log Z_{n+1}(x)
\end{aligned}$$

$$= \int_{\Omega} \log Z_{n+1}(x)\, \mathrm{d}\mathbb{P}.$$

By Theorem 3.2.3 (3), we have

$$\lim_{n\to\infty} A_n = S(\mathbb{P}) = \lim_{n\to\infty} \int_{\Omega} \log Z_n(x)\, \mathrm{d}\mathbb{P}. \tag{3.4}$$

We shall keep this relation for later use in the proofs in Chapter 4.

We also have Z_n is integrable for all $n \geq 2$:

$$\int_{\Omega} Z_n \,\mathrm{d}\mathbb{P} = \sum_{x_1^n \in \mathcal{A}^n} Z_n(x)\mathbb{P}([x_1^n]) = \sum_{x_1^n \in \mathcal{A}^n} \frac{\mathbb{P}_{n-1}(x_2^n)}{\mathbb{P}_n(x_1^n)} \mathbb{P}_n(x_1^n)$$
$$= \sum_{x_1^n \in \operatorname{supp} \mathbb{P}_n} \frac{\mathbb{P}_{n-1}(x_2^n)}{\mathbb{P}_n(x_1^n)} \mathbb{P}_n(x_1^n) = \sum_{x_1^n \in \operatorname{supp} \mathbb{P}_n} \mathbb{P}_{n-1}(x_2^n) \leq \sum_{x_1^n \in \mathcal{A}^n} \mathbb{P}_{n-1}(x_2^n) = \ell,$$

so

$$\int_{\Omega} Z_n \,\mathrm{d}\mathbb{P} \leq \ell, \quad \forall\, n \geq 2. \tag{3.5}$$

Remark. In the above step of deriving (3.5), it is important to do the transition from "$x_1^n \in \mathcal{A}^n$" to "$x_1^n \in \operatorname{supp} \mathbb{P}_n$" for the summation, because it is possible that for some $x_1^n \in \mathcal{A}^n \backslash \operatorname{supp} \mathbb{P}_n$, $\mathbb{P}_n(x_1^n) = 0$ while $\mathbb{P}_{n-1}(x_2^n) \neq 0$.

Another important subsequent construction given our already established $\{Z_n\}_{n\geq 2}$ is function $Z_{\max}$, which is set as below:

$$Z_{\max}(x) := \sup_{n\geq 2} Z_n(x) = \sup_{n\geq 2} \frac{\mathbb{P}_{n-1}(x_2^n)}{\mathbb{P}_n(x_1^n)}.$$

Since Z_n is well-defined $\mathbb{P}$-a.s. for all $n \geq 2$, $Z_{\max}$ is also well-defined $\mathbb{P}$-a.s. We next introduce two lemmas regarding $Z_{\max}$.

Lemma 3.2.4. *For any $t > 0$,*

$$\mathbb{P}\left(\{x \in \Omega : Z_{\max}(x) > t\}\right) \leq \frac{\ell}{t}.$$

Proof. Fix arbitrary $t > 0$. Denote

$$B = \{x \in \Omega : Z_{\max}(x) > t\}$$

and set

$$B_2 = \{x \in \Omega : Z_2(x) > t\} \text{ and } B_n = \{x \in \Omega : Z_2(x) \leq t, ..., Z_{n-1}(x) \leq t, Z_n(x) > t\},$$

for all $n > 2$. We then can write

$$B = \bigcup_{n=2}^{\infty} B_n.$$

We can see the sets B_n's are disjoint. For every $n \geq 2$, if we denote $\mathcal{F}_n$ to be the sub-σ-algebra generated by the standard cylinder sets $[x_1^n]$ for $x_1^n \in \mathcal{A}^n$, then B_n is measurable with respect to $\mathcal{F}_n$ and if $x \in B_n$, then $[x_1^n] \subseteq B_n$. This is due to the definitions of Z_n and B_n for each $n \geq 2$.

To proceed, we construct a probability measure $\tilde{\mathbb{P}}$ on Ω defined by the marginals:

$$\tilde{\mathbb{P}}_1(x_1) = \frac{1}{\ell}, \quad \forall\, x_1 \in \mathcal{A},$$

and

$$\tilde{\mathbb{P}}_n(x_1, ..., x_n) = \tilde{\mathbb{P}}_1(x_1)\mathbb{P}_{n-1}(x_2, ..., x_n), \quad \forall\, x_1^n \in \mathcal{A}^n,$$

for any $n \geq 2$.

One can easily check the above marginals $\{\tilde{\mathbb{P}}_n\}_{n\in\mathbb{N}}$ are well-defined probability measures on $\mathcal{A}^n$, and by Kolmogorov's consistency theorem (it suffices to use the simple version of the special case of one-sided shift, namely Theorem C.1), $\tilde{\mathbb{P}}$ is a uniquely defined probability measure on Ω by these marginals. The introduction of $\tilde{\mathbb{P}}$ is to give an upper bound to the measure of each B_n, and we will see how this is done below.

For rigorous use of languages, we set

$$\tilde{B}_n = \{x_1^n \in \mathcal{A}^n : [x_1^n] \subseteq B_n\}, \quad \forall\, n \geq 2.$$

On each B_n, we know $Z_n > t$. With an application of Chebyshev's inequality, we have

$$\mathbb{P}(B_n) \leq \frac{1}{t}\int_{B_n} Z_n(x)\,\mathrm{d}\mathbb{P} = \frac{1}{t}\int_{B_n} \frac{\mathbb{P}_{n-1}(x_2^n)}{\mathbb{P}_n(x_1^n)}\,\mathrm{d}\mathbb{P} = \frac{1}{t}\sum_{\substack{x_1^n \in \tilde{B}_n \\ x_1^n \in \operatorname{supp}\mathbb{P}_n}} \mathbb{P}_n(x_1^n)\frac{\mathbb{P}_{n-1}(x_2^n)}{\mathbb{P}_n(x_1^n)}$$

$$= \frac{1}{t}\sum_{\substack{x_1^n \in \tilde{B}_n \\ x_1^n \in \operatorname{supp}\mathbb{P}_n}} \mathbb{P}_{n-1}(x_2^n) \leq \frac{1}{t}\sum_{x_1^n \in \tilde{B}_n} \mathbb{P}_{n-1}(x_2^n) = \frac{\ell}{t}\sum_{x_1^n \in \tilde{B}_n} \frac{1}{\ell}\cdot\mathbb{P}_{n-1}(x_2^n)$$

$$= \frac{\ell}{t} \sum_{x_1^n \in \tilde{B}_n} \tilde{\mathbb{P}}_1(x_1)\mathbb{P}_{n-1}(x_2^n) = \frac{\ell}{t} \sum_{x_1^n \in \tilde{B}_n} \tilde{\mathbb{P}}_n(x_1^n) = \frac{\ell}{t} \int_{B_n} \mathrm{d}\tilde{\mathbb{P}} = \frac{\ell}{t}\, \tilde{\mathbb{P}}(B_n).$$

Therefore, we have obtained this bounded from above relation

$$\mathbb{P}(B_n) \leq \frac{\ell}{t}\, \tilde{\mathbb{P}}(B_n) \tag{3.6}$$

for each $n \geq 2$.

Besides, B_n's are disjoint. Then with (3.6), it gives

$$\mathbb{P}(B) = \mathbb{P}\left(\bigcup_{n=2}^{\infty} B_n\right) = \sum_{n=2}^{\infty} \mathbb{P}(B_n) \leq \frac{\ell}{t} \sum_{n=2}^{\infty} \tilde{\mathbb{P}}(B_n) = \frac{\ell}{t}\, \tilde{\mathbb{P}}\left(\bigcup_{n=2}^{\infty} B_n\right) = \frac{\ell}{t}\, \tilde{\mathbb{P}}(B) \leq \frac{\ell}{t}.$$

Hence, we get $\mathbb{P}(B) \leq \frac{\ell}{t}$ and this finishes the proof. □

Lemma 3.2.5. $\log Z_{\max} \in L^1(\Omega, \mathrm{d}\mathbb{P})$.

Proof. Fix arbitrary $t > 0$. By Lemma 3.2.4,

$$\mathbb{P}\left(\{x \in \Omega : \log Z_{\max} > t\}\right) = \mathbb{P}\left(\{x \in \Omega : Z_{\max} > e^t\}\right) \leq \frac{\ell}{e^t} = \ell e^{-t}.$$

Note that $Z_{\max} \geq 1$ $\mathbb{P}$-a.s., and thus, $\log Z_{\max} \geq 0$ $\mathbb{P}$-a.s. To check integrability of $Z_{\max}$, we apply Theorem 3.2.2:

$$\int_{\Omega} \log Z_{\max}\, \mathrm{d}\mathbb{P} = \int_0^{\infty} \mathbb{P}\left(\{x \in \Omega : \log Z_{\max} > t\}\right) \mathrm{d}t \leq \int_0^{\infty} \ell e^{-t}\, \mathrm{d}t = \ell.$$

Hence, $\log Z_{\max} \in L^1(\Omega, \mathrm{d}\mathbb{P})$. □

Chapter 4

Three proofs of the SMB theorem

We give three different proofs of the SMB theorem in this chapter, which are referred as the subadditive proof, the martingale proof, and the Ornstein–Weiss proof respectively. Among these three proofs, the subadditive proof and the martingale proof will fully justify Theorem 3.1.1, while the Ornstein–Weiss proof will only show the ergodic case of the theorem.

4.1 The subadditive proof

In the given setup of the SMB theorem, we would like to undergo some constructions and manipulations to create subadditive condition so that Kingman's subadditive ergodic theorem (Theorem 2.2.1) can be applied to give the desired conclusion. Due to the central involvement of the subadditive ergodic theorem, we refer this first proof as "the subadditive proof".

The subadditive proof was originally given by Derriennic [Der83], in the same paper that he proved a generalized version of Kingman's subadditive ergodic theorem, which we presented as Theorem 2.2.1 in Section 2.2. Here we give a refined version of the proof under the one-sided shift setting, and it will involve the extension from one-sided shift to two-sided shift.

Proof. First, we would like to extend our one-sided shift setting to two-sided shift one, as it will turn out that the two-sided shift case will be more convenient to work on for our purpose. Please see Remark 4.1.1 for a more detailed explanation. The related standard extension has been introduced in Subsection 3.2.1.

Let $\widehat{\Omega} = \mathcal{A}^{\mathbb{Z}}$ and our given $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$ can be uniquely extended to some $\widehat{\mathbb{P}} \in \mathcal{P}_{\text{inv}}(\widehat{\Omega})$. If we establish the conclusion of the SMB theorem for the two-sided shift case, namely the convergence of $-\frac{1}{n}\log \widehat{\mathbb{P}}([x_1^n])$, then we are done since the $\widehat{\mathbb{P}}$-a.s. convergence and the L^1 convergence of $-\frac{1}{n}\log \widehat{\mathbb{P}}([x_1^n])$ are essentially the same as the ones of $-\frac{1}{n}\log \mathbb{P}([x_1^n])$ and the conclusion can be smoothly translated to the one-sided shift case. We will elaborate this in the end of the proof.

We begin with letting

$$\widehat{X}_n(x) := -\log \widehat{\mathbb{P}}_1^n(x_1^n) = -\log \widehat{\mathbb{P}}([x_1^n]), \quad \forall\, n \in \mathbb{N}.$$

Each $\widehat{X}_n$ is well-defined for $\widehat{\mathbb{P}}$-almost all $x \in \widehat{\Omega}$ and clearly each $\widehat{X}_n$ is non-negative. We then want to show that $\widehat{X}_n$ is in $L^1(\widehat{\Omega}, \widehat{\mathbb{P}})$ for each $n \in \mathbb{N}$.

Recall the function $\log Z_{\max}$ constructed and analyzed in Subsection 3.2.2. We can define a version of this function on $\widehat{\Omega}$ (denoted as $\log \widehat{Z}_{\max}$) analogously to the definition given in the one-sided shift case and this can be done by repeating exactly the same definitions given to Z_n for $n \geq 2$ and $Z_{\max}$ in Section 3.2, except substituting Ω with $\widehat{\Omega}$. We would have $\log \widehat{Z}_{\max}(x) = \log Z_{\max}(x|_\Omega)$ for every $x \in \widehat{\Omega}$, where $x|_\Omega$ stands for the restriction of x on Ω. This is because the definition of every Z_n only depends on $x_1, \ldots, x_n$, whose indices are all positive. Then all the subsequent results related to $\log Z_{\max}$ also translate to the two-sided shift case and hold true for $\log \widehat{Z}_{\max}$, especially Lemma 3.2.5, so we have $\log \widehat{Z}_{\max} \in L^1(\widehat{\Omega}, \widehat{\mathbb{P}})$.

We then observe the following manipulation and derive an *a priori* bound for $\widehat{X}_n$:

$$\begin{aligned}\widehat{X}_n(x) = \log \frac{1}{\widehat{\mathbb{P}}_1^n(x_1^n)} &= \log \left(\frac{\widehat{\mathbb{P}}_1^{n-1}(x_2^n)}{\widehat{\mathbb{P}}_1^n(x_1^n)} \cdot \frac{\widehat{\mathbb{P}}_1^{n-2}(x_3^n)}{\widehat{\mathbb{P}}_1^{n-1}(x_2^n)} \cdots \frac{\widehat{\mathbb{P}}_1^1(x_n)}{\widehat{\mathbb{P}}_1^2(x_{n-1}^n)} \cdot \frac{1}{\widehat{\mathbb{P}}_1^1(x_n)} \right) \\ &\leq \log \left(\sup_{n \geq 2} \frac{\widehat{\mathbb{P}}_1^{n-1}(x_2^n)}{\widehat{\mathbb{P}}_1^n(x_1^n)} \right)^{n-1} + \log \frac{1}{\widehat{\mathbb{P}}_1^1(x_n)} \leq (n-1) \log \widehat{Z}_{\max}(x) + K,\end{aligned}$$

where

$$K = \sup_{a \in \operatorname{supp} \widehat{\mathbb{P}}_1^1} \left(\log \frac{1}{\widehat{\mathbb{P}}_1^1(a)} \right) < \infty.$$

Since K is a finite constant and $\log \widehat{Z}_{\max} \in L^1(\widehat{\Omega}, \widehat{\mathbb{P}})$, then we have $\widehat{X}_n \in L^1(\widehat{\Omega}, \widehat{\mathbb{P}})$ for each $n \in \mathbb{N}$.

We now want to establish the desired weak subadditivity with respect to our $\{\widehat{X}_n\}_{n\in\mathbb{N}}$. For

$n, m \in \mathbb{N}$, set

$$D_{n,m}(x) := \log\left(\frac{\widehat{\mathbb{P}}([x_1^n])\widehat{\mathbb{P}}([x_{n+1}^{n+m}])}{\widehat{\mathbb{P}}([x_1^{n+m}])}\right),$$

then with easy computation, it can be checked that

$$\widehat{X}_{n+m} = \widehat{X}_n + \widehat{X}_m \circ \widehat{T}^n + D_{n,m}. \tag{4.1}$$

If in addition, for any integer $j \leq 0$ and any $m \in \mathbb{N}$, we define

$$R_{j,m}(x) := \log\left(\frac{\widehat{\mathbb{P}}([x_j^0])\widehat{\mathbb{P}}([x_1^m])}{\widehat{\mathbb{P}}([x_j^m])}\right),$$

then it can be computationally checked that for any $n, m \in \mathbb{N}$,

$$R_{-(n-1),m}\left(\widehat{T}^n(x)\right) = D_{n,m}(x). \tag{4.2}$$

A good thing about the definition of $R_{j,m}$ is that in the relation (4.2) when connecting with $D_{n,m}$, it involves composition with $\widehat{T}^n$. This provides a sign that $R_{j,m}$ can be used to construct the desired error terms so that the weak subadditive condition is created. Meanwhile, we also want to generate an upper bound for (4.1), so supremum of $R_{j,m}$'s over $j \leq 0$ is a good choice to realize that.

Let us denote the desired subadditive error terms as $\widehat{Y}_n$ for $n \in \mathbb{N}$, which is the same notation as in the subadditive condition (2.5). Based on the above analysis, we set

$$\widehat{Y}_n(x) := \sup_{j \leq 0} \max\left(0, R_{j,n}(x)\right)$$

for all $n \in \mathbb{N}$. Note that this definition guarantees each $\widehat{Y}_n$ is non-negative. Together with (4.2), we further derive the following estimate from (4.1):

$$\begin{aligned}
\widehat{X}_{n+m}(x) &= \widehat{X}_n(x) + \widehat{X}_m \circ \widehat{T}^n(x) + D_{n,m}(x) \\
&= \widehat{X}_n(x) + \widehat{X}_m \circ \widehat{T}^n(x) + R_{-(n-1),m} \circ \widehat{T}^n(x) \\
&\leq \widehat{X}_n(x) + \widehat{X}_m \circ \widehat{T}^n(x) + \widehat{Y}_m \circ \widehat{T}^n(x)
\end{aligned}$$

$$\Rightarrow \quad \widehat{X}_{n+m} \leq \widehat{X}_n + \widehat{X}_m \circ \widehat{T}^n + \widehat{Y}_m \circ \widehat{T}^n, \quad \forall\, n, m \in \mathbb{N}. \tag{4.3}$$

Hence, our defined error terms $\widehat{Y}_n$'s give the desired weak subadditivity. It remains to check that $\{\widehat{Y}_n\}_{n\in\mathbb{N}}$ satisfies the two properties as outlined in the statement of Theorem 2.2.1, namely

$$\sup_{n\geq 1}\left\|\widehat{Y}_n\right\|_1 < \infty \quad \text{and} \quad \lim_{n\to\infty}\frac{\widehat{Y}_n}{n} = 0 \;\; \widehat{\mathbb{P}}\text{-a.s.} \tag{4.4}$$

To show these two properties are met, we instead prove an *a priori* estimate first: Let $t > 0$, then

$$\widehat{\mathbb{P}}\left(\left\{x \in \widehat{\Omega} : \widehat{Y}_n(x) > t\right\}\right) \leq e^{-t}, \quad \forall\, n \in \mathbb{N}. \tag{4.5}$$

Then the two properties (4.4) will follow from (4.5).

To show the *a priori* estimate (4.5), we fix arbitrary $t > 0$ and let

$$A = \left\{x \in \widehat{\Omega} : \widehat{Y}_n(x) > t\right\}.$$

By the definition of each $\widehat{Y}_n$, we can write

$$A_k = \left\{x \in \widehat{\Omega} : \max\left(0, R_{-k,n}(x)\right) > t \;\text{ and }\; \max\left(0, R_{-i,n}(x)\right) \leq t, \;\; \forall\, 0 \leq i \leq k-1\right\}$$

for every $k \in \mathbb{N}$ and then it is a basic measure-theoretic fact that

$$A = \bigsqcup_{k=1}^{\infty} A_k,$$

where A_k's are disjoint by construction and for each $k \in \mathbb{N}$, the indicator function $\mathbb{1}_{A_k}$ only depends on $x_{-k}^n \in \mathcal{A}^{n+k+1}$.

For each $x \in A_k$, we have $\max\left(0, R_{-k,n}(x)\right) > t$ by its definition and this implies that we must have $R_{-k,n}(x) > t$. Besides, based on how $R_{-k,n}$ is defined, we would then have

$$R_{-k,n}(x) > t \;\;\Rightarrow\;\; \log\left(\frac{\widehat{\mathbb{P}}([x_{-k}^0])\widehat{\mathbb{P}}([x_1^n])}{\widehat{\mathbb{P}}([x_{-k}^n])}\right) > t \;\;\Rightarrow\;\; \frac{\widehat{\mathbb{P}}([x_{-k}^0])\widehat{\mathbb{P}}([x_1^n])}{\widehat{\mathbb{P}}([x_{-k}^n])} > e^t,$$

for all $x \in A_k$. Then by Chebyshev's inequality,

$$\widehat{\mathbb{P}}(A_k) \leq \frac{1}{e^t}\int_{A_k}\frac{\widehat{\mathbb{P}}([x_{-k}^0])\widehat{\mathbb{P}}([x_1^n])}{\widehat{\mathbb{P}}([x_{-k}^n])}\,\mathrm{d}\widehat{\mathbb{P}} = e^{-t}\sum_{x_{-k}^n\in\mathcal{A}^{n+k+1}}\mathbb{1}_{A_k}(x)\,\frac{\widehat{\mathbb{P}}([x_{-k}^0])\widehat{\mathbb{P}}([x_1^n])}{\widehat{\mathbb{P}}([x_{-k}^n])}\cdot\widehat{\mathbb{P}}([x_{-k}^n])$$

$$= e^{-t} \sum_{x_{-k}^{n} \in \mathcal{A}^{n+k+1}} \mathbb{1}_{A_k}(x)\, \widehat{\mathbb{P}}([x_{-k}^{0}])\widehat{\mathbb{P}}([x_1^n]). \tag{4.6}$$

Let $\mathbb{Z}_{\leq} = \{n \in \mathbb{Z} : n \leq 0\}$ and set $\Omega_- = \mathcal{A}^{\mathbb{Z}_{\leq}}$ and $\Omega_+ = \Omega = \mathcal{A}^{\mathbb{N}}$. For our probability measure $\widehat{\mathbb{P}}$ on $\widehat{\Omega}$, by Kolmogorov's consistency theorem, if we consider the restrictions of $\widehat{\mathbb{P}}$ on Ω_- and Ω_+ respectively, then these two restrictions are probability measures on Ω_- and Ω_+ respectively. We denote them as $\widehat{\mathbb{P}}_-$ and $\widehat{\mathbb{P}}_+$ respectively.

Now take $\tilde{\mathbb{P}} = \widehat{\mathbb{P}}_- \times \widehat{\mathbb{P}}_+$, which is the product measure obtained by taking the product between $\widehat{\mathbb{P}}_-$ and $\widehat{\mathbb{P}}_+$. $\tilde{\mathbb{P}}$ is then a probability measure on $\Omega_- \times \Omega_+ = \widehat{\Omega}$. Note that $\tilde{\mathbb{P}}$ is not necessarily shift-invariant. We can see that the term (4.6) looks exactly like the decomposed form of $\tilde{\mathbb{P}}$, so given this probability measure $\tilde{\mathbb{P}}$ on $\widehat{\Omega}$ we just constructed, (4.6) can be rewritten as

$$e^{-t} \sum_{x \in A_k} \widehat{\mathbb{P}}([x_{-k}^{0}])\widehat{\mathbb{P}}([x_1^n]) = e^{-t} \sum_{x \in A_k} \tilde{\mathbb{P}}([x_{-k}^{n}]) = e^{-t}\tilde{\mathbb{P}}(A_k),$$

and we obtain

$$\widehat{\mathbb{P}}(A_k) \leq e^{-t}\tilde{\mathbb{P}}(A_k),$$

which holds for all $k \in \mathbb{N}$. This then implies

$$\widehat{\mathbb{P}}(A) = \widehat{\mathbb{P}}\left(\bigsqcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \widehat{\mathbb{P}}(A_k) \leq e^{-t} \sum_{k=1}^{\infty} \tilde{\mathbb{P}}(A_k) = e^{-t}\tilde{\mathbb{P}}\left(\bigsqcup_{k=1}^{\infty} A_k\right) = e^{-t}\tilde{\mathbb{P}}(A) \leq e^{-t},$$

which gives the desired *a priori* estimate (4.5).

With (4.5) proven, we now show that $\{\widehat{Y}_n\}_{n\in\mathbb{N}}$ satisfies the two properties (4.4). First, by Theorem 3.2.2 and the estimate (4.5), we have

$$\int_{\widehat{\Omega}} \widehat{Y}_n \, \mathrm{d}\widehat{\mathbb{P}} = \int_0^{\infty} \widehat{\mathbb{P}}\left(\left\{\widehat{Y}_n > t\right\}\right) \mathrm{d}t \leq \int_0^{\infty} e^{-t} \, \mathrm{d}t = 1, \quad \forall\, n \in \mathbb{N},$$

so trivially

$$\sup_{n\geq 1} \left\|\widehat{Y}_n\right\|_1 \leq 1 < \infty.$$

On the other hand, note that for any $k \in \mathbb{N}$, (4.5) gives

$$\sum_{n=1}^{\infty} \widehat{\mathbb{P}}\left(\left\{\frac{\widehat{Y}_n}{n} > \frac{1}{k}\right\}\right) \leq \sum_{n=1}^{\infty} e^{-\frac{n}{k}} < \infty,$$

since $e^{-\frac{1}{k}} < 1$ for every $k \in \mathbb{N}$. Then by the Borel–Cantelli lemma, we have

$$\frac{1}{n}\widehat{Y}_n(x) \leq \frac{1}{k}$$

eventually always for $\widehat{\mathbb{P}}$-almost all $x \in \widehat{\Omega}$. Since this holds for all $k \in \mathbb{N}$, then

$$\lim_{n\to\infty} \frac{\widehat{Y}_n}{n} = 0 \quad \widehat{\mathbb{P}}\text{-a.s.}$$

Hence, (4.4) is justified and we have both $\{\widehat{X}_n\}_{n\in\mathbb{N}}$ and $\{\widehat{Y}_n\}_{n\in\mathbb{N}}$ meet the conditions to apply Theorem 2.2.1. By Theorem 2.2.1, there exists a limit function, denoted as $h_{\widehat{\mathbb{P}}}$, such that

$$h_{\widehat{\mathbb{P}}} = \lim_{n\to\infty} \frac{1}{n}\widehat{X}_n \quad \widehat{\mathbb{P}}\text{-a.s.}, \tag{4.7}$$

and the convergence also holds in L^1 since $\{\widehat{X}_n\}_{n\in\mathbb{N}}$ is non-negative. Moreover, $h_{\widehat{\mathbb{P}}}$ is non-negative and $\widehat{T}$-invariant, and

$$\int_{\widehat{\Omega}} h_{\widehat{\mathbb{P}}} \,\mathrm{d}\widehat{\mathbb{P}} = \lim_{n\to\infty} \frac{1}{n}\int_{\widehat{\Omega}} \widehat{X}_n \,\mathrm{d}\widehat{\mathbb{P}}. \tag{4.8}$$

As we mentioned earlier in the beginning, this result translates smoothly to the one-sided shift case. A key reason is that, by its definition, we have the functions $\widehat{X}_n$ only depend on entries with positive indices for elements in $\widehat{\Omega}$. Moreover, for each $n \in \mathbb{N}$,

$$\widehat{X}_n = -\log \widehat{\mathbb{P}}([x_1^n]) = -\log \mathbb{P}([x_1^n]), \tag{4.9}$$

so it directly relates to our studied subject in the one-sided shift setting, namely $-\frac{1}{n}\log \mathbb{P}([x_1^n])$.

Then the limit function $h_{\widehat{\mathbb{P}}}$ essentially also depends on entries with positive indices for $\widehat{\mathbb{P}}$-almost all elements in $\widehat{\Omega}$, and we can define a corresponding function $h_{\mathbb{P}}$ on Ω by simply restricting $h_{\widehat{\mathbb{P}}}$ on Ω. Then $h_{\mathbb{P}}$ is non-negative and shift-invariant and we have $h_{\mathbb{P}}(x|_{\Omega}) = h_{\widehat{\mathbb{P}}}(x)$ for $\widehat{\mathbb{P}}$-almost all $x \in \widehat{\Omega}$. Combining this with (4.7) and (4.9), we get

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}([x_1^n]) = h_{\mathbb{P}}(x)$$

for $\mathbb{P}$-almost all $x \in \Omega$. Convergence in L^1 also follows immediately, given $h_{\mathbb{P}}$ is obtained by simply restricting $h_{\widehat{\mathbb{P}}}$ on Ω and by the fact that the only variables these quantities depend on are the entries with positive indices.

Finally, plugging (4.9) into (4.8) and replacing $h_{\widehat{\mathbb{P}}}$ with $h_{\mathbb{P}}$, we have

$$\begin{aligned}\int_\Omega h_{\mathbb{P}}(x)\,\mathrm{d}\mathbb{P} &= \lim_{n\to\infty} -\frac{1}{n}\int_\Omega \log \mathbb{P}([x_1^n])\,\mathrm{d}\mathbb{P}\\ &= \lim_{n\to\infty} -\frac{1}{n}\sum_{x_1^n\in\mathcal{A}^n} \mathbb{P}_n(x_1^n)\log\mathbb{P}_n(x_1^n)\\ &= \lim_{n\to\infty}\frac{1}{n}S(\mathbb{P}_n) = S(\mathbb{P})\\ \Rightarrow \int_\Omega h_{\mathbb{P}}(x)\,\mathrm{d}\mathbb{P} &= S(\mathbb{P}).\end{aligned}$$

For the ergodic case, namely if $\mathbb{P}\in\mathcal{P}_{\text{erg}}(\Omega)$, it follows as a special case of Theorem 2.2.1 (just like the ergodic version of Theorem 2.1.1) that $h_{\mathbb{P}}$ is a constant $\mathbb{P}$-a.s. and this constant is exactly $S(\mathbb{P})$ by the relation we just derived above. □

Remark 4.1.1. We can see in the proof the convenience of extending the setting to two-sided shift, when the definition of $R_{j,m}$ for integers $j\le 0$ and $m\in\mathbb{N}$ was introduced. Deeper motivation for such extension can be understood from the functions $\widehat{X}_n$ we were concerned with and the relation (4.1), which we would like to manipulate to incorporate some error terms to satisfy the weak subadditive property. Note that in the condition of weak subadditivity, the error terms are composed with the transformation T raised to some power n. This means that in the shift setting, especially in the relation of our interest (4.1) regarding the term $D_{n,m}$ and its potential subsequent constructions, we need to preserve n entries in the elements so that they could be shifted to the left to have the desired error term match the weak subadditivity. This cannot be realized in the one-sided shift setting, but it can be done in the two-sided shift, as the left shift is bijective.

This extension is also convenient in the way that the conclusions derived in the two-sided shift setting can be smoothly translated back to the one-sided shift case. It therefore shows evidence that such extension can be an effective strategy to be used in information theory and dynamical systems when shift spaces are involved.

4.2 The martingale proof

The second proof to be presented in this section is based on the information established in Subsection 3.2.2 about the functions $\{Z_n\}_{n\ge 2}$. In fact, the entire technical input for this proof is given with respect to $\{Z_n\}_{n\geq 2}$, where Doob's martingale convergence theorem (Theorem 2.3.1) is our central technical ingredient, ensuring some nice convergence behavior of $\{Z_n\}_{n\geq 2}$. For this reason, we refer this proof of the SMB theorem as “the martingale proof”.

The use of martingales first appeared in the proofs of a series of generalizations of the early result of the SMB theorem, which were mostly done in the 1960s and 1970s. These generalizations include the extension of the conclusion to L^1 convergence and to different measures such as Lebesgue measure, counting measure, and Markov measure. The 1985 paper by Andrew R. Barron [Ba85] gives a quick historical account on these works and readers may consult it for an overview and for the listed papers that present these generalizations. The paper [Ba85] itself also gives a further generalization of the SMB theorem to non-discrete processes and extends the L^1 convergence conclusion obtained from preceding works. Its proof involves logarithms of supermartingales and it gives some insights to the second proof we are about to present, despite we are going to work in the specific setting of one-sided shift.

Proof. We start by showing more facts about the functions $\{Z_n\}_{n\geq 2}$ that will be of central importance in our martingale proof.

For each $n \in \mathbb{N}$, denote $\mathcal{F}_n$ to be the σ-algebra generated by the standard cylinder sets $[x_1^n]$, where $x_1^n \in \mathcal{A}^n$. Fix some $m, n \in \mathbb{N}$ such that $2 \leq m < n$, let us consider the conditional expectation $\mathbb{E}[Z_n|\mathcal{F}_m]$. Let φ be an $\mathcal{F}$-measurable function on Ω such that $0 \leq \varphi \leq 1$ and φ depends only on the first m entries of $x \in \Omega$, namely $x_1, x_2, ..., x_m$. Then clearly, φ is $\mathcal{F}_m$-measurable.

In the following manipulation of integrals that involve the conditional expectation $\mathbb{E}[Z_n|\mathcal{F}_m]$, we take use of Property (xii) as stated and proven in [AB19, Proposition 11.1]:

$$\begin{aligned}
\int_\Omega \mathbb{E}[Z_n|\mathcal{F}_m]\,\varphi\,\mathrm{d}\mu &= \int_\Omega \mathbb{E}[Z_n\varphi|\mathcal{F}_m]\,\mathrm{d}\mu = \int_\Omega Z_n\varphi\,\mathrm{d}\mu \\
&= \sum_{x_1^n \in \operatorname{supp}\mathbb{P}_n} \frac{\mathbb{P}_{n-1}(x_2^n)}{\mathbb{P}_n(x_1^n)}\,\varphi(x_1, ..., x_m)\mathbb{P}_n(x_1^n) \\
&= \sum_{x_1^n \in \operatorname{supp}\mathbb{P}_n} \mathbb{P}_{n-1}(x_2^n)\,\varphi(x_1, ..., x_m) \\
&\leq \sum_{x_1^{n-1} \in \operatorname{supp}\mathbb{P}_{n-1}} \sum_{x_n \in \mathcal{A}} \mathbb{P}_{n-1}(x_2^n)\,\varphi(x_1, ..., x_m) \\
&= \sum_{x_1^{n-1} \in \operatorname{supp}\mathbb{P}_{n-1}} \mathbb{P}_{n-2}(x_2^{n-1})\,\varphi(x_1, ..., x_m).
\end{aligned}$$

For the above derivation, we repeat the last two steps until it hits m:

$$\sum_{x_1^{n-1} \in \operatorname{supp} \mathbb{P}_{n-1}} \mathbb{P}_{n-2}(x_2^{n-1})\, \varphi(x_1, ..., x_m) \leq \cdots \leq \sum_{x_1^m \in \operatorname{supp} \mathbb{P}_m} \mathbb{P}_{m-1}(x_2^m)\, \varphi(x_1, ..., x_m).$$

Thus, we obtain the following:

$$\begin{aligned}
\int_\Omega \mathbb{E}[Z_n|\mathcal{F}_m]\, \varphi \, \mathrm{d}\mu = \int_\Omega Z_n \varphi \, \mathrm{d}\mu &\leq \sum_{x_1^m \in \operatorname{supp} \mathbb{P}_m} \mathbb{P}_{m-1}(x_2^m)\, \varphi(x_1, ..., x_m) \\
&= \sum_{x_1^m \in \operatorname{supp} \mathbb{P}_m} \frac{\mathbb{P}_{m-1}(x_2^m)}{\mathbb{P}_m(x_1^m)}\, \varphi(x_1, ..., x_m) \mathbb{P}_m(x_1^m) \\
&= \int_\Omega Z_m \varphi \, \mathrm{d}\mu \\
\Rightarrow \quad \int_\Omega \mathbb{E}[Z_n|\mathcal{F}_m]\, \varphi \, \mathrm{d}\mu &\leq \int_\Omega Z_m \varphi \, \mathrm{d}\mu.
\end{aligned}$$

This relation holds for any φ that depends only on the first m entries of $x \in \Omega$ and takes values between 0 and 1, so it follows that

$$\mathbb{E}[Z_n|\mathcal{F}_m] \leq Z_m \quad \mathbb{P}\text{-a.s.} \tag{4.10}$$

for any $2 \leq m < n$ (one may see this by taking $\varphi = \mathbb{1}_E$ for arbitrary $E \in \mathcal{F}_m$). Moreover, it is obvious that Z_n is $\mathcal{F}_n$-measurable for every $n \geq 2$, and we have shown (3.5) which says Z_n is L^1 for all $n \geq 2$. Thus, $\{Z_n\}_{n \geq 2}$ is a supermartingale with respect to $\{\mathcal{F}_n\}_{n \geq 2}$ by definition.

Remark. If the probability measure $\mathbb{P}$ is fully supported, then $\mathbb{P}_n(x_1^n) > 0$ for all $x_1^n \in \mathcal{A}^n$ and for all $n \in \mathbb{N}$, and all the inequalities in the above derivation would become equality, resulting in a martingale for us.

Now that $\{Z_n\}_{n \geq 2}$ is a non-negative supermartingale with respect to $\{\mathcal{F}_n\}_{n \geq 2}$, then by Theorem 2.3.1, there exists a non-negative function Z such that

$$Z(x) = \lim_{n \to \infty} Z_n(x)$$

for $\mathbb{P}$-almost all $x \in \Omega$. Clearly $Z \geq 1$ $\mathbb{P}$-a.s. Z is also L^1, which can shown by Fatou's lemma and (3.5):

$$\int_\Omega Z(x)\,\mathrm{d}\mathbb{P} = \int_\Omega \lim_{n\to\infty} Z_n(x)\,\mathrm{d}\mathbb{P} \le \liminf_{n\to\infty} \int_\Omega Z_n(x)\,\mathrm{d}\mathbb{P} \le \ell.$$

We next turn to the subsequent construction $\{\log Z_n\}_{n\ge 2}$ by taking logarithm on both sides of (4.10). Since logarithmic function is concave, we apply conditional Jensen's inequality[1] and get

$$\mathbb{E}[\log Z_n|\mathcal{F}_m] \le \log \mathbb{E}[Z_n|\mathcal{F}_m] \le \log Z_n \;\Rightarrow\; \mathbb{E}[\log Z_n|\mathcal{F}_m] \le \log Z_n \;\; \mathbb{P}\text{-a.s.}$$

The integrability of each $\log Z_n$ is in fact already justified, given by (3.1) and Theorem 3.2.3 (1):

$$0 \le \int_\Omega \log Z_n\,\mathrm{d}\mathbb{P} \le \log \ell.$$

Then $\{\log Z_n\}_{n\ge 2}$ is also a non-negative supermartingale with respect to $\{\mathcal{F}_n\}_{n\ge 2}$. Again by Theorem 2.3.1, there exists $\mathbb{P}$-a.s. a limit function for $\{\log Z_n\}_{n\ge 2}$. In fact, this limit function can be directly granted by the $\mathbb{P}$-a.s. convergence of $\{Z_n\}_{n\ge 2}$ to Z, and we have

$$\lim_{n\to\infty} \log Z_n(x) = \log Z(x)$$

for $\mathbb{P}$-almost all $x \in \Omega$. We have $\log Z \ge 0$ $\mathbb{P}$-a.s. and it is L^1 by Fatou's lemma and (3.4):

$$\int_\Omega \log Z\,\mathrm{d}\mathbb{P} = \int_\Omega \lim_{n\to\infty} \log Z_n\,\mathrm{d}\mathbb{P} \le \lim_{n\to\infty} \int_\Omega \log Z_n\,\mathrm{d}\mathbb{P} = S(\mathbb{P}) \le \log \ell.$$

However, we do not only have an inequality here, but actually an equality. This can be shown using the dominated convergence theorem, where the dominant L^1 function is $\log Z_{\max}$.

By the definition of $Z_{\max}$, we have $1 \le Z_n \le Z_{\max}$ for any $n \ge 2$. Then $0 \le \log Z_n \le \log Z_{\max}$ for any $n \ge 2$ as well. $\log Z_{\max}$ is L^1 by Lemma 3.2.5. We can then apply the dominated convergence theorem and naturally get

$$\begin{aligned} S(\mathbb{P}) = \lim_{n\to\infty} \int_\Omega \log Z_n\,\mathrm{d}\mathbb{P} &= \int_\Omega \lim_{n\to\infty} \log Z_n\,\mathrm{d}\mathbb{P} = \int_\Omega \log Z\,\mathrm{d}\mathbb{P} \\ \Rightarrow \quad S(\mathbb{P}) &= \int_\Omega \log Z\,\mathrm{d}\mathbb{P}. \end{aligned} \tag{4.11}$$

Relation (4.11) will be of central importance in this proof. In addition, the dominated conver-

[1] See Property (viii) in [AB19, Proposition 11.1].

gence theorem can also be applied to show that $\{\log Z_n\}_{n\geq 2}$ converges to $\log Z$ in L^1. We have $\{\log Z_n\}_{n\geq 2}$ converges to $\log Z$ $\mathbb{P}$-a.s., and we have $\log Z_n$'s and $\log Z$ are all L^1 functions as well. Moreover, $0 \leq \log Z_n \leq \log Z_{\max}$ for any $n \geq 2$ also gives $0 \leq \log Z \leq \log Z_{\max}$, then

$$0 \leq |\log Z_n - \log Z| \leq \log Z_n + \log Z \leq 2\log Z_{\max},$$

where $\log Z_{\max}$ is L^1. Hence, by the dominated convergence theorem,

$$\lim_{n\to\infty} \int_\Omega |\log Z_n(x) - \log Z(x)| \, \mathrm{d}\mathbb{P} = 0. \tag{4.12}$$

Returning to our second proof, we observe the following manipulations:

$$\begin{aligned} -\log \mathbb{P}([x_1^n]) &= \log \frac{1}{\mathbb{P}_n(x_1^n)} \\ &= \log \left(\frac{\mathbb{P}_{n-1}(x_2^n)}{\mathbb{P}_n(x_1^n)} \cdot \frac{\mathbb{P}_{n-2}(x_3^n)}{\mathbb{P}_{n-1}(x_2^n)} \cdots \frac{\mathbb{P}_1(x_n)}{\mathbb{P}_2(x_{n-1}^n)} \cdot \frac{1}{\mathbb{P}_1(x_n)} \right) \\ &= \log \frac{1}{\mathbb{P}_1(x_n)} + \sum_{k=0}^{n-2} \log Z_{n-k}(T^k(x)) \\ &= \log \frac{1}{\mathbb{P}_1(x_n)} + \sum_{k=0}^{n-2} \big(\log Z_{n-k}(T^k(x)) - \log Z(T^k(x))\big) + \sum_{k=0}^{n-2} \log Z(T^k(x)). \end{aligned}$$

Dividing both sides by n, it yields

$$-\frac{1}{n}\log \mathbb{P}([x_1^n]) = \frac{1}{n}\log \frac{1}{\mathbb{P}_1(x_n)} + \tag{4.13}$$

$$\frac{1}{n}\sum_{k=0}^{n-2} \big(\log Z_{n-k}(T^k(x)) - \log Z(T^k(x))\big) + \tag{4.14}$$

$$\frac{1}{n}\sum_{k=0}^{n-2} \log Z(T^k(x)), \tag{4.15}$$

which decomposes our primary object of interest $-\frac{1}{n}\log \mathbb{P}([x_1^n])$ into three manageable parts, upon which are worked to prove the SMB theorem.

(4.13) gives an elementary estimate. Note that if we denote

$$K = \max\left\{ \log \frac{1}{\mathbb{P}_1(a)} : a \in \mathcal{A} \text{ and } \mathbb{P}_1(a) > 0 \right\},$$

then K is a finite constant and

$$\log \frac{1}{\mathbb{P}_1(x_n)} \leq K \quad \mathbb{P}\text{-a.s.}$$

Thus, as $n \to \infty$, (4.13) goes to 0 $\mathbb{P}$-a.s., and we are done with this term.

As for (4.14) and (4.15), we can see the term (4.15) is in the form of an ergodic sum and a limit function would exist $\mathbb{P}$-a.s. by Birkhoff's ergodic theorem (Theorem 2.1.1). On the other hand, regarding the limit behavior of (4.14), which is represented by terms involving differences between $\log Z_n$'s and $\log Z$, an estimate could be given to be 0 $\mathbb{P}$-a.s. This is exactly due to the convergence of $\log Z_n$'s to $\log Z$. Hence, by setting

$$\begin{aligned} X_n(x) &:= \frac{1}{n} \sum_{k=0}^{n-2} \left|\log Z_{n-k}(T^k(x)) - \log Z(T^k(x))\right| \\ &= \frac{1}{n} \sum_{k=2}^{n} \left|\log Z_k(T^{n-k}(x)) - \log Z(T^{n-k}(x))\right| \end{aligned}$$

for all $n \geq 2$, it then suffices for us to show that X_n converges to 0 $\mathbb{P}$-a.s.

Let $m \geq 2$ and define

$$Y_m(x) := \sup_{k \geq m} \left|\log Z_k(x) - \log Z(x)\right|.$$

For any $m \geq 2$, Y_m is well-defined $\mathbb{P}$-a.s. and $Y_m \geq 0$. Since $\{\log Z_n\}_{n \geq 2}$ converges to $\log Z$ $\mathbb{P}$-a.s., we have

$$\lim_{m \to \infty} Y_m(x) = 0 \quad \mathbb{P}\text{-a.s.}$$

For well-defined value x, it is also clear that $\{Y_m(x)\}_{m \geq 2}$ is decreasing, and

$$Y_2(x) = \sup_{k \geq 2} \left|\log Z_k(x) - \log Z(x)\right| \leq \log \left(\sup_{k \geq 2} Z_k(x) \right) + \log Z \leq 2 \log Z_{\max},$$

so $Y_m \in L^1(\Omega, \mathrm{d}\mathbb{P})$ for all $m \geq 2$.

Now fix some $m > 2$. For any $n \geq m$, we rewrite the expression of X_n and derive an upper bound for it as follows:

$$\begin{aligned}
X_n(x) &= \frac{1}{n}\sum_{k=2}^{m-1}\left|\log Z_k(T^{n-k}(x)) - \log Z(T^{n-k}(x))\right| + \\
&\quad \frac{1}{n}\sum_{k=m}^{n}\left|\log Z_k(T^{n-k}(x)) - \log Z(T^{n-k}(x))\right| \\
&\le \frac{2}{n}\sum_{k=2}^{m-1}\log Z_{\max}(T^{n-k}(x)) + \frac{1}{n}\sum_{k=m}^{n} Y_m(T^{n-k}(x)) \\
&\le \frac{2}{n}\sum_{k=2}^{m-1}\log Z_{\max}(T^{n-k}(x)) + \frac{1}{n}\sum_{k=0}^{n-1} Y_m(T^{k}(x)). \qquad (4.16)
\end{aligned}$$

We end up getting the bound (4.16) by adding $m-1$ more non-negative terms to the previous step. We focus on this estimate and show $\{X_n\}_{n\geq 2}$ goes to 0 $\mathbb{P}$-a.s. by showing this asymptotic behavior holds for (4.16).

We initially deal with the first term in (4.16). Note that with respect to our n, by elementary probabilistic arguments,

$$\left\{x \in \Omega : \sum_{k=2}^{m-1}\log Z_{\max}(T^{n-k}(x)) \geq \sqrt{n}\right\} \subseteq \bigcup_{k=2}^{m-1}\left\{x \in \Omega : \log Z_{\max}(T^{n-k}(x)) \geq \frac{\sqrt{n}}{m-2}\right\}.$$

Then

$$\begin{aligned}
\mathbb{P}\left(\left\{\sum_{k=2}^{m-1}\log Z_{\max}\circ T^{n-k} \geq \sqrt{n}\right\}\right) &\le \sum_{k=2}^{m-1}\mathbb{P}\left(\left\{\log Z_{\max}\circ T^{n-k} \geq \frac{\sqrt{n}}{m-2}\right\}\right) \\
&\le \sum_{k=2}^{m-1}\mathbb{P}\left(\left\{\log Z_{\max}\circ T^{n-k} \geq \frac{\sqrt{n}}{m}\right\}\right) \\
\text{(By shift-invariance of } \mathbb{P}\text{)} \quad &= (m-2)\,\mathbb{P}\left(\left\{\log Z_{\max} \geq \frac{\sqrt{n}}{m}\right\}\right) \\
\text{(By Lemma 3.2.4)} \quad &\le (m-2)\,\ell\, e^{-\frac{\sqrt{n}}{m}}.
\end{aligned}$$

Hence,

$$\sum_{n=m}^{\infty}\mathbb{P}\left(\left\{\sum_{k=2}^{m-1}\log Z_{\max}\circ T^{n-k} \geq \sqrt{n}\right\}\right) \le (m-2)\,\ell\sum_{n=1}^{\infty} e^{-\frac{\sqrt{n}}{m}} < \infty$$

by integral test. It then follows from the Borel–Cantelli lemma, that for $\mathbb{P}$-almost all $x \in \Omega$, there exists some $n_0(x) \geq m$, such that for all $n \geq n_0(x)$,

$$\sum_{k=2}^{m-1} \log Z_{\max} \circ T^{n-k}(x) \leq \sqrt{n}.$$

Thus, for $\mathbb{P}$-almost all $x \in \Omega$, we have

$$0 \leq \lim_{n\to\infty} \frac{2}{n} \sum_{k=2}^{m-1} \log Z_{\max}(T^{n-k}(x)) \leq \lim_{n\to\infty} \frac{2\sqrt{n}}{n} = 0,$$

and the first term in (4.16) goes to 0 asymptotically $\mathbb{P}$-a.s.

As for the second term in (4.16), we notice that it takes the form of an ergodic sum. Then by Theorem 2.1.1, there exists a limit function $\tilde{Y}_m$ such that

$$\tilde{Y}_m(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Y_m(T^k(x)) \quad \mathbb{P}\text{-a.s.}$$

Therefore, we have so far derived (4.16) to

$$\begin{aligned} 0 \leq \limsup_{n\to\infty} X_n(x) \leq 0 + \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Y_m(T^k(x)) = \tilde{Y}_m(x) \quad \mathbb{P}\text{-a.s.} \\ \Rightarrow \quad \limsup_{n\to\infty} X_n(x) \leq \tilde{Y}_m(x) \quad \mathbb{P}\text{-a.s.} \end{aligned} \tag{4.17}$$

Since $\{Y_m\}_{m\geq 2}$ is decreasing, non-negative, and converging to 0 $\mathbb{P}$-a.s., and $Y_2 \leq 2\log Z_{\max}$ where $\log Z_{\max}$ is integrable, then by the dominated convergence theorem,

$$\lim_{m\to\infty} \int_\Omega Y_m \, \mathrm{d}\mathbb{P} = 0.$$

Besides, as shown in the end of the proof of Theorem 2.1.1, we also have

$$\int_\Omega Y_m \, \mathrm{d}\mathbb{P} = \int_\Omega \tilde{Y}_m \, \mathrm{d}\mathbb{P} \tag{4.18}$$

for every $m > 2$. This gives

$$\lim_{m\to\infty} \int_\Omega Y_m \, \mathrm{d}\mathbb{P} = \lim_{m\to\infty} \int_\Omega \tilde{Y}_m \, \mathrm{d}\mathbb{P} = 0. \tag{4.19}$$

On the other hand, for every $m > 2$, $\tilde{Y}_m$ is clearly non-negative. It is also true that the sequence of the limit functions $\{\tilde{Y}_m\}_{m>2}$ is decreasing. This can be observed in a pointwise sense: Fix arbitrary value x on the set with full measure where all $\tilde{Y}_m$'s are well-defined. For each pair of $m_1, m_2 \in \mathbb{N} \cap (2, \infty)$ such that $m_1 < m_2$, and for any $n \geq m_2$, we have

$$\frac{1}{n} \sum_{k=0}^{n-1} Y_{m_1}(T^k(x)) \geq \frac{1}{n} \sum_{k=0}^{n-1} Y_{m_2}(T^k(x)).$$

This is because we used the fact that $\{Y_m\}_{m \geq 2}$ is decreasing: For every $k \in [0, n-1] \cap \mathbb{N}$, $Y_{m_1}(T^k(x)) \geq Y_{m_2}(T^k(x))$. Then taking limit of $n \to \infty$ on both sides, we get $\tilde{Y}_{m_1}(x) \geq \tilde{Y}_{m_2}(x)$.

As $\{\tilde{Y}_m\}_{m>2}$ is non-negative and decreasing, then a limit function for $\{\tilde{Y}_m\}_{m>2}$ exists for $\mathbb{P}$-almost all $x \in \Omega$:

$$\tilde{Y}(x) := \inf_{m>2} \tilde{Y}_m(x).$$

Since (4.17) holds for every $m > 2$, then we have

$$\limsup_{n\to\infty} X_n(x) \leq \tilde{Y}(x) \quad \mathbb{P}\text{-a.s.}, \tag{4.20}$$

Moreover, every $\tilde{Y}_m$ is integrable by (4.18). Then another application of the dominated convergence theorem and (4.19) give

$$\int_\Omega \tilde{Y}\, \mathrm{d}\mathbb{P} = \lim_{m\to\infty} \int_\Omega \tilde{Y}_m\, \mathrm{d}\mathbb{P} = 0.$$

However, $\tilde{Y}$ is clearly non-negative $\mathbb{P}$-a.s., so $\tilde{Y} = 0$ $\mathbb{P}$-a.s. This, together with (4.20) and the fact that X_n is non-negative for all $n \geq 2$, implies $\lim_{n\to\infty} X_n = 0$ $\mathbb{P}$-a.s.

This in turn shows the representation (4.14) asymptotically goes to 0 $\mathbb{P}$-a.s.

Finally, we are left with (4.15), whose limit behavior is determined by Theorem 2.1.1: For $\mathbb{P}$-almost all $x \in \Omega$, there exists a limit function, which we denote as $h_{\mathbb{P}}$, such that

$$h_{\mathbb{P}}(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-2} \log Z(T^k(x)).$$

Combing the limit behavior of all three terms (4.13), (4.14), and (4.15), we conclude that

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}([x_1^n]) = h_{\mathbb{P}}(x) \quad \mathbb{P}\text{-a.s.}$$

This finishes the part for showing $\mathbb{P}$-a.s. convergence.

To show the convergence also holds in L^1, we go back to the three decomposed parts (4.13), (4.14), and (4.15), where it is trivial to see that the estimate (4.13) converges to 0 in L^1 by the $\mathbb{P}$-a.s. upper bound of $\frac{K}{n}$. Theorem 2.1.1 also tells that the convergence of (4.15) to $h_{\mathbb{P}}$ holds in L^1. It remains to show that (4.14) converges to 0 in L^1 by showing $X_n \to 0$ in L^1.

We know X_n is non-negative for every $n \geq 2$. Then plugging in its expression and by shift-invariance of $\mathbb{P}$, we get

$$\begin{aligned}\int_\Omega |X_n - 0| \,\mathrm{d}\mathbb{P} = \int_\Omega X_n \,\mathrm{d}\mathbb{P} &= \frac{1}{n}\sum_{k=2}^{n}\int_\Omega |\log Z_k - \log Z| \circ T^{n-k} \,\mathrm{d}\mathbb{P} \\ &= \frac{1}{n}\sum_{k=2}^{n}\int_\Omega |\log Z_k - \log Z| \,\mathrm{d}\mathbb{P}.\end{aligned}$$

For the integral terms in the summation on the right hand side, they are all non-negative and finite, and the whole term (the sum divided by n) can be viewed as the Cesàro mean. Since we have shown (4.12), and recall how we justified the equality (3.3) regarding the Cesàro mean, we have

$$\lim_{n\to\infty}\int_\Omega X_n \,\mathrm{d}\mathbb{P} = \lim_{n\to\infty}\frac{1}{n}\sum_{k=2}^{n}\int_\Omega |\log Z_k - \log Z| \,\mathrm{d}\mathbb{P} = \lim_{n\to\infty}\int_\Omega |\log Z_n - \log Z| \,\mathrm{d}\mathbb{P} = 0.$$

This completes the proof of convergence to $h_{\mathbb{P}}$ in L^1.

For the remaining properties of $h_{\mathbb{P}}$, it is obvious that $h_{\mathbb{P}} \geq 0$ $\mathbb{P}$-a.s., and shift-invariance ($h_{\mathbb{P}} \circ T = h_{\mathbb{P}}$) also follows from Theorem 2.1.1.

By the convergence of the term (4.15) to $h_{\mathbb{P}}$ via Theorem 2.1.1, we also have

$$\int_\Omega h_{\mathbb{P}} \,\mathrm{d}\mathbb{P} = \int_\Omega \log Z \,\mathrm{d}\mathbb{P},$$

and by (4.11), we obtain

$$\int_\Omega h_{\mathbb{P}} \,\mathrm{d}\mathbb{P} = S(\mathbb{P}).$$

Finally, if $\mathbb{P} \in \mathcal{P}_{\text{erg}}(\Omega)$, then by the ergodic version of Theorem 2.1.1, we have $h_{\mathbb{P}}$ is constant

$\mathbb{P}$-a.s., and this constant value is exactly $S(\mathbb{P})$, so in this case,

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}([x_1^n]) = S(\mathbb{P})$$

for $\mathbb{P}$-almost all $x \in \Omega$ and the convergence also holds in $L^1(\Omega, \mathrm{d}\mathbb{P})$. □

In this martingale proof, we can see the absolute central involvement of the constructed functions $\log Z_n$ and $\log Z$ as well as their relations with entropy, such as (3.1), (3.4), and (4.11). These functions, as established earlier in Subsection 3.2.2, exhibit excellent convergence behavior, motivating the applications of the martingale convergence theorem and Birkhoff's ergodic theorem and nicely leading to the desired conclusions of convergence to a limit function.

These constructions themselves, especially the function $\log Z$, play fundamental roles not only in information theory, but also in the theory of dynamical systems and statistical mechanics. We will conclude the technical aspects and more about the martingale proof in Chapter 5.

4.3 The Ornstein–Weiss proof

For the third proof, it is self-contained in an information-theoretic way, where packing ideas are used, and it will only prove the ergodic case of the SMB theorem. The very original proof, given by Donald Ornstein and Benjamin Weiss dating back to 1983 [OW83], is part of their works to extend ideas from ergodic theory and dynamical systems to random fields and amenable groups. For this reason, the third proof is credited as "the Ornstein–Weiss proof". The reference, which our third proof is closely based upon, is [Sh96, Section I.5].

We begin with a very quick introduction to covering and packing. For simplicity, in this specific section, we use the usual interval notations to denote intervals in $\mathbb{N}$, since there will be no appearance of continous interval in the sense of real line. That is, for any pairwise natural numbers $n \leq m$, we denote $[n, m] = \{j \in \mathbb{N} : n \leq j \leq m\}$. Similarly, we have $(n, m] = \{j \in \mathbb{N} : n < j \leq m\}$ and so on. When we mention intervals in this section, we mean intervals referred and denoted in the above sense.

Definition 4.3.1 (Strong cover). *A strong cover C of $\mathbb{N}$ is a collection of intervals $\{[n, m(n)]\}_{n\in\mathbb{N}}$, where $m : \mathbb{N} \longrightarrow \mathbb{N}$ such that $m(n) \geq n$ for all $n \in \mathbb{N}$.*

By cover, we clearly mean the property that the union of the given intervals is $\mathbb{N}$. A strong cover C of $\mathbb{N}$ has a subcover $\{[n_i, m(n_i)]\}_{i\in\mathbb{N}}$ such that its members are disjoint: $\{n_i\}_{i\in\mathbb{N}}$ is a sequence of natural numbers satisfying $n_1 = 1$ and $n_{i+1} = 1 + m(n_i)$ for all $i \geq 1$. This is referred as the "packing property" in [Sh96].

Definition 4.3.2. *Let C be a strong cover of $\mathbb{N}$. We say that the interval $[1, K]$ is (L, δ)-strongly covered by C if*

$$\frac{|\{n \in [1, K] : m(n) - n + 1 > L\}|}{K} \leq \delta.$$

Definition 4.3.3. *A collection C' of subintervals of $[1, K]$ is called a $(1 - \delta)$-packing of $[1, K]$ if the intervals in C' are disjoint and their union has cardinality at least $(1 - \delta)K$.*

Here is a classical and important result concerning these defined properties:

Lemma 4.3.4 (The packing lemma). *Let C be a strong cover of $\mathbb{N}$ and let $L \in \mathbb{N}$ and $0 < \delta < \frac{1}{2}$ be given. Suppose $K > \frac{L}{\delta}$. If $[1, K]$ is (L, δ)-strongly covered by C, then there exists a subcollection $C' \subseteq C$ that is a $(1 - 2\delta)$-packing of $[1, K]$.*

Proof. We define a sequence of natural numbers $\{n_i\}_{i \in \mathbb{N}}$ inductively as follows: Set $n_0 = m(0) = 0$ and

$$n_i = \min\{j \in [1 + m(n_{i-1}), K - L] : m(j) - j + 1 \leq L\}.$$

In other words, we proceed from small values to large ones, picking the first interval with length at most L that is disjoint from the previous selected intervals. Our inductive assumption stops after $I < \infty$ steps, where in the end, either $m(n_I) \geq K - L$ or there is no $j \in [1 + m(n_I), K - L]$ for which $m(j) - j + 1 \leq L$.

Now we claim

$$C' = \{[n_i, m(n_i)]\}_{i=1}^{I}$$

is a $(1 - 2\delta)$-packing of $[1, K]$.

We check Definition 4.3.3. By construction, C' is automatically disjoint. We also have all the intervals in C' are contained in $[1, K]$. To show this, we need to show $m(n_I) \leq K$. By the above inductive assumption, we have $m(n_I) - n_I + 1 \leq L$ and $n_I \leq K - L$. Then trivially,

$$m(n_I) \leq L + n_I - 1 \leq L + K - L - 1 = K - 1.$$

Finally, we check the cardinality of the union of all intervals in C' is at least $(1 - \delta)K$. Let us denote that

$$U = \bigsqcup_{i \in [1,I]} [n_i, m(n_i)].$$

By our assumption, $|(K-L,K]| \leq L < \delta K$. Then,

$$|(K-L,K] \backslash U| \leq |(K-L,K]| < \delta K.$$

On the other hand, if $j \in [1, K-L] \backslash U$, then we must have $m(j) - j + 1 > L$ (otherwise, our construction would give $j \in U$). The assumption that $[1, K]$ is (L, δ)-strongly covered by C then gives $|[1, K-L] \backslash U| \leq \delta K$. Combing both estimates, we get

$$|[1,K] \backslash U| \leq 2\delta K \quad \Rightarrow \quad |U| \geq (1-2\delta)K.$$

Thus, by Definition 4.3.3, we have found a subcollection C' of C, which is a $(1-2\delta)$-packing of $[1, K]$. □

Another useful related result which will provide some estimate of upper bound in our third proof is a combinatorial one. Given some $0 < \delta < 1$, we denote by

$$H(\delta) = -\delta \log \delta - (1-\delta) \log (1-\delta), \tag{4.21}$$

where we adapt the notation H of entropy with respect to a random variable from Appendix D, since if we set some probability measure $\mathbb{P}$ on $\mathcal{B} = \{0,1\}$ by letting $\mathbb{P}(0) = \delta$ and $\mathbb{P}(1) = 1-\delta$, then (4.21) could be regarded as the entropy of some random variable taking values in $\mathcal{B}$ whose probability distribution is $\mathbb{P}$. Then simply $H(\delta) = S(\mathbb{P})$.

Proposition 4.3.5 (Combinatorial bound). *Let* $0 < \delta < \frac{1}{2}$*, then*

$$\sum_{0 \leq k \leq n\delta} \binom{n}{k} \leq e^{nH(\delta)}.$$

Proof. We start by defining a function $f : \mathbb{R} \longrightarrow \mathbb{R}$ as below:

$$f(x) := -x \log \delta - (1-x) \log (1-\delta).$$

Then since $0 < \delta < \frac{1}{2}$,

$$f'(x) = \log\left(\frac{1-\delta}{\delta}\right) > 0$$

for all $x \in \mathbb{R}$, so f is increasing.

For $x \leq \delta$, we have $f(x) \leq f(\delta) = H(\delta)$. Then for any k such that $0 \leq k \leq n\delta$, we would have $0 \leq \frac{k}{n} \leq \delta$ and

$$\begin{aligned} f\left(\frac{k}{n}\right) = -\frac{k}{n}\log\delta - \left(\frac{n-k}{n}\right)\log(1-\delta) &\leq H(\delta) \\ \Rightarrow \quad -\frac{1}{n}\log\left(\delta^k(1-\delta)^{n-k}\right) &\leq H(\delta) \\ \Rightarrow \quad e^{-nH(\delta)} &\leq \delta^k(1-\delta)^{n-k}. \end{aligned}$$

We now do the following manipulations based on what we obtained above:

$$\begin{aligned} e^{-nH(\delta)}\sum_{0\leq k\leq n\delta}\binom{n}{k} &= \sum_{0\leq k\leq n\delta} e^{-nH(\delta)}\binom{n}{k} \\ &\leq \sum_{0\leq k\leq n\delta} \delta^k(1-\delta)^{n-k}\binom{n}{k} \\ &\leq \sum_{k=0}^{n} \delta^k(1-\delta)^{n-k}\binom{n}{k} \\ \text{(Binomial expansion)} \quad &= (\delta + (1-\delta))^n = 1 \\ \Rightarrow \quad \sum_{0\leq k\leq n\delta}\binom{n}{k} &\leq e^{nH(\delta)}, \end{aligned}$$

which completes the proof. □

We now introduce the other ingredients for the Ornstein–Weiss proof. A measurable function $\tau : \Omega \longrightarrow \mathbb{N} \cup \{\infty\}$ is called a generalized stopping time. If $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$ and

$$\mathbb{P}(\{x \in \Omega : \tau(x) = \infty\}) = 0,$$

then τ is said to be $\mathbb{P}$-a.s. finite. By shift-invariance of $\mathbb{P}$,

$$\mathbb{P}(\{x \in \Omega : \tau(x) = \infty\}) = \mathbb{P}(\{x \in \Omega : \tau \circ T^{n-1}(x) = \infty\}), \;\; \forall\, n \in \mathbb{N},$$

and we shall have $\tau \circ T^{n-1}$ is $\mathbb{P}$-a.s. finite for all $n \in \mathbb{N}$. Then in this case, the so-called stopping interval starting at n, $[n, \tau(T^{n-1}(x)) + n - 1]$, is well-defined for $\mathbb{P}$-almost all $x \in \Omega$, and from here, we can define a strong cover in a pointwise sense for $\mathbb{N}$:

$$C(x, \tau) = \{[n, \tau(T^{n-1}(x)) + n - 1] : n \geq 1\},$$

which is well-defined for $\mathbb{P}$-almost all $x \in \Omega$.

We then introduce a lemma that will be of central importance in the Ornstein–Weiss proof:

Lemma 4.3.6 (The ergodic stopping time packing lemma). *Let $\mathbb{P} \in \mathcal{P}_{\text{erg}}(\Omega)$ and τ be a $\mathbb{P}$-a.s. finite stopping time. For any $0 < \delta < 1$ and $\mathbb{P}$-almost all $x \in \Omega$, there exists $N(\delta, x) \in \mathbb{N}$ such that for all $n \geq N(\delta, x)$, the interval $[1, n]$ is $(1 - \delta)$-packed by intervals from $C(x, \tau)$.*

Proof. Fix some arbitrary $0 < \delta < 1$. Since τ is $\mathbb{P}$-a.s. finite, then

$$\mathbb{P}(\{x \in \Omega : \tau(x) = \infty\}) = \lim_{L \to \infty} \mathbb{P}(\{x \in \Omega : \tau(x) > L\}) = 0,$$

and there exists $L \in \mathbb{N}$ large enough such that

$$\mathbb{P}(\{x \in \Omega : \tau(x) > L\}) \leq \frac{\delta}{2}.$$

We set $D = \{x \in \Omega : \tau(x) > L\}$ and consider the corresponding indicator function $\mathbb{1}_D$. By Theorem 2.1.1, we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} \mathbb{1}_D(T^j(x)) = \int_\Omega \mathbb{1}_D(x) \, \mathrm{d}\mathbb{P} = \mathbb{P}(D) \leq \frac{\delta}{2}$$

for $\mathbb{P}$-almost all $x \in \Omega$.

Hence, if we let

$$G_n = \left\{ x \in \Omega : \frac{1}{n} \sum_{j=0}^{n-1} \mathbb{1}_D(T^j(x)) \leq \frac{\delta}{2} \right\}$$

for all $n \in \mathbb{N}$, then we have $x \in G_n$ eventually almost-surely (by this, we mean that for $\mathbb{P}$-almost all $x \in \Omega$, it holds that there exists some $n(x) \in \mathbb{N}$ such that for all $n \geq n(x)$, $x \in G_n$).

Now pick such an $x \in \Omega$ and let $n > \max\left(n(x), \frac{2L}{\delta}\right)$. The definition of G_n then gives that $\tau \circ T^{j-1}(x) \leq L$ for at least $\left(1 - \frac{\delta}{2}\right) n$[2] indices $j \in [1, n]$, because otherwise, there would be more

[2] As we take the value of n to be greater than $\frac{2L}{\delta}$, it guarantees the value $\left(1 - \frac{\delta}{2}\right) n$ is strictly positive.

than $n - \left(1 - \frac{\delta}{2}\right) n$ indices from $[1, n]$ such that $\tau \circ T^{j-1}(x) > L$, implying

$$\frac{1}{n} \sum_{j=0}^{n-1} \mathbb{1}_D(T^j(x)) > \frac{1}{n} \left(n - \left(1 - \frac{\delta}{2}\right) n\right) \cdot 1 = \frac{\delta}{2},$$

leading to a contradiction.

Hence,

$$\begin{aligned} &\left|\{j \in [1, n] : \tau \circ T^{j-1}(x) > L\}\right| \leq n - \left(1 - \frac{\delta}{2}\right) n = \frac{\delta}{2} n \\ \Rightarrow \quad &\frac{\left|\{j \in [1, n] : \tau \circ T^{j-1}(x) > L\}\right|}{n} \leq \frac{\delta}{2}, \end{aligned}$$

and it means $[1, n]$ is $\left(L, \frac{\delta}{2}\right)$-strongly covered by $C(x, \tau)$. By Lemma 4.3.4, there exists a subcollection of $C(x, \tau)$ that is $(L, 1 - \delta)$-packing of $[1, n]$. Note that the existence of such $N(\delta, x)$ is already indicated above, where we can set $N(\delta, x) = \max\left(n(x), \frac{2L}{\delta}\right) + 1$, and then this conclusion holds for all $n \geq N(\delta, x)$. □

We now turn to the proof of the SMB theorem, which is referred as "the entropy theorem" in [Sh96]. In this book, the theorem is given to state that for $\mathbb{P} \in \mathcal{P}_{\text{erg}}(\Omega)$, $-\frac{1}{n} \log \mathbb{P}([x_1^n])$ converges $\mathbb{P}$-a.s. to a non-negative constant, denoted as $h(\mathbb{P})$. We will show this statement first, and from there, further derive that this constant $h(\mathbb{P})$ is actually $S(\mathbb{P})$, and the convergence also holds in L^1.

Proof. Let $\mathbb{P} \in \mathcal{P}_{\text{erg}}(\Omega)$. Set

$$h(x) := \liminf_{n \to \infty} -\frac{1}{n} \log \mathbb{P}([x_1^n]),$$

which is $\mathbb{P}$-a.s. well-defined on Ω. Clearly, $h \geq 0$. By Fatou's lemma, h is L^1:

$$\int_\Omega h(x) \, \mathrm{d}\mathbb{P} \leq \liminf_{n \to \infty} -\frac{1}{n} \int_\Omega \log \mathbb{P}([x_1^n]) \, \mathrm{d}\mathbb{P} = \liminf_{n \to \infty} \frac{S(\mathbb{P}_n)}{n} = S(\mathbb{P}). \tag{4.22}$$

Note that for all $x \in \Omega$,

$$\mathbb{P}([x_2^{n+1}]) \geq \mathbb{P}([x_1^{n+1}]) \quad \Rightarrow \quad \mathbb{P}([T(x)]_1^n) \geq \mathbb{P}([x_1^{n+1}]),$$

so $h(T(x)) \leq h(x)$ for all well-defined x. Let $G = \{h \circ T < h\}$. We want to show that $\mathbb{P}(G) = 0$

so that $h \circ T = h$ $\mathbb{P}$-a.s. Assume $\mathbb{P}(G) > 0$, then with shift-invariance of $\mathbb{P}$,

$$\begin{aligned} \int_{\Omega \setminus G} h \circ T \, \mathrm{d}\mathbb{P} + \int_G h \circ T \, \mathrm{d}\mathbb{P} &< \int_{\Omega \setminus G} h \, \mathrm{d}\mathbb{P} + \int_G h \, \mathrm{d}\mathbb{P} \\ \Rightarrow \quad \int_\Omega h \circ T \, \mathrm{d}\mathbb{P} &< \int_\Omega h \, \mathrm{d}\mathbb{P} \\ \Rightarrow \quad \int_\Omega h \, \mathrm{d}\mathbb{P} &< \int_\Omega h \, \mathrm{d}\mathbb{P}, \end{aligned}$$

which is a contradiction.

Thus, $\mathbb{P}(G) = 0$ and $h \circ T = h$ $\mathbb{P}$-a.s. Besides, since $\mathbb{P}$ is ergodic, then there exists some non-negative constant $h(\mathbb{P})$, such that $h(x) = h(\mathbb{P})$ for $\mathbb{P}$-almost all $x \in \Omega$. (Similar reasoning has been mentioned in the conclusion part of Section 2.1.)

Hence, we have shown that

$$\liminf_{n \to \infty} -\frac{1}{n} \log \mathbb{P}([x_1^n]) = h(\mathbb{P}) \quad \mathbb{P}\text{-a.s.}, \tag{4.23}$$

and by (4.22), $h(\mathbb{P}) \leq S(\mathbb{P})$. We would then like to show this also holds for limit superior.

Let $0 < \delta < 1$ and $M \in \mathbb{N}$. We will make further restrictions on δ and M later on. Fix arbitrary $\varepsilon > 0$. Let $K \geq M$ and let $J_K(\delta, M)$ be the set of all $x_1^K \in \mathcal{A}^K$ such that the following holds: There is a collection $S(x_1^K) = \{[n_i, m_i]\}_{i \in I_K}$ of disjoint subintervals of $[1, K]$ (I_K is a finite index set depending on K) with the following properties:

(i). $m_i - n_i + 1 \geq M$ for all $i \in I_K$;

(ii).

$$\mathbb{P}([x_{n_i}^{m_i}]) \geq e^{-(m_i - n_i + 1)(h(\mathbb{P}) + \varepsilon)}, \quad \forall \, i \in I_K;$$

(iii).

$$\sum_{i \in I_K} (m_i - n_i + 1) \geq (1 - \delta) K.$$

We give the following lemmas to characterize the existence and size of $J_K(\delta, M)$.

Lemma 4.3.7. *$x_1^K \in J_K(\delta, M)$ eventually almost-surely.*

Proof. We would like to show that for $\mathbb{P}$-almost all $x \in \Omega$, there exists some $K_0 = K_0(x, \delta, M) \geq M$ such that for all $K \geq K_0$, $x_1^K \in J_K(\delta, M)$.

For any $x \in \Omega$, let

$$\tau(x) := \min\left\{n \geq M : \mathbb{P}([x_1^n]) \geq e^{-n(h(\mathbb{P})+\varepsilon)}\right\}.$$

Apparently τ is measurable and is a generalized stopping time. By (4.23), we have $\tau(x) < \infty$ $\mathbb{P}$-a.s., and τ is $\mathbb{P}$-a.s. finite. Applying Lemma 4.3.6, for $\mathbb{P}$-almost all $x \in \Omega$, there would exist $K_0 = K_0(x, \delta, M) \geq M$ such that for all $K \geq K_0$, the interval $[1, K]$ is $(1-\delta)$-packed by the intervals from

$$C(x, \tau) = \{[n, \tau(T^{n-1}(x)) + n - 1] : n \geq 1\},$$

and we take these intervals to form the desired collection $S(x_1^K)$.

By the definition of τ, we have $\tau(x) \geq M$ for $\mathbb{P}$-almost all $x \in \Omega$, so Property (i) holds as we would have $m_i = \tau(T^{n_i-1}(x)) + n_i - 1$ for all $i \in I_K$. The definition of τ also implies Property (ii), since for every $i \in I_K$, $m_i = \tau(T^{n_i-1}(x)) + n_i - 1$ gives $\tau(T^{n_i-1}(x)) = m_i - n_i + 1$, and we also have

$$\left[(T^{n_i-1}(x))_1^{m_i-n_i+1}\right] = [x_{n_i}^{m_i}],$$

implying

$$\begin{aligned} &\mathbb{P}\left(\left[(T^{n_i-1}(x))_1^{\tau(T^{n_i-1}(x))}\right]\right) \geq e^{-\tau(T^{n_i-1}(x))(h(\mathbb{P})+\varepsilon)} \\ &\qquad \Rightarrow \quad \mathbb{P}([x_{n_i}^{m_i}]) \geq e^{-(m_i-n_i+1)(h(\mathbb{P})+\varepsilon)}. \end{aligned}$$

Finally, Property (iii) follows immediately from $(1-\delta)$-packing. □

Lemma 4.3.8. *There exist* $0 < \delta < 1$ *and* $M \in \mathbb{N}$ *such that*

$$|J_K(\delta, M)| \leq e^{K(h(\mathbb{P})+2\varepsilon)}.$$

Proof. We adapt terminologies and notations from [Sh96]. A collection $S = \{[n_i, m_i]\}_{i \in I_K}$ of disjoint subintervals of $[1, K]$ is called a skeleton if Properties (i) and (iii) hold. A string x_1^K is compatible with the skeleton S if Property (ii) holds.

Technically, we will get the bound on $|J_K(\delta, M)|$ by first bounding the number of possible skeletons, and then bounding the number of strings compatible with each skeleton. $|J_K(\delta, M)|$ is then bounded by the number of possible skeletons multiplied by the number of compatible strings. Finally, we shall choose δ and M such that this lemma holds.

Suppose S is a skeleton. Since S is a collection of disjoint subintervals of $[1, K]$ and satisfies Property (i), we trivially have $M|S| \leq K \Rightarrow |S| \leq \frac{K}{M}$. Hence, there are at most $\frac{K}{M}$ points in $[1, K]$ that could be the starting points of the intervals in S.

An upper bound for the number of possible skeletons (if we denote this number simply as N_1) is then given by Proposition 4.3.5:

$$N_1 \leq \sum_{0 \leq j \leq \frac{K}{M}} \binom{K}{j} \leq e^{KH\left(\frac{1}{M}\right)}, \tag{4.24}$$

where $H\left(\frac{1}{M}\right)$ is of the form shown in (4.21).

Now for each skeleton S, let us consider the number of possible strings x_1^K that are compatible with it. First of all, for each $i \in I_K$, if we look at the number of possible distinct substrings $x_{n_i}^{m_i}$ for which Property (ii) holds, then it is bounded from above by $e^{(m_i - n_i + 1)(h(\mathbb{P}) + \varepsilon)}$, because otherwise, we can see that $\mathbb{P}$ would not be a probability measure by summing over these possible choices of distinct $x_{n_i}^{m_i}$: Denoting the set of all possible distinct substrings $x_{n_i}^{m_i}$ here as Q_i,

$$\sum_{x_{n_i}^{m_i} \in Q_i} \mathbb{P}([x_{n_i}^{m_i}]) > e^{(m_i - n_i + 1)(h(\mathbb{P}) + \varepsilon)} \cdot e^{-(m_i - n_i + 1)(h(\mathbb{P}) + \varepsilon)} = 1.$$

For the other indices in x_1^K that fall outside of $[n_i, m_i]$'s, there will be at most δK of them, due to Property (iii). Then for each such index, or each entry with such an index, it can be filled by letters from $\mathcal{A}$ and there are ℓ letters. Then this number of choices for the remaining entries is bounded above by $\ell^{\delta K}$.

Summing up what we have concluded above, for a given skeleton $S = \{[n_i, m_i]\}_{i \in I_K}$, the number of possible strings x_1^K that are compatible with S (denoted as N_2) is bounded above by

$$N_2 \leq \ell^{\delta K} \prod_{i \in I_K} e^{(m_i - n_i + 1)(h(\mathbb{P}) + \varepsilon)} \leq \ell^{\delta K} e^{K(h(\mathbb{P}) + \varepsilon)}, \tag{4.25}$$

where we used the fact that all $[n_i, m_i]$'s here are subintervals of $[1, K]$.

Combining (4.24) and (4.25), we obtain the estimate

$$|J_K(\delta, M)| \leq N_1 N_2 \leq e^{KH\left(\frac{1}{M}\right)} \ell^{\delta K} e^{K(h(\mathbb{P}) + \varepsilon)}.$$

Note that regardless what value ℓ is, we always have $\ell \leq e^{\ell}$. Then the estimate becomes

$$|J_K(\delta, M)| \leq e^{KH\left(\frac{1}{M}\right)} e^{\ell\delta K} e^{K(h(\mathbb{P})+\varepsilon)}.$$

Now we control the values of δ and M. Since by the expression (4.21), we can see that

$$\lim_{M\to\infty} H\left(\frac{1}{M}\right) = 0,$$

then we can choose M large enough such that $H\left(\frac{1}{M}\right) < \frac{\varepsilon}{2}$. On the other hand, choose $\delta < \frac{\varepsilon}{2\ell}$. We finally obtain

$$|J_K(\delta, M)| \leq e^{K\frac{\varepsilon}{2}} e^{K\frac{\varepsilon}{2}} e^{K(h(\mathbb{P})+\varepsilon)} = e^{K(h(\mathbb{P})+2\varepsilon)},$$

as desired. □

Let us proceed to finish the third proof. We choose δ and M as the ones from Lemma 4.3.8. Let

$$R_K = \left\{x_1^K \in \mathcal{A}^K : \mathbb{P}_K(x_1^K) < e^{-K(h(\mathbb{P})+3\varepsilon)}\right\}.$$

Then with $J_K = J_K(\delta, M)$, we have

$$\begin{aligned}\mathbb{P}_K(R_K \cap J_K) &= \sum_{x_1^K \in R_K \cap J_K} \mathbb{P}_K(x_1^K) < \sum_{x_1^K \in R_K \cap J_K} e^{-K(h(\mathbb{P})+3\varepsilon)} = |R_K \cap J_K|\, e^{-K(h(\mathbb{P})+3\varepsilon)} \\ &\leq |J_K|\, e^{-K(h(\mathbb{P})+3\varepsilon)} \leq e^{K(h(\mathbb{P})+2\varepsilon)} e^{-K(h(\mathbb{P})+3\varepsilon)} = e^{-K\varepsilon},\end{aligned}$$

where we applied Lemma 4.3.8.

It then follows that

$$\sum_{K=M}^{\infty} \mathbb{P}_K(R_K \cap J_K) < \infty,$$

which by the Borel–Cantelli lemma, means for $\mathbb{P}$-almost all $x \in \Omega$, $x_1^K \notin R_K \cap J_K$ eventually always[3]. However, by Lemma 4.3.7, we have $x_1^K \in J_K$ eventually almost-surely. Thus, we must have that $x_1^K \notin R_K$ eventually almost-surely. In other words, for $\mathbb{P}$-almost all $x \in \Omega$, there exists

[3] Note that there is a subtle difference between the terms "eventually always" and "eventually almost-surely", where the latter also encompasses $\mathbb{P}$-almost all $x \in \Omega$.

some $K'(x) \geq M$, such that for all $K \geq K'(x)$,

$$\mathbb{P}([x_1^K]) = \mathbb{P}_K(x_1^K) \geq e^{-K(h(\mathbb{P})+3\varepsilon)} \quad \Rightarrow \quad -\frac{1}{K}\log \mathbb{P}([x_1^K]) \leq h(\mathbb{P}) + 3\varepsilon.$$

Hence, for $\mathbb{P}$-almost all $x \in \Omega$,

$$\limsup_{K\to\infty} -\frac{1}{K}\log \mathbb{P}([x_1^K]) \leq h(\mathbb{P}) + 3\varepsilon.$$

Since $\varepsilon > 0$ is taken arbitrarily, we obtain that

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{P}([x_1^n]) \leq h(\mathbb{P}) \quad \mathbb{P}\text{-a.s.} \tag{4.26}$$

Combining (4.23) and (4.26), we get

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}([x_1^n]) = h(\mathbb{P}) \quad \mathbb{P}\text{-a.s.} \tag{4.27}$$

This completes the proof of "the entropy theorem" as presented in [Sh96]. We continue to show that $h(\mathbb{P}) = S(\mathbb{P})$ and the convergence also holds in L^1.

As stated earlier, we have got $h(\mathbb{P}) \leq S(\mathbb{P})$ by (4.22). It remains to show $h(\mathbb{P}) \geq S(\mathbb{P})$. Fix arbitrary $\varepsilon > 0$ and set

$$E_n = \left\{x_1^n \in \mathcal{A}^n : \mathbb{P}([x_1^n]) > e^{-n(h(\mathbb{P})+\varepsilon)}\right\}, \quad \forall\, n \in \mathbb{N}.$$

Let $B_n = \mathcal{A}^n \backslash E_n$ and we can do the following manipulations:

$$\begin{aligned} S(\mathbb{P}_n) &= -\sum_{x_1^n \in B_n} \mathbb{P}([x_1^n])\log \mathbb{P}([x_1^n]) - \sum_{x_1^n \in E_n} \mathbb{P}([x_1^n])\log \mathbb{P}([x_1^n]) \\ &\leq -\sum_{x_1^n \in B_n} \mathbb{P}([x_1^n])\log \mathbb{P}([x_1^n]) + n(h(\mathbb{P})+\varepsilon)\mathbb{P}_n(E_n). \end{aligned} \tag{4.28}$$

If $\mathbb{P}(B_n) = 0$, then we have a bound

$$S(\mathbb{P}_n) \leq n(h(\mathbb{P})+\varepsilon)\mathbb{P}_n(E_n) \tag{4.29}$$

right away; Otherwise, we can further observe that:

$$
-\sum_{x_1^n \in B_n} \frac{\mathbb{P}_n(x_1^n)}{\mathbb{P}_n(B_n)} \log\left(\frac{\mathbb{P}_n(x_1^n)}{\mathbb{P}_n(B_n)}\right) \leq \log|B_n| \leq \log|\mathcal{A}_n| = n\log\ell,
$$

where it is trivial to see that $\frac{\mathbb{P}_n(x_1^n)}{\mathbb{P}_n(B_n)}$ is a probability measure on B_n, and then Proposition D.1 (2) is applied.

Then the first term as in (4.28) can be further bounded:

$$
-\sum_{x_1^n \in B_n} \mathbb{P}([x_1^n]) \log \mathbb{P}([x_1^n]) \leq n\log\ell\, \mathbb{P}_n(B_n) - \mathbb{P}_n(B_n)\log\mathbb{P}_n(B_n),
$$

and from (4.28), the bound would further become

$$
S(\mathbb{P}_n) \leq n\log\ell\, \mathbb{P}_n(B_n) - \mathbb{P}_n(B_n)\log\mathbb{P}_n(B_n) + n(h(\mathbb{P})+\varepsilon)\mathbb{P}_n(E_n). \tag{4.30}
$$

A basic probabilistic fact is that $\mathbb{P}$-a.s. convergence implies convergence in probability. Hence, given how E_n is defined for each $n \in \mathbb{N}$, (4.27) implies that $\mathbb{P}_n(E_n) \to 1$ as $n \to \infty$. At the same time, $\mathbb{P}_n(B_n) \to 0$ as $n \to \infty$ as well.

Therefore, for either cases, if dividing n on both sides and passing it to infinity, we would have both (4.29) and (4.30) give that

$$
S(\mathbb{P}) = \lim_{n\to\infty} \frac{S(\mathbb{P}_n)}{n} \leq h(\mathbb{P}) + \varepsilon,
$$

which holds true for any $\varepsilon > 0$. We then obtain that $S(\mathbb{P}) \leq h(\mathbb{P})$ and this completes the proof of $h(\mathbb{P}) = S(\mathbb{P})$ and

$$
\lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}([x_1^n]) = S(\mathbb{P}) \quad \mathbb{P}\text{-a.s.} \tag{4.31}
$$

Finally, we have each $-\frac{1}{n}\log\mathbb{P}([x_1^n])$ is L^1, since its integral is simply $S(\mathbb{P}_n)$. As we also have both $\mathbb{P}$-a.s. convergence (4.31) and integral convergence as shown in (4.22), then by Scheffé's lemma (Lemma B.3), we have $-\frac{1}{n}\log\mathbb{P}([x_1^n]) \to S(\mathbb{P})$ in L^1. □

Chapter 5

Conclusion

The three proofs we have given to the SMB theorem elucidates three different facets of the theorem and their comparisons can be done on both technical and conceptual levels.

The subadditive proof is done in the extended setting of one-sided shift, namely two-sided shift, making it easier for us to construct the desired subadditivity. The key technical ingredient that this proof is based on is Kingman's subadditive ergodic theorem, together with two key *a priori* estimates which share similar forms: One is

$$\mathbb{P}\left(\{\log Z_{\max} > t\}\right) \leq \ell e^{-t},$$

which is used to show the integrability of $\log Z_{\max}$ and in turn, implies that the defined non-negative functions $\{\widehat{X}_n\}_{n\in\mathbb{N}}$ are integrable; the other one is

$$\widehat{\mathbb{P}}\left(\left\{\widehat{Y}_n > t\right\}\right) \leq e^{-t},$$

which is applied to verify the conditions for the error terms $\{\widehat{Y}_n\}_{n\in\mathbb{N}}$ as specified in Kingman's subadditive ergodic theorem. Note that the proofs of these two estimates are essentially identical and the two *a priori* estimates play a key role in verifying the assumptions of Kingman's subadditive ergodic theorem.

On the other hand, the martingale proof centers around the formula

$$S(\mathbb{P}_{n+1}) - S(\mathbb{P}_n) = \int_{\Omega} \log Z_{n+1} \, \mathrm{d}\mathbb{P}$$

and relies on the basic facts and properties of entropy and the functions Z_n's. It then employs the martingale convergence theorem to control the limit of Z_n's. Together with Birkhoff's ergodic theorem and *a priori* estimates, the final result follows. Some of the constructions and key relations involved in the martingale proof, for example, the so-called entropy formula

$$S(\mathbb{P}) = \int_{\Omega} \log Z \, \mathrm{d}\mathbb{P},$$

also have far-reaching implications in other topics of information theory and statistical mechanics, a sister field of information theory. One example is that it links with important notions of weak Gibbsianity, which is closely related to the theory of equilibrium measures of spin systems on Ω.

Finally, the Ornstein–Weiss proof is rather self-contained and it did not use any major theorem. A key ingredient is the ergodic stopping time packing lemma, which relates to Doob's upcrossing inequality, a result that is essential to the proof of the martingale convergence theorem. In this sense, the notion martingale is appearing in this proof, but not directly. Comparing with the other two, the flavor of the Ornstein–Weiss proof appears to be the most information-theoretic among the three. This is important in the sense that the arguments or technicality involved in the Ornstein–Weiss proof can also be sketched for deriving or proving some relevant subsequent results in information theory. This aspect cannot be seen in the other two proofs.

Despite the Ornstein–Weiss proof presented in this thesis does not cover general shift-invariant measures, this generalization can be realized through "ergodic decomposition" of any $\mathbb{P} \in \mathcal{P}_{\mathrm{inv}}(\Omega)$ (see [Sh96, Section I.4.c]). This can be fully detailed and conducted as an independent project.

Appendix

A Conditional expectation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and T be a measure-preserving transformation. For $X \in L^1(\Omega, \mathrm{d}\mathbb{P})$, if we define a measure $\mathbb{P}^X$ as follows,

$$\forall\, E \in \mathcal{F} : \mathbb{P}^X(E) := \int_E X \,\mathrm{d}\mathbb{P},$$

then naturally $\mathbb{P}^X \ll \mathbb{P}$ and X is the Radon–Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{P}^X$.

For the sub-σ-algebra $\mathcal{F}_T$ (see Section 2.1 for its definition), we know that X is $\mathcal{F}$-measurable but not necessarily $\mathcal{F}_T$-measurable. To construct an approximation of X which only preserves the information carried in $\mathcal{F}_T$, we can follow the same idea on viewing X as the Radon–Nikodym derivative $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}^X$.

For the measurable space $(\Omega, \mathcal{F}_T)$, we can define a probability measure on it by simply restricting $\mathbb{P}$ on the sub-σ-algebra $\mathcal{F}_T$. In other words, let $\mathcal{I} : \mathcal{F}_T \longrightarrow \mathcal{F}$ be the natural injection, then the probability measure we are defining on $(\Omega, \mathcal{F}_T)$ is $\mathbb{P}|_{\mathcal{F}_T} := \mathbb{P} \circ \mathcal{I}$.

A more natural way to see this is to consider the identity mapping $\mathrm{id} : (\Omega, \mathcal{F}) \longrightarrow (\Omega, \mathcal{F}_T)$. id is clearly $\mathcal{F}/\mathcal{F}_T$-measurable and its distribution $\mathbb{P} \circ (\mathrm{id})^{-1}$ defines a probability measure on $(\Omega, \mathcal{F}_T)$, which is exactly $\mathbb{P}|_{\mathcal{F}_T}$, so the natural injection $\mathcal{I}$ here is $(\mathrm{id})^{-1}$.

Similarly, define $\mathbb{P}^X|_{\mathcal{F}_T} := \mathbb{P}^X \circ (\mathrm{id})^{-1}$, and it is a measure on $(\Omega, \mathcal{F}_T)$. It is trivial to check that $\mathbb{P}^X|_{\mathcal{F}_T} \ll \mathbb{P}|_{\mathcal{F}_T}$ and by the Lebesgue–Radon–Nikodym theorem, the Radon–Nikodym derivative $X_T := (\mathrm{d}\mathbb{P}|_{\mathcal{F}_T})/(\mathrm{d}\mathbb{P}^X|_{\mathcal{F}_T})$ exists. We call this function X_T the conditional expectation of X conditioning on $\mathcal{F}_T$ and it is usually denoted as $X_T = \mathbb{E}[X|\mathcal{F}_T]$ in probability theory.

It can be checked that the integrals of X and X_T on any set $E \in \mathcal{F}_T$ agree:

$$\int_E X_T \,\mathrm{d}\mathbb{P}|_{\mathcal{F}_T} = \mathbb{P}^X|_{\mathcal{F}_T}(E) = \mathbb{P}^X \circ (\mathrm{id})^{-1}(E) = \mathbb{P}^X(E) = \int_E X \,\mathrm{d}\mathbb{P},$$

and on the other hand,

$$\int_E X_T \,\mathrm{d}\mathbb{P}|_{\mathcal{F}_T} = \int_E X_T \,\mathrm{d}\mathbb{P} \circ (\mathrm{id})^{-1} = \int_{(\mathrm{id})^{-1}(E)} X_T \circ \mathrm{id} \,\mathrm{d}\mathbb{P} = \int_E X_T \,\mathrm{d}\mathbb{P},$$

so $\int_E X \,\mathrm{d}\mathbb{P} = \int_E X_T \,\mathrm{d}\mathbb{P}$ for all $E \in \mathcal{F}_T$.

We also have X_T is T-invariant. Consider arbitrary $E \in \mathcal{F}_T$, then

$$\int_E X_T \,\mathrm{d}\mathbb{P} = \int_E X_T \,\mathrm{d}\mathbb{P}_T = \int_{T^{-1}(E)} X_T \circ T \,\mathrm{d}\mathbb{P} = \int_E X_T \circ T \,\mathrm{d}\mathbb{P},$$

so $X_T \circ T = X_T$ $\mathbb{P}$-a.s. and X_T is T-invariant.

We list some basic properties of conditional expectation below:

Similar to $\mathcal{F}_T$, for any sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$ and arbitrary $X \in L^1(\Omega, \mathrm{d}\mathbb{P})$, we can always define the conditional expectation of X conditioning on $\mathcal{G}$: $\mathbb{E}[X|\mathcal{G}]$, and it could be regarded as the best approximation of X on the sub-σ-algebra $\mathcal{G}$. $\mathbb{E}[X|\mathcal{G}]$ is $\mathcal{G}$-measurable and it exists uniquely $\mathbb{P}$-a.s.

(i). If $Y = \mathbb{E}[X|\mathcal{G}]$, then

$$\int_\Omega Y \,\mathrm{d}\mathbb{P} = \int_\Omega X \,\mathrm{d}\mathbb{P};$$

(ii). If X is already $\mathcal{G}$-measurable, then $\mathbb{E}[X|\mathcal{G}] = X$;

(iii) (Linearity). For any $a, b \in \mathbb{R}$ and $X, Y \in L^1(\Omega, \mathrm{d}\mathbb{P})$,

$$\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}];$$

(iv) (Monotonicity). If $X_1 \leq X_2$ $\mathbb{P}$-a.s., then $\mathbb{E}[X_1|\mathcal{G}] \leq \mathbb{E}[X_2|\mathcal{G}]$ $\mathbb{P}$-a.s.

For more properties and their proofs, please see Sections 9.7 and 9.8 of [Wil91].

B Minor mathematical theorems

The mathematical results given and proven in this section of the appendix are applied as necessary steps in the proofs involved in Chapter 2 and Chapter 4. Although as a mere measure-theoretic result, Lemma B.3 may be not as deep and important as the others, it is not that common in the usual mathematical literature and thus, is also included here.

Lemma B.1 (Fekete). *Let $\{a_n\}_{n\in\mathbb{N}}$ be a sequence of real numbers satisfying subadditive property:*

$$a_{n+m} \leq a_n + a_m, \quad \forall\, n, m \in \mathbb{N}.$$

Then the limit exists for $\frac{a_n}{n}$ as $n \to \infty$ and we have

$$\lim_{n\to\infty} \frac{a_n}{n} = \inf_{n\in\mathbb{N}} \frac{a_n}{n} \in [-\infty, \infty).$$

Proof. Fix an arbitrary $k \in \mathbb{N}$ and consider any $n \in \mathbb{N}$ such that $n \geq m$. Then there exist some $q_n \in \mathbb{N}$ (as quotient) and $r_n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ (as remainder), such that $n = kq_n + r_n$.

If for some $n \in \mathbb{N}$, we have $r_n = 0$, then $n = kq_n$ and by subadditivity (for q_n times),

$$\frac{a_n}{n} = \frac{a_{kq_n}}{kq_n} \leq \frac{q_n a_k}{kq_n} = \frac{a_k}{k}.$$

Then

$$\limsup_{n\to\infty} \frac{a_n}{n} \leq \frac{a_k}{k}.$$

Otherwise, if $r_n \neq 0$, then still, through subadditivity,

$$\frac{a_n}{n} = \frac{a_{kq_n+r_n}}{kq_n + r_n} \leq \frac{a_{kq_n} + a_{r_n}}{kq_n + r_n} = \frac{a_{kq_n}}{kq_n + r_n} + \frac{a_{r_n}}{n} \leq \frac{a_{kq_n}}{kq_n} + \frac{a_{r_n}}{n} \leq \frac{q_n a_k}{kq_n} + \frac{a_{r_n}}{n}. \tag{b.1}$$

Since in this case, $r_n \in \{1, 2, ..., k-1\}$ as the remainder is always smaller than the divisor, then we can bound a_{r_n} by the maximum among a_1, a_2, ..., a_{k-1}:

$$a_{r_n} \leq \max\{a_1, a_2, ..., a_{k-1}\} =: K.$$

As k is a fixed value, then so is the above maximum quantity, which we denote as K.

Then (b.1) becomes

$$\frac{a_n}{n} \le \frac{a_k}{k} + \frac{K}{n},$$

and by taking limit superior on both sides, we are getting

$$\limsup_{n\to\infty} \frac{a_n}{n} \le \frac{a_k}{k} + \lim_{n\to\infty} \frac{K}{n} = \frac{a_k}{k} + 0 = \frac{a_k}{k}.$$

Hence, we get

$$\limsup_{n\to\infty} \frac{a_n}{n} \le \frac{a_k}{k}$$

either way, and this holds for any $k \in \mathbb{N}$. Then, by taking infimum over $k \in \mathbb{N}$, we get

$$\limsup_{n\to\infty} \frac{a_n}{n} \le \inf_{k\in\mathbb{N}} \frac{a_k}{k}.$$

On the other hand, for every $n \in \mathbb{N}$, we naturally have

$$\inf_{k\in\mathbb{N}} \frac{a_k}{k} \le \frac{a_n}{n},$$

so

$$\inf_{k\in\mathbb{N}} \frac{a_k}{k} \le \liminf_{n\to\infty} \frac{a_n}{n}.$$

Then due to the relation

$$\limsup_{n\to\infty} \frac{a_n}{n} \le \inf_{k\in\mathbb{N}} \frac{a_k}{k} \le \liminf_{n\to\infty} \frac{a_n}{n},$$

we have shown that the limit exists for $\frac{a_n}{n}$ and

$$\lim_{n\to\infty} \frac{a_n}{n} = \inf_{n\in\mathbb{N}} \frac{a_n}{n}.$$

The fact that the limit equals the infimum of $\frac{a_n}{n}$ guarantees that the limit value cannot be positive infinity. □

Remark. It is possible that the limit of $\frac{a_n}{n}$ is $-\infty$. One trivial example is $a_n = -n^2$ for all $n \in \mathbb{N}$.

Clearly subadditivity holds:

$$a_{n+m} = -(n+m)^2 = -n^2 - 2nm - m^2 \leq -n^2 - m^2 = a_n + a_m,$$

and the limit for $\frac{a_n}{n}$ is trivially $-\infty$.

Theorem B.2 (Poincaré). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and T be a measure-preserving transformation on Ω. If $E \in \mathcal{F}$ and $\mathbb{P}(E) > 0$, then for $\mathbb{P}$-almost all $x \in E$, the sequence $\{T^n(x)\}_{n\in\mathbb{N}}$ revisits E infinitely often.*

Proof. Fix an arbitrary $E \in \mathcal{F}$ such that $\mathbb{P}(E) > 0$. We may focus on the subset of E such that the sequence $\{T^n(x)\}_{n\in\mathbb{N}}$ will never go back to E eventually always. We characterize this subset as follows.

For every $m \in \mathbb{N}_0$, set

$$F_m := \bigcup_{k=m}^{\infty} T^{-k}(E).$$

We then immediately have $E \subseteq F_0$, and we also have the nested relation $F_j \subseteq F_i$ whenever $i \leq j$. Moreover, it is easy to check that for any $i \leq j$, $F_j = T^{i-j}(F_i)$. Then by T-invariance of $\mathbb{P}$, we have $\mathbb{P}(F_j) = \mathbb{P}(F_i)$ for all $i, j \in \mathbb{N}_0$.

Now for any $x \in E$ such that $\{T^n(x)\}_{n\in\mathbb{N}}$ will eventually always be outside E, it means that there exists some $N \in \mathbb{N}$ such that for all $n \geq N$, $T^n(x) \notin E$, or equivalently $x \notin T^{-n}(E)$. In other words, for this specific x, we have

$$x \notin \bigcup_{n=N}^{\infty} T^{-n}(E) = F_N,$$

which is equivalent to

$$x \in E \backslash F_N.$$

We would like to consider all possible such x's. In other words, the threshold index N as appeared above might be any natural number. Then the set

$$G := \bigcup_{N=1}^{\infty} (E \backslash F_N) = E \backslash \left(\bigcap_{N=1}^{\infty} F_N \right)$$

characterizes all possible $x \in E$ such that for some $N \in \mathbb{N}$ and for all $n \geq N$, $T^k(x) \notin E$. We shall now show that G is a $\mathbb{P}$-null set.

Since $E \subseteq F_0$, then $E \backslash F_n \subset F_0 \backslash F_n$ and by monotonicity of a probability measure,

$$\mathbb{P}(E \backslash F_n) \leq \mathbb{P}(F_0 \backslash F_n).$$

Since for any $n \in \mathbb{N}$, $F_n \subseteq F_0$ and $\mathbb{P}(F_n) = \mathbb{P}(F_0)$, then

$$\mathbb{P}(E \backslash F_n) \leq \mathbb{P}(F_0 \backslash F_n) = \mathbb{P}(F_0) - \mathbb{P}(F_n) = 0.$$

This shows $\mathbb{P}(E \backslash F_n) = 0$ for all $n \in \mathbb{N}$, and

$$\mathbb{P}(G) = \mathbb{P}\left(\bigcup_{N=1}^{\infty} (E \backslash F_N)\right) \leq \sum_{N=1}^{\infty} \mathbb{P}(E \backslash F_N) = 0.$$

Hence, $\mathbb{P}(G) = 0$, meaning the subset of E such that $\{T^n(x)\}_{n \in \mathbb{N}}$ will eventually always be outside E has zero measure, so for $\mathbb{P}$-almost all $x \in E$, the sequence $\{T^n(x)\}_{n \in \mathbb{N}}$ revisits E infinitely often. □

Lemma B.3 (Scheffé). *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, suppose $\{X_n\}_{n \in \mathbb{N}}$ is a sequence of L^1 functions, $X \in L^1(\Omega, \mathrm{d}\mathbb{P})$, and X_n converges to X $\mathbb{P}$-a.s. Then X_n converges to X in L^1 if and only if*

$$\lim_{n \to \infty} \int_\Omega |X_n| \, \mathrm{d}\mathbb{P} = \int_\Omega |X| \, \mathrm{d}\mathbb{P}.$$

Proof. We begin with the "only if" direction. Suppose $X_n \to X$ in L^1. Then

$$\lim_{n \to \infty} \|X_n - X\|_1 = 0.$$

On the other hand,

$$\|X_n - X\|_1 = \int_\Omega |X_n - X| \, \mathrm{d}\mathbb{P} \geq \int_\Omega ||X_n| - |X|| \, \mathrm{d}\mathbb{P} \geq \left| \int_\Omega |X_n| \, \mathrm{d}\mathbb{P} - \int_\Omega |X| \, \mathrm{d}\mathbb{P} \right| \geq 0.$$

Hence,

$$\lim_{n \to \infty} \left| \int_\Omega |X_n| \, \mathrm{d}\mathbb{P} - \int_\Omega |X| \, \mathrm{d}\mathbb{P} \right| = 0$$

and

$$\lim_{n \to \infty} \int_\Omega |X_n| \, \mathrm{d}\mathbb{P} = \int_\Omega |X| \, \mathrm{d}\mathbb{P}.$$

For the “if” direction, we set

$$Y_n := |X_n| + |X| - |X_n - X|$$

for every $n \in \mathbb{N}$. It is easy to see that $\{Y_n\}_{n\in\mathbb{N}}$ is a sequence of non-negative L^1 functions.

Then we apply Fatou’s lemma to $\{Y_n\}_{n\in\mathbb{N}}$:

$$\int_\Omega \liminf_{n\to\infty} Y_n \,\mathrm{d}\mathbb{P} \leq \liminf_{n\to\infty} \int_\Omega Y_n \,\mathrm{d}\mathbb{P}. \tag{b.2}$$

Plug in the expression of each Y_n, we have

$$\begin{aligned}
\liminf_{n\to\infty} Y_n &= \liminf_{n\to\infty} \left(|X_n| + |X| - |X_n - X|\right) \\
&= \lim_{n\to\infty} |X_n| + |X| + \liminf_{n\to\infty} \left(-|X_n - X|\right) \\
&= |X| + |X| - \underbrace{\limsup_{n\to\infty} |X_n - X|}_{=0},
\end{aligned}$$

where we applied $\mathbb{P}$-a.s. convergence of X_n to X twice. Thus,

$$\int_\Omega \liminf_{n\to\infty} Y_n \,\mathrm{d}\mathbb{P} = \int_\Omega 2|X| \,\mathrm{d}\mathbb{P}.$$

On the other hand,

$$\begin{aligned}
\liminf_{n\to\infty} \int_\Omega Y_n \,\mathrm{d}\mathbb{P} &= \liminf_{n\to\infty} \int_\Omega \left(|X_n| + |X| - |X_n - X|\right) \mathrm{d}\mathbb{P} \\
&= \liminf_{n\to\infty} \left(\int_\Omega |X_n| \,\mathrm{d}\mathbb{P} + \int_\Omega |X| \,\mathrm{d}\mathbb{P} - \int_\Omega |X_n - X| \,\mathrm{d}\mathbb{P} \right) \\
&= \lim_{n\to\infty} \int_\Omega |X_n| \,\mathrm{d}\mathbb{P} + \int_\Omega |X| \,\mathrm{d}\mathbb{P} + \liminf_{n\to\infty} \left(-\int_\Omega |X_n - X| \,\mathrm{d}\mathbb{P} \right) \\
&= \int_\Omega |X| \,\mathrm{d}\mathbb{P} + \int_\Omega |X| \,\mathrm{d}\mathbb{P} - \limsup_{n\to\infty} \int_\Omega |X_n - X| \,\mathrm{d}\mathbb{P} \\
&= 2 \int_\Omega |X| \,\mathrm{d}\mathbb{P} - \limsup_{n\to\infty} \|X_n - X\|_1,
\end{aligned}$$

where we used the assumption that

$$\lim_{n\to\infty}\int_\Omega |X_n| \,\mathrm{d}\mathbb{P} = \int_\Omega |X| \,\mathrm{d}\mathbb{P}.$$

Then (b.2) becomes

$$2\int_\Omega |X| \,\mathrm{d}\mathbb{P} \leq 2\int_\Omega |X| \,\mathrm{d}\mathbb{P} - \limsup_{n\to\infty} \|X_n - X\|_1 \;\Rightarrow\; \limsup_{n\to\infty} \|X_n - X\|_1 \leq 0.$$

Hence, we get

$$\lim_{n\to\infty} \|X_n - X\|_1 = 0,$$

namely X_n converges to X in L^1. □

Theorem B.4 (The optional stopping theorem). *Let $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n\in\mathbb{N}_0}, \mathbb{P})$ be a filtered probability space and let $\{X_n\}_{n\in\mathbb{N}_0}$ be a supermartingale with respect to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$. Then for any bounded stopping times S and T with $0 \leq S \leq T$ $\mathbb{P}$-a.s., we have $\mathbb{E}[X_T] \leq \mathbb{E}[X_S]$.*

Proof. Suppose we have two bounded stopping times S and T such that $0 \leq S \leq T$ for $\mathbb{P}$-almost all on $x \in \Omega$. Then there exists some $n \in \mathbb{N}$ such that $T(x) \leq n$ for all $x \in \Omega$, namely

$$\mathbb{P}\left(\{T \leq n\}\right) = 1,$$

and $S \wedge T = \min(S, T)$ is also a stopping time.

For each $x \in \Omega$, we have either $S(x) \geq T(x)$ or $S(x) < T(x)$, and for the case that $S(x) < T(x)$, we would have at least one integer $k \in [0, n-1]$ such that $S(x) \leq k < T(x)$. With this in mind, we can then write the following:

$$X_T = X_{S\wedge T} + \sum_{k=0}^{n} (X_{k+1} - X_k)\, \mathbb{1}_{\{S\leq k<T\}}. \tag{b.3}$$

Recall that for a stopping time R, we define

$$\mathcal{F}_R := \{A \in \mathcal{F} : A \cap \{R \leq k\} \in \mathcal{F}_k, \forall\, 0 \leq k \leq \infty\}.$$

Then for any $A \in \mathcal{F}_S$ and $k \in \mathbb{N}_0$, we have $A \cap \{S \leq k\} \in \mathcal{F}_k$. Besides, $\{T > k\} \in \mathcal{F}_k$ for any

$k \in \mathbb{N}_0$ as well. Hence,

$$(X_{k+1} - X_k)\,\mathbb{1}_{\{S \le k < T\}}\mathbb{1}_A = (X_{k+1} - X_k)\,\mathbb{1}_{\{T > k\}}\mathbb{1}_{A \cap \{S \le k\}}$$

is $\mathcal{F}_k$-measurable for any $A \in \mathcal{F}_S$ and $k \in \mathbb{N}_0$.

Together with the properties of conditional expectation, as formulated in [AB19, Proposition 11.1][1], we have that for any $A \in \mathcal{F}_S$ and $k \in \mathbb{N}_0$,

$$\begin{aligned}\mathbb{E}\left[(X_{k+1} - X_k)\,\mathbb{1}_{\{T>k\}}\mathbb{1}_{A\cap\{S\le k\}}\right] &= \mathbb{E}\left[\mathbb{E}\left[(X_{k+1} - X_k)\,\mathbb{1}_{\{T>k\}}\mathbb{1}_{A\cap\{S\le k\}}|\mathcal{F}_k\right]\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[(X_{k+1} - X_k)\,|\mathcal{F}_k\right]\mathbb{1}_{\{T>k\}}\mathbb{1}_{A\cap\{S\le k\}}\right],\end{aligned}$$

and by the fact that $\{X_n\}_{n\in\mathbb{N}_0}$ is a supermartingale,

$$\mathbb{E}[X_{k+1}|\mathcal{F}_k] \le X_k = \mathbb{E}[X_k|\mathcal{F}_k] \;\Rightarrow\; \mathbb{E}\left[(X_{k+1} - X_k)\,|\mathcal{F}_k\right] \le 0,$$

so

$$\begin{aligned}\mathbb{E}\left[(X_{k+1} - X_k)\,\mathbb{1}_{\{S\le k<T\}}\mathbb{1}_A\right] &= \mathbb{E}\left[\mathbb{E}\left[(X_{k+1} - X_k)\,|\mathcal{F}_k\right]\mathbb{1}_{\{T>k\}}\mathbb{1}_{A\cap\{S\le k\}}\right]\\ &\le \mathbb{E}\left[0\cdot\mathbb{1}_{\{T>k\}}\mathbb{1}_{A\cap\{S\le k\}}\right]\\ &= 0.\end{aligned}$$

Combine this with (b.3), where we multiply $\mathbb{1}_A$ for arbitrary $A \in \mathcal{F}_S$ and integrate both sides, we get

$$\mathbb{E}[X_T\cdot\mathbb{1}_A] = \mathbb{E}[X_{S\wedge T}\cdot\mathbb{1}_A] + \sum_{k=0}^{n}\mathbb{E}\left[(X_{k+1} - X_k)\,\mathbb{1}_{\{S\le k<T\}}\mathbb{1}_A\right] \le \mathbb{E}[X_{S\wedge T}\cdot\mathbb{1}_A].$$

Then again by the properties of conditional expectation, we can write

$$\mathbb{E}\left[\mathbb{E}[X_T|\mathcal{F}_S]\cdot\mathbb{1}_A\right] = \mathbb{E}\left[\mathbb{E}[X_T\cdot\mathbb{1}_A|\mathcal{F}_S]\right] = \mathbb{E}[X_T\cdot\mathbb{1}_A] \le \mathbb{E}[X_{S\wedge T}\cdot\mathbb{1}_A].$$

Now that both $\mathbb{E}[X_T|\mathcal{F}_S]$ and $X_{S\wedge T}$ are $\mathcal{F}_S$-measurable, it follows that $\mathbb{E}[X_T|\mathcal{F}_S] \le X_{S\wedge T}$ $\mathbb{P}$-a.s.

[1] More specifically, see Properties (i) and (xii) in [AB19, Proposition 11.1].

and

$$\mathbb{E}[X_T] = \mathbb{E}\left[\mathbb{E}[X_T|\mathcal{F}_S]\right] \leq \mathbb{E}[X_{S\wedge T}].$$

In addition, since $S \leq T$ $\mathbb{P}$-a.s., then $X_{S\wedge T} = X_S$ $\mathbb{P}$-a.s. and

$$\mathbb{E}[X_{S\wedge T}] = \mathbb{E}[X_S].$$

Therefore, we have obtained $\mathbb{E}[X_T] \leq \mathbb{E}[X_S]$. □

Remark. The proof of Theorem B.4 given above follows closely to the proofs of different directions of implication involved in [AB19, Theorem 12.4].

C Shift-invariant measures on one-sided shift spaces

Given any probability measure $\mathbb{P}$ defined on $\Omega = \mathcal{A}^{\mathbb{N}}$, we can associate it with a sequence of functions $\{\mathbb{P}_n\}_{n\in\mathbb{N}}$, with each $\mathbb{P}_n$ defined on $\mathcal{A}^n$ in the following way:

$$\mathbb{P}_n(x_1, ..., x_n) := \mathbb{P}([x_1^n]).$$

One can see that each $\mathbb{P}_n$ essentially defines a probability measure on $\mathcal{A}^n$, with the σ-algebra being the power set of $\mathcal{A}^n$. For simplicity, we shall use the same notation $\mathbb{P}_n$ for this probability measure, and we refer $\mathbb{P}_n$'s as the marginals of the probability measure $\mathbb{P}$.

A nice thing about the marginals of a probability measure $\mathbb{P}$ is that they possess a good compatible property. Observe that

$$\bigsqcup_{x_{n+1}\in\mathcal{A}} [x_1^{n+1}] = [x_1^n],$$

and we apply $\mathbb{P}$ on both sides:

$$\mathbb{P}\left(\bigsqcup_{x_{n+1}\in\mathcal{A}} [x_1^{n+1}]\right) = \sum_{x_{n+1}\in\mathcal{A}} \mathbb{P}\left([x_1^{n+1}]\right) = \mathbb{P}\left([x_1^n]\right)$$

$$\Rightarrow \sum_{x_{n+1}\in\mathcal{A}} \mathbb{P}_{n+1}(x_1, x_2, ..., x_n, x_{n+1}) = \mathbb{P}_n(x_1, x_2, ..., x_n). \tag{c.1}$$

This holds true for all $(x_1, x_2, ..., x_n) \in \mathcal{A}^n$ and for all $n \in \mathbb{N}$. We refer (c.1) as "compatibility with marginals".

On the other hand, if we further assume $\mathbb{P}$ is shift-invariant, another compatibility condition will be satisfied. Suppose $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$. Observe that for our one-sided shift map T,

$$T^{-1}([x_1^n]) = \bigsqcup_{a\in\mathcal{A}} [(a, x_1, x_2, ..., x_n)].$$

Apply $\mathbb{P}$ on both sides and by its shift-invariance,

$$\mathbb{P}\left(T^{-1}([x_1^n])\right) = \mathbb{P}([x_1^n]) = \sum_{a\in\mathcal{A}} \mathbb{P}([(a, x_1, x_2, ..., x_n]) = \mathbb{P}\left(\bigsqcup_{a\in\mathcal{A}} [(a, x_1, x_2, ..., x_n)]\right)$$

$$\Rightarrow \quad \mathbb{P}_n(x_1, x_2, ..., x_n) = \sum_{a \in \mathcal{A}} \mathbb{P}_{n+1}(a, x_1, x_2, ..., x_n),$$

which can be rewritten as

$$\mathbb{P}_n(x_2, x_3, ..., x_{n+1}) = \sum_{x_1 \in \mathcal{A}} \mathbb{P}_{n+1}(x_1, x_2, ..., x_{n+1}). \tag{c.2}$$

This also holds true for all $(x_2, x_3, ..., x_{n+1}) \in \mathcal{A}^n$ and for all $n \in \mathbb{N}$, and we call the condition as shown in (c.2) "compatibility implied by shift-invariance".

In contrast, for a sequence of probability measures $\{\mathbb{P}_n\}_{n \in \mathbb{N}}$ with $\mathbb{P}_n$ defined on $\mathcal{A}^n$ for each $n \in \mathbb{N}$, if they fulfill the first compatibility (c.1), a unique probability measure $\mathbb{P}$ on Ω can be constructed from them, with each $\mathbb{P}_n$ acting as a marginal of $\mathbb{P}$. In addition, if $\{\mathbb{P}_n\}_{n \in \mathbb{N}}$ satisfies the second compatibility (c.2), then $\mathbb{P}$ is also shift-invariant.

This inverse direction of implication is a fundamental consequence of Kolmogorov's consistency theorem (also known as Kolmogorov's extension theorem or Kolmogorov's existence theorem), which is a very important theorem in probability theory for constructing probability measures on infinite product spaces. Let us formulate it in this very specific context of one-sided shift:

Theorem C.1 (Kolmogorov). *Let $\{(\mathcal{A}^n, \mathbb{P}_n)\}_{n \in \mathbb{N}}$ be a sequence of probability spaces such that for all $n \in \mathbb{N}$ and for all $x_1^n \in \mathcal{A}^n$, if the compatibility condition (c.1) is satisfied, then there exists a unique probability measure $\mathbb{P}$ on Ω such that for all $n \in \mathbb{N}$ and for all $x_1^n \in \mathcal{A}^n$, $\mathbb{P}([x_1^n]) = \mathbb{P}_n(x_1^n)$. If in addition, the compatibility condition (c.2) is also met by $\{\mathbb{P}_n\}_{n \in \mathbb{N}}$, then $\mathbb{P}$ is shift-invariant.*

In the two-sided shift setting, as discussed in Subsection 3.2.1, for any $\widehat{\mathbb{P}} \in \mathcal{P}(\widehat{\Omega})$, its marginals also satisfy a consistency condition, which is trivial to check: If four integers satisfy $n' \leq n \leq m \leq m'$, then

$$\widehat{\mathbb{P}}_n^m(x_n, ..., x_m) = \sum_{x_{n'}, ..., x_{n-1}, x_{m+1}, ..., x_{m'} \in \mathcal{A}} \widehat{\mathbb{P}}_{n'}^{m'}(x_{n'}, ..., x_{n-1}, x_n, ..., x_m, x_{m+1}, ..., x_{m'}). \tag{c.3}$$

There is also a version of Kolmogorov's consistency theorem for two-sided shift:

Theorem C.2. *Let $\{\mathcal{A}^{m-n+1}, \widehat{\mathbb{P}}_n^m\}_{n \leq m}$ be a family of probability spaces such that the probability measures $\widehat{\mathbb{P}}_n^m$ satisfy the consistency condition (c.3). Then there exists a unique Borel probability measure $\widehat{\mathbb{P}} \in \mathcal{P}(\widehat{\Omega})$ such that for all integers $n \leq m$, $\widehat{\mathbb{P}}_n^m(x_n^m) = \widehat{\mathbb{P}}([x_n^m])$.*

For the proof of general Kolmogorov's consistency theorem, we refer to [Bill95, Section 36].

D Entropy

This section of the appendix is dedicated to review basic aspects of entropy and give coherent notations as used in the main body of the thesis. For comprehensive preliminaries and materials on entropy with full details, one can consult [Jak19].

We introduce the notions of entropy progressively in the following order:

- Entropy of probability distribution on a finite set;
- Entropy of random variables taking values in finite sets;
- Entropy of shift-invariant measures on the shift space $\Omega = \mathcal{A}^{\mathbb{N}}$.

Entropy of probability distribution on a finite set

Let $\mathcal{A}$ be a finite set with $|\mathcal{A}| = \ell$. Denote $\mathcal{P}(\mathcal{A})$ to be the collection of all probability measures on $\mathcal{A}$. In other words, $\mathcal{P}(\mathcal{A})$ is the collection of all maps $\mathbb{P} : \mathcal{A} \longrightarrow [0, \infty)$ with $\sum_{a\in\mathcal{A}} \mathbb{P}(a) = 1$.

We have $\mathcal{P}(\mathcal{A})$ is a convex set with an obvious notion of convergence: For a sequence of probability measures $\{\mathbb{P}_n\}_{n\in\mathbb{N}}$ on $\mathcal{A}$, $\mathbb{P}_n$ converges to $\mathbb{P}$ if

$$\forall\, a \in \mathcal{A} : \lim_{n\to\infty} \mathbb{P}_n(a) = \mathbb{P}(a).$$

An element $\mathbb{P} \in \mathcal{P}(\mathcal{A})$ is called pure if $\mathbb{P}(a) = 1$ for some $a \in \mathcal{A}$, and is called uniform if $\mathbb{P}(a) = \frac{1}{\ell}$ for all $a \in \mathcal{A}$.

The entropy of a probability measure $\mathbb{P} \in \mathcal{P}(\mathcal{A})$ is defined to be

$$S(\mathbb{P}) = -\sum_{a\in\mathcal{A}} \mathbb{P}(a) \log \mathbb{P}(a).$$

Remark. We take the base for the logarithm in the definition of entropy to be e in our context. In many other contexts of information theory, however, it could be more often to use base 2 for the logarithm. Moreover, based on different contexts, the entropy $S(\mathbb{P})$ is defined up to a positive multiplicative constant, which amounts to specify units.

From this given definition, one can think about entropy as the measure of the randomness of a probability distribution. The more random the probability distribution is, the larger its entropy is. Note that we adapt the convention that “$0 \cdot \log 0 = 0 \cdot \infty = 0$” for the case that $\mathbb{P}(a) = 0$ for some $a \in \mathcal{A}$. Here are some basic properties of entropy:

Proposition D.1. *Let $\mathbb{P} \in \mathcal{P}(\mathcal{A})$ for some finite set $\mathcal{A}$ with $|\mathcal{A}| = \ell$.*

(1). $S(\mathbb{P}) \geq 0$, and $S(\mathbb{P}) = 0$ if and only if $\mathbb{P}$ is pure;

(2). $S(\mathbb{P}) \leq \log \ell$, and $S(\mathbb{P}) = \log \ell$ if and only if $\mathbb{P}$ is uniform;

(3). The map $\mathcal{P}(\mathcal{A}) \ni \mathbb{P} \mapsto S(\mathbb{P})$ (known as entropy map) is continuous and concave[2].

(4). The concavity property as shown in (3) has the following "almost convexity" counter-part:

$$S\left(\sum_{k=1}^{\ell} p_k \mathbb{P}_k\right) \leq \sum_{k=1}^{\ell} p_k S(\mathbb{P}_k) + S(p_1, ..., p_\ell) \tag{d.1}$$

with equality if and only if the supports[3] of pairwise probability distributions have empty intersection:

$$\operatorname{supp} \mathbb{P}_i \cap \operatorname{supp} \mathbb{P}_j = \varnothing, \quad \forall\, i \neq j. \tag{d.2}$$

Remark. Property (3) is called stability property of entropy. This concavity follows from the concavity of logarithm. The term $S(p_1, ..., p_\ell)$ as in (d.1) means that we set a new probability distribution by assigning the probability values to be $p_1, ..., p_\ell$ and then take its entropy. For the equality case of Property (4), namely when (d.2) holds, it is called split-additivity property of entropy.

For the proof of Proposition D.1, we refer to [Jak19, Proposition 3.1].

Now we consider the case of product spaces. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two finite sets and set $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$. Given $\mathbb{P} \in \mathcal{P}(\mathcal{A})$, denote $\mathbb{P}_1$ and $\mathbb{P}_2$ to be its marginals. We have

$$\mathbb{P}_1(a) = \sum_{a' \in \mathcal{A}_2} \mathbb{P}(a, a'),\ \forall\, a \in \mathcal{A}_1, \ \text{ and } \ \mathbb{P}_2(a') = \sum_{a \in \mathcal{A}_1} \mathbb{P}(a, a'),\ \forall\, a' \in \mathcal{A}_2.$$

For $a \in \operatorname{supp} \mathbb{P}_1$, we define the conditional probability measure on $\mathcal{A}_2$ conditioning on a by setting

$$\mathbb{P}^a_{2|1}(a') = \frac{\mathbb{P}(a, a')}{\mathbb{P}_1(a)}.$$

An identity then follows from this definition:

$$\sum_{a \in \operatorname{supp} \mathbb{P}_1} \mathbb{P}_1(a) \mathbb{P}^a_{2|1}(a') = \sum_{a \in \operatorname{supp} \mathbb{P}_1} \mathbb{P}(a, a') = \sum_{a \in \mathcal{A}_1} \mathbb{P}(a, a') = \mathbb{P}_2(a'),$$

[2] That is, for ℓ scalars $p_1, ..., p_\ell$ such that $p_k > 0$ for all $k \in \{1, ..., \ell\}$ and $\sum_{k=1}^{\ell} p_k = 1$ and $\mathbb{P}_1, ..., \mathbb{P}_\ell \in \mathcal{P}(\mathcal{A})$, we have $\sum_{k=1}^{\ell} p_k S(\mathbb{P}_k) \leq S\left(\sum_{k=1}^{\ell} p_k \mathbb{P}_k\right)$ with equality if and only if $\mathbb{P}_1 = \cdots = \mathbb{P}_\ell$.

[3] The support of a probability measure $\mathbb{P} \in \mathcal{P}(\mathcal{A})$ is defined to be: $\operatorname{supp} \mathbb{P} = \{a \in \mathcal{A} : \mathbb{P}(a) > 0\}$.

which holds for any $a' \in \mathcal{A}_2$.

Proposition D.2. *With the same setting as above, we have*

(1).

$$S(\mathbb{P}) = S(\mathbb{P}_1) + \sum_{a \in \operatorname{supp} \mathbb{P}_1} \mathbb{P}_1(a) S\left(\mathbb{P}^a_{2|1}\right);$$

(2). Strict subadditivity of entropy:

$$S(\mathbb{P}) \leq S(\mathbb{P}_1 \times \mathbb{P}_2) = S(\mathbb{P}_1) + S(\mathbb{P}_2), \tag{d.3}$$

with equality holds if and only if $\mathbb{P} = \mathbb{P}_1 \times \mathbb{P}_2$.

Remark. An intuitive view of (d.3) is that when taking the product between the marginals $\mathbb{P}_1$ and $\mathbb{P}_2$, the correlation between the two subsystems is completely erased, resulting in greater randomness.

Both split-additivity (d.2) and strict subadditivity (d.3) extend and define properties of $S(\mathbb{P})$ naturally, and it goes to the subject of axiomatization of entropy. An elegant discussion on the axiomatic characterizations of entropy is given in [Jak19, Section 3.4].

We refer Proposition D.2 together with its proof to [Jak19, Proposition 3.2].

Entropy of random variables taking values in finite sets

We adapt finite set $\mathcal{A}$ with $|\mathcal{A}| = \ell$. The notion of entropy related to a random variable X defined on $\mathcal{A}$ is defined via its probability distribution, denoted as $\mathbb{P}_X$.

Given an underlying probability space $(\tilde{\Omega}, \tilde{\mathbb{P}})$ and a random variable $X : \tilde{\Omega} \longrightarrow \mathcal{A}$ with its probability distribution

$$\mathbb{P}_X(a) = \tilde{\mathbb{P}}\left(\{x \in \tilde{\Omega} : X(x) = a\}\right),$$

the entropy of X is defined as

$$H(X) = S(\mathbb{P}_X) = -\sum_{a \in \mathcal{A}} \mathbb{P}_X(a) \log \mathbb{P}_X(a).$$

The basic properties of $H(X)$ are then the same as those of $S(\mathbb{P}_X)$ with change of terminologies. The properties presented in Proposition D.1 and Proposition D.2 would trivially be translated to the setting of random variables.

Full discussions on entropy of random variables, including the one involved when conditioning one random variable on another, can be found in [Sh96, Section I.6.a].

Entropy of shift-invariant measures on the shift space

Now we extend the setting to the shift space $\Omega = \mathcal{A}^{\mathbb{N}}$ and discuss the entropy of shift-invariant measures, which we have introduced in Appendix C.

For any $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$, its n-th marginal $\mathbb{P}_n$ is a probability measure on $\mathcal{A}^n$, which is a finite set. Hence, the entropy of $\mathbb{P}_n$ for each $n \in \mathbb{N}$ is

$$S(\mathbb{P}_n) = -\sum_{x_1^n \in \mathcal{A}^n} \mathbb{P}_n(x_1^n) \log \mathbb{P}_n(x_1^n).$$

Let $n, m \in \mathbb{N}$ and write $\mathcal{A}^{n+m} = \mathcal{A}^n \times \mathcal{A}^m$. Let μ_1 and μ_2 be the marginals of $\mathbb{P}_{n+m}$ on $\mathcal{A}^{n+m}$ with respect to this decomposition. We show that $\mu_1 = \mathbb{P}_n$ and $\mu_2 = \mathbb{P}_m$.

For any $x_1^n \in \mathcal{A}^n$, we have, by the definition of a marginal,

$$\begin{aligned} \mu_1(x_1, ..., x_n) &= \sum_{x_{n+1}, ..., x_{n+m} \in \mathcal{A}} \mathbb{P}_{n+m}(x_1, ..., x_n, x_{n+1}, ..., x_{n+m}) \\ \text{(By (c.1))} \quad &= \mathbb{P}_n(x_1, ..., x_n), \end{aligned}$$

so μ_1 and $\mathbb{P}_n$ agree on $\mathcal{A}^n$ and $\mu_1 = \mathbb{P}_n$.

For any $x_1^m \in \mathcal{A}^m$, since $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$, then by shift-invariance,

$$\begin{aligned} \mu_2(x_1, ..., x_m) &= \sum_{y_1, ..., y_n \in \mathcal{A}} \mathbb{P}_{n+m}(y_1, ..., y_n, x_1, ..., x_m) \\ \text{(By (c.2))} \quad &= \mathbb{P}_m(x_1, ..., x_m), \end{aligned}$$

so μ_2 and $\mathbb{P}_m$ agree on $\mathcal{A}^m$ and $\mu_2 = \mathbb{P}_m$.

Together with the subadditivity of entropy (d.3), we obtain that

$$S(\mathbb{P}_{n+m}) \leq S(\mu_1) + S(\mu_2) = S(\mathbb{P}_n) + S(\mathbb{P}_m).$$

Hence, the non-negative real sequence $\{S(\mathbb{P}_n)\}_{n \in \mathbb{N}}$ is subadditive and by Lemma B.1, it has a limit

$$S(\mathbb{P}) := \lim_{n\to\infty} \frac{S(\mathbb{P}_n)}{n} = \inf_{n\geq 1} \frac{S(\mathbb{P}_n)}{n}.$$

The quantity $S(\mathbb{P})$ is then called the entropy of $\mathbb{P} \in \mathcal{P}_{\text{inv}}(\Omega)$. Sometimes it is also called the specific entropy of $\mathbb{P}$. In more sophisticated literature, it is the special one-sided shift case of the more general notion of the Kolmogorov–Sinai entropy.

It then follows trivially from (1) and (2) of Proposition D.1 and the fact that $\frac{S(\mathbb{P}_n)}{n}$ converges to $S(\mathbb{P})$ as $n \to \infty$, that $0 \leq S(\mathbb{P}) \leq \log \ell$.

Bibliography

[AB] Artur Avila and Jairo Bochi, *On the subadditive ergodic theorem*, preprint (January, 2009). URL: https://personal.science.psu.edu/jzd5895/docs/kingbirk.pdf.

[AB19] Louigi Addario-Berry, *MATH 587/589 course notes*, lecture notes, McGill University, 2019. URL: https://problab.ca/louigi/courses/2020/math589/probnotes.pdf.

[Ba85] Andrew R. Barron, *The strong ergodic theorem for densities: generalized Shannon–McMillan–Breiman theorem*, Ann. Probab. **13** (1985), pp. 1292–1303. ISSN: 0091-1798. DOI: 10.1214/aop/1176992813. MR806226.

[Bill95] Patrick Billingsley, *Probability and Measure* (3rd ed.), John Wiley & Sons, Inc., New York, 1995. ISBN: 0-471-00710-2. MR1324786.

[Bir31] George D. Birkhoff, *Proof of the ergodic theorem*, Proc. Nat. Acad. Sci. USA **17** (1931), no. 12, pp. 656–660. ISSN: 00278424. URL: http://www.jstor.org/stable/86016.

[Bre57] Leo Breiman, *The individual ergodic Theorem of information theory*, Ann. Math. Stat. **28** (1957), no. 3, pp. 809–811. ISSN: 0003-4851. DOI: 10.1214/aoms/1177706899. MR92710.

[Der83] Yves Derriennic, *Un théorème ergodique presque sous-additif*, Ann. Probab. **11** (1983), no. 3, pp. 669–677. ISSN: 0091-1798. MR704553.

[Doob] Joseph L. Doob, *Stochastic Processes*, John Wiley & Sons, New York, 1953, pp. 292–390. MR26286.

[Fo99] Gerald B. Folland, *Real Analysis: Modern Techniques and Their Applications* (2nd ed.), John Wiley & Sons, New York, 1999. ISBN: 0-471-31716-0. MR1681462.

[Jak19] Vojkan Jakšić, *Lectures on Entropy. I: Information-Theoretic Notions*, Bahns et al (Eds): Dynamical Methods in Open Quantum Systems, Tutorials, Schools and Workshops in the Mathematical Sciences, Springer, 2019, pp. 141–268. MR3965239.

[KH95] Anatole Katok and Boris Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, vol. 54, Encyclopedia of Mathematics and its Applications, with a supplementary chapter by Katok and Leonardo Mendoza, Cambridge University Press, Cambridge, 1995. ISBN: 0-521-34187-6. DOI: 10.1017/CBO9780511809187. MR1326374.

[King68] John F. C. Kingman, *The ergodic theory of subadditive stochastic processes*, J. Roy. Statist. Soc. Ser. B **30** (1968), pp. 499–510. ISSN: 0035-9246. MR254907.

[Mc53] Brockway McMillan, *The basic theorems of information theory*, Ann. Math. Stat. **24** (1953), pp. 196–219. ISSN: 0003-4851. DOI: 10.1214/aoms/1177729028. MR55621.

[OW83] Donald Ornstein and Benjamin Weiss, *The Shannon–McMillan–Breiman theorem for a class of amenable groups*, Israel J. Math. **44** (1983), no. 1, pp. 53–60. ISSN: 0021-2172. DOI: 10.1007/BF02763171. MR693654.

[Raq23] Renaud Raquépas, *A gapped generalization of Kingman's subadditive ergodic theorem*, J. Math. Phys. **64** (2023), no. 6. ISSN: 0022-2488. DOI: 10.1063/5.0142431. MR4605876.

[Ru87] Walter Rudin, *Real and Complex Analysis* (3rd ed.), McGraw–Hill Book Co., New York, 1987. ISBN: 0-07-054234-1. MR924157.

[Sha48] Claude E. Shannon, *A mathematical theory of communication*, Bell System Tech. J. **27** (1948), pp. 379–423, 623–656. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1948.tb01338.x. MR26286.

[Sh96] Paul C. Shields, *The Ergodic Theory of Discrete Sample Paths*, American Mathematical Society, Providence, RI, 1996. ISBN: 0-8218-0477-4. DOI: 10.1090/gsm/013. MR1400225.

[Wa82] Peter Walters, *An Introduction to Ergodic Theory*, Springer–Verlag, New York–Berlin, 1982. ISBN: 0-387-90599-5. MR0648108.

[Wil91] David Williams, *Probability with Martingales*, Cambridge University Press, Cambridge, 1991. ISBN: 0-521-40455-X; 0-521-40605-6. DOI: 10.1017/CBO9780511813658. MR1155402.