Errata for
Entropy and Information Theory:
Second Edition

Robert M. Gray

Formatted June 4, 2023

1 Introduction

This document collects errata of the Second Edition of *Entropy and Information Theory* with occasional reference to the free updated hard cover original First Edition. Some of the errors and typos are inherited from the First Edition, but were not caught in time to be fixed in the 2011 publication of the Second Edition. Errors that were caught in time have been corrected in the final version of the First Edition, Corrected, and are included here as corrections.

Added citations and references are given local numbers. Equations relating to existing numbered equations in the book are given the numbers used in the book with the exception of the two complete proofs that are included.

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, and Jun Muramatsu. Dr Maramatsu in particular motivated my revisiting the First Edition and belatedly publishing an updated Errata for the Seccond Edition 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.*

2 Simple typos

page xx Equation (2) 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 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

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 the Second Edition. During our email exchange, she found that I had omitted an important detail used by Paul in [2] 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 the Second Edition. 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 the Second Edition 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: 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 [1], Lemma 7.12 in the Second Edition). 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 [1] and Section 8.1 of the Second Edition.

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 $n \geq 1$ for which $n^{-1}h_n(x) \leq \underline{h} + \epsilon$. 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

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

that is, redefine

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

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 δ 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* ℓ-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 $\{\ell_n; n \in \mathcal{Z}_+\}$ (ℓ 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 ℓ_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 ℓ_n is also stationary and ergodic.

The process ℓ_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 ℓ_{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 δ 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 ℓ-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.

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

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

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

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

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

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

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 $\delta(L-N)/3 \le \delta L/3$ time indices n for which $x_n^N \in C$; that is, $\ell_n \ge N$. From (4) the sample average must converge to $m(B) \le \delta/3$ 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

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

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 δ 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 $M \ge 3/\delta$ 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 $L \ge 3N/\delta$ 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

$$\sum_{k \le \delta L} \binom{L}{k} \le e^{L h_2(\delta)} \tag{7}$$

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

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

Given ℓ_i probabilities sum to one:

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

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 δ 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 [1]) $m(x : x^L \in B_L \bigcap G_L \text{ i.o.}) = 0$. 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 ϵ 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 and hence also proves Theorem 4.1 stationary ergodic measures. □

4 Variational Description of Divergence

Section 7.1 pp. 187-9 inherited problems from the First Edition, which were partially fixed in 3/3/2013 Final First Edition, Corrected. A more complete and clarified correction of the entire section was provided in the May 2023 *Final First Edition, Corrected.* The key problem was with Theorem 5.2.1 in the First Edition, corresponding to Theorem 7.1 in the Second Edition. The entire subsection should be replaced by the following version.

4.1 Section 7.1, subsection on Variational Description of Divergence

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 Φ 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^Φ by

$$M^\Phi(F) = \int_F \frac{e^\Phi}{E_M(e^\Phi)}dM \tag{10}$$

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^Φ 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

$$D(P\|M) = \sup_\Phi \left(E_P\Phi - \ln(E_M(e^\Phi)) \right), \tag{11}$$

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

Proof: First consider the random variable Φ 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 Φ is no smaller than the divergence $D(P\|M)$ since the divergence is achievable with the given choice of Φ, 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 Φ 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 the Second Edition was in print and hence my incorrect proof propagated to the Second Edition. The correct proof is included into the May 2023 errata list for the Second Edition. 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.

5 Information for General Alphabets

Section 7.4, p.205.

Replace the paragraph:

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

by

Letting the generating field be the field generated by all rectangles of the form $F \times G$, $F \in \mathcal{B}_{A_X}$ and $G \in \mathcal{B}_{A_Y}$, we have the following lemma which is often used as a definition for mutual information (e.g., in Pinsker's *Information and Information Stability*, p. 9).

The mistake lies in "the field of all rectangles" was caught by John Ducci in February 2012 and was a carryover from the First Edition. It was corrected in the 3/3/2013 First Edition, Corrected Version, but is appropriate with this collection of errata in the Second Edition.

6 Section 4.3: Nonergodic Sources

p. 107

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_ψ is the distribution of ψ (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 has 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_ψ as a conditional expectation given ψ) 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_ψ 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

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

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.

Remove the extra "Thus"

References

[1] R. M. Gray. *Probability, Random Processes, and Ergodic Properties.* Springer-Verlag, New York, 1988.

[2] P. C. Shields. The ergodic and entropy theorems revisited. *IEEE Trans. Inform. Theory*, IT-33:263–266, 1987.