# Errata & Comments
*Entropy and Information Theory:*
*Second Edition (2011)*

Robert M. Gray

Spring 2026

Formatted April 14, 2026

## Contents



# 1 Introduction

This document collects errata of *Entropy and Information Theory, Second Edition* (2011) [6] (hereafter abbreviated to *E&IT2*) along with a few technical and historical comments on the statements and the proofs of the results being corrected or clarified.

Some of the errors and typos were inherited from the *First Edition*, but were not caught in time to be fixed in the 2011 publication of *E&IT2*. Errors that were caught in time have been corrected in the final version of the *First Edition, Corrected* and are also included here as corrections to *E&IT2*.

Added citations and references are given local numbers. Equations relating to existing numbered equations in *E&IT2* are given the numbers used in the

book with the exception of a few complete proofs that are included here as revisions or replacements. Notation here follows that of *E&IT2* except for a few cases where proofs are revised for correction or clarity and slight changes in notation add to clarity.

I thank the several readers who have pointed out the errors, suggested corrections, and reported simple typos. These include David Rosenberg, Yevgeny Seldin, John Duchi, Wei Mao, Segismundo Izquierdo, Raul Caram de Assis, Weiying Wang, Jun Muramatsu. and David Neuhoff. Dr Maramatsu in particular motivated my revisiting the *First Edition* and belatedly publishing an updated Errata for *E&IT2* when he provided me with numerous corrections and comments on both *Entropy and Information Theory* and on my earlier book *Probability, Random Processes, and Ergodic Properties* [7].

This spring 2026 update incorporates comments from Dave Neuhoff begun in October 2025 and subsequent email exchanges regarding mutually agreeable edits to repair problems in my proofs of Dobrushin's theorem (Lemma 7.3) and a derivative result (Lemma 7.14) relating maximization over partitions and maximization over product quantizers. Resolution of the problems led to revisiting the early evolution of generalizations of Shannon's concepts of entropy and information and the addition of some historical notes to these errata.

Revisiting the errata has also led to some rewriting and reorganizing.

Easily repaired typos are collected together in the second section.

## 2 Simple Typos

**page xx** Equation (2)

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(X|Y)$$

should read

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

**page 4** Eq. (1.12)

$\int_G ghdP = m(G \bigcap H)$; all $G \in \mathcal{G}$.

should be

$\int_G gdP = m(G \bigcap H)$; all $G \in \mathcal{G}$.

**page 23** Final paragraph.

Line 4:

$(B^{\mathbb{T}}, \mathcal{B}_A{}^{\mathbb{T}})$. should be $(A^{\mathbb{T}}, \mathcal{B}_A{}^{\mathbb{T}})$.

Line 7:

$(A^{\mathbb{T}}, \mathcal{B}_B^{\mathbb{T}})$ should be $(B^{\mathbb{T}}, \mathcal{B}_B^{\mathbb{T}})$

**page 85** In the bottom equation of the first group of equations change $f(X,Y))$ to $f(X,Y)$; that is, remove the extra right paren. The same error occurs twice in the equation at the bottom of the page.

**page 126** Just after "Proof:" following statement of Lemma 5.1, $\|A\|$ should be $\|\mathcal{Q}\|$.

**page 175** Near the middle of the page the statement "simultaneously to $M$ and $G$" should read "simultaneously to $M$ and $P$". Although correcting a typo, the statement containing it does not make sense. The first paragraph of the proof of Lemma 7.3 is replaced in Section 5

**page 205** Lead-up to Lemma 7.14, near top of page: "From Dobrushin's theorem (Lemma 7.3), the supremum can be taken over partitions whose elements are contained in generating field." should be ""From Dobrushin's theorem (Lemma 7.3), the supremum can be taken over partitions whose elements are contained in a generating field. " (missing article)

Lemma 7.14 as published has more serious problems which are dealt with in these errata in Section 6.

**page 240** Lemma 9.2 (10.6.2 in *First Edition*) Replace the first equation

$$D(R,\mu) = \lim_{N\to\infty} D_N(R,\mu) = \inf_N \frac{1}{N} D_N(R,\mu).$$

by

$$D(R,\mu) = \lim_{N\to\infty} \frac{1}{N} D_N(R,\mu) = \inf_N \frac{1}{N} D_N(R,\mu).$$

**page 249** The two citations of Lemma 9.2 below the middle of the page should be to Lemma 9.4.

**page 333** Third line from bottom. $B$ should be $R$.

# 3 Proof of the Entropy Ergodic Theorem

## 3.1 Stationary & Ergodic Sources (Lemma 4.2)

**2023 Notes on proof:** The suggestion to use the Ornstein and Weiss approach for the entropy ergodic theorem for discrete stationary and ergodic sources was made to me by Paul C. Shields during the writing of the original version of the first edition of this book during the late 1980s. The First Edition of the book was published in 1990. My original proof however, had a critical technical error (pointed out in these notes). Some time in the early 2000s Paul informed me that he had noted an error in my proof, but that he knew how to fix it and that we should discuss it. Unfortunately we never did. Paul suffered a brain aneurism in fall 2006 and his mathematics activity diminished steadily after

that. We were in touch by email until 2008, but the error was never brought up.

As a result, my error propagated into the second edition of the book published in 2011. In September 2012 Wei Mao, then a Ph.D. student at Cal Tech, wrote to me regarding a mistake in a counting argument I made in my proof as given in *E&IT2*. During our email exchange, she found that I had omitted an important detail used by Paul in [16] in his 1987 proof of the result for binary sources, and that the addition of two constraints on the construction used in the proof would fix the problem she found with my proof. I had intended in 2013 to correct the *First Edition* and incorporate the corrected version into an *Errata for the Second Edition*. But it did not get done at that time, likely because I retired from Stanford that year and moved twice before settling in Rockport, Massachusetts. I forgot the corrections and Errata until in April 2023 when Dr. Jun Muramatsu of NTT pointed out several typos and mistakes in *E&IT2*. He had earlier reported a collection of suggested corrections in my earlier book, *Probability, Random Processes, and Ergodic Processes* which motivated me to return to correct the online *First Edition* of that book and to update the online *Errata* for the *Probability* book. So I decided to do the same with the *Entropy and Information Theory* Book. Dr. Muramatsu provided a collection of typos and errors for the *Entropy* book as well.

Scouring my notes, correspondence, and email for the *Entropy* book I realized that I had never published the Errata for *E&IT2* as intended in 2013 and I had not updated the *First Edition* to fix the reported mistakes and a few I had found since its 2011 publication.. Hence I have finally in spring 2023 made an effort to update the *First Edition* and to post online the Errata for the Second.

The most important error was in the entropy ergodic theorem, Lemma 3.2.1 in the First Edition, Lemma 4.2 in the second. Many other proofs of the result exist, but the point here was to present a version of the Ornstein-Weiss approach proof of the result consistent with the context of the book as inspired by Paul Shields.

The following proof follows my original notation and construction reasonably closely with some changes made for clarity based on hindsight. I missed two key constraints, which are now incorporated into the proof given here. I have also tried to improve the clarity of the development which involved slight modifications in the notation and the addition of several comments. Revisiting the math after a decade has been a challenge, but it has been fun to rekindle fond memories of Paul Shields.

I am indebted to Dr. Wei Mao for subsequently bringing the problem and the corrections to my attention. I apologize for taking so long to respond and acknowledge her contribution.

---

*Proof:* (of Lemma 4.2 in *Entropy and Information Theory*, Second Edition) Define

$$h_n(x) = -\ln m(X^n)(x) = -\ln m(x^n)$$

and

$$\underline{h}(x) = \liminf_{n\to\infty} \frac{1}{n} h_n(x) = \liminf_{n\to\infty} \frac{-\ln m(x^n)}{n}.$$

Since $m((x_0, \cdots, x_{n-1})) \le m((x_1, \cdots, x_{n-1}))$, we have that

$$h_n(x) \ge h_{n-1}(Tx).$$

Dividing by $n$ and taking the limit infimum of both sides shows that $\underline{h}(x) \ge \underline{h}(Tx)$. Since the $n^{-1}h_n$ are nonnegative and uniformly integrable (Lemma 3.7,we can use Fatou's lemma to deduce that $\underline{h}$ and hence also $\underline{h}T$ are integrable with respect to $m$. Integrating with respect to the stationary measure $m$ yields

$$\int dm(x)\underline{h}(x) = \int dm(x)\underline{h}(Tx)$$

which can only be true if

$$\underline{h}(x) = \underline{h}(Tx); m - \text{a.e.},$$

that is, if $\underline{h}$ is an invariant function with $m$-probability one. If $\underline{h}$ is invariant almost everywhere, however, it must be a constant with probability one since $m$ is ergodic (Lemma 6.7.1 of [7], Lemma 7.12 in *E&IT2*). Since it has a finite integral (bounded by $\bar{H}_m(X)$), $\underline{h}$ must also be finite. Henceforth we consider $\underline{h}$ to be a finite constant.

The Lemma will be proved by demonstrating that the limit supremum of $h_n/n$ equals the limit infimum $\underline{h}$ with probability 1. We proceed with steps that resemble those of the proof of the ergodic theorem in Section 7.2 of [7] and Section 8.1 of *E&IT2*.

Fix $\epsilon > 0$. We also choose for later use a $\delta > 0$ small enough to have the following properties: If $A$ is the alphabet of $X_0$ and $||A||$ is the finite cardinality of the alphabet, then

$$\delta \ln ||A|| < \epsilon, \tag{1}$$

and

$$-\delta \ln \delta - (1-\delta)\ln(1-\delta) \equiv h_2(\delta) < \epsilon. \tag{2}$$

The latter property is possible since $h_2(\delta) \to 0$ as $\delta \to 0$.

Tentatively define the random variable $n(x)$ to be the smallest integer $n \ge 1$ for which $n^{-1}h_n(x) \le \underline{h} + \epsilon$. By definition of the limit infimum there must be infinitely many $n$ for which this is true and hence with probability one $n(x)$ is everywhere finite.

For later use the definition of $n(x)$ is modified to force a minimum value

$$M \ge \frac{\delta}{3};$$

that is, redefine

$$n(x) = \min\{n \ge M : n^{-1}h_n(x) \le \underline{h} + \epsilon\}$$

This modification does not effect the finiteness of $n$.

The random variable $n$ maps single-sided sequences of the form $x = (x_0, x_1, \cdots)$ with $x_i \in A$, a finite alphabet, into a collection of positive integers. Since $n(x)$ is finite with probability 1 and since $\sum_k \Pr(n = k) = 1$, given $\delta$ there must be an $N = N(\delta)$ so large that

$$\Pr(n \geq N) \leq \frac{\delta}{2}.$$

Define a set of "bad" infinite sequences $B = \{x : n(x) \geq N\}$ with indicator function

$$1_B(x) = \begin{cases} 1 & x \in B \\ 0 & \text{otherwise} \end{cases}.$$

The inequality for the bad set $B$ can be stated as

$$m(B) = E_m(1_B) \leq \frac{\delta}{2}.$$

From the definition of $n$, membership of an infinite sequence $x$ in $B$ can be determined from its first $N$ samples $x^N$ since if $n(x)$ does not find an $n \geq M$ for which the inequality $n^{-1}h_n(x^n) \leq \epsilon$ by the time $N$ when it sees all of $x^N = (x_0, \ldots, x_{N-1})$, then it must be true that $n(x) \geq N$ and hence $x \in B$.

Define the set $C$ of $N$-tuples $x^N$ which are prefixes of $x \in B$ so that

$$I_B(x) = 1_C(x^N)$$

$C$ (and $B$) can be characterized by defining a set $S(\ell) \subset A^\ell$ of *good sample entropy* $\ell$-tuples by

$$S(\ell) = \{a^\ell : m(a^l) \geq e^{-\ell(\underline{h}+\epsilon)} \text{ or } -\frac{1}{\ell}\ln m(a^l) \leq \underline{h} + \epsilon\} \tag{3}$$

and observing that

$$I_B(x) = 1_C(x^N) = \begin{cases} 1 & x^\ell \notin S(\ell); \ell = 1, 2, \ldots, N-1 \\ 0 & \text{otherwise} \end{cases}.$$

A random process $\{\ell_n; n \in \mathcal{Z}_+\}$ ($\ell$ for "length") with alphabet the positive integers is defined by applying $n$ to shifts of $x$; that is,

$$\ell_n(x) = n(T^n x) = \ell(x_n, x_{n+1}, \ldots); n = 0, 1, \ldots$$

In particular $\ell_0(x) = n(x)$. The process $\ell_n$ is a sliding-block (stationary) coding of the process $X = \{X_n\}$ described by a stationary and ergodic process distribution $m$ and hence the process $\ell_n$ is also stationary and ergodic.

The process $\ell_n$ provides a means of carving up or parsing an infinite sequence $x$ into consecutive non-overlapping variable length blocks which have good sample entropy; that is, finding a sequence of time indices $n_i; i \in \mathcal{Z}_+$ and a sequence of source sample vectors $x_{n_i}^{\ell_{n_i}}; i = 1, 2, \ldots$. This parsing of the sequence into consecutive contiguous blocks of the source implies a partition of

the time indices $\mathcal{Z}_+$ into a collection of disjoint sets $I_i = \{n_i, \ldots, n_i + \ell_{n_i} - 1\}$ of length $\ell_{n_i}$ having good sample entropy; that is,

$$x_{n_i}^{\ell_{n_i}} \in S(\ell_{n_i})$$

as in (3):

$$m(x_{n_i}^{\ell_{n_i}}) \geq e^{-\ell_{n_i}(\underline{h}+\epsilon)} \text{ or } -\frac{1}{\ell_{n_i}} \ln m(x_{n_i}^{\ell_{n_i}}) \leq \underline{h} + \epsilon.$$

As a simplistic example of the partition of time indices consider

$$\underbrace{0,1,2,3,4}, \underbrace{5,6,7,8,9,10,11,12,13,14}, \underbrace{15,16,17,18,19,20}, \underbrace{21,22,23}\ldots$$

Here the minimum length is $M = 4$ and only the beginning of a possibly infinite length sequence is given. Here, also, the atoms of the partition are adjacent in the sequence. All of the short blocks correspond to good sample entropy blocks and there are no gaps between the blocks. Unfortunately this simple structure is insufficient for the proof of lemma.

The overall goal of proving the entropy ergodic theorem following the Ornstein-Weiss-Sheilds approach is based on a finite version of the above parsing of an infinite sequence and the corresponding partition of the time indices. This can be achieved a block decomposition of an $L$-dimensional sample vector $X^L$ into good sample entropy blocks with block lengths constrained to be neither too large or too small and by inserting gap indices following each good block and indicating when no acceptable good blocks are available at a particular time index.

Given $\delta$ and $N$, choose $L$ so that

$$L \geq \frac{N}{\delta/3} \gg N.$$

A long block $x^L \in A^L$ is parsed into a sequence of non-overlapping relatively short blocks of length no greater than $N$ of the form $x_{n_i}^{\tilde{\ell}_i} = (x_{n_i}, \ldots, x_{n_i+\tilde{\ell}_i-1})$ for which either

$$\tilde{\ell}_i = \ell_{n_i} \leq N, \text{ hence } x_{n_i}^{\tilde{\ell}_i} \in S(\tilde{\ell}) \text{ and } \tilde{\ell} \geq M,$$

or

$$\tilde{\ell}_i = 1, \text{ hence } i \text{ is a gap index and } x_{n_i}^1 \in A.$$

Blocks with $M \leq \ell_i < N$ are called acceptable good sample entropy blocks or simply good blocks (or good $\ell$-blocks). Blocks with $\ell_i = 1$ are called a "gap blocks."

The parsing of $x^N$ induces a partition of the time index set $\mathcal{Z}_L$ into sets

$$\mathcal{Z}_L = \bigcup_i I_i$$

$$I_i = [n_i, n_i + \tilde{\ell}_i - 1].$$

Gap indices occur in three types:

**Gap type 1** $n_i$ is the first time index *following* a good block, that is, $M \leq \tilde{\ell}_{i-1} = \ell_{n_{i-1}} < N$. These blocks ensure that good blocks are separated by at least one gap block. [1]

**Gap type 2** No good block is available at time $n_i$, that is $\ell_{n_i} \geq N$. (by definition $\ell_n \geq M$ for all $n$). Equivalently, $x_{n_i}^N \in C$.

**Gap type 3** $n_i > L - N$; that is, $x_{n_i}^N$ is no longer a sub-vector of $x^L$ so membership $x_{n_i}^N \in C$ can not be tested.

A simplistic example of the partition of time indices for the modified construction is

$$\underbrace{0,1,2,3,4}\ \underbrace{5}\ \underbrace{6,7,8,9,10,11,12}\ \underbrace{13}\ \underbrace{14}\ \underbrace{15,16,17,18,19,20}\ \underbrace{21}\ \underbrace{22}\ \underbrace{23}.$$

In the example, the total blocklength is 23 and the remaining blocks have length 1 (gap blocks) or a length between $M = 4$ and $N = 7$. The non-gap blocks have the good sample entropy property. In addition to the stated constraints, the above picture and the construction show a gap index at the end of each non-gap block. Thus good blocks are always separated by at least on unit length gap index. Gap indices also occur when for a specified initial index no satisfactory length meeting the constraints can be found. Indices at the end of the block are gap indices when there are insufficient indices left to see a full $N$ samples of the end of the $L$-block.

A block decomposition of $x^L$ with the desired properties can be obtained by induction:

**Step 1 Initialize**

$$n_0 = 0$$
$$\tilde{\ell}_0 = \begin{cases} \ell_0 & \text{if } M \leq \ell_0 \leq N \\ 1 & \text{otherwise, } x^N \in C \end{cases}$$

**Step 2 Loop** Given $(n_i, \tilde{\ell}_i)$, find $(n_{i+1}, \tilde{\ell}_{i+1})$. ,

$$n_{i+1} = \begin{cases} n_i + 1 & n_i + \tilde{\ell}_i = n_i + 1 \text{ if } \tilde{\ell}_i = 1 \\ n_i + \tilde{\ell}_i + 1 & \text{otherwise, index } n_i \text{ follows a good block ending at } n_i + \tilde{\ell}_i - 1 \end{cases}$$

If $n_i + 1 > L - N$, go to Step 3. Otherwise

$$\tilde{\ell}_{i+1} = \begin{cases} \ell_{n_{i+1}} & \text{if } M \leq \ell_{n_{I+1}} \leq N \\ 1 & \text{otherwise, } x_{n_{i+1}}^N \in C \end{cases}$$

**Step 3 Finish** For $k = 1 \ldots, L - n_i$ set $n_{i+k} = n_i + k$, $\tilde{\ell}_{i+k} = 1$.

[1] This important constraint was missing from my original proof.

Recall that $\ell_n$ is stationary and ergodic and hence with probability 1 the relative frequency of of $\ell_n \geq N$ will be small.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} 1_B(T^i x) = \frac{1}{n} \sum_{k=0}^{n-1} 1_C(x_k^N) = m(B) \leq \frac{\delta}{2}. \tag{4}$$

Define a set $G_L$ of "good" $L$-tuples

$$G_L = \{x^L : \frac{1}{L-N} \sum_{n=0}^{L-N-1} 1_C(x_n^N) \leq \frac{\delta}{3}\}.$$

$G_L$ is a collection of $L$-tuples which have fewer than $\delta(L-N)/3 \leq \delta L/3$ time indices $n$ for which $x_n^N \in C$; that is, $\ell_n \geq N$. From (4) the sample average must converge to $m(B) \leq \delta/3$ as $L \to \infty$ with probability one and hence also in probability. Thus with probability 1 there is an $L_0 = L_0(x)$ such that

$$\frac{1}{L-N} \sum_{i=0}^{L-N-1} 1_C(x_i^N) \leq \frac{\delta}{3}; \text{ for all } L > L_0(x). \tag{5}$$

This follows simply because if the limit is less than $\delta/2$, there must be an $L_0$ so large that for larger $L$ the time average is at least no greater than $2\delta/2 = \delta$. We can restate (5) as follows: with probability 1 $x^L \in G_L$ for all but a finite number of $L$. Stating this in negative fashion, we have one of the key properties required by the proof: If $x^L \in G_L$ for all but a finite number of $L$, then $x^L$ cannot be in the complement $G_L^c$ infinitely often, that is,

$$m(x : x^L \in G_L^c \text{ i.o.}) = 0 \tag{6}$$

## Counting

The next step is to count the number $\|G_L\|$ of $L$-tuples in $G_L$, which will allow a specification of how large $L$ or how small $\delta$ must be chosen to complete the proof. This involves counting the number of possible gap indices and the number of good (acceptable sample entropy) vectors whose location in time and length are determined by the type 1 gap indices.

For an $x^L \in G_L$ there can be no more than $L/M$ good blocks in the block decomposition and hence no more than $L/M$ type 1 gap indices. The choice of $M \geq 3/\delta$ ensures that the number of type 1 gap indices is no greater than $L\delta/3$.

By construction, there can be no more than $L\delta/3$ type two gap indices.

There can be no more than $N$ type 3 gap indices. The choice of $L \geq 3N/\delta$ bounds above the number of type 3 indices by $L\delta/3$.

Thus the the number of gap indices is bound above by $L\delta$. These $L\delta$ indices can occur in any of at most

$$\sum_{k \leq \delta L} \binom{L}{k} \leq e^{Lh_2(\delta)} \tag{7}$$

where we have used Lemma 3.6Eq. (7) provides an upper bound on the number of ways that a sequence in $G_L$ can be parsed by the given rules.

Each pattern specifies the type two indices which in turn specify the location of the good blocks $x_{n_i}^{\ell_i} \in S(\ell_i)$ for which

$$m(x_{n_i}^{\ell_i}) \geq e^{-\ell_i(\underline{h}+\epsilon)}.$$

Given $\ell_i$ probabilities sum to one:

$$1 = \sum_{a^{\ell_i} \in A^{\ell_i}} m(a^{l_i}) \geq \sum_{a^{\ell_i} \in S(\ell_i)} m(a^{l_i}) \geq \|S(\ell_i)\| e^{-\ell_i(\underline{h}+\epsilon)}$$

whence

$$\|S(\ell_i)\| \leq e^{\ell_i(\underline{h}+\epsilon)}.$$

Each of the fewer than $e^{Lh_2(\delta)}$ patterns has no more than

$$\prod_i \|S(\ell_i)\| \leq e^{\sum_i \ell_i(\underline{h}+\epsilon)} \leq e^{L(\underline{h}+\epsilon)}$$

possible patterns of good blocks.

Combining the counts for the number of patterns of gap indices and the number of possibilities for gap indices and good blocks yields

$$||G_L|| \leq e^{h_2(\delta)L} ||A||^{\delta L} e^{L(\underline{h}+\epsilon)} = e^{L(h_2(\delta)+\delta \ln ||A||+\underline{h}+\epsilon)}$$

Since $\delta$ satisfies (1)–(2),

$$||G_L|| \leq e^{L(\underline{h}+3\epsilon)}. \tag{8}$$

This bound provides the second key result in the proof of the lemma. We now combine (8) and (6) to complete the proof.

Let $B_L$ denote a collection of $L$-tuples that are bad in the sense of having too large a sample entropy or, equivalently, too small a probability; that is if $x^L \in B_L$, then

$$m(x^L) \leq e^{-L(\underline{h}+5\epsilon)}$$

or, equivalently, for any $x$ with prefix $x^L$

$$h_L(x) \geq \underline{h} + 5\epsilon.$$

The upper bound on $||G_L||$ provides a bound on the probability of $B_L \bigcap G_L$:

$$m(B_L \bigcap G_L) = \sum_{x^L \in B_L \bigcap G_L} m(x^L) \leq \sum_{x^L \in G_L} e^{-L(\underline{h}+5\epsilon)}$$

$$\leq ||G_L|| e^{-L(\underline{h}+5\epsilon)} \leq e^{-\epsilon L}.$$

Recall now that the above bound is true for a fixed $\epsilon > 0$ and for all $L \geq L_1$. Thus

$$\sum_{L=1}^{\infty} m(B_L \bigcap G_L) = \sum_{L=1}^{L_1-1} m(B_L \bigcap G_L) + \sum_{L=L_1}^{\infty} m(B_L \bigcap G_L)$$

$$\leq L_1 + \sum_{L=L_1}^{\infty} e^{-\epsilon L} < \infty$$

and hence from the Borel-Cantelli lemma (Lemma 4.6.3 of [7]) $m(x : x^L \in B_L \bigcap G_L \text{ i.o.}) = 0$. We also have from (6), however, that $m(x : x^L \in G_L^c \text{ i.o. }) = 0$ and hence $x^L \in G_L$ for all but a finite number of $L$. Thus $x^L \in B_L$ i.o. if and only if $x^L \in B_L \bigcap G_L$ i.o. As this latter event has zero probability, we have shown that $m(x : x^L \in B_L \text{ i.o.}) = 0$ and hence

$$\limsup_{L\to\infty} h_L(x) \leq \underline{h} + 5\epsilon.$$

Since $\epsilon$ is arbitrary we have proved that the limit supremum of the sample entropy $-n^{-1} \ln m(X^n)$ is less than or equal to the limit infimum and therefore that the limit exists and hence with $m$-probability 1

$$\lim_{n\to\infty} \frac{-\ln m(X^n)}{n} = \underline{h}. \tag{9}$$

Since the terms on the left in (9) are uniformly integrable from Lemma 3.7 we can integrate to the limit and apply Lemma 3.8 to find that

$$\underline{h} = \lim_{n\to\infty} \int dm(x) \frac{-\ln m(X^n(x))}{n} = \bar{H}_m(X),$$

which completes the proof of the lemma □

---

## 3.2 Stationary Sources (Lemma 4.4))

Lemma 4.4 (Lemma 3.3.2 in the *First Edition*) has an error in its proof on p. 108. Insert the following text following the line "where $P_\psi$ is the distribution of $\psi$ (which follows the third displayed equation):

It was pointed out by Weiying Wang in 2017 that the evaluation above applies the ergodic decomposition of Theorem 1.5 (Theorem 1.8.3 in the *First Edition*) which requires that $m(X^n)/m_\psi(X^n)$ have a finite integral (be in $L^1(m)$), but this was not shown. The following paragraph fills in the details and shows that $f \equiv m(X^n)/m_\psi(X^n) \in L^1(m)$ and bounds the integral independently of $n$.

---

Define for all $M > 0$ the non-negative bounded function $f_M$ by

$$f_M = \max(\frac{m(X^n)}{m_\psi(X^n)}, M)$$

or, pointwise

$$f_M(x) = \max(\frac{m(X^n(x))}{m_{\psi(x)}(X^n(x))}, M) = \max(\frac{m(x^n)}{m_{\psi(x)}(x^n)}, M)$$

The truncated functions $f_M$ converge monotonically to $f$ as $M \to \infty$. Since $f_M$ is a nonnegative integrable function it is in $L^1(m)$ and hence the ergodic decomposition of Theorem 1.6. (or iterated expectation by identifying expectation over $m_\psi$ as a conditional expectation given $\psi$) can be applied to obtain

$$E_m f_M = E[E[f_M|\psi]].$$

The conditional expectation given $\psi = \lambda$ can be bounded as

$$\begin{aligned} E[f_M|\psi=\lambda] &= \int dm_\lambda(x) \max\left(\frac{m(X^n(x))}{m_\lambda X^n(x))}, M\right) \\ &= \sum_{a^n} m_\lambda(a^n) \max\left(\frac{m(a^n)}{m_\lambda(a^n)}, M\right) \end{aligned}$$

where the sums are over all possible $a^n \in A^n$, the $n$-tuple source alphabet. As noted, with $P_\psi$ probability 1, $m_\lambda(a^n)$ cannot be 0 unless $m(a^n)$ is, in which case the ratio is taken to be 0. Defining the set $F_n = \{a^n : m(a^n)/m_\lambda(a^n) \le M\}$

$$\begin{aligned} E[f_M|\psi=\lambda] &= \sum_{a^n \in F_n} m_\lambda(a^n) \max\left(\frac{m(a^n)}{m_\lambda(a^n)}, M\right) + \sum_{a^n \notin F_n} m_\lambda(a^n) \max\left(\frac{m(a^n)}{m_\lambda(a^n)}, M\right) \\ &\le \sum_{a^n \in F_n} m_\lambda(a^n) \frac{m(a^n)}{m_\lambda(a^n)} + \sum_{a^n \notin F_n} m_\lambda(a^n) M \\ &= m(F_n) + M m_\lambda(F_n^c) \end{aligned}$$

For $a^n \notin F_n$, however, $m_\lambda(a^n) \le m(a^n)/M$, whence

$$m_\lambda(F_n^c) = \sum_{a^n \notin F_n} m_\lambda(a^n) \le \sum_{a^n \notin F_n} \frac{1}{N} m(a^n) = \frac{m(F_n^c)}{M}$$

so that

$$E[f_M|\psi=\lambda] \le m(F_n) + m(F_n^c) = 1$$

for all $n$ and $T$. Thus from the dominated convergence theorem, the monotone nondecreasing integrable function $f_T$ has expectations which converge to a limit which equals the expectation of the limit of $f_T$ as $T$ goes to infinity. Thus $f \in L^1(m)$ as required and its integral is bound above by 1,

---

Continue the proof of Lemma 4.4 as in the book (removing the extra "Thus" following the replaced material).

# 4 Historical Detour: Entropy, Mutual Information & Relative Entropy

(Added April 2026) This section serves as a prologue to the corrections and discussion of the next two sections which are devoted to fixing problems in Chapter 7 "Relative Entropy" of *E&IT2*. The focus in this section is on historical context along with some hindsight prior to dealing with the details of the statements and proofs of two interdependent results: Lemmas 7.3 (a variation of Pinsker's Theorem 2.4.1 [14], which he calls "Dobrushin's theorem"), and 7.14 (a variation of Pinsker's Theorem 2.1.1, which he calls "a theorem of R.L. Dobrushin" ).

These two lemmas describe and relate two different generalizations of Shannon's *entropy* of a single discrete random variable and his *rate of transmission of information* between two discrete random variables, which is now commonly referred to as *mutual information*.

## 4.1 Shannon Entropy & Information

Claude E. Shannon in his original 1948 publication of "A Mathematical Theory of Communication" in the *Bell Systems Technical Journal* [15] proposed several probabilistic definitions of information for random variables or vectors which depended only on their probability distribution as described by probability mass functions (pmfs) on discrete spaces or probability density functions (pdfs) on the real line or Euclidean space. Two of these definitions are of primary importance here: *entropy* (as named by Shannon with acknowledgement to Hartley) and *average mutual information*, which Shannon called "rate of transmission" but soon after became generally known as average mutual information, the terminology used by Shannon's MIT colleague Robert M. Fano in his notes for his 1952-53 MIT course 6.574: *Transmission of Information*, likely the first formal class on Information Theory. These notes were an early draft for Fano's book *Transmission of Information: A Statistical Theory of Communications*. [3] The name is commonly abbreviated to simply *mutual information*, as in *Wikipedia*.

**Discrete Random Variables**

**Entropy:** For a single discrete random variable $X$ Shannon's entropy $H$ was defined in terms of the probability mass function (pmf) of a random variable $X$, $p_X(x) = \Pr(X = x); x \in A_X = \{a_i; i = 1, \ldots, N_X\}$ by

$$H(X) \triangleq \sum_{x \in A_X} p_X(x) \ln \frac{1}{p_X(x)} = \sum_{i=1}^{N_X} p_X(a_i) \ln \frac{1}{p_X(a_i)}$$

where the two expressions emphasize that the entropy does not depend upon the values the random variable $X$ takes on, only on their probabilities. Entropy is often written as $H(p_X)$ to emphasize this fact. The indeterminant expression $0 \ln 0$ is defined to be 0.

Shannon developed properties and applications for this measurement of the quantity of information (or uncertainty) in a discrete-valued random variable, but here the focus is on the definition itself and generalizations which preserve its properties such as non-negativity, scale invariance, and usefulness for solving mathematical problems.

Two jointly distributed random variables $X$ and $Y$ described by a joint pmf $p_{XY}(x,y) = \Pr(X = x, Y = y)$ $x \in A_X, y \in A_Y$ can be considered as a single 2-dimensional random vector and the joint pmf implies the marginal pmfs $p_X$ and $p_Y$, so one can use use the basic Shannon formula to define the joint entropy

$$H(X,Y) = \sum_{(x,y)\in A_X \times A_Y} p_{XY}(x,y) \ln \frac{1}{p_{XY}(x,y)}$$

or $H(p_{XY})$ as well as the second individual entropy $H(Y)$ or $H(p_Y)$. Shannon also defined various conditional entropies using conditional probabilities, but these are not needed here

**Mutual Information**

Following the definition and development of entropy, Shannon introduced a different notion of information $R$ he called a measure of the "rate of transmission of information." Shannon's initial definition was in terms of conditional entropies between two random variables, but he included in his development the identity

$$R = H(X) + H(Y) - H(X,Y)$$

which best fits the current discussion as a definition of the Shannon mutual information $I(X,Y)$ or $I(p_{XY})$ for discrete random variables:

$$I(X,Y) = I(p_{XY}) \stackrel{\Delta}{=} \sum_{x\in A_X, y\in A_Y} p_{XY}(x,y) \ln \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \tag{10}$$

**Continuous Random Variables**

Shannon considered extensions of his definitions to real-valued random variables and vectors. He accomplished this by analogy with the discrete case rather than by derivation from the discrete case. He replaced probability mass functions by probability density functions (pdfs) in Riemann integral analogs of the sums involving pmfs.

The continuous analog worked for the case of mutual information, where Shannon's definition for two real-valued random variables $X, Y$ with joint pdf $f_{XY}(x,y)$ was defined by the integral of pdfs mimicking the sum involving pmfs of Shannon's definition:

$$I(X,Y) = I(f_{XY}) \stackrel{\Delta}{=} \int\int f_{XY}(x,y) \ln \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)} dxdy. \tag{11}$$

Shannon's integral definition with pdfs preserved the properties of the discrete case and played a similar role in developing theorems in in communications and coding theory.

The approach of a continuous imitation of the sum did not work well for entropy as important intuitive properties of information such as non-negativity and scale invariance failed, but the continuous analog of Shannon entropy did have its uses and eventually became known as "differential entropy" to avoid confusion with Shannon entropy. Differential entropy was not a generalization or extension of Shannon's discrete entropy to the continuous case.

Many generalizations, extensions, and variations of Shannon's entropy and mutual information were developed following publication of his 1948 paper. [15] Two of these play the most important role in this document: The generalization of mutual information $I(X, Y)$ and the implied definition of entropy as a special case $H(X) = I(X, X)$ of Kolmogorov and his students and colleagues, and the generalization of Kullback and Leibler [11] of entropy to a relative entropy, the entropy of one probability distribution with respect to another. Both approaches used measure theory, incorporating the general Lebesgue integral and Kolmogorov's axiomatic development of probability to extend Shannon's definitions to general abstract probability spaces providing rigorous tools for probabilities and integrals. Kolmogorov's goal was twofold: rigorous generalizations and variations of Shannon's results for communications involving discrete or real-valued models and the application of Shannon entropy to the solution of long-standing problems in Ergodic Theory. Kullback's focus was on mathematical statistics. Both approaches began with Shannon.

## 4.2 Early Generalizations of Shannon: Kolmogorov & Kullback

Generalizations of Shannon definitions began to appear in the early 1950s. Both Kolmogorov et al. and Kullback et al. extended Shannon's definitions to the general abstract probability spaces of measure theory, but they did so in different ways. Kolmogorov began with a focus on the underlying probability distributions, deriving general definitions from discretized approximations for which the Shannon discrete models held — the general definitions were maximizations or limits over quantized versions of the abstract probability spaces and Shannon's definitions for the discrete case. From their definitions they showed that in general mutual information could be expressed in a general form resembling Shannon's continuous result — as an integral involving probability density functions, but now the integral was Lebesgue and the density functions were Radon-Nikodym derivatives of probability distributions. The abstract versions yielded Shannon's formulas in his special cases and resembled them in the general case. Many of the basic results first appeared together in print in a 1959 article by Roland Dobrushin [1].

Kullback and Leibler (1951) [11] on the other hand, began with a general measure theoretic definition of a variation of Shannon entropy in terms of Lebesgue integrals and Radon-Nikodym derivatives of probability distributions.

In his classic 1959 book *Information and Statistics* [10] Kullback showed that the Kullback-Leibler information of one distribution with respect to another — known by many names including Kullback-Leibler information, discrimination information, Kullback-Leibler divergence, and relative entropy — includes Shannon mutual information as a special case.

Mark Pinsker effectively merged the two approaches to extending Shannon's notions of information in his book *Information and information stability of random variables and processes* published in 1960 (in Russian). [14] The English translation by Amiel Feinstein was published with corrections by Pinsker and comments and added proofs by Feinstein in 1964.

**Andrei N. Kolmogorov** with his students Roland L. Dobrushin and Mark S. Pinsker and colleagues Israel M. Gelfand and Akiva M. Yaglom developed a measure theoretic definition for the "information" $I(X,Y)$ of two random variables $X$ and $Y$ [1], also named " the average amount of information about the variable $X$ conveyed by specifying the value of the variable $Y$" [5] and "the information of one of these variables with respect to the other." [14]

The development of the general definition went through several stages before culminating in the concise description in Dobrushin's 1959 paper *A general formulation of the fundamental Shannon theorem in information theory* [1], but the common goal was to directly generalize Shannon's discrete mutual information by finitizing the general case so that Shannon's discrete formulas held, and then maximizing the mutual information of the finite versions by taking the the supremum over a collection. The mathematical foundation used was that of probability spaces, a concept for which Kolmogorov had a foundational role.

Mutual information between two random objects $X$ and $Y$ can be described by a joint probability space $(A_{XY}, \mathcal{B}_{XY}, P_{XY})$ where $(A_{XY}, \mathcal{B}_{XY})$ is a *measurable space* consisting of an *alphabet* or *sample space* $A_{XY} = A_X \times A_Y$ (the Cartesian product of $A_X$ and $A_Y$) of all the possible values the random vector $(X,Y)$ can assume, and $\mathcal{B}_{XY}$ is a $\sigma$-field (or $\sigma$-algebra or *event space*) of subsets of $A_{XY}$; and $P_{XY}$ is a probability measure defined on the measurable space.

The joint probability space implies two separate probability spaces for the individual random variables: $(A_X, \mathcal{B}_X, P_X)$ and $(A_Y, \mathcal{B}_Y, P_Y)$. Conversely, the two measurable spaces $(A_X, \mathcal{B}_X)$ and $(A_Y, \mathcal{B}_Y)$ imply the product measurable space $(A_X \times A_Y, \mathcal{B}_{XY})$, where $\mathcal{B}_{XY} = \sigma(RECT_{XY})$ — the $\sigma$-field generated by the *rectangles*, that is by the collection $RECT_{XY}$ of all sets of the form $F \times G$; $F \in \mathcal{B}_X, G \in \mathcal{B}_Y$. This is the mathematical environment in Kolmogorov, Dobrushin, Pinsker, and *E&IT2*.

Given two marginal probability spaces $(A_X, \mathcal{B}_X, P_X)$ and $(A_Y, \mathcal{B}_Y, P_Y)$, a simple but important joint probability space consistent with the marginal spaces is $(A_X \times A_Y, \mathcal{B}_{XY}, P_X \times P_Y)$ where $P_X \times P_Y$ is the product distribution specified by its values on the rectangles:

$$P_X \times P_Y(F \times G) = P_X(F)P_Y(G); F \in \mathcal{B}_X, G \in \mathcal{B}_Y.$$

A basic tool in the development is a *partition* of a measurable space $(\Omega, \mathcal{B})$, a

finite collection $\mathcal{Q} = \{Q_i; i = 1, \ldots, \|\mathcal{Q}\|\}$ of disjoint (mutually exclusive) events $Q_i \in \mathcal{B}$ called *atoms* for which $\bigcup_i Q_i = \Omega$, which can be abbreviated to $\mathcal{Q} \subset \mathcal{B}$. Dobrushin [1] calls partitions *dissections*.

The general definition of mutual information $I(X, Y)$ as given in Dobrushin [1] and Pinsker [14] is:

$$I(X, Y) \stackrel{\Delta}{=} \sup_{\mathcal{Q} \subset \mathcal{B}_X, \mathcal{R} \subset \mathcal{B}_Y} \sum_{i,j} P_{XY}(Q_i \times R_j) \ln \frac{P_{XY}(Q_i \times R_j)}{P_X(Q_i) P_Y(R_j)} \tag{12}$$

where the supremum is over all (finite, measurable) partitions $\mathcal{Q} = \{Q_i\} \subset \mathcal{B}_X$ and $\mathcal{R} = \{R_j\} \subset \mathcal{B}_Y$.

Dobrushin called $I(X, Y)$ the "information of these random variables" and Pinsker called it "the information of one of these variables with respect to the other." It is useful to add a definition for the mutual information between two partitions to abbreviate the overall definition:

$$\begin{aligned} I(X, Y) &= \sup_{\mathcal{Q}, \mathcal{R}} I_{XY}(\mathcal{Q}, \mathcal{R}), \text{ where} \\ I_{XY}(\mathcal{Q}, \mathcal{R}) &= \sum_{i,j} P_{XY}(Q_i \times R_j) \ln \frac{P_{XY}(Q_i \times R_j)}{P_X(Q_i) P_Y(R_j)}. \end{aligned} \tag{13}$$

Although this general definition was not published in its final form until 1959, Gelfand and Yaglom [5] credit a presentation by Kolmogorov at the All-Union Conference on Probability Theory and its Applications in May 1955 as inspiring the work and state that some of his earlier results were incorporated into their joint paper. [4] That paper, however, did not explicitly deal with the complete notion of probability spaces, it dealt instead with a related idea which also interested Kolmogorov — Boolean algebras and sub-algebras instead of $\sigma$-fields and generating fields.

During the late 1950s the approach led to results showing that the general definition of mutual information inherited all of the properties of Shannon's discrete case and it also yielded a general integral representation in terms of densities consistent with Shannon's continuous mutual information, but which required significantly more machinery of measure and integration theory — general Lebesgue integrals with respect to general probability measures and probability density functions defined as Radon-Nikodym derivatives of probability distributions. These general integral/density properties are treated in *E&IT2* but not in these notes.

The generalization of mutual information by Kolmogorov et al. also provided a rigorous extension of Shannon entropy to continuous and abstract alphabets as a special case of mutual information: $H(X) = I(X, X)$, a formula which also holds in Shannon's original discrete case definitions in terms of pmfs. This generalization of Shannon entropy came to be called the "Kolmogorov-Sinai Invariant" in Ergodic Theory and led to significant advances in solving problems regarding isomorphisms in dynamical systems.

Shannon began with entropy and then used it to define mutual information, but Kolmogorov et al. reversed the process, making mutual information the basic quantity of information and using it to define entropy.

The Kolmogorov et al. definition of mutual information is given in the language of probability theory involving $\sigma$-fields and partitions, but it has a simple and intuitive equivalent statement in the language of quantization. A *quantizer* is simply a mapping of a general input space into a finite collection, which might be a collection of reproduction symbols for data compression, of integer indices indicating a classification of the input, or any finite collection of abstract items. Mathematically, it is a measurable mapping of an abstract measurable space into a set with a finite number of elements.

For example, for a general random variable $X$ defined on a measurable space $(A_X, \mathcal{B}_X)$, a quantizer $q$ is described by a (measurable) partition $\mathcal{Q} = \{Q_i; i = 1, \ldots, N\}$ and a collection of distinct labels $\{q_i; i = 1, \ldots, N\}$ and the mapping

$$q(x) = q_i \text{ if } x \in Q_i; i = 1, \ldots, N$$

The $q_i$ are required to be distinct so their inverse images are disjoint and not empty. Given a partition $\mathcal{Q}$ a useful choice of the labels is $q_i = i$; that is, the index of the atom $Q_i$ of the partition containing $x$. Another common choice arises in data compression or Shannon's source coding with a fidelity criterion: Set $q_i$ to be the generalized centroid of the atom $Q_i$.

Given any partition $\mathcal{Q}$ there is a quantizer $q$ having that partition, e.g., $q_i = i$. Conversely, by definition any quantizer $q$ has an associated partition $\mathcal{Q} = \{Q_i = q^{-1}(q_i)\}$ with atoms the inverse images under $q$ of the finite collection of quantizer outputs $q_i$.

Similarly, a quantizer $r$ for $Y$ is implied by any partition of $A_Y$ and vice versa.

Given any partitions $\mathcal{Q} \subset \mathcal{B}_X$, $\mathcal{R} \subset \mathcal{B}_Y$ there exist quantizers $q$ and $r$ for which

$$I_{XY}(\mathcal{Q}, \mathcal{R}) = \sum_{i,j} P_{XY}(Q_i \times R_j) \ln \frac{P_{XY}(Q_i \times R_j)}{P_X(Q_i)P_Y(R_j)} = I(q(X), r(Y)),$$

where the rightmost term is Shannon's mutual information between the discrete random variables $q(X)$ and $r(Y)$. Furthermore any pair of quantizers $q$ and $r$ are defined by measurable partitions, and their mutual information does not depend on the actual values of $q_i$ or $r_i$, only on the probabilities of the partition atoms.

Hence

$$I(X, Y) = \sup_{\mathcal{Q},\mathcal{R}} I_{XY}(\mathcal{Q}, \mathcal{R}) = \sup_{q,r} I(q(X), r(Y)). \tag{14}$$

This is not a lemma or theorem since it is it is merely an observation of the fact that any quantizer implies a partition and any partition has a corresponding partition with labels or indices that do not influence mutual information, only the probabilities matter. The interpretation of the quantization view is

that the mutual information between two abstract random variables is simply the supremum of the achievable mutual information over all possible quantized approximations, where Shannon's definition for discrete random variables holds with all of its properties and applications. This property yields useful asymptotic results and intuition based on simple discrete cases.

---

The second approach to the extension of Shannon's entropy also came from mathematicians using measure theory to extend Shannon's entropy, but instead of developing it for generalizing Shannon's coding theorems, the goal was applications to statistics.

**Solomon Kullback** and Richard Leibler published "On information and sufficiency" in 1951. [11] The paper begins its introduction with "This note generalizes to the abstract case Shannon's definition of information" and cites both Shannon's BSTJ paper [15] and the Shannon and Weaver book including it. By "abstract case" is meant in terms of a probability space, the same general approach used by Kolmogorov et al. Unlike the Kolmogorov approach which begins with an emphasis on the simple finite discrete case, Kullback and Leibler provide a definition in terms of general integrals and density functions defined by Radon-Nikodym derivatives of probability distributions. In their introduction they state that

> We are also concerned with the statistical problem of discrimination by considering a measure of the"distance" or"divergence" between statistical populations in terms of our measure of information.

Rather than introducing the necessary mathematical machinery to describe the Kullback-Leibler information in terms of general integrals and densities as they do, it is simpler to begin with Shannon's environments of discrete and real-valued continuous examples. Kullback and Leibler's measure of information reduces to Shannon-like formulas for discrete random variables with common alphabets. Given two pmfs $p = \{p_i; 1 = 1, \ldots, N\}$ and $m = \{m_i; 1 = 1, \ldots, N\}$ using the notation of *E&IT2* the Kullback-Leibler *divergence* (one of its many names)

$$D(p\|m) = \sum_i p_i \ln \frac{p_i}{m_i}$$

and for two pdfs $f$ and $g$ on the real line (or more generally finite-dimensional Euclidean space) the abstract definition reduces to

$$D(f\|g) = \int dx f(x) \ln \frac{f(x)}{g(x)}.$$

A technicality that is always assumed is that in the discrete case pmfs are only allowed for which whenever $m_i = 0$, then also $p_i = 0$ (in which case the indeterminant $0/0$ is taken as 1 and hence its logarithm is 0). The analogous

condition is required in the continuous case and pdfs. Both are are special cases of the probability distribution in the numerator and being averaged over must be *absolutely continuous* with respect the the other probability distribution.

The discrete formula relates directly to Shannon entropy. If $m$ is uniform over an alphabet of size $N$, then

$$D(p\|m) = \ln N - H(p).$$

In Shannon's continuous case with a single random variable and two pdfs, then there is no natural comparison of Kullback-Leibler information with entropy, but it does relate to *differential* entropy if the pdf $g$ is a constant $c$ over a region of finite volume $V$, then

$$D(f\|g) = \ln V - h(f),$$

but this is not or much interest since differential entropy plays a minor role in information theory in comparison to entropy and mutual information.

Of much more interest in the continuous case is an example not in Kullback and Leibler's article [11], but which is explicitly contained in Kullback's later classic 1959 book *Information and Statistics* [10] as Example 4.3. Kullback observes that if in the case of real-valued random variables $X$,$Y$ with joint pdf $f(x,y) = f_{XY}(x,y)$ and reference measure the product density with the same marginals $g(x,y) = f_X(x)f_Y(y)$, then the Kullback-Leibler information formula becomes using (11)

$$\begin{aligned} D(f_{XY}\|f_X f_Y) &= \int dxdy f_{XY}(x,y) \ln \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)} \\ &= I(X,Y) \end{aligned} \tag{15}$$

matching Shannon's definition of mutual information for the same case. It is also immediately true in the discrete case using (10):

$$\begin{aligned} D(p_{XY}\|p_X p_Y) &= \sum_{x,y} p_{XY}(x,y) \ln \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \\ &= I(X,Y) \end{aligned}$$

and hence in both the cases considered Shannon of discrete and continuous real valued random variables mutual information is a special case of divergence.

Kullback cited Shannon along with Kolmogorov's 1956 *IRE Transactions Transactions on Information Theory* article "On the Shannon theory of information in the case of continuous signals" [8], which uses definitions in terms of integrals involving densities (not suprema over quantizers or partitions) and the Gelfand, Kolmogorov, and Yaglom 1956 paper [4] which used a variation of the maximization over discretized models approach and developed the integral/density formulas. [9]

It is in fact true in the general case, as is stated in Pinsker and is proved in the next section with corrections and a clarification of Lemma 7.3 in Section 7.4 of *E&IT2*. The result is obvious using the generalized definitions of mutual information and relative entropy given in terms of Lebesgue integrals and Radon-Nikodym derivatives, which which takes the form of Shannon's Eq. (15) with the more general versions of integration and densities.

The Kullback-Leibler divergence was extensively developed by Kullback for statistical applications described in his book using his original definition in terms of Lebesgue integrals over probability spaces of general probability densities defined as Radon-Nikodym derivatives of probability distributions. Divergence also has a general Kolmogorov-style definition as a supremum over partitions which is considered by Pinkser (1960) [14] but not by Dobrushin (1959) [1]. The timing of the inclusion of relative entropy into the Kolmogorov school aligned with the 1959 publication of Kullback's book, the same year as Dobrushin's article and a year before Pinsker's book was published in Russian.

In 1958 Amiel Feinstein published his book *Foundations of Information Theory*, which appears to be the first book treatment of Shannon's work using measure theory and Lebesgue integration to deal with continuous random variables. Kullback's book was then not yet published and Feinstein did not cite the earlier 1951 Kullback and Leibler article. Feinstein did not formally develop the idea of mutual information in his book, but he like Kullback [10] cited Kolmogorov's "On the Shannon theory of information in the case of continuous signals" [8] and Gelfand, Kolmogorov, and Yaglom. [4] Feinstein also cites Gelfand and Yaglom (1956) [5] while Kullback does not.

One might infer that when Kullback's book appeared in 1959, Kullback was aware of the Kolmogorov et al. work on generalizing Shannon's mutual information using the Lebesgue integral/ Radon-Nikodym derivative definitions, but not with the fact that they did not *define* mutual information in the general case via abstract integration and probability densities — the Kolmogorov definitions were based on maximizing discretized versions of the problem which involved only the simple Shannon sums and probability mass functions. Kolmogorov et al. *derived* the complicated generalized integral formulas from the more basic definition. On the other hand, Pinsker in his 1960 book provided a development of the extensions Shannon's mutual information by himself, his doctoral supervisor Kolmogorov, his fellow doctoral student Dobrushin, and colleagues of Kolmogorov including Gelfond and Yaglom and then proceeded to incorporate Kullback's divergence as a relative entropy and show it to be a generalization of mutual information and its special case of Shannon entropy. This seems a remarkably rapid development in mathematical information theory in such a short time. Mathematical statistics was a field of strong interest to Pinsker, so it seems likely he was one of the first and possibly the first in the Kolmogorov circle to appreciate Kullback's contributions and spread them within the information theory community as a generalization of Shannon mutual information. It was fortunate that his translator, Ameil Feinstein, was proficient in both Russian and measure theory as well as information theory. Feinstein did not mention

Kullback in his 1958 book *Foundations of Information Theory*, which appears to be the first book treatment in English of Shannon's work which used measure theory and Lebesgue integration to deal with continuous random variables, so he might have learned about Kullback divergence/relative entropy either from Kullback's 1959 book or from Pinsker's 1960 book citing Kullback and proposing Kullback's information measure as a useful generalization of Shannon's mutual Information.

Regardless of these conjectures, Feinstein chose to first prove the more general theorem for relative entropy rather than follow the historical development and first prove the earlier theorem for mutual information.

## 4.3 Relative Entropy

It is not known who in the Kolmogorov school first noticed the Kullback-Leibler information, that it reduced to Shannon's mutual information as a special case in Shannon's simple examples, and that it could be analyzed with the same derivation from the simple finite models that had proven successful for mutual information. It is plausible that it was Mark Pinsker since the first published treatment of these topics together with a citation of Kullback was Pinsker's 1960 book (in Russian) as translated by Feinstein (1964) [14]. Pinsker's book followed and credited Dobrushin's methods, but Dobrushin made no reference to Kullback's work while Pinsker cited both Dobrushin's 1959 paper and Kullback's 1959 book and integrated them into the methods and derivations of his own work. None of Pinsker's citations that I have been able to track down cite Kullback and Leibler's earlier 1951 work, only the 1959 book.

Chapter 2 on "Information" in Pinsker's book begins in Section 2.1 with the general definition of mutual information using the approach of the Kolmogorov school of maximizing mutual information over finite partitions, then developed its properties in Section 2.3 and derived general general representations for mutual information in terms of general integrals involving general densities in Section 2.3.

In Section 2.4 Pinsker begins with the general definition for entropy as a special case of mutual information as $H(X) = I(X, X)$ and then introduces a generalization of Shannon entropy that includes mutual information as a special case. The generalization is the Kullback-Leibler divergence defined in the same manner as the Kolmogorov school initially defined mutual information — as a supremum over partitions. Relative entropy is introduced not as a generalization of Shannon entropy, but rather as a variation of Shannon entropy which turns out to be a generalization or mutual information.

The first of the cited references in Pinsker for Section 2.4 is Kullback's 1959 book [10], but Pinsker also includes earlier references to Perez (1957) [13] and (1959) [12] and Rosenblatt-Rot (1959). Like Kullback and Leibler, Perez (1957) used definitions in terms of abstract integrals and densities, and I suspect that Rosenblatt-Rot did likewise, but I have not seen his paper. Pinsker, however, used the partition approach of Kolmogorov and Dobrushin to *define* the *entropy of a probability measure $P$ with respect to another probability measure $M$*, both

measures defined on a common measurable space $(\Omega, \mathcal{B})$, as

$$H_M(P) = \sup_{\mathcal{S}\subset\mathcal{B}} \sum_{i=1}^{\|\mathcal{S}\|} P(S_i) \ln \frac{P(S_i)}{M(S_i)} \tag{16}$$

where the supremum is over all measurable finite partitions $\mathcal{S} = \{S_i; i = 1, \ldots, \|\mathcal{S}\|\}$.

A technical detail in the definition pointed out by Feinstein is the definition for the indeterminant case $0 \ln 0/a = 0$ for $a \geq 0$.

In *E&IT2* Pinsker's relative entropy for probability measures is called the *divergence* and following Pinsker is expressed as a supremum over partitions of the relative entropy of the partition:

$$\begin{aligned} D(P\|M) &= \sup_{\mathcal{S}\subset\mathcal{B}} H_{P\|M}(\mathcal{S}) \text{ where} \\ H_{P\|M}(\mathcal{S}) &= \sum_{i=1}^{\|\mathcal{S}\|} P(S_i) \ln \frac{P(S_i)}{M(S_i)}. \end{aligned} \tag{17}$$

Pinsker states that "obviously" with this definition of relative entropy, mutual information becomes a special case by choosing $P = P_{XY}$ and $M = P_X \times P_Y$, the product measure. It is obvious for the discrete case with pmfs from (10) and the real-valued continuous case from (11). It is also obvious given the generalized integral and density formulations for relative entropy developed later in Pinsker. On p. 19 of [14], however, Pinsker has only given the definition of relative entropy as a the supremum of the relative entropy over *all* partitions $\mathcal{S} \subset \mathcal{B}_{XY}$ and the conclusion is not obvious: The two quantities of interest can be written terms of suprema over partitions, but for the divergence definition of mutual information it is over *all* partitions in $\mathcal{B}_{XY}$, but for $I(X, Y)$ only product partitions of the form

$$\mathcal{Q} \times \mathcal{R} = \{Q_i \times R_j; Q_i \in \mathcal{B}_X, R_j \in \mathcal{B}_Y\}$$

are permitted. While it is obvious that

$$\begin{aligned} D(P_{XY}\|P_X \times P_Y) &= \sup_{\mathcal{S}\subset\mathcal{B}_{XY}} H_{P_{XY}\|P_X\times P_Y}(\mathcal{S}) \\ &\geq \sup_{\mathcal{Q}\subset\mathcal{B}_X, \mathcal{R}\subset\mathcal{B}_Y} H_{P_{XY}\|P_X\times P_Y}(\mathcal{Q}\times\mathcal{R}) \\ &= \sup_{\mathcal{Q}\subset\mathcal{B}_X, \mathcal{R}\subset\mathcal{B}_Y} I(\mathcal{Q}, \mathcal{R}) = I(X, Y), \end{aligned} \tag{18}$$

the revere inequality is not obvious.

The result that $I(X, Y) = D(P_{XY}\|P_X \times P_Y)$ follows from Pinsker's theorems 2.1.2 for mutual information and 2.4.1 for relative entropy, which largely correspond to Lemmas 7.14 for mutual information and 7.3 for relative entropy in *E&IT2*. The order of the presentation of the two results in the two books is reversed because the development here follows the the order in which Amiel Feinstein, Pinsker's translator, proved the two results in his translator's remarks.

Feinstein knew that mutual information was a special case of relative entropy, so it made sense to prove the most general result first.

This is a good point to provide historical information Pinsker's translator. Amiel Feinstein was born in January 1930 to Russian immigrant parents in Brooklyn, New York. He received his Bachelor of Science at Brooklyn College 1950. In 1954 he received his Ph. D. from MIT, his supervisor was Robert M. Fano and his dissertation was titled *A New Basic Theorem of Information Theory*. It was considered by some mathematicians to provide the first rigorous proofs of Shannon's basic results, a view that did not win friends in the engineering community. Nonetheless, Feinstein was intimately familiar with both the Russian language and with the measure theoretic tools that Kolmogorov and his students and colleagues were using to extend Shannon theory at the time. He was exceptionally qualified to translate Pinsker's book and he provided proofs which were often lacking in the Soviet literature along with explanatory comments and observations. He should be considered as a member of Kolmogorov and his school. His translation of Pinsker's book was written in communication with Pinsker who provided corrections and comments on the translation and Feinstein's additions.

The lemma statements of *E&IT2* differ from the corresponding theorems of Pinsker, which incorporate language limiting the collections of partitions considered. The importance of refining partitions remains, but here it is confined to the proof of Lemma 7.14, leading to simplified statements of the lemmas at the expense of some added work in the proof of Lemma 7.14. The appropriate definitions and properties are collected here for later use.

A partition $\mathcal{R}$ is said to *refine* or be finer than a partition $\mathcal{Q}$ if every atom in $\mathcal{Q}$ is a union of atoms of $\mathcal{R}$, in which case we write $\mathcal{Q} < \mathcal{R}$ or $\mathcal{R} > \mathcal{Q}$. Refinements are called *subpartitions* in Pinsker, but the name also has other meanings in the literature and is not used here.

A key property of relative entropy and refinement is provided by Lemma 3.3. of *E&IT2*: Suppose that $P$ and $M$ are two measures defined on a common measurable space $(\Omega, \mathcal{B})$ and that we are given finite partitions $\mathcal{Q} < \mathcal{R}$. Then

$$H_{P\|M}(\mathcal{Q}) \leq H_{P\|M}(\mathcal{R}) \tag{19}$$

A refinement of a partition has equal or larger relative entropy than the original.

---

This concludes the historical technical context and the prologue to the revision of Lemmas 7.3 and 7.14 in *E&IT2*.

# 5 Lemma 7.3 Dobrushin's Theorem for Relative Entropy

Lemma 7.3 (p. 175) has a confusing statement in its proof which is here replaced by an expanded explanation.

> **Lemma 7.3.** Suppose that $(\Omega, \mathcal{B})$ is a measurable space where $\mathcal{B}$ is generated by a field $\mathcal{F}$, $\mathcal{B} = \sigma(\mathcal{F})$. Then if $P$ and $M$ are two probability measures on this space,
>
> $$D(P\|M) = \sup_{\mathcal{Q}\subset\mathcal{F}} H_{P\|M}(\mathcal{Q}).$$

The following statement occurs in the latter half of the first paragraph in the proof of Lemma 7.3.

> Approximating this event by a field element $F_0$ by applying Theorem 1.1 simultaneously to $M$ and $G$ will yield a partition $\{F_0, F_0^c\}$ for which the right hand side of the previous equation is arbitrarily large.

It is not clear what "simultaneously to $M$ and $G$" means (and $G$ should be $P$). Replace the offending paragraph with the following:

> The event $F$ being approximated is an event for for which $M(F) = 0$ but $P(F) > 0$, which event must exist because the case under consideration is that where $P$ is not absolutely continuous with respect to $M$. Form the mixture measure $m = (P + M)/2$; that is, $m$ is a probability measure on a measurable space $(\Omega, \mathcal{B})$ with $\mathcal{B} = \sigma(\mathcal{F})$ for a field $\mathcal{F}$ as stated on the lemma 7.3 and $m(G) = (P(G) + M(G))/2$ for all events $G \in \mathcal{B}$. Applying Theorem 1.1 to the mixture measure $m$ implies that given $\epsilon > 0$ and $F \in \sigma(\mathcal{F})$ there is a $F_0 \in \mathcal{F}$ for which $m(F\Delta F_0) \leq \epsilon$ and hence that
>
> $$P(F\Delta F_0) + M(F\Delta F_0) \leq 2\epsilon$$
>
> whence $P(F\Delta F_0) \leq 2\epsilon$ and $M(F\Delta F_0) \leq 2\epsilon$. For any probability measure $P$ and events $G_1$ and $G_2$ $P(G_1\Delta G_2) \geq |P(G_1) - P(G_2)|$, whence
>
> $$\begin{aligned} |P(F) - P(F_0)| &\leq 2\epsilon \\ M(F_0) &\leq 2\epsilon \end{aligned}$$
>
> which implies that as $\epsilon$ gets smaller, the relative entropy with respect to the partion $\{F_0, F_0^c\}$ grows without bound and hence the divergence $D(P\|M)$ is infinite and the Lemma holds for the case where $P$ is not absolutely continuous wrt $M$. The remainder of the proof in the book follows as given, with the mixture trick being used again.

---

The statement of Dobrushin's theorem in Theorem 2.4.1 in Pinsker as translated by Feinstein [14] differs from that of Lemma 7.3 in *E&IT2*. Following Dobrushin [1], Pinsker imposes additional conditions involving refinements of the partitions allowed which are not included in Lemma 7.3.

Dobrushin's theorem states that the supremum over partitions defining mutual information can be limited to a generating field, but he adds a constraint in the theorem statement that the family of partitions allowed be such that every partition must have a refinement with atoms that are also contained in the field. Pinsker adopts the same explicit assumption and then later extends the same idea to relative entropy.

Feinstein points out the that the extra constraint on the family of partitions is not needed. I agree. Confining partitions to a generating field is sufficient for the relative entropy result (Lemma 7.3 in *E&IT2*), but refinements must be dealt with in proving that $I(X,Y) = D(P_{XY} \| P_X \times P_Y)$ in Lemma 7.14, as will be done in Section 6.

I am indebted to Dave Neuhoff for pointing out the problem and discussions leading to a clarified proof.

A note in hindsight: Although Pinsker describes Theorem 2.4.1 (Lemma 7.3 in *E&IT2*) as "Dobrushin's theorem," Dobrushin [1] dealt with mutual information, not relative entropy, but the key idea of finding a supremum over a family of partitions by restricting the supremum to a generating field is the same. Pinsker states the result first for mutual information in Theorem 2.1.1 which he describes as a "theorem due to Dobrushin." Pinsker's translator Feinstein proves the results in reverse order taking advantage of the fact that mutual information is a special case of relative entropy. *E&IT2* and these errata follow Feinstein. No inclusion of conditions on the family of partitions other than that they form a generating field has been required. When proving the equivalence of the direct and relative entropy definitions, however, it will be necessary in the proof to consider the field generated by the rectangles and refinements of partitions.

# 6 Lemma 7.14: $I(X,Y) = D(P_{XY} \| P_X \times P_Y)$

The following lines precede Lemma 7.14 on p. 205:

> Letting the generating field be the field of all rectangles of the form $F \times G$, $F \in \mathcal{B}_X$ and $G \in \mathcal{B}_Y$, we have the following lemma which is often used as a definition for mutual information.

These lines should be replaced by

> Letting the generating field be the field generated by all rectangles of the form $F \times G$, $F \in \mathcal{B}_X$ and $G \in \mathcal{B}_Y$, we have the following lemma which is often used as a definition for mutual information.

The error is the wording "the field of all rectangles" since the collection of all rectangles is not a field. It should be "the field generated by the rectangles." The error was was caught by John Duchi in February 2012 and was a carryover from the *First Edition*. It was corrected in the 3/3/2013 *First Edition, Corrected Version*, and the subsequent errata for *E&IT2*. Unfortunately I did not there explain how simply mending the incorrect wording of the lemma statement proved the lemma. The remainder of this section is devoted to the statement of the first part of the lemma and providing its missing proof. Only the first part of the lemma in the book is of concern here because the remaining parts follow from the first part as in the book.

The first part of Lemma 7.14 using the notation of *E&IT2* is stated as follows:

**Lemma 7.14** (first part. Lemma 5.5.1 in *First Edition*)

$$I(X,Y) = \sup_{q,r} I(q(X); r(Y)), \tag{20}$$

where the supremum is over all quantizers $q$ and $r$ of $A_X$ and $A_Y$.

The notation is consistent with *E&IT2*, but not with the history of definitions of mutual information considered in the previous section. To be consistent with the historical development, the right side is the quantization version of the generalization by Kolmogorov et al. of mutual information which up to this point of this section was denoted $I(X,Y)$, but in the context of *E&IT2* $I(X,Y)$ was defined by $D(P_{XY}\|P_X \times P_Y)$ so that the lemma statement in the current context should read

**Lemma 7.14**

$$D(P_{XY}\|P_X \times P_Y) = I(X,Y); \tag{21}$$

that is,

$$\sup_{\mathcal{S}\subset\mathcal{B}_{XY}} H_{P_{XY}\|P_X\times P_Y}(\mathcal{S}) = \sup_{\mathcal{Q}\subset\mathcal{B}_X\mathcal{R}\subset\mathcal{B}_Y} H_{P_{XY}\|P_X\times P_Y}(\mathcal{Q}\times\mathcal{R}) \tag{22}$$

This has already been seen to hold for Shannon's examples of discrete random variables and real-valued random variables or vectors with ordinary pdfs. It has been stated that the result also follows immediately in the general case by using the definitions involving general integrals and density functions. Here the goal is to provide a proof based only on the definitions for both quantities as extrema over partitions or quantizers.

As argued in (18), the restatement makes clear that that

$$D(P_{XY}\|P_X \times P_Y) \geq I(X,Y)$$

because both quantities maximize the same relative entropy of partitions, but the divergence does so over all finite measurable partitions while the mutual information maximizes over the subset of product partitions.

To prove the reverse inequality and thereby the lemma, begin by invoking Lemma 7.3 applied to the probability space $(A_X \times A_Y, \mathcal{B}_{XY}, P_{XY})$ where $\mathcal{B}_{XY} = \sigma(RECT_{XY})$, which event space is also generated by the field $\mathcal{F} = \mathcal{F}(RECT_{XY})$ — the field generated by the rectangles. Lemma 7.3 implies that

$$D(P_{XY}\|P_X \times P_Y) = I(X,Y) = \sup_{\mathcal{S}\subset\mathcal{F}} H_{P_{XY}\|P_X\times P_Y}(\mathcal{S});$$

that is, the supremum is unchanged by restricting the partition to have atoms in the generating field $\mathcal{F}$. Hence for any $\epsilon > 0$ there exists a partition $\mathcal{S} = \{S_i; i = 1, \ldots, N\}$ with $S_i \in \mathcal{F}$, the field generated by the rectangles, for which

$$H_{P_{XY}\|P_X\times P_Y}(\mathcal{S}) \geq D(P_{XY}\|P_X \times P_Y) - \epsilon. \tag{23}$$

The field generated by the rectangles in the abstract setting has a useful structure noted by Dobrushin [1] and which is also described in Lemma 2.2 of [7], p. 48: the field generated by the rectangles, $\mathcal{F}(RECT_{XY})$, is precisely the collection of all finite *disjoint* unions of rectangles. The result is usually developed for the special case of real alphabets and Borel fields[2], but it holds for the more general alphabets considered in Dobrushin [1] and *E&IT2*.

Since the finite number of atoms $S_i$ of $\mathcal{S}$ are disjoint and each atom is a finite union of disjoint rectangles, the collection of all of the rectangles in all of the atoms is itself a finite collection of disjoint rectangles whose union is the union of all of the atoms of $\mathcal{S}$, which is the full space $A_X \times A_Y$. Call this collection $\mathcal{S}' = \{S'_i; i = 1, \ldots, N\}$ where $S'_i = F_i \times G_i$, $F_i \in \mathcal{B}_X$, $G_j \in \mathcal{B}_Y$, $i = 1, \ldots, N$. $\mathcal{S}'$ is by construction a partition into a finite number of rectangles and that and that it is a refinement of $\mathcal{S}$ since membership in one of the rectangles in the collection implies membership in an atom of $\mathcal{S}$: $\mathcal{S}' > \mathcal{S}$ and hence

$$H_{P_{XY}\|P_X\times P_Y}(\mathcal{S}') \geq H_{P_{XY}\|P_X\times P_Y}(\mathcal{S}) \tag{24}$$

The final stage is to demonstrate that there is a product partition of the form $\mathcal{P}_X \times \mathcal{P}_Y$, where $\mathcal{P}_X = \{\Xi_i; i = 1, \ldots, \|\mathcal{P}_X\|\}$, $\mathcal{P}_Y = \{\Upsilon_j; j = 1, \ldots, \|\mathcal{P}_X\|\}$, $F_i \in \mathcal{B}_X$, $G_j \in \mathcal{B}_Y$, which refines $\mathcal{S}' = \{F_i \times G_i; i = 1, \ldots, N\}$.

Consider first the $X$ partition $\mathcal{P}_X$. Begin with the collection of all of $X$ components of the disjoint rectangles in $\mathcal{S}'$, that is, the sets $\{F_i; i = 1, \ldots, N\}$, $F_i \subset \mathcal{B}_\mathcal{X}$, all $i$. Although the rectangles containing these $X$ components are disjoint, the $F_i$ need not be disjoint, but $\bigcup_i F_i = A_X$ since the collection $\mathcal{S}_\rangle{}' = \{F_i \times G_i\}$ is a partition of $A_X \times A_Y$. For every $x \in A_X$ there must be at least one $y \in A_Y$ and hence at least one pair $(x, y)$ which falls in one of the atoms of a partition of $F_i \times G_i$.

The collection $\{F_i; i = 1, \ldots, N\}$ can be converted into a partition from which membership in $F_i$ can be recovered using set differencing:

$$A - B \stackrel{\Delta}{=} A \bigcap B^c.$$

[2]See, e.g., Dudley (2004) [2], p. 257.

Define the collection $\{\Xi_i\}$ iteratively as follows:

$$\begin{aligned}
\Xi_1 &= F_1 \\
\Xi_2 &= F_2 - F_1 = F_2 \bigcap F_1^c = F_2 - \Xi_1 \\
\Xi_3 &= F_3 - \bigcup_{i\leq 2} F_i = F_3 - \bigcup_{i\leq 2} \Xi_i \\
&\vdots \\
\Xi_N &= F_N - \bigcup_{i\leq N-1} F_i = F_N - \bigcup_{i\leq N-1} \Xi_i
\end{aligned}$$

Along the way an $\Xi_n$ may be empty because $F_n$ may have no new points not already found a previous $\Xi_j$. It is simpler to allow some of the $\Xi_i$ to be empty than to reindex the procedure to ensure all atoms are nonempty. Doing this provides a convenient means of recovering the sets $F_n$ from the disjoint $\Xi_n$ by

$$F_n = \bigcup_{i\leq n} \Xi_i. \tag{25}$$

where some of the $\Xi_i$ may be empty. The construction yields a partition of $A_X$: $\mathcal{P}_X = \{\Xi_i; i = 1, \ldots \|\mathcal{P}_X\|\}$, where $\|\mathcal{P}_X\| \leq N = \|\mathcal{S}'\|$ such that $F_i$ can be reconstructed as finite unions of the disjoint atoms of the partition.

The same construction can be used for the $Y$ components $G_i \in \mathcal{B}_Y, i = 1, \ldots, N\}$ of the partition $\mathcal{S}'$ of rectangles to obtain a partition $\mathcal{P}_Y = \{\Upsilon_i; i = 1, \ldots \|\mathcal{P}_Y\|\}$ of $A_Y$ with $\|\mathcal{P}_Y\| \leq N$.

$$G_n = \bigcup_{i\leq n} \Upsilon_i. \tag{26}$$

Consider the product partition $\mathcal{P}_X \times \mathcal{P}_Y$ which consists of all disjoint nonempty rectangles of the form $\Xi_i \times \Upsilon_j$ and consider a single atom $S'_n = F_n \times G_n \in \mathcal{S}'$ which can from (25-26) is

$$S'_n = F_n \times G_n = \left[\bigcup_{i\leq n} \Xi_i\right] \times \left[\bigcup_{j\leq n} \Upsilon_j\right]$$

A point $(x, y) \in S'_n$ if $x \in \Xi_i$ for any $i \leq n$ and $y \in \Upsilon_j$ for any $j \leq n$. Since the $\Xi_i$ are disjoint, $x$ can be in only one of the $\Xi_i$ and similarly for the $y$ and the $\Upsilon_j$. This means that $(x, y)$ must be in exactly one rectange of the form $\Xi_i \times \Upsilon_j$ for $i \leq n$ and $j \leq n$ and therefore

$$S'_n = \bigcup_{i\leq n} \bigcup_{j\leq n} (\Xi_i \times \Upsilon_j),$$

which is a finite union of disjoint rectangles $(\Xi_i \times \Upsilon_j)$, atoms of the product partition $\mathcal{P}_X \times \mathcal{P}_Y$. As this is true for all atoms of $S'_n \in \mathcal{S}'$, $\mathcal{P}_X \times \mathcal{P}_Y$ refines $\mathcal{S}'$ and hence with Eq. (23)

$$\begin{aligned} H_{P_{XY}\|P_X\times P_Y}(\mathcal{P}_X \times \mathcal{P}_Y) &\geq H_{P_{XY}\|P_X\times P_Y}(\mathcal{S}') \geq H_{P_{XY}\|P_X\times P_Y}(\mathcal{S}) \\ &\geq D(P_{XY}\|P_X \times P_Y) - \epsilon. \end{aligned}$$

□

# 7 Variational Description of Divergence

The unnumbered subsection of Section 7.1 with the title above has several errors and should be replaced by the following material:

## Variational Description of Divergence

As in the discrete case, divergence has a variational characterization that is a fundamental property for its applications to large deviations theory [182] [31]. We again take a detour to state and prove the property without delving into its applications.

Suppose now that $P$ and $M$ are two probability measures on a common probability space, say $(\Omega, \mathcal{B})$, such that $M \gg P$ and hence the density

$$f = \frac{dP}{dM}$$

is well defined. Suppose that $\Phi$ is a real-valued random variable defined on the same space. which has finite cumulant generating function:

$$E_M(e^{\Phi}) < \infty.$$

Then we can define a probability measure $M^{\Phi}$ by

$$M^{\Phi}(F) = \int_F \frac{e^{\Phi}}{E_M(e^{\Phi})} dM \tag{27}$$

and observe immediately that by construction $M \gg M^{\Phi}$ and

$$\frac{dM^{\Phi}}{dM} = \frac{e^{\Phi}}{E_M(e^{\Phi})}.$$

The measure $M^{\Phi}$ is called a "tilted" or "exponentially tilted" distribution in statistics and in information theory. Furthermore, by construction $dM^{\Phi}/dM \neq 0$ and hence we can write

$$\int_F \frac{f}{e^{\Phi}/E_M(e^{\Phi})}\, dM^{\Phi} = \int_F \frac{f}{e^{\Phi}/E_M(e^{\Phi})} \frac{dM^{\Phi}}{dM}\, dM = \int_F f dM = P(F)$$

and hence $P \ll M^\Phi$ and

$$\frac{dP}{dM^\Phi} = \frac{f}{e^\Phi / E_M(e^\Phi)}$$

which implies that $M \gg M^\Phi \gg P$.

We are now ready to state and prove the principal result of this section, a variational characterization of divergence.

**Theorem 7.1** (Theorem 5.2.1 in the *First Edition*)

Suppose that $M \gg P$. Then

$$D(P\|M) = \sup_\Phi \left( E_P \Phi - \ln(E_M(e^\Phi)) \right), \tag{28}$$

where the supremum is over all random variables $\Phi$ for which $e^\Phi$ is $M$-integrable and $E_P(\Phi)$ is well-defined.

*Proof:* First consider the random variable $\Phi$ defined by $\Phi = \ln dP/dM$. This choice meets the constraints required by the theorem since

$$\begin{aligned} \int e^\Phi dM &= \int dM \frac{dP}{dM} = \int dP = 1 \\ \int \Phi dP &= \int dP \ln \frac{dP}{dM} = D(P\|M) \end{aligned}$$

and hence for this choice

$$E_P \Phi - \ln(E_M(e^\Phi)) = D(P\|M) - \ln 1 = D(P\|M).$$

This proves that the supremum over all $\Phi$ is no smaller than the divergence $D(P\|M)$ since the divergence is achievable with the given choice of $\Phi$, Note that this is true even if the divergence $D(P\|M)$ is infinite, which is possible even if $M \gg P$.

To prove the other half of the theorem observe that for any $\Phi$ satisfying the constraints of the theorem, we have as above that $M \gg M^\Phi \gg P$ and hence from Corollary 7.1 with $Q = M^\Phi$ and the divergence inequality

$$\begin{aligned} D(P\|M) &= D(P\|M^\Phi) + E_P \left( \ln \frac{dM^\Phi}{dM} \right) \\ &= D(P\|M^\Phi) + E_P \left( \ln \frac{e^\Phi}{E_M(e^\Phi)} \right) \\ &\geq E_P \left( \ln \frac{e^\Phi}{E_M(e^\Phi)} \right) = E_P \Phi - \ln E_M(e^\Phi) \end{aligned}$$

which completes the proof. Note that equality holds and the supremum is achieved if and only if $M^\Phi = P$. □

---

The author thanks David Rosenberg for finding the errors in the First Edition in February 2011 and suggesting how to repair the proof. His correction arrived when *E&IT2* was in print and hence my incorrect proof propagated to *E&IT2*. The correct proof is included in the May 2023 errata list for *E&IT2*. The above proof is a slight modification of the one that appeared in the 3 March 2013 Corrected Version of the *First Edition*. The errors in the proof of the theorem were also pointed out by Yevgeny Seldin in May 2012. I am indebted to both for finding and reporting and helping to repair the proof.

# References

[1] R. L. Dobrushin. "A general formulation of the fundamental Shannon theorem in information theory". In: *Uspehi Mat. Akad. Nauk. SSSR* 14 (1959). Translation in *Transactions Amer. Math. Soc*, series 2,vol. 33,323–438. `http://dx.doi.org/10.1090/trans2/033/11`, pp. 3–104.

[2] R.M. Dudley. *Real Analysis and Probability*. Cambridge, UK: Cambridge University Press, 2004.

[3] R.M. Fano. *Transmission of Information: A Statistical Theory of Communications*. M.I.T. Press, 1961. URL: `https://books.google.com/books?id=dNADsB7eOI8C`.

[4] I. M. Gelfand, A. N. Kolmogorov, and A. M. Yaglom. "On the general definitions of the quantity of information". In: *Dokl. Akad. Nauk* 111 (1956). (In Russian.), pp. 745–748.

[5] I.M. Gelfand and A.M. Yaglom. *Calculation of the Amount of Information about a Random Function Contained in Another Such Function*. American Mathematical Society translations, 2, 12 (1959), 199-246. 1959 Translation from a paper presented at the Conference on Functional Analysis and its Applications which was held in Moscow in January 1956 and published in Russian in *Uspekhi Matem. Nauk*, XII, No. 1, (73), pp. 3-52 (1957). Some of the results were included in a joint paper with A. N. Kolmogorov which was presented at the Third All-Union Mathematics convention (Moscow , June —July, 1956) . American Mathematical Society, 1956. URL: `https://www.ams.org/books/trans2/012/09/trans2012-09.pdf?t=1774278538113`.

[6] R. M. Gray. *Entropy and Information Theory*. Second Edition, 2011. New York: Springer-Verlag, 1990.

[7] R. M. Gray. *Probability, Random Processes, and Ergodic Properties*. Second Edition, 2009. New York: Springer-Verlag, 1988.

[8] A. N. Kolmogorov. "On the Shannon theory of information in the case of continuous signals". In: *IRE Transactions Inform. Theory* IT-2 (1956), pp. 102–108.

[9] Samuel Kotz. "Recent Results in Information Theory". In: *Journal of Applied Probability* 3.1 (June 1966). Also published as a book in *Methuin Monographs on Applied Probability and Statistics*, M.S. Bartlett, General Editor, London, Methuen & Co., pp. 1–93. URL: https://www.jstor.org/stable/3212039.

[10] S. Kullback. *Information Theory and Statistics*. New York: Wiley, 1959.

[11] S. Kullback and R.A. Leibler. "On Information and Sufficiency". In: *Annals of Math. Stat.* (1951).

[12] A. Perez. "Information Theory with Abstract Alphabets". In: *Theory of Probability and its Applications* 4.1 (1959).

[13] A. Perez. "Sur la théorie de l'information dans le cas d'un alphabet abstrait". In: *Transactions First Prague Conf. on Information Theory, Stat. Decision Functions,Random Processes*. Czech. Acad. Sci. Publishing House, 1957, pp. 209–244.

[14] M. S. Pinsker. *Information and information stability of random variables and processes*. Translated by A. Feinstein from the Russian edition *Informatsiya I Informatsionnaya Ustiochivost' Sluchainykh Velichin I Protsessov* published in 1960 by Izd. Akad. Nauk. SSSR. San Francisco: Holden Day, 1964.

[15] C. E. Shannon. "A mathematical theory of communication". In: *Bell Syst. Tech. J.* 27 (1948), pp. 379–423, 623–656.

[16] P. C. Shields. "The ergodic and entropy theorems revisited". In: *IEEE Trans. Inform. Theory* IT-33 (1987), pp. 263–266.