

PROCEEDINGS

**THIRTIETH ANNUAL ALLERTON CONFERENCE
ON COMMUNICATION, CONTROL, AND COMPUTING**

**Paul Van Dooren
Mark W. Spong
Conference Co-Chairs**

**Conference held
September 30 - October 2, 1992
Allerton House
Monticello, Illinois**

**Sponsored by
The Coordinated Science Laboratory
and
The Department of Electrical and Computer Engineering
of the
UNIVERSITY OF ILLINOIS
at
Urbana-Champaign**

On Generating Random Bits from an Arbitrary Random Source*

Sridhar Vembu Sergio Verdú

Department of Electrical Engineering
Princeton University
Princeton, NJ 08544

Abstract

We address the following question: Suppose we are given a random source and want to use it as a random number generator, how many approximately fair bits can we generate from it? We answer this question with two different measures of approximation between probability distributions: the variational distance and the normalized divergence. The results of this paper provide an operational characterization of the *inf-entropy rate* of a source, defined in Han and Verdú, [1]. We also relate the inf and sup entropy rates to the quantity called *resolvability* of channels, also defined in [1].

1 Introduction

The recent results of Han and Verdú, [1], have opened some new questions in information theory. We focus on the source coding results of [1] in this article. Specifically, we address the following question: suppose we are given a random source and we want to use it as a random number generator, how many fair bits can we generate from it? Of course, we may not be able to get too many “pure” random bits - for example, if the probability masses of the source are irrational. Instead we will focus on synthesizing an “approximately” uniform distribution from the source and address the issue of the largest asymptotic exponential rate at which this can be done. We will call this rate the *Intrinsic Randomness rate* (IR rate) of the source. In this paper we use two measures of approximation of probability distributions - variational distance and divergence. In both cases the answer to the question turns out to be the same: it will be the *inf-entropy rate* as defined in Han and Verdú, [1]. The results of this article nicely complement the source coding results of [1] where the minimum fixed length encoding rate is shown to be the *sup-entropy rate* for an arbitrary finite alphabet source. Thus we see (Figure 1) that intrinsic randomness plays the counterpart to the minimum source coding rate, analogously to resolvability and channel capacity.

We will also relate the inf-entropy and the sup-entropy rates to the *resolvability* of channels defined in [1].

*This work was supported in part by the National Science Foundation under PYI Grant ECSE-8857689, and the Office of Naval Research under Grant N00014-90-J-1734. The work of the first author was also supported by an IBM Graduate Fellowship.

Source	Channel
$\overline{H}(X)$ Fixed length source coding	$\sup_X \overline{I}(X;Y)$ Resolvability
$\underline{H}(X)$ Intrinsic Randomness	$\sup_X \underline{I}(X;Y)$ Channel Capacity

Figure 1: Operational Characterization

Throughout this paper, by source we mean a sequence of finite dimensional distributions $P = \{P^n\}$. Note that we do not impose any consistency requirements - i. e. a higher dimensional distribution need not have the lower dimensional distributions as its marginals. We will denote the finite dimensional random variables by X^n, Y^n etc and their sequences by X, Y etc. In a slight abuse of notation, in order to denote the distance between two probability distributions, we use the corresponding random variables. All logarithms in this paper have an arbitrary base greater than 1 and \exp refers to that base.

In Cover and Thomas, [2], page 43, problem 7, the question of generating fair coin flips from biased coins is mentioned. The results of this paper could be thought of answering this question in a very general setup, with the relaxation in the requirement of pure coin flips.

In [1] it is shown that the resolvability of a source i. e. the number of bits required in the worst case to approximate the source is shown to be its sup-entropy rate. In this paper we prove that the number of bits we can usefully extract from the source is characterized by its inf-entropy rate.

2 Main Results

2.1 Approximation in Variational Distance

In the spirit of [1] we will adopt the following definitions.

Definition 1 Given a discrete source X with alphabet A . Fix $\epsilon > 0$. R is said to be an ϵ -achievable Intrinsic Randomness (IR) rate if there exists a sequence of mappings $\phi_n : A^n \rightarrow \{1, 2, 3, \dots, M\}$ such that for all $\gamma > 0$ and for all sufficiently large n ,

$$\frac{\log M}{n} > R - \gamma$$

and

$$d(\phi_n(X^n), U_M) < \epsilon$$

where U_M is uniformly distributed on $\{1, 2, 3, \dots, M\}$ and d is the variational distance.

The variational distance between two probability distributions P_1 and P_2 on an alphabet A is defined to be

$$d(P_1, P_2) = \sum_{\{x \in A\}} |P_1(x) - P_2(x)|$$

By a slight abuse of notation, we will use $d(X, Y)$ to denote the variational distance between the probability distributions that correspond to the random variables X and Y respectively. It is easy to show that

$$d(P_1, P_2) = 2 \max_{\{E: E \subset A\}} |P_1(E) - P_2(E)|$$

We will find both characterizations of the variational distance useful.

Remark 1: In the above we can use the normalized divergence $\frac{1}{n} D(\phi(X^n) || U_M)$ instead of the variational distance and in order to distinguish this we will call it achievability in the sense of divergence. The results for that approximation measure are proved separately.

Definition 2 *If a rate is ϵ -achievable for all $\epsilon > 0$ then it is called an achievable IR rate.*

Definition 3 *The maximum achievable IR rates is called the MIR rate of the source X . We will denote it by $U(X)$.*

Intuitively, the MIR rate refers to the rate of the the maximal uniform distribution we can synthesize from a given source asymptotically.

Now we will define the inf-entropy rate as in [1].

Definition 4 *The inf-entropy rate of X , $\underline{H}(X)$ is defined as the largest α that satisfies, for each $\epsilon > 0$,*

$$\lim_{n \rightarrow \infty} P_{X^n} \{x^n \in A^n : \frac{1}{n} \log \frac{1}{P(x^n)} \leq \alpha - \epsilon\} = 0$$

Now we are ready to state our main result: The MIR rate of a source is equal to its inf-entropy rate. We will first prove the converse part and then the direct (achievability) part.

Lemma 1

$$U(X) \leq \underline{H}(X)$$

where $U(X)$ is the MIR rate and $\underline{H}(X)$ is the inf-entropy rate of X .

Proof:

By contradiction. Suppose $U > \underline{H}$. Then there exists a $\delta > 0$ such that $U > \underline{H} + 2\delta$. By definition of U it follows that there exists a sequence of mappings $\phi_n : A^n \rightarrow \{1, 2, 3, \dots, M\}$ such that for all sufficiently large n ,

$$\frac{\log M}{n} > \underline{H} + \delta$$

and

$$d(\phi_n(X^n), U_M) \rightarrow 0 \text{ as } n \rightarrow \infty \quad (1)$$

Now by definition of \underline{H} it follows that

$$P_{X^n}(D) \geq \alpha > 0 \text{ infinitely often in } n \quad (2)$$

where

$$D = \{x^n \in A^n : \frac{1}{n} \log \frac{1}{P(x^n)} \leq \underline{H} + \frac{\delta}{2}\}$$

We will focus on those blocklengths n for which (2) is true.

It is also clear that the size of D is upper bounded as

$$|D| \leq \exp\{n(\underline{H} + \frac{\delta}{2})\}$$

Hence we get, for any mapping ϕ_n ,

$$\begin{aligned} d(\phi_n(X^n), U_M) &\geq P_{X^n}\{\phi_n(X^n) \in \phi_n(D)\} - P_{U_M}(\phi_n(D)) \\ &\geq P_{X^n}(D) - \frac{|\phi_n(D)|}{M} \\ &\geq \alpha - \frac{|\phi_n(D)|}{M} \\ &\geq \alpha - \exp\{n(\underline{H} + \frac{\delta}{2})\} \exp\{-n(\underline{H} + \delta)\} \\ &= \alpha - \exp\{-\frac{n\delta}{2}\} > \frac{\alpha}{2} > 0 \text{ infinitely often in } n \end{aligned}$$

where we have used the fact that $|\phi_n(D)| \leq |D| \leq \exp\{n(\underline{H} + \frac{\delta}{2})\}$. Hence we get a contradiction of (1) establishing the converse result. ■

Lemma 2

$$U(\mathbf{X}) \geq \underline{H}(\mathbf{X})$$

Proof: We will show that $\underline{H} - 3\gamma$ is an achievable IR rate for any $\gamma > 0$. From the definition of \underline{H} it follows that, for any $\gamma > 0$,

$$\lim_{n \rightarrow \infty} P_{X^n}\{x^n : \frac{1}{n} \log \frac{1}{P(x^n)} \leq \underline{H} - \gamma\} = 0$$

Hence the set $B_n = \{x^n : P(x^n) < \exp\{-n(\underline{H} - \gamma)\}\}$ is such that $\lim_{n \rightarrow \infty} P_{X^n}(B_n) = 1$. So for any $0 < \epsilon < \frac{1}{2}$ and for sufficiently large n ,

$$P_{X^n}(B_n) \geq 1 - \epsilon$$

Hence

$$|B_n| \geq (1 - \epsilon) \exp\{n(\underline{H} - \gamma)\} \geq [(1 - \epsilon) \exp\{n(\underline{H} - \gamma)\}] = N \text{ (say)}$$

It is easy to see that, for large n ,

$$\exp\{n(\underline{H} - 2\gamma)\} \leq N \leq \exp\{n(\underline{H} - \gamma)\}$$

We will define a sequence of mappings $\{\phi_n : A^n \rightarrow \{1, 2, 3, \dots, M\}\}$ as follows. For each n , define a partition of A^n into a collection of sets $\{B(i)\}_{i=1}^M$ as follows. First of all $B(1) = B_n^c$ (complement of B_n). The rest of the $B(i)$'s are defined by the following procedure.

Step 1: $j = 2$;

Step 2: Initialize $B(j) = \{\}$.

Step 3: Pick an arbitrary $x^n \in B_n - \bigcup_{i=2}^j B(i)$. Stop if this set is empty.

Step 4: $B(j) \leftarrow B(j) \cup \{x^n\}$. We know that $P_{X^n}(x^n) \leq \frac{1}{N}$.

Step 5: Repeat steps 3 and 4 until

$$\frac{n-1}{N} \leq P_{X^n}(B(j)) \leq \frac{n}{N} \tag{3}$$

Step 6: $j \leftarrow j + 1$. Go to step 2.

This procedure results in a partition of B_n into $\{B(i)\}_2^M$ with each of these subsets of B_n satisfying (3). We call B_n^c as $B(1)$. This procedure basically aggregates elements of B_n into subsets that have strict bounds on their probability. Now the number of partitions M is bounded as

$$\frac{(1-\epsilon)N}{n} \leq M-1 \leq \frac{N}{n-1} \tag{4}$$

We now define $\{\phi_n\}$ as:

$$\phi_n(x^n) = i \quad \text{if } x^n \in B(i)$$

Immediately we see that

$$\frac{\log M}{n} \geq \frac{\log N}{n} + \frac{\log(1-\epsilon)}{n} - \frac{\log n}{n} > \underline{H} - 3\gamma \quad \text{for sufficiently large } n$$

Now the variational distance can be bounded as

$$d(\phi_n(X^n), U_M) \leq P(B(1)) + \frac{1}{M} + \sum_{i=2}^M |P(B(i)) - \frac{1}{M}| \tag{5}$$

$$\leq \epsilon + \frac{1}{M} + (M-1) \left(\frac{n\epsilon}{(1-\epsilon)N} + \frac{1}{N} \right) \tag{6}$$

$$\leq \epsilon + \frac{n\epsilon}{(n-1)(1-\epsilon)} + \frac{1}{n-1} \tag{7}$$

$$\leq 5\epsilon \quad \text{for sufficiently large } n \tag{8}$$

where (6) follows from (3) and (4). Since the bound on the variational distance is true for any $0 < \epsilon < \frac{1}{2}$, we have established the result. ■

Remark 2: The above is a statement of the *existence* of a close to uniform distribution on a set of size $N \geq n(\underline{H} - \delta)$ for any $\delta > 0$. From this we can easily show that we can approximate an arbitrary M -type distribution for any $M \leq \exp\{n(\underline{H} - 2\delta)\}$. This would involve aggregating the probabilities just as we did in the proof of Lemma 2.

2.2 Approximation in the sense of Normalized Divergence

We will now prove analogous results with normalized divergence rather than variational distance as our measure of approximation of probability distributions. Divergence between two probability distributions P_1 and P_2 is defined as

$$D(P_1||P_2) = \sum_{\{x \in A\}} P_1(x) \log \frac{P_1(x)}{P_2(x)}$$

We will also use the notation $D(X||Y)$ to denote the divergence between the distributions from which X and Y are drawn. Note that divergence is not a metric in the space of probability distributions and is neither stronger nor weaker than the variational distance.

As noted in Remark 1, will replace the variational distance in definition 1 with the normalized divergence $\frac{1}{n}D(\phi(X^n)||U_M)$. Even with this new definition of achievability it turns out that the same result is true, i. e. the largest achievable IR rate is equal to the inf-entropy rate of the source \underline{H} .

Before proving this statement we will first explain what the new definition means. Note that

$$D(\phi(X^n)||U_M) = \log M - H(\phi(X^n))$$

where $H(\phi(X^n))$ is the entropy of the M -valued random variable $\phi(X^n)$. Therefore, ϵ -achievability implies that

$$\frac{1}{n} \log M - \epsilon \leq H(\phi(X^n)) \leq \frac{1}{n} \log M$$

where the first inequality is true for large n and the second one is true for any n .

Lemma 3 *The largest achievable IR rate in the sense of divergence of a source X , $U_d(X)$ satisfies, $U_d(X) \geq \underline{H}(X)$.*

Proof:

The proof is by the same aggregation procedure followed in lemma 2 with a slight modification and so will not be repeated here. ■

The converse part is proved below.

Lemma 4 *If R is an achievable IR rate in the sense of divergence, then $R \leq \underline{H}$.*

Proof: The proof is by contradiction. Suppose there is a $R > \underline{H}$ that is achievable. Hence for some $\delta > 0$, $R > \underline{H} + 3\delta$. Let

$$D_n = \{x^n \in A^n : P(x^n) \geq \exp\{-n(\underline{H} + \delta)\}\}$$

and

$$P(D_n) = \alpha_n$$

By definition of \underline{H} , there exists $\alpha > 0$ such that $\alpha_n \geq \alpha$ infinitely often in n . We will focus on those block lengths n only. The size of D_n is bounded by

$$|D_n| \leq \alpha_n \exp\{n(\underline{H} + \delta)\}$$

Let us choose $\epsilon = \frac{\alpha\delta}{4}$. Now achievability of R implies that for we can find a sequence of mappings $\{\phi_n\}$ where $\phi_n : A^n \rightarrow \{1, 2, 3, \dots, M\}$ such that for sufficiently large n ,

$$\frac{1}{n} \log M > R - \delta \tag{9}$$

and

$$\frac{1}{n} H(\phi_n(X^n)) \geq \frac{\log M}{n} - \epsilon \tag{10}$$

Now we bound the entropy of $H(\phi_n(X^n))$ as follows. Note that $\phi_n(X^n)$ takes values in the set $\{1, 2, \dots, M\}$. Now define a new transformation ψ_n that takes values in a set of size $e^{n(\underline{H}+\delta)} + M$ as follows: for $x^n \in D_n$, $\psi_n(x^n)$ takes value in the first $e^{n(\underline{H}+\delta)}$ elements and for $x^n \notin D_n$, $\psi_n(x^n)$ takes value in the remaining M elements. It is clear that

$$H(\phi_n(X^n)) \leq H(\psi_n(X^n)) \tag{11}$$

because we can always construct ϕ from ψ through another deterministic transformation. Now we get

$$\begin{aligned} H(Y_n) &= H(\psi_n(X^n)) \\ &= \sum_{i=1}^{e^{n(\underline{H}+\delta)}} P_{Y_n}(i) \log \frac{1}{P_{Y_n}(i)} + \sum_{i=e^{n(\underline{H}+\delta)+1}}^{e^{n(\underline{H}+\delta)}+M} P_{Y_n}(i) \log \frac{1}{P_{Y_n}(i)} \end{aligned}$$

Now $P_{Y_n}(i) \geq e^{-n(\underline{H}+\delta)}$ or $P_{Y_n}(i) = 0$ for the i 's in the first term in the right hand side and also the total probability of this set of i 's is $P(D_n) = \alpha_n$, and hence the first term is bounded above by $\alpha_n n(\underline{H} + \delta)$. The second term is easily bounded by $(1 - \alpha_n) \log M - (1 - \alpha_n) \log(1 - \alpha_n)$. Putting together all these we obtain,

$$H(\phi_n(X^n)) \leq n\alpha_n(\underline{H} + \delta) + (1 - \alpha_n) \log(M - 1) - (1 - \alpha_n) \log(1 - \alpha_n) \tag{12}$$

Now by assumption on R and from (9), we see that $M > \exp\{n(\underline{H} + 2\delta)\}$ and using this and (12) we get,

$$\begin{aligned} \frac{1}{n} H(\phi_n(X^n)) &\leq \alpha_n \left(\frac{\log M}{n} - \delta \right) + (1 - \alpha_n) \frac{\log M}{n} + \frac{(1 - \alpha_n)}{n} \log \frac{1}{1 - \alpha_n} \\ &\leq \frac{\log M}{n} - \frac{\alpha\delta}{2} \\ &= \frac{\log M}{n} - 2\epsilon \end{aligned}$$

for sufficiently large n , infinitely often in n . This contradicts (10) thereby establishing the result. ■

The above results giving an operational characterization of \underline{H} complement the fixed length source coding result of [1]. In the finite alphabet case if $\underline{H} = \overline{H}$ then the following limit exists and the above two quantities are also equal to $H = \lim_{n \rightarrow \infty} \frac{1}{n} E \log \frac{1}{P(X^n)}$. This is true for example in the stationary, ergodic case. The result presented in the paper may be viewed as a generalization of the asymptotic equipartition property (AEP).

As a nonergodic example, consider independent flips of a coin which is fair with probability $\frac{1}{2}$ and has probability of heads equal to $\frac{1}{\pi}$. Then the minimum achievable source coding rate is 1 bit/symbol, whereas its intrinsic randomness is $h(\frac{1}{\pi})$ and its entropy rate is $\frac{1}{2} + \frac{1}{2} h(\frac{1}{\pi})$, where $h(p) = -p \log p - (1 - p) \log(1 - p)$.

3 Connection with Resolvability

We will now provide a connection between resolvability and inf-entropy. Resolvability is concerned with the following question: what is the minimum M required so that if the n -dimensional input distribution is replaced by an M -type, the output distribution is arbitrarily close to the original output distribution in the sense of variational distance. The asymptotic exponential rate of this M , i. e. $\limsup \frac{\log M}{n}$ is called the resolvability of the channel for that particular input. If we supremize over all input distributions we get the resolvability of the channel. In [1], resolvability of a channel is shown to be $S = \sup_{\mathbf{X}} \bar{I}(\mathbf{X}, \mathbf{Y})$ where \mathbf{X} and \mathbf{Y} refer to the channel input and the channel output distributions respectively.

Motivated by the famous source-channel separation theorem in information theory, we ask the following question: we are given a source S with inf-entropy rate \underline{H} . We want to use S to emulate an arbitrary source S' in the following sense: apply a deterministic transformation on the source S so that when this is applied to the channel, the output is close in variational distance to that of the source S' . We can show that this is possible if $S < \underline{H}$ and it is not possible if $S > \bar{H}$. If we could show that if $S > \underline{H}$, the source cannot be an approximating source, it would very nicely complement the separation theorem that relates the minimum source coding rate and channel capacity. Unfortunately, we have been unable to answer the question of what happens if $\underline{H} < S < \bar{H}$.

Before we state the result formally we need some definitions. We will denote a sequence of random transformations $\{W^n\}$ by \mathbf{W} .

Definition 5 Suppose we are given a channel \mathbf{W} and a source $\mathbf{Z} = \{Z^n\}$ with probability law $\bar{P} = \{P_{Z^n}\}$. Fix $\epsilon > 0$. We call \mathbf{Z} an ϵ -approximating source for the channel if for any arbitrary input source \mathbf{X} with law P , there exists a sequence of mappings $\psi = \{\psi_n\}$ with $\tilde{\mathbf{X}} = \{\psi_n(Z^n)\}$ having the same alphabet as the input alphabet of the channel such that for sufficiently large n ,

$$d(Y^n, \tilde{Y}^n) < \epsilon$$

where \mathbf{Y} and $\tilde{\mathbf{Y}}$ are the outputs of the channel due to the inputs \mathbf{X} and $\tilde{\mathbf{X}}$ respectively.

Definition 6 We call a source an approximating source for a channel if it is ϵ -approximating for all $\epsilon > 0$.

Note that the above definitions do not require an approximating source \mathbf{Z} to have the same alphabet as the channel input alphabet.

Now the connection between resolvability of a channel and the inf-entropy and sup-entropy rates of a source is provided by the following lemma.

Lemma 5 A source \mathbf{Z} is an approximating source for a channel with resolvability S if the inf-entropy rate of the source \underline{H} satisfies

$$\underline{H} > S \tag{13}$$

(ii) If $\bar{H} < S$, the source cannot be an approximating source for this channel.

Proof: (Outline) (i) (13) implies that $\underline{H} > S + 2\gamma$ for some $\gamma > 0$. Now by definition of resolvability, we can replace the source \mathbf{X} by a source \mathbf{X}_M whose distribution is of M -type with $M < \exp\{n(S + \gamma)\}$ and the output distributions will be close in variational

distance. On the other hand, by the Remark following Lemma 2, we can approximate any N type distribution using the given source distribution of Z if $N < \exp\{n(\underline{H} - \gamma)\}$. Since $\underline{H} > S + 2\gamma$, we see immediately that $M < \exp\{n(\underline{H} - \gamma)\}$. Now the result will follow from noting that if two input distributions are close in variational distance, the corresponding channel output distributions will be close as well.

(ii) Assume the converse, i. e. $\bar{H} + 2\delta < S$ for some $\delta > 0$ and it is an approximating source for the channel. Now a source with sup-entropy rate \bar{H} can be approximated arbitrarily closely in the sense of variational distance by a M -type distribution where $M = \exp\{n(\bar{H} + \delta)\}$ when n is sufficiently large. Hence this approximating source can be replaced by an M -type source which is also an approximating source where $\frac{\log M}{n} < S - \delta$ and this will contradict the definition of resolvability establishing the desired result. ■

4 Conclusion

In this paper, we have addressed the question of using an arbitrary random source as a random number generator. This gives an operational characterization to the inf-entropy rate. This complements the characterization of the sup-entropy rate found in [1]. Some further directions in this area could be to extend these results in a rate distortion framework. Also our results leave a gap in the connection between resolvability and inf and sup entropy rates. It would be interesting to find out if this gap can be bridged.

Another interesting problem to be considered is to extend these and the results of [1] with another measure of distance between probability distributions called the \bar{d} distance introduced by Ornstein and discussed in [3], [4]. In [5], it is shown that the \bar{d} distance is over bounded by half the variational distance. Hence the direct part of the results hold for the \bar{d} distance. We recently have succeeded in proving the converse part as well and this will be presented in a subsequent paper.

References

- [1] T. S. Han and S. Verdú, "Approximation Theory of Output Statistics," submitted for publication. Also, presented at the 1992 *IEEE Information Theory Workshop*, Salvador, Brazil, June 1992.
- [2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, pp. 43, problem 7. John Wiley & sons, 1991.
- [3] D. S. Ornstein, "An application of ergodic theory to probability theory," *Annals of Probability*, vol. I, pp. 43-65, 1973.
- [4] R. M. Gray, D. L. Neuhoff, P. C. Shields, "A generalization of Ornstein's distance with applications to Information Theory," *Annals of Probability*, vol. III, pp. 315-328, April 1975.
- [5] R. M. Gray, D. S. Ornstein, "Block coding for discrete stationary d-continuous noisy channels," *IEEE Transactions on Information Theory*, IT-25, pp. 292-306, 1979.