Entropy and Compression: A Simple Proof of an Inequality of Khinchin–Ornstein–Shields

R. Aragona<sup>a,∗</sup>, F. Marzi<sup>a,∗∗</sup>, F. Mignosi<sup>a,b,∗∗∗</sup>, and M. Spezialetti<sup>c,∗∗∗∗</sup>

<sup>a</sup>DISIM, University of L’Aquila, Coppito, L’Aquila, Italy
<sup>b</sup>ICAR-CNR, Palermo, Italy
<sup>c</sup>University of Naples “Federico II,” Napoli, Italy

e-mail: ∗riccardo.aragona@univaq.it, ∗∗francesca.marzi@univaq.it,
∗∗∗filippo.mignosi@univaq.it, ∗∗∗∗matteo.spezialetti@guest.univaq.it

Received December 12, 2019; revised January 8, 2020; accepted January 15, 2020

Abstract—This paper concerns the folklore statement that “entropy is a lower bound for compression.” More precisely, we derive from the entropy theorem a simple proof of a pointwise inequality first stated by Ornstein and Shields and which is the almost-sure version of an average inequality first stated by Khinchin in 1953. We further give an elementary proof of the original Khinchin inequality, which can be used as an exercise for information theory students, and we conclude by giving historical and technical notes of such inequality.

Key words: ergodic sources, entropy, lossless data compression, one-to-one code sequence, Shannon–McMillan theorem.

DOI: 10.1134/S0032946020010020

In memoriam of Professor Aldo de Luca

1. INTRODUCTION AND NOTATION

This paper concerns the folklore statement that “entropy is a lower bound for compression.” Whilst almost every expert in the field knows this statement, both the history and the most general mathematical formulation of it, that is an inequality due to Ornstein and Shields in 1990 [1], are much less known. The main objective of this note is to give a simple proof of the Ornstein–Shields inequality. Such a simple proof, analogously as other proofs of same result known in literature, assumes that the well-known Shannon–McMillan–Breiman theorem holds. The name of this latter theorem refers to the entropy theorem for three different kinds of convergence. Since we would like to get straight to the point of the matter, in this section we simply introduce the notation and main results that will be used in the following, and then we will dedicate a whole section to a historical survey for this inequality that we consider one of the main contributions of this paper. We just point out here that Shields in [2, Section II.1] gives two proofs of the Ornstein and Shields inequality: one of them assumes, as we do, that the entropy theorem holds; the second proof is quite long but uses no previous deep result. Shields in his book also shows that the entropy theorem can easily be deduced from both the 1990 Ornstein and Shields inequality together with its complementary result that claims that universal codes reaching the entropy exist.

For any notation not explicitly defined in this paper, we refer the reader to [2].

We recall the definition of a typical set and, since we will make use of it in the following, we state Shannon’s Theorem 3 in [3], proved therein by Shannon for independent and identically distributed (i.i.d. for short) sources, as being presented in [4]. McMillan in [5] called it the asymptotic equipartition property, or AEP for short. For any set $A$, $|A|$ denotes the cardinality of $A$ and $X^n$ is the set of all the sequences of length $n$ of elements in $X$. 
Now we give the definition of a typical set and we state Shannon’s Theorem 3 for the case of i.i.d. sources.

**Definition 1.** The typical set \(A_ε^{(n)}\) with respect to \(p(x)\) is the set of sequences \((x_1, x_2, \ldots, x_n) \in X^n\) with the property

\[
2^{-n(H(X)+ε)} ≤ p(x_1, x_2, \ldots, x_n) ≤ 2^{-n(H(X)−ε)}.
\]

**Theorem 1** [3, Theorem 3]. 1) If \((x_1, \ldots, x_n) \in A_ε^{(n)}\), then

\[
H(X) - ε ≤ -\frac{1}{n} \log_2 p(x_1, \ldots, x_n) ≤ H(X) + ε;
\]

2) \(P\{A_ε^{(n)}\} > 1 - ε\) for \(n\) sufficiently large;
3) \(|A_ε^{(n)}| ≤ 2^{n(H(X)+ε)}\) for \(n\) sufficiently large;
4) \(|A_ε^{(n)}| ≥ (1 - ε)2^{n(H(X)−ε)}\) for \(n\) sufficiently large.

It is worth noting that points 1) and 2) of the previous theorem represent a statement of [3, Theorem 3], which will be exactly stated in the historical section, Section 3, and points 3) and 4) are direct consequences of points 1) and 2).

Shannon’s theorem is proved in [3] for convergence in probability: in the body of the paper only for i.i.d. sources, and in Appendix 3 only for Markov-like sources that are ergodic and stationary. Later McMillan [5] stated and proved the entropy theorem for stationary (not necessarily ergodic) processes with mean \(L_1\)-convergence. Finally, Breiman [6,7] proved the same result for stationary ergodic processes and finite alphabets in the case of almost surely convergence; this latter result is extended by Chung [8,9] to countably infinite alphabets. As is mentioned above, these statements of the entropy theorems for different types of convergence are called the Shannon–McMillan–Breiman theorem or the entropy theorem as in [2].

The following statement is the entropy theorem in its almost surely version, or pointwise version, of Theorem 1 as being presented in [2, Theorem I.7.1].

**Theorem 2.** For any stationary and ergodic source and for each \(ε > 0\), \((x_1, \ldots, x_n) \in A_ε^{(n)}\) eventually almost surely.

Clearly, Theorem 2 implies Theorem 1.

For the sake of completeness, anyhow, we report from [2] the following definition and the Borel–Cantelli lemma (cf. [2, Lemma I.1.14]).

**Definition 2.** A property \(P\) is said to be measurable if the set of all \(x\) for which \(P(x)\) is true is a measurable set. If \(\{P_n(x)\}\) is a sequence of measurable properties, then \(P_n(x)\) holds eventually almost surely if for almost every \(x\) there is an \(N = N(x)\) such that \(P_n(x)\) is true for \(n ≥ N\).

In Theorem 2 the property \(P\) of \(x = (x_1, \ldots, x_n)\) is the membership in \(A_ε^{(n)}\).

**Lemma** (Borel–Cantelli). If \(\{C_n\}\) is a sequence of measurable sets in a probability space \((X, Σ, μ)\) such that \(\sum μ(C_n) < ∞\), then for almost every \(x\) there is an \(N = N(x)\) such that \(x ∉ C_n\) for \(n ≥ N\).

We now give the main definitions of this note.

First, we recall that, given a finite set of symbols (or words) \(χ\), the set \(χ^*\) is the set of all sequences of symbols (or words) of any length \(χ\), i.e., \(χ^* = ∪_{n=0}^{+∞} χ^n\).

**Definition 3.**

1) A binary faithful-code sequence or one-to-one code sequence is any function \(γ\) from \(χ^*\) to \(\{0, 1\}^*\) such that for any integer \(n\) its restriction to \(χ^n\) is injective.
2) A binary prefix-code sequence is a one-to-one code sequence $\gamma$ such that for any integer $n$ its restriction to $\chi^n$ is a prefix code.

3) A binary lossless compressor is any injective function $\gamma$ from $\chi^*$ to $\{0,1\}^*$.

Note that the term faithful-code sequence in part 1) of Definition 3 is used in [1] and in [2], and one-to-one code sequence is used in [10] and in many other articles (see [11,12], references therein, and citing articles). The same notion is also classically called a non-singular code sequence (see, for example, [4] and again [10]).

For the sake of simplicity, we restrict our attention to binary codes. All the notations and results presented here can be extended to general finite alphabets, analogously as in Khinchin [13].

By definition, the class of one-to-one code sequences strictly includes the class of prefix-code sequences and the class of lossless compressors, and therefore, any result that holds for the class of one-to-one code sequences holds also for the other two classes and it is, from a logical view point, a stronger result. Anyhow, by adding some further reasoning and proofs it is sometimes possible to pass from a weaker to a stronger result. This has been done from an historical point of view and it is explained in this note in Section 3.2 below.

As the reader can notice, for us a lossless compressor is just an injective function, which grants for unique decodability. If the length of the message to encode is known to the decoder, also one-to-one code sequences grant for unique decodability. Note also that a compression, in the usual meaning of this term, is not granted, since a “compressor,” following Definition 3 above, can even expand texts in average.

The paper is organized as follows. In Section 2, we derive from the Shannon–McMillan–Breiman theorem an original proof of Ornstein–Shields inequality, which we think to be simpler than the ones given in literature until now to our best knowledge. Even if the inequality first stated by Khinchin [13], called the Khinchin inequality, follows from the Ornstein–Shields inequality, in Section 2.1 we give, starting from Shannon’s Theorem 3 in [3], an original and elementary proof of such inequality that avoids the use of measure theory tools such as the lemma of Borel–Cantelli and that can be used as an exercise for information theory students.

Section 3 is entirely dedicated to a historical survey and some technical observations of such inequality. Once given an overview on the proofs of Ornstein–Shields inequality found in literature, we show in Section 3.3 why we think that our proof is simpler than the others.

The last two sections are dedicated to the acknowledgements and to some memories of the third author concerning Professor Aldo de Luca.

2. SIMPLE PROOFS

The following theorem is due to Ornstein and Shields in [1], and we give a simpler original proof. It holds for any stationary and ergodic source of entropy $H$.

To avoid confusion in notation, we denote by $\|\cdot\|$ the length of a sequence of symbols (or words) instead of the more common $|\cdot|$ that is used here to denote the cardinality of a set.

Theorem 3. For any one-to-one code sequence $\gamma$ almost surely,

$$\liminf_{n \in \mathbb{N}} \frac{\|\gamma((x_1,\ldots,x_n))\|}{n} \geq H.$$

Proof. For any $\varepsilon > 0$ and any $n$, define the sets $C_{\varepsilon}^{(n)}$ as

$$C_{\varepsilon}^{(n)} = \{(x_1,\ldots,x_n) \in A_{\varepsilon}^{(n)} : \|\gamma(x_1,\ldots,x_n)\| \leq \log_2(|A_{\varepsilon}^{(n)}|) - 3\varepsilon n - 1\}.$$

Since $\gamma$ is one-to-one, $|C_{\varepsilon}^{(n)}| = |\gamma(C_{\varepsilon}^{(n)})|$. Since $\gamma(C_{\varepsilon}^{(n)})$ is a subset of $\{0,1\}^*$, for any $t$ it must contain less than $2^{t+1}$ strings of length at most $t$. Thus, $|C_{\varepsilon}^{(n)}| \leq |A_{\varepsilon}^{(n)}|2^{-3\varepsilon n}$. PROBLEMS OF INFORMATION TRANSMISSION Vol. 56 No. 1 2020
Since $C^{(n)}_{\varepsilon} \subseteq A^{(n)}_{\varepsilon}$, each element in $C^{(n)}_{\varepsilon}$ has a bound on its probability given in Definition 1. Using it and the bound on $|A^{(n)}_{\varepsilon}|$ given in part 3) of Theorem 1, we have that $P(C^{(n)}_{\varepsilon}) \leq 2^{-\varepsilon n}$ for $n$ sufficiently large. For each fixed $\varepsilon > 0$ we apply the Borel–Cantelli lemma to the sequence $C^{(n)}_{\varepsilon}$ and, by Theorem 2, eventually almost surely $(x_1, \ldots, x_n)$ belongs to $A^{(n)}_{\varepsilon}$ and not to $C^{(n)}_{\varepsilon}$, i.e., $\|\gamma(x_1, \ldots, x_n)\| \geq \log_2(|A^{(n)}_{\varepsilon}|) - 3\varepsilon n$.

Using the bound on $|A^{(n)}_{\varepsilon}|$ given in part 4) of Theorem 1, for each $\varepsilon$, almost surely

$$\liminf_{n \to \infty} \frac{\|\gamma(x_1, \ldots, x_n)\|}{n} \geq \liminf_{n \to \infty} \frac{\log_2(|A^{(n)}_{\varepsilon}|) - 3\varepsilon n}{n} \geq H(X) - 4\varepsilon,$$

where the value $\frac{\log_2(1 - \varepsilon)}{n}$ disappears in the lim inf because $\varepsilon$ is fixed and $\log_2(1 - \varepsilon)$ is a constant. Since this holds for any $\varepsilon$, and in particular for the enumerable sequence $\varepsilon_m = \frac{1}{m}$, $m = 1, \ldots, +\infty$, a simple exercise in measure theory completes the proof.

### 2.1. Average Case

Here we give an elementary proof of the Khinchin result, i.e., that entropy is a lower bound for the average compression, without using the Borel–Cantelli lemma (Section 1). We derive it directly from the Shannon theorem (Section 1, Theorem 1), which is present in all textbooks of information theory (cf., for instance, [4,14–16]).

Clearly, any average result follows from the analogous pointwise result and, in particular, the result of this subsection follows from Theorem 3. Anyway, we have decided to keep both proofs, since the proof of the average result uses only elementary mathematical notions and it could be used as an exercise for information theory students analogously as in the case of Shannon’s [3, Theorem 4] (cf. [4, Chapter 3, Exercise 11]). The average result implies as corollaries classical information theory results such as, for instance, the fact that the entropy is a lower bound of the average length of uniquely decodable block codes or a lower bound for the compression ratio of arithmetic compressors.

Note that we make use of the same idea of the proof of Theorem 3, and, indeed, the first six lines of both proofs are exactly the same.

**Theorem 4.** For all i.i.d. sources of entropy $H$ and any one-to-one code sequence $\gamma$

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{x \in X^n} \|\gamma(x)\| \cdot p(x) \geq H.$$

**Proof.** Proceed as in the proof of Theorem 3 up to the fact that $P(C^{(n)}_{\varepsilon}) \leq 2^{-\varepsilon n} \leq \varepsilon$ for $n$ sufficiently large.

For any $n$,

$$\frac{1}{n} \sum_{x \in X^n} \|\gamma(x)\| \cdot p(x) \geq \frac{1}{n} \sum_{x \in A^{(n)}_{\varepsilon} \setminus C^{(n)}_{\varepsilon}} \|\gamma(x)\| \cdot p(x)$$

$$\geq \frac{\log_2(|A^{(n)}_{\varepsilon}|) - 3\varepsilon n}{n} \sum_{x \in A^{(n)}_{\varepsilon} \setminus C^{(n)}_{\varepsilon}} p(x)$$

$$= \left[ \frac{\log_2(|A^{(n)}_{\varepsilon}|)}{n} - 3\varepsilon \right] [P(A^{(n)}_{\varepsilon}) - P(C^{(n)}_{\varepsilon})].$$
For $n$ sufficiently large, using part 4) of Theorem 1, one has $\frac{\log_2(|A_{e(n)}|)}{n} \geq H - 2\varepsilon$. For $n$ sufficiently large, using part 2) of Theorem 1 and the previously obtained bound on $\mathbf{P}(C_{e(n)}^{(n)})$, one has that for any $\varepsilon < \frac{1}{2}$

$$\liminf_{n \in \mathbb{N}} \frac{1}{n} \sum_{x \in \chi^n} \|\gamma(x)\| \cdot p(x) \geq (H - 5\varepsilon)(1 - 2\varepsilon).$$

Since this holds for any $\varepsilon < \frac{1}{2}$, a simple exercise in calculus completes the proof.

From the previous proof we can derive, by Theorem 1, the desired average lower bound for compression. It is worth to highlight that the previous proof can be used to prove the same average lower bound also for stationary and ergodic sources, analogously as the original Khinchin proof, since Theorem 1 can be generalized to these kind of sources. Note finally that the original Khinchin’s result was proved for stationary Markov chains.

3. HISTORICAL AND TECHNICAL NOTES

3.1. Historical Survey

Theorem 3 has been proved for the first time in [1] in 1990 by Ornstein and Shields. More precisely, it is a part of their Theorem 1 in the invertible case, which is proved in [1, Section 2]. Another simple proof is given in the book of Shields [2] in 1996, together with a second proof, which is similar to the original proof given in [1, Section 2] and does not make use of the entropy theorem. The inequality is stated in Theorem II.1.2 of [2], and the two proofs are given, respectively, in Sections II.1.b and II.1.c. The third proof is given in [12] in 2014 by Kontoyiannis and Verdú in part ii) of their Theorem 12. Those three above are all the proofs of Theorem 3 that are present in the literature to our best knowledge.

As was noted by Kontoyiannis and Verdú in [12] before their Theorem 12, the weaker corresponding result for prefix code sequences instead of that for one-to-one sequences was established in [11,17,18]. More precisely, in [17], which is the 1985 PhD thesis of Barron, there is proved a lemma that has, as an easy consequence, the analog of Theorem 3 for prefix code sequences. Kontoyiannis uses it in his 1997 paper [11] and claims that: “It is an unpublished result that appeared in [17], and also, in a more general form in [19],” and then he gives a proof of it in the Appendix. Indeed, it seems a bit earlier Barron’s lemma was stated, again as an unpublished result of Barron, and proved in [2]. It is worth noting that Barron’s PhD thesis [17] and Algoet’s PhD thesis [19] are of the same year, both from Stanford University and with the same supervisor, Prof. Thomas M. Cover.

We further note here that the weaker corresponding result for lossless compressors was stated and proved for the first time by Khinchin in [13] in 1953 (see also [20]). It is a weaker result not just because of the fact that it is proved only for lossless compressors, but also because it is a result that is stated “in average” instead of being pointwise. Note that none of the previously cited paper that we examined cites, in turn, Khinchin. Also for this reason we consider this historical section as one of the main contributions of this paper.

It is worth highlighting that, due to the analogy with the Shannon–McMillan–Breiman theorem, we decided to call Theorem 3 the “Khinchin–Ornsten–Shields inequality” even if some credits have to be given to Imre Csiszár, as it will be explained in Section 3.2 below.

Khinchin’s result comes from the golden age of the beginning of information theory. Indeed, Khinchin relates it to one of the first theorems of Shannon’s seminal paper [3]. Indeed, the following
Theorem 5 [3, Theorem 3]. Given any \( \varepsilon > 0 \) and \( \delta > 0 \), we can find \( N_0 \) such that the sequences of any length \( N \geq N_0 \) fall into two classes:

1. A set whose total probability is less than \( \varepsilon \);
2. The remainder, all of whose members have probabilities satisfying the inequality

\[
\left| \frac{\log p^{-1}N}{N} - H \right| < \delta.
\]

Note that Theorem 1 is the same statement of the previous theorem considering \( \delta = \varepsilon \).

Theorem 5 is proved in [3] for convergence in probability: in the body of the paper only for i.i.d. sources, and in Appendix 3 only for Markov-like sources that are ergodic and stationary. Later McMillan [5] showed for stationary (not necessarily ergodic) processes with mean \( L_1 \)-convergence a simple corollary to the entropy ergodic theorem relating codebook sizes to the appropriate entropies, using measure theory and martingale theory. Finally, Breiman [6,7] proved the same result for stationary and ergodic processes and finite alphabets in the case of almost everywhere convergence, using again the martingale theory. Breiman’s result was extended by Chung [8,9] to countably infinite alphabets.

As was already mentioned, these three statements of the entropy ergodic theorems for different kinds of convergence are called the Shannon–McMillan–Breiman theorem.

There is wide literature on developing simple proofs for the entropy ergodic theorems and their generalizations. Usually these proofs are not short but often they use fairly elementary mathematics. Just to give an example of a simple proof of a generalization of the Shannon–McMillan theorem, we may cite a result proved by Kieffer in [21], where a short proof is provided of the most general then-known related result which did not involve martingale theory using an extension of a proof presented in [22, Theorem 3.5.3]. These results successfully tackled the overall problem for the \( L_1 \) convergence case.

However, it is far beyond the scope of this section to give a historical survey of the Shannon–McMillan–Breiman theorem and its generalizations, extensions, and consequences, but the interested reader can also see [23–25] and citing articles.

In general, it is far beyond the scope of this paper to give a simple proof of the entropy theorem even if, as is noted by Shields in [2] and reported by us in Section 1, the Ornstein–Shields inequality is deeply connected to the entropy theorem. We want to give a new simple proof of the Ornstein–Shields inequality, which we believe to be simpler than the other existing proofs of the same kind that make use of the entropy theorem, as we will discuss in Section 3.3 below.

Coming back to Shannon’s fourth theorem, consider for any \( n \) the sequences of length \( n \) to be arranged in order of decreasing probability. For any \( q, 0 < q < 1 \), define \( n(q) \) to be the number that we must take from this ordered set, starting with the most probable one, in order to accumulate a total probability \( q \) for those taken. [3, Theorem 4] states that \( \lim_{n \to \infty} \frac{\log_2 n(q)}{n} = H \).

Shannon, after stating his Theorem 4, claims, without proving anything, that: “We may interpret \( \log_2 (n(q)) \) as the number of bits required to specify the sequence when we consider only the most probable sequences with a total probability \( q \). Then \( \frac{\log_2 n(q)}{n} = H \) is the number of bits per symbol for the specification. The theorem say that for large \( n \) this will be independent of \( q \) and equal to \( H \).”

This Shannon’s interpretation is correct, and starting from this seminal claim several formal consequences have been proved that usually concern compressors that code blocks of uniform length \( n \) into blocks of uniform length \( k \) (see, for instance, [15,26]).
Instead, following the above Shannon’s interpretation as a research direction and indeed exploiting [3, Theorem 4], Khinchin in [13] gives a formal definition of the average compression for sequences of fixed length and of the compression coefficient as lim sup of the average compression. Then he proves in [13, Theorem 4], for the first time to our best knowledge, that the entropy is a lower bound for the compression coefficient of any injective function. Khinchin’s proof works also when lim inf is used in the place of lim sup in the definition of the average compression, as we did in Theorem 4. The lim inf gives indeed a stronger result from a logical point of view.

This Khinchin’s inequality is also proved in another context, i.e., in the case of linguistic sources, by Hansel, Perrin, and Simon in [27].

3.2. From Weaker to Stronger Inequality

In Section 2 of the original 1990 paper of Ornstein and Shields [1] where it is given the first written statement and proof of Theorem 3, it is described a technique of transforming one-to-one code sequences into prefix-code sequences adding a “small” overhead header. The description of this technique takes a good part of their Section 2, and it is also described in the 1996 book of Shields [2] in Section I.7.d; the “small” overhead header is $O(\log(\|\gamma(w)\|))$ size for any $w \in \chi^*$ and gives “no change in asymptotic performance” as said in the book of Shields.

There is no room in this short subsection to describe in detail this technique that mainly consists in prepending to $\gamma(w)$ the Elias delta coding of $\|w\|$ [28]. Here we want to note that in [1] this technique is credited, in their Section 2, to Imre Csiszár and, moreover, the authors say in the acknowledgements: “We wish to give special thanks to Imre Csiszár, who corrected several of our errors and made many suggestions for improvement of our discussion.”

In the 1996 book of Shields [2] the simpler proof among the two proofs contained therein consists exactly in linking the technique of Csiszár and Barron’s 1985 result [17], and it will be discussed in Section 3.3.

3.3. Analogies and Differences between Proofs

What is a “simple” proof? Can we say that a simple and elegant proof that is given at the end of a mathematical book and that makes use of all previous results of that book is really “simple”? We think that the second question has “no” as a right answer, and we have no answer to the first question. Maybe a good attempt is given by Occam’s principle or “razor” discussed also in Barron’s PhD thesis [17], which can give suggestions to decide when a proof is “simpler” than another. In this subsection we analyze the three proofs that were known before our proof and compare them all.

As we said above, in [2, Sections II.1.b and II.1.c] there are reported two proofs of Theorem 3, one short and elegant, and the second longer, “which was developed in [1], and does not make use of the Shannon–McMillan–Breiman theorem [...]” as Shields wrote.

Let us first discuss the second longer proof. It is more than four pages long in the book of Shields and also refers to some previous notation and results, but it does not use previous deep results. Even if it does not make use of the Shannon–McMillan–Breiman theorem, we cannot consider it simpler than our proof. We think that it is not possible to claim that one of these two proofs is simpler than the other. We emphasize that the beauty of this long proof resides also in the fact that just after it, in the book of Shields, there is given a short proof that logically derives the Shannon–McMillan–Breiman theorem from Theorem 3 and from the converse inequality stated in Theorem II.1.1 of [2]. This reasoning shows the generality and logical power of Theorem 3.

Let us now analyze the three remaining proofs: the first elegant proof written in the book of Shields in 1996, the 2014 Kontoyiannis and Verdú proof, and our proof of Theorem 3. All three use the Borel–Cantelli lemma (Section 1) and the Shannon–McMillan–Breiman theorem.
Concerning the first elegant proof in [2], Shields claims that Theorem 3 “will follow from the entropy theorem, together with a surprisingly simple lower bound on prefix-code word length.” This proof indeed makes use of several components:

1) For any \( n \), there is a conversion of one-to-one code into a prefix code “with no change in asymptotic performance” by using the Imre Csiszár technique (cf. Section 3.2);

2) The use of Barron’s lemma proved in [17] (see also [2, Lemma II.1.3]), which, in turn, uses
   2a) the Kraft inequality for prefix codes [29];
   2b) the Borel–Cantelli lemma (Section 1);

3) The Shannon–McMillan–Breiman theorem (Theorem 2).

We want to emphasize here that our simple proof of Theorem 3 makes use only of the above points 2b) and 3) analogously as the Kontoyiannis and Verdú proof, and it is overall shorter even if the length of the proof of the Kraft inequality is not considered. Therefore, we think that our proof is simpler than the first of the two proofs contained in the book of Shields. A last argument in favor of our thinking is explained in the last part of this subsection.

Let us now examine the proof of Kontoyiannis and Verdú of part ii) of [12, Theorem 12], that is exactly our Theorem 3. Their proof makes use in turns of [12, Theorem 11] that, again in turn, uses [12, Theorem 5], which “is a natural analog of the corresponding converse established for prefix compressors in [17]” by Barron, as the authors say. Their elegant proof of [12, Theorem 5] uses a counting argument to generalize Barron’s lemma and avoids the use of the Kraft inequality. The resulting proof of part ii) of [12, Theorem 12], even including all these backpointers to previous theorems and their proofs, turns out to be overall just a bit longer than the elegant proof reported in the Shields book [2] but at least it does not make use of the Kraft inequality.

As a final argument we note that both above proofs [2, Section II.1.b] and [12, Theorem 12], but not our proof, in the first part show that the probability of the set of sequences of length \( n \) that have a “small” compressed length is summable in \( n \), and this allows the use of the Borel–Cantelli lemma. Then such proofs use the Shannon–McMillan–Breiman theorem to obtain the result. Somehow this procedure is analog to what Khinchin does in the average case by exploiting [3, Theorem 4].

Our proof in the first part shows instead that the probability of the set of sequences of length \( n \) that 1) have a small compressed length, and 2) are typical, is summable in \( n \), and this fact allows us to use the bounds on the probability of each element in the typical set and, consequently, to simplify the proof.

Clearly, we prove a weaker result in a simpler way, but this weaker result still allows us to obtain the desired pointwise inequality. We think that maybe Shannon’s interpretation reported at the end of Section 3.1 directed the other two proofs along the lines of a stronger result, lines that were also followed by Khinchin in the proof of his average result.

4. CONCLUSION

In this paper we derived from the entropy theorem for i.i.d. sources in the case of convergence in probability, i.e., Shannon’s Theorem 3 in [3], and from the more general entropy theorem for stationary and ergodic processes in the case of almost surely convergence, also called the McMillan–Breiman theorem [5–7], respectively, an elementary proof for the Khinchin inequality and a simple proof for the Ornstein–Shields inequality. In particular, we use a deeper classical result in order to prove the latter inequality, which is the almost-sure version of the first one, i.e., an average inequality.

5. IN MEMORIAM OF PROFESSOR ALDO DE LUCA

The third author remembers the discussions and explanations of Aldo de Luca given to him around thirty years ago while walking in Boulevard Saint Michel in Paris. Aldo was very fond
of Khinchin’s formalization effort and, indeed, in his research paper [30] he uses Khinchin’s term “standard sequences” reported in the English translation of Khinchin’s work, instead of the more common “typical sequences.”

Aldo’s voice had a seducing sound, similar to the sound of a father reading a beautiful fairy tale to his sons, or an history of brave knights fighting for honor and mathematical rigor.

ACKNOWLEDGEMENTS

The authors thank Professor D. Perrin for pointing out the reference [2] during a conference in Rome, July 11–12, 2019, in memoriam of Professor de Luca, where they presented a preliminary version of the above results. The authors are also grateful to the referees for their suggestions.

ADDITIONAL INFORMATION

R. Aragona is a member of INdAM-GNSAGA, Italy.

REFERENCES

1. Ornstein, D. and Shields, P.C., Universal Almost Sure Data Compression, *Ann. Probab.*, 1990, vol. 18, no. 2, pp. 441–452.
2. Shields, P.C., *The Ergodic Theory of Discrete Sample Paths*, Providence, R.I.: Amer. Math. Soc., 1996.
3. Shannon, C.E., A Mathematical Theory of Communication, *Bell Syst. Tech. J.*, 1948, vol. 27, no. 3, pp. 379–423.
4. Cover, T.M. and Thomas, J.A., *Elements of Information Theory*, Hoboken, NJ: Wiley, 2006, 2nd ed.
5. McMillan, B., The Basic Theorems of Information Theory, *Ann. Math. Statist.*, 1953, vol. 24, no. 2, pp. 196–219.
6. Breiman, L., The Individual Ergodic Theorem of Information Theory, *Ann. Math. Statist.*, 1957, vol. 28, no. 3, pp. 809–811.
7. Breiman, L., Correction Notes: Correction to “The Individual Ergodic Theorem of Information Theory,” *Ann. Math. Statist.*, 1960, vol. 31, no. 3, pp. 809–810.
8. Chung, K.L., A Note on the Ergodic Theorem of Information Theory, *Ann. Math. Statist.*, 1961, vol. 2, no. 2, pp. 612–614.
9. Chung, K.L., The Ergodic Theorem of Information Theory, *Recent Developments in Information and Decision Processes (Proc. 3rd Sympos. on Information and Decision Processes, Purdue Univ., Lafayette, IN, USA, April 12–13, 1961)*, Machol, R.E. and Gray, P.E., Eds., New York: Macmillan, 1962, pp. 141–148.
10. Blundo, C. and De Prisco, R., New Bounds on the Expected Length of One-to-One Codes, *IEEE Trans. Inform. Theory*, 1996, vol. 42, no. 1, pp. 246–250.
11. Kontoyiannis, I., Second-Order Noiseless Source Coding Theorems, *IEEE Trans. Inform. Theory*, 1997, vol. 43, no. 4, pp. 1339–1341.
12. Kontoyiannis, I. and Verdú, S., Optimal Lossless Data Compression: Non-asymptotics and Asymptotics, *IEEE Trans. Inform. Theory*, 2014, vol. 60, no. 2, pp. 777–795.
13. Khinchin, A.Ya., The Entropy Concept in Probability Theory, *Uspekhi Mat. Nauk*, 1953, vol. 8, no. 3 (55), pp. 3–20.
14. Yeung, R.W., *Information Theory and Network Coding*, Boston: Springer, 2008.
15. Csiszár, I. and Körner, J., *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Cambridge, UK: Cambridge Univ. Press, 2011, 2nd ed.
16. MacKay, D.J.C., *Information Theory, Inference and Learning Algorithms*, New York: Cambridge Univ. Press, 2003.

17. Barron, A.R., Logically Smooth Density Estimation, *PhD Thesis*, Stanford Univ., CA, USA, 1985.

18. Kieffer, J.C., Sample Convereses in Source Coding Theory, *IEEE Trans. Inform. Theory*, 1991, vol. 37, no. 2, pp. 263–268.

19. Algoet, P.H., Log-Optimum Investment, *PhD Thesis*, Stanford Univ., CA, USA, 1985.

20. Khinchin, A.I., *Mathematical Foundations of Information Theory*, New York: Dover, 1957.

21. Kieffer, J.C., A Simple Proof of the Moy–Perez Generalization of the Shannon–McMillan Theorem, *Pacific J. Math.*, 1974, vol. 51, no. 1, pp. 203–206.

22. Gallager, R.G., *Information Theory and Reliable Communication*, New York: Wiley, 1968.

23. Algoet, P.H. and Cover, T.M., A Sandwich Proof of the Shannon–McMillan–Breiman Theorem, *Ann. Probab.*, 1988, vol. 16, no. 2, pp. 899–909.

24. Barron, A.R., The Strong Ergodic Theorem for Densities: Generalized Shannon–McMillan–Breiman Theorem, *Ann. Probab.*, 1985, vol. 13, no. 4, pp. 1292–1303.

25. Bjelaković, I., Krüger, T., Siegmund-Schultze, R., and Szkola, A., The Shannon–McMillan Theorem for Ergodic Quantum Lattice Systems, *Invent. Math.*, 2004, vol. 155, no. 1, pp. 203–222.

26. Longo, G. and Sgarro, A., The Source Coding Theorem Revisited: A Combinatorial Approach, *IEEE Trans. Inform. Theory*, 1979, vol. 25, no. 5, pp. 544–548.

27. Hansel, G., Perrin, D., and Simon, I., Compression and Entropy, *Proc. 9th Annual Sympos. on Theoretical Aspects of Computer Science (STACS’92)*, Cachan, France, Feb. 13–15, 1992, Finkel, A. and Jantzen, M., Eds., Lect. Notes Comp. Sci., vol. 577, Berlin: Springer, 1992, pp. 515–528.

28. Elias, P., Universal Codeword Sets and Representations of the Integers, *IEEE Trans. Inform. Theory*, 1975, vol. 21, no. 2, pp. 194–203.

29. Kraft, L.G., A Device for Quantizing, Grouping, and Coding Amplitude-Modulated Pulses, *PhD Thesis*, MIT, Cambridge, USA, 1949.

30. de Luca, A., On the Entropy of a Formal Language, *Automata Theory and Formal Languages (Proc. 2nd GI Conf., Kaiserslautern, Germany, May 20–23, 1975)*, Brakhage, H., Ed., Lect. Notes Comp. Sci., vol. 33, Berlin: Springer, 1975, pp. 103–109.