

# Channel Simulation and Coding With Side Information

Yossef Steinberg and Sergio Verdú, *Fellow, IEEE*

**Abstract**—We study the minimum random bit rate required to simulate a random system (channel), where the simulator operates with a given external input. As measures of simulation accuracy we use both the variational distance and the  $\bar{d}$  distance between joint input–output distributions. We find the asymptotic number of random bits per input sample required for accurate simulation, as a function of the distribution of the input process. These results hold for arbitrary channels and input processes, including nonstationary and nonergodic processes and do not hinge on a specific simulation scheme. A by-product of our analysis is a general formula for the minimal achievable source coding rate with side information.

**Key Words**—Channel simulation, channel sup-entropy, conditional sup-entropy rate, conditional resolvability, source coding with side information.

## I. INTRODUCTION

THIS paper studies the minimum randomness necessary to simulate an arbitrary given random system (channel). The problem we solve is complementary to that of approximation of output statistics introduced in [1] by Han and Verdú. For a given channel and input process [1] studies the minimum randomness of those input processes whose output statistics approximate the original output statistics arbitrarily accurately. The maximum randomness over all input processes is called the resolvability of the channel, which is shown in [1] to equal the Shannon capacity.<sup>1</sup> Han and Verdú show that the problem of channel resolvability has strong connections with the problems of source coding, channel coding, and identification via channels [2].

If in [1] the focus is on the complexity of generating random inputs to a given channel (which need not be simulated), here we consider the dual problem where the input process is given and the channel random transformation has to be simulated. As an example, consider a telephone-channel simulator which operates with real external speech signals. Since the channel is noisy, a ran-

domness source is needed for its simulation. In this paper we determine the minimum complexity of the channel simulator so that the resulting joint input–output statistics are arbitrarily close to the desired ones. As in [1], we measure the complexity of the simulator by the number of fair bits per input sample required to generate every realization of the simulated process, and we adopt the variational (or  $l_1$ ) distance as a measure of similarity between probability distributions. By arbitrarily close approximation of the desired input–output statistics we mean that the variational distance goes to zero as the block-length goes to infinity. If we were to insist that the distributions are identical (zero variational distance), then the problem would have a very different nature. For example, we could only simulate channels whose conditional probabilities have finite binary representations. The conditional resolvability of a channel given a specific input process  $X$  is defined as the minimum number of random bits per input sample required to simulate a random transformation of  $X$ , so that the joint input–output statistics are simulated arbitrarily accurately.

The problem considered in [1] and this paper share the special case of approximating a given source statistics (by considering an identity channel in [1] and a deterministic input in the present setting). It was shown in [1] that the resolvability of any finite-alphabet source is equal to its minimum achievable fixed-length source coding rate. Furthermore [1] shows that this quantity is equal to the *sup-entropy rate* of the source, without assuming any restrictions such as ergodicity or stationarity. Those two results are generalized in this paper: we show that the conditional resolvability is equal to both the *conditional sup-entropy rate* and the minimum achievable fixed-length source coding rate with side information.

As in [1], we deal with arbitrary input processes and channels. In particular, we do not impose stationarity or ergodicity conditions. In this work, the purpose of channel simulation is to approximate the channel so that the joint input–output distributions are accurately approximated. It may seem that this is a point of departure from [1]. However, that kind of approximation criterion can also be incorporated in the problem of [1] by considering a channel whose output is equal to  $(X, Y)$ , the input and output of the original channel. On the other hand, we would arrive at different, and less interesting results, had we adopted the output approximation criterion in this work.

The related problem of channel approximation has been studied by Neuhoff and Shields in [3], [4]. They focus

<sup>1</sup>For all finite input channels that satisfy the strong converse.

Manuscript received December 11, 1992; revised manuscript received May 18, 1993.

This work was partially supported by the U.S. Office of Naval Research under Grant N00014-90-J-1734 and the National Science Foundation under Grant ECSE-8857689. The material in this paper was presented in part at the Sixth Joint Swedish-Russian International Workshop on information Theory, Mölle, Sweden, August 22–27, 1993.

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544.

IEEE Log Number 9402030.

attention on stationary channels whose input and output memories satisfy conditions called  $\bar{d}$  continuity and conditional almost block independence (CABI), respectively. A class of those channels, called primitive channels (nonlinear, time-invariant sliding-block transformations of input and iid noise) is shown in [3] to be dense in the class of  $\bar{d}$ -continuous CABI channels. This result suggests a question related to the one treated in this paper: quantify the complexity necessary to simulate a  $\bar{d}$ -continuous CABI channel by the minimum noise entropy among its approximating primitive channels. Neuhoff and Shields [4] show that the maximum complexity (over all input processes) is the so-called channel entropy: the supremum of the conditional output entropy over all stationary input sources.

Motivated by [3] and [4], we consider the  $\bar{d}$ -distance as an approximation measure in addition to the variational distance and we show that the conditional channel resolvability is the same in both cases. Since we do not restrict ourselves to stationary,  $\bar{d}$ -continuous or CABI channels, in contrast to [3], [4] we do not restrict ourselves to the specific simulator structure suggested by the primitive channel which need not be sufficient to approximate a channel not encompassed by the class considered in [3], [4]. Moreover, we adopt a worst-case randomness measure instead of the average measure (entropy) used in [3], [4] (cf. [1, Section 3]). Another difference with [1] and [4] is that in this paper we find an expression for the conditional resolvability as a function of the statistics of the input process. Specifically, we do not need to take the worst case over all input processes. It should be noted that the resolvability of a channel with a given input is known only within specific channel models such as discrete memoryless channels with full rank [5].

To fix ideas, let us consider the following examples (solved in Section V) where it is desired to simulate a binary symmetric channel:

$$Y_i = X_i \oplus Z_i,$$

where all the sequences are binary, and  $\{Z_i\}$  is an independent sequence conditioned on  $\{X_i\}$ , such that the crossover probability may depend on the input sequence:

$$P[Z_n = 1] = \alpha_n(X_1, \dots, X_n).$$

*Example 1:*

$$\alpha_n(X_1, \dots, X_n) = \alpha.$$

In this case, the conditional resolvability is equal to  $h(\alpha)$ , the binary entropy of  $\alpha$ , regardless of the actual input statistics. This means that no matter how complex the input process is, the most economical way to simulate the channel is actually to implement it directly. Note that we would have arrived at a different result had the objective been the reproduction of output statistics rather than joint input-output statistics. For example, no random bits are necessary to obtain a Bernoulli-(1/2) process at the output if the input is itself Bernoulli-(1/2).

*Example 2:*

$$\alpha_n(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i.$$

This channel is not encompassed by the channels considered in [3], [4]. Now the conditional resolvability of the channel does depend on the input process. If the input is Bernoulli( $\theta$ ), then the conditional resolvability is  $h(\theta)$ . If the input is deterministic, then the conditional resolvability is

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n h(\alpha_i).$$

*Example 3:*

$$\alpha_n(X_1, \dots, X_n) = f\left(\frac{1}{n} \sum_{i=1}^n X_i\right),$$

where

$$f(u) = 2u \bmod 1 \triangleq \begin{cases} 2u, & u \in [0, \frac{1}{2}], \\ 2u - 1, & u \in (\frac{1}{2}, 1]. \end{cases}$$

At first glance one might have expected that the conditional resolvability given  $X$  plus the resolvability of  $X$  (i.e., the complexity of generating  $X$ ) be equal to the resolvability of the joint process  $XY$ . If this were the case, then conditional resolvability would follow straightforwardly from the results in [1]. This channel disproves that behavior. Let the input be a nonergodic process which is Bernoulli-(1/4) with probability 1/2 and Bernoulli-(1/2) with probability 1/2. The resolvability of  $X$  is equal to 1 bit, and the conditional resolvability of the channel given  $X$  is also equal to 1 bit. However the resolvability of the pair  $XY$  is equal to  $1 + h(1/4)$  bits. Thus, it is harder to simulate  $X$  and then  $Y$  given  $X$ , than  $XY$ .

Section II presents the main definitions and the statement of our central problem. Section III gives the main results characterizing the conditional resolvability as the conditional sup-entropy rate. Section IV examines the canonical Slepian-Wolf setting of source coding with side information. It shows that for an arbitrary source (not necessarily stationary/ergodic) the minimum achievable fixed-length source coding rate with side information is equal to the conditional sup-entropy rate. Finally, in Section V we solve Examples 1–3 of the Introduction.

## II. BASIC DEFINITIONS

Let  $A, B$  be finite sets. We denote by  $M_1(A)$  [resp.  $M_1(B)$ ] the set of all probability distributions on  $A$  (resp.  $B$ ). Throughout, by a source  $X$  with alphabet  $A$  we mean a sequence  $P_X = \{P_X^n(\cdot)\}_{n \geq 1}$  of finite dimensional distributions  $P_X^n \in M_1(A^n)$ , where as in [1], we do not impose any consistency requirements on this sequence. To denote a source we will use both notations  $X$  and  $P_X$  interchangeably. Similarly, a channel  $W_{Y|X}$  with input alphabet  $A$  and output alphabet  $B$  is a sequence of conditional distributions  $\{W_{Y|X}^n(\cdot|\cdot)\}_{n \geq 1}$  such that  $W_{Y|X}^n(\cdot|a^n) \in M_1(B^n)$  for every  $a^n \in A^n$ . A joint input-output process  $\{P_{XY}^n \in M_1(A^n \times B^n)\}_{n \geq 1}$  will be denoted by  $XY$  and  $P_{XY}$ . All

logarithms in this paper have an arbitrary base and  $\exp(\cdot)$  refers to that base. We start with a few definitions in the spirit of [1].

**Definition 1:** Given a joint distribution  $P_{XY}^n(a^n, b^n) = P_X^n(a^n)W_{Y|X}^n(b^n|a^n)$ , the *conditional entropy density* is the function

$$i_{Y|X}^n(b^n|a^n) = -\log W_{Y|X}^n(b^n|a^n).$$

The distribution of the random variable  $(1/n)i_{Y|X}^n(Y^n|X^n)$  where  $X^n, Y^n$  have joint distribution  $P_{XY}^n$  is referred to as the *conditional entropy spectrum*, and the expected value of the conditional entropy spectrum is the normalized conditional entropy  $(1/n)H(Y^n|X^n)$ .

**Definition 2:** Let  $P_{XY} = \{P_{XY}^n\}_{n \geq 1}$  be given. The *conditional sup-entropy rate*  $\bar{H}(Y|X)$  of  $Y$  given  $X$  is defined as the lim sup in probability of the sequence of random variables  $\{(1/n)i_{Y|X}^n(Y^n|X^n)\}_{n \geq 1}$ , i.e.,

$$\bar{H}(Y|X) = \inf \left\{ h: \lim_{n \rightarrow \infty} P_{XY}^n \left( a^n b^n: \frac{1}{n} i_{Y|X}^n(b^n|a^n) > h \right) = 0 \right\}.$$

Analogously, the *conditional inf-entropy rate*  $\underline{H}(Y|X)$  is the lim inf in probability of the sequence  $\{(1/n)i_{Y|X}^n(Y^n|X^n)\}_{n \geq 1}$ , i.e.,

$$\underline{H}(Y|X) = \sup \left\{ h: \lim_{n \rightarrow \infty} P_{XY}^n \left( a^n b^n: \frac{1}{n} i_{Y|X}^n(b^n|a^n) < h \right) = 0 \right\}.$$

In case that the input process  $X$  is deterministic, or in case that  $Y^n$  is independent of  $X^n$  for every  $n$ ,  $\bar{H}(Y|X)$  and  $\underline{H}(Y|X)$  coincide with the sup-entropy rate  $\bar{H}(Y)$  and inf-entropy rate  $\underline{H}(Y)$  of  $Y$  as defined in [1], respectively.

The conditional entropy rate of  $Y$  given  $X$  is defined as

$$H(Y|X) = \lim_{n \rightarrow \infty} \frac{1}{n} H(Y^n|X^n) \quad (1)$$

provided that the limit exists. The following lemma states that convergence in probability of the conditional entropy density implies that the limit in (1) exists.

**Lemma 1:** If  $|A| < \infty$ ,  $|B| < \infty$ , then  $\bar{H}(Y|X) = \underline{H}(Y|X)$  implies that  $H(Y|X)$  exists and

$$H(Y|X) = \bar{H}(Y|X).$$

*Proof:* The proof follows the lines of the proof of [1, Lemma 1] and is therefore omitted.  $\square$

**Definition 3:** (e.g., [6]). The *variational distance* or  $l_1$  distance between two distributions  $P$  and  $Q$  on  $A$  is

$$d(P, Q) = \sum_{a \in A} |P(a) - Q(a)| = 2 \max_{A' \subset A} |P(A') - Q(A')|.$$

In a slight abuse of notation, the variational distance between any pair of distributions  $P$  and  $\tilde{P}$  will be denoted by  $d(P, \tilde{P})$  and  $d(X, \tilde{X})$  interchangeably, where  $X, \tilde{X}$  are the corresponding random variables.

**Definition 4:** [1] The *resolution*  $R(P)$  of a distribution  $P$  on  $A$  is the minimum log  $M$  such that  $P$  is an  $M$ -type (i.e., its masses are integer multiples of  $1/M$ ). If such  $M$  does not exist,  $R(P) = \infty$ .

**Definition 5:** The *resolution*  $R_c(W_{Y|X})$  of a conditional distribution  $W_{Y|X}(\cdot|\cdot)$  is defined as

$$R_c(W_{Y|X}) = \max_{a \in A} R(W_{Y|X}(\cdot|a)).$$

**Definition 6:** Let  $W_{Y|X}$  and  $P_X$  be given and let  $P_{XY}$  be the sequence of joint input-output distributions  $\{P_{XY}^n W_{Y|X}^n\}_{n \geq 1}$ .  $R$  is an  $\epsilon$ -achievable resolution rate of  $XY$  given  $X$  if for every  $\gamma > 0$  there exists a channel  $\tilde{W}_{Y|X}$  satisfying

$$\frac{1}{n} R_c(\tilde{W}_{Y|X}^n) < R + \gamma$$

and

$$d(P_X^n W_{Y|X}^n, P_X^n \tilde{W}_{Y|X}^n) < \epsilon \quad (2)$$

for all sufficiently large  $n$ .

**Definition 7:**  $\sigma_\epsilon(XY|X)$  is the minimal  $\epsilon$ -achievable resolution rate of  $XY$  given  $X$ .

$R$  is an *achievable resolution rate* for  $XY$  given  $X$  if it is  $\epsilon$ -achievable for every  $\epsilon > 0$ . The minimal achievable resolution rate of  $XY$  given  $X$  is called the *conditional resolvability* of  $XY$  given  $X$ , and is denoted by  $\sigma(XY|X)$ . Note that  $\sigma_\epsilon(XY|X)$  is monotone nonincreasing in  $\epsilon$ , and that

$$\sigma(XY|X) = \lim_{\epsilon \rightarrow 0} \sigma_\epsilon(XY|X) = \sup_{\epsilon > 0} \sigma_\epsilon(XY|X). \quad (3)$$

The operational meaning of  $\sigma(XY|X)$  can be summarized as follows: For a given channel  $W_{Y|X}$  and an input process  $X$ ,  $\sigma(XY|X)$  is the minimal number of pure random bits per input sample required to simulate the random transformation  $W_{Y|X}^n$ , so that the variational distance between the resulting joint input-output statistics and the desired statistics vanishes as  $n \rightarrow \infty$ . This statement follows from Definitions 4-7 and the following lemma in a straightforward fashion.

**Lemma 2:** Let  $\tilde{W}_{Y|X}$  be a channel such that for all  $n > n_0$  and every  $a^n \in A^n$  the normalized resolution of  $\tilde{W}_{Y|X}^n(\cdot|a^n)$  is less than  $\alpha$ . Then for every  $\delta > 0$  there exists a sequence of mappings

$$\phi_n: A^n \times \{1, 2, \dots, M'\} \rightarrow B^n$$

such that for all sufficiently large  $n$  and every input process  $X$

$$d(X^n Y^n, X^n \phi_n(X^n, Z^n)) < \delta \quad (4)$$

where  $Z^n$  is uniformly distributed over  $\{1, 2, \dots, M'\}$ , and  $M' = \exp(\lceil n\alpha + n\delta \rceil)$ .

*Proof:* Let  $n > n_0$ . Define

$$C_n(a^n) = \left\{ b^n: -\frac{1}{n} \log \tilde{W}_{Y|X}^n(b^n|a^n) \leq \alpha \right\}.$$

Then  $\tilde{W}_{Y|X}^n(C_n(a^n)|a^n) = 1$ , and  $\tilde{W}_{Y|X}^n(b^n|a^n) \geq \exp(-n\alpha)$  for every  $b^n \in C_n(a^n)$ , implying

$$|C_n(a^n)| \leq \exp(n\alpha) \quad \forall a^n.$$

By assumption on the type of  $\tilde{W}_{Y|X}^n$ , there exist integers  $\{k(b^n|a^n)\}_{b^n, a^n}$  such that

$$\tilde{W}_{Y|X}^n(b^n|a^n) = \frac{k(b^n|a^n)}{M(a^n)} \quad \forall a^n, b^n,$$

where

$$\sum_{b^n} k(b^n|a^n) = M(a^n) \leq \exp(n\alpha).$$

Define

$$k'(b^n|a^n) = \left\lfloor \frac{M'}{M(a^n)} k(b^n|a^n) \right\rfloor.$$

It is easy to verify that

$$\frac{k(b^n|a^n)}{M(a^n)} - \frac{1}{M'} \leq \frac{k'(b^n|a^n)}{M'} \leq \frac{k(b^n|a^n)}{M(a^n)}. \quad (5)$$

and therefore

$$1 - \exp\{n\alpha - [n(\alpha + \delta)]\} \leq \frac{1}{M'} \sum_{b^n \in C_n(a^n)} k'(b^n|a^n) \leq 1.$$

Choose a default element  $b' \in B^n$  and define a distribution  $V^n(\cdot|a^n)$  on  $B^n$ :

$$V^n(b^n|a^n) = \begin{cases} 1 - \sum_{\beta \in B^n, \beta \neq b'} \frac{k'(\beta|a^n)}{M'} & \text{for } b^n = b', \\ \frac{k'(b^n|a^n)}{M'}, & \text{otherwise.} \end{cases} \quad (6)$$

By (5), for sufficiently large  $n$

$$d(\tilde{W}_{Y|X}^n(\cdot|a^n), V^n(\cdot|a^n)) \leq \exp(-n\delta) < \delta \quad \forall a^n. \quad (7)$$

Now the distributions  $\{V(\cdot|a^n)\}_{a^n}$  can be precisely synthesized by taking a uniform distribution over a set of size  $M'$ , aggregating its elements into bins according to the values of  $k'(b^n|a^n)$ , and identifying each bin with the corresponding  $b^n$ . The maps  $\phi_n(\cdot, \cdot)$  in (4) stand for this aggregation procedure. Since the right-hand side in (7) is independent of  $a^n$ , the lemma is proved.  $\square$

Note that a uniform distribution over a set of size  $M' = \exp([n\alpha + n\delta])$  is nothing more than a sequence of  $[n(\alpha + \delta)]$  iid pure random bits (in case that the exponent base is taken as 2). Thus Lemma 2 states that for  $n$  large enough,  $\tilde{W}_{Y|X}^n(\cdot|a^n)$  can be approximated within distance  $\delta$  by mapping a sequence of  $[n(\alpha + \delta)]$  pure random bits, and that this approximation is uniformly good with respect to the input [although the mappings  $\phi_n(\cdot, \cdot)$  depend on  $a^n$ ].

If we quantify the complexity of the simulation scheme by the number of random bits per input sample needed for the simulation, then in view of its operational meaning,  $\sigma(\mathbf{XY}|\mathbf{X})$  serves as a measure of the *worst case complexity* of the most efficient channel simulator acting with “real” input  $\mathbf{X}$ . The term *worst case* is with respect to both the input and the output realizations.

Another measure of accuracy we shall consider in this work is the  $\bar{d}$  distance, introduced in ergodic theory by Ornstein [7], [8] and used extensively by Gray and Ornstein [9] and Neuhoff and Shields [3], [4]. The definition of this distance is as follows. Let  $A$  be a finite set. Denote by  $d_H(\cdot, \cdot)$  the (per letter) Hamming distance on  $A$  and by  $d_n(\cdot, \cdot)$  the  $n$ th order normalized Hamming distance on  $A^n$

$$d_n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^n d_H(x_i, \hat{x}_i) \quad x^n \in A^n, \hat{x}^n \in A^n,$$

where  $x_i$  is the  $i$ th coordinate of  $x^n$ . We have

*Definition 8:* (e.g., [3]). The  $\bar{d}$  distance between two distributions  $P^n \in M_1(A^n)$ ,  $\hat{P}^n \in M_1(A^n)$ , is defined as

$$\bar{d}(P^n, \hat{P}^n) = \inf_{\bar{P} \in \mathcal{P}(P^n, \hat{P}^n)} E_{\bar{P}} d_n(X^n, \hat{X}^n)$$

where  $\mathcal{P}(P^n, \hat{P}^n)$  stands for the collection of all distributions on  $A^n \times A^n$  having  $P^n$  and  $\hat{P}^n$  as marginals, and  $E_{\bar{P}}$  denotes expectation according to  $\bar{P}$ .

Now the conditional resolvability  $\sigma(\mathbf{XY}|\mathbf{X})$  can be defined also in the  $\bar{d}$  sense. Thus, an  $\epsilon$ -achievable  $\bar{d}$  resolution rate of  $\mathbf{XY}$  given  $\mathbf{X}$  is defined exactly as in Definition 6, but with  $\bar{d}$  replacing the variational distance in (2). Accordingly,  $\bar{\sigma}_\epsilon(\mathbf{XY}|\mathbf{X})$  stands for the minimal  $\epsilon$ -achievable  $\bar{d}$  resolution rate for  $\mathbf{XY}$  given  $\mathbf{X}$ , and  $R$  is an *achievable  $\bar{d}$  resolution rate for  $\mathbf{XY}$  given  $\mathbf{X}$*  if it is  $\epsilon$ -achievable (in the  $\bar{d}$  sense) for every  $\epsilon > 0$ . Finally, the minimal achievable  $\bar{d}$  resolution rate for  $\mathbf{XY}$  given  $\mathbf{X}$  is called *the conditional  $\bar{d}$  resolvability of  $\mathbf{XY}$  given  $\mathbf{X}$*  and is denoted by  $\bar{\sigma}(\mathbf{XY}|\mathbf{X})$ . It is clear that a relation analogous to (3) holds also for the corresponding  $\bar{d}$ -quantities, and the same type of operational significance holds for  $\bar{\sigma}(\mathbf{XY}|\mathbf{X})$ .

It is interesting to compare  $\sigma(\mathbf{XY}|\mathbf{X})$  with  $\bar{\sigma}(\mathbf{XY}|\mathbf{X})$  in view of the relation between the variational and the  $\bar{d}$  distances. It is shown in [9] that the  $\bar{d}$  distance is upper bounded by half the variational distance, i.e., for every  $n \geq 1$  and every pair of distributions  $P^n, \hat{P}^n$  on  $A^n$ ,

$$\bar{d}(P^n, \hat{P}^n) \leq \frac{1}{2} d(P^n, \hat{P}^n). \quad (8)$$

This immediately implies that

$$\bar{\sigma}(\mathbf{XY}|\mathbf{X}) \leq \sigma(\mathbf{XY}|\mathbf{X}). \quad (9)$$

The bound in (8) is far from being tight. In fact, it is easy to construct sequences of distributions  $\{P^n\}_{n \geq 1}, \{\hat{P}^n\}_{n \geq 1}$  for which the variational distance is 2 for every  $n$  whereas the  $\bar{d}$  distance vanishes. Therefore one would expect it be possible to find a channel  $\mathcal{W}_{Y|X}$  and an input process  $\mathbf{X}$  for which strict inequality holds in (9). Surprisingly, this is not the case: As shown in Section 3, (9) is satisfied with equality for any joint input-output process  $\mathbf{XY}$ .

### III. CONDITIONAL RESOLVABILITY AND SYSTEM SIMULATION

In this section we state and prove our result on worst-case complexity of channel simulation. We show that the conditional resolvability associated with a channel and an input process is equal to the conditional sup-entropy rate  $\bar{H}(Y|X)$ , for both the variational and the  $\bar{d}$  accuracy measures. In view of (8), it is enough to prove the direct (achievability) part with the variational distance, and the converse part with  $\bar{d}$  distance. To this end, we need some more definitions and notations. Let  $\mathcal{S}_A$  stand for the set of all sequences of strings  $\{a^n\}_{n \geq 1}$ . For an element  $\alpha$  of  $\mathcal{S}_A$ , we denote by  $\alpha_i$  the string of length  $i$  in the sequence  $\alpha$ ; thus,  $\alpha_i \in A^i$ . Observe that  $\alpha_i$  need not be a prefix of  $\alpha_{i+1}$ .

For every  $\alpha \in \mathcal{S}_A$ , define the quantity

$$\bar{H}(Y|\alpha) \triangleq \inf \left\{ h: \lim_{n \rightarrow \infty} W_{Y|X}^n(b^n: -\frac{1}{n} \log W_{Y|X}^n(b^n|\alpha_n) > h | X^n = \alpha_n) = 0 \right\}. \quad (10)$$

For a fixed  $\alpha \in \mathcal{S}_A$ , the sequence of finite-dimensional distributions  $\{W_{Y|X}^n(\cdot|\alpha_n)\}_{n \geq 1}$  can be viewed as a source whose resolvability is exactly  $\bar{H}(Y|\alpha)$  bits [1, Theorem 3]. Intuitively, the conditional resolvability of  $XY$  given  $X$  should equal the minimal  $h$  such that  $\bar{H}(Y|\alpha) \leq h$  "with probability 1  $P_X$ ," and the assertion that the conditional resolvability of  $XY$  given  $X$  is equal to the conditional sup-entropy rate should follow by investigating the relation between  $\bar{H}(Y|\alpha)$  and  $\bar{H}(Y|X)$ . However, since  $X$  is an arbitrary sequence of finite-dimensional distributions and  $\alpha_i$  need not be a prefix of  $\alpha_{i+1}$ , we do not have a probability measure on  $\mathcal{S}_A$ . Therefore, to prove achievability (and later also converse part), we first express  $\bar{H}(Y|\alpha)$  and  $\bar{H}(Y|X)$  as limits of quantities defined pointwise in  $n$ .

For every  $\epsilon > 0$ ,  $n \geq 1$  and  $a^n \in A^n$ , define

$$h_{n,\epsilon} \triangleq \inf \left\{ h: P_{XY}^n(a^n b^n: -\frac{1}{n} \log W_{Y|X}^n(b^n|a^n) > h) < \epsilon \right\},$$

$$g_{n,\epsilon}(a^n) \triangleq \inf \left\{ h: W_{Y|X}^n(b^n: -\frac{1}{n} \log W_{Y|X}^n(b^n|a^n) > h | X^n = a^n) < \epsilon \right\},$$

and let  $g_\epsilon$  be the lim sup in probability of  $g_{n,\epsilon}(a^n)$ ; that is,

$$g_\epsilon \triangleq \inf \left\{ h: \lim_{n \rightarrow \infty} P_X^n(a^n: g_{n,\epsilon}(a^n) > h) = 0 \right\}.$$

We have

*Lemma 3:* (a)  $\lim_{\epsilon \rightarrow 0} \limsup_{n \rightarrow \infty} h_{n,\epsilon} = \bar{H}(Y|X)$ .

(b) For  $\alpha \in \mathcal{S}_A$ ,  $\lim_{\epsilon \rightarrow 0} \limsup_{n \rightarrow \infty} g_{n,\epsilon}(\alpha_n) = \bar{H}(Y|\alpha)$ .

(c)  $\lim_{\epsilon \rightarrow 0} g_\epsilon = \bar{H}(Y|X)$ .

*Proof:* (a) Observe that

$$P_{XY}^n(a^n b^n: -\frac{1}{n} \log W_{Y|X}^n(b^n|a^n) > h_{n,\epsilon}) \leq \epsilon. \quad (11)$$

To see this, let  $h_k \searrow h_{n,\epsilon}$  and define the sets

$$\beta_k = \left\{ a^n b^n: -\frac{1}{n} \log W_{Y|X}^n(b^n|a^n) > h_k \right\},$$

$$\beta = \left\{ a^n b^n: -\frac{1}{n} \log W_{Y|X}^n(b^n|a^n) > h_{n,\epsilon} \right\}.$$

Since  $\beta_k \nearrow \beta$  (that is,  $\beta_k \subseteq \beta_{k+1}$  and  $\beta = \bigcup_{k=1}^{\infty} \beta_k$ ) (11) follows by the continuity of probability measures (continuity of  $P_{XY}^n$ ).

To prove (a), assume first that

$$\lim_{\epsilon \rightarrow 0} \limsup_{n \rightarrow \infty} h_{n,\epsilon} < \bar{H}(Y|X).$$

Then, by the monotonicity of  $h_{n,\epsilon}$  in  $\epsilon$ , for any  $\epsilon > 0$  and  $n$  sufficiently large

$$h_{n,\epsilon} < \bar{H}(Y|X) - \delta \quad (12)$$

for some  $\delta > 0$ , independent of  $\epsilon$ . But by definition of  $\bar{H}(Y|X)$ , there exists  $\epsilon_\delta > 0$  such that

$$P_{XY}^n(a^n b^n: -\frac{1}{n} \log W_{Y|X}^n(b^n|a^n) > \bar{H}(Y|X) - \delta) > \epsilon_\delta \quad \text{infinitely often in } n. \quad (13)$$

Choosing  $\epsilon < \epsilon_\delta$ , (13) together with (12) contradict (11).

Assume now that the reverse inequality holds, i.e.,

$$\lim_{\epsilon \rightarrow 0} \limsup_{n \rightarrow \infty} h_{n,\epsilon} > \bar{H}(Y|X).$$

Then there exist  $\delta > 0$ ,  $\epsilon_0 > 0$ , such that for every  $\epsilon < \epsilon_0$  there exists a subsequence  $\{n_k(\epsilon)\}_{k \geq 1}$ , denoted by  $J_\epsilon$ , such that

$$h_{m,\epsilon} > \bar{H}(Y|X) + \delta \quad \forall m \in J_\epsilon. \quad (14)$$

By definition of  $h_{m,\epsilon}$

$$P_{XY}^m(a^m b^m: -\frac{1}{m} \log W_{Y|X}^m(b^m|a^m) > h_{m,\epsilon} - \frac{\delta}{2}) \geq \epsilon \quad \forall m \in J_\epsilon.$$

But this contradicts the definition of  $\bar{H}(Y|X)$  since, according to (14),  $h_{m,\epsilon} - \delta/2 > \bar{H}(Y|X) + \delta/2$ . Thus part (a) follows.

(b) The proof follows exactly the lines of the proof of part (a), and is omitted.

(c) For every  $h > 0$ ,  $n \geq 1$ , define the sets

$$\beta_n(a^n, h) \triangleq \left\{ b^n: -\frac{1}{n} \log W_{Y|X}^n(b^n|a^n) > h \right\}$$

$$\gamma_n(h) \triangleq \left\{ a^n b^n: -\frac{1}{n} \log W_{Y|X}^n(b^n|a^n) > h \right\}$$

$$= \bigcup_{a^n \in A^n} a^n \times \beta_n(a^n, h).$$

Clearly,

$$g_{n,\epsilon}(a^n) = \inf \{h: W_{Y|X}^n(\beta_n(a^n, h)|a^n) < \epsilon\}, \quad (15)$$

$$h_{n,\epsilon} = \inf \{h: P_{XY}^n(\gamma_n(h)) < \epsilon\}$$

$$= \inf \left\{ h: \sum_{a^n \in A^n} P_X^n(a^n) W_{Y|X}^n(\beta_n(a^n, h)|a^n) < \epsilon \right\}. \quad (16)$$

Assume now that

$$\lim_{\epsilon \rightarrow 0} g_\epsilon < \bar{H}(Y|X). \quad (17)$$

Then  $g_\epsilon < \bar{H}(Y|X)$  for every  $\epsilon$ . Moreover, there exists  $\epsilon_1 > 0$  such that for every  $\nu < \epsilon_1$  there exists a subsequence  $\{n_k(\nu)\}_{k \geq 1}$  denoted by  $J_\nu$ , such that

$$g_\epsilon < h_{m,\nu} \quad \forall m \in J_\nu$$

and this inequality is strict, uniformly in  $\epsilon$ . Choose  $h'$  satisfying

$$\lim_{\epsilon \rightarrow 0} g_\epsilon < h' < h_{m,\nu} \quad \forall m \in J_\nu. \quad (18)$$

By (18) and the definition of  $g_\epsilon$ , for every  $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P_X^n(a^n: g_{n,\epsilon}(a^n) > h') = 0, \quad (19)$$

and by (16)

$h_{m,\nu}$

$$= \inf \left\{ h: \sum_{a^m: g_{m,\epsilon}(a^m) > h'} P_X^m(a^m) W_{Y|X}^m(\beta_m(a^m, h)|a^m) \right. \\ \left. + \sum_{a^m: g_{m,\epsilon}(a^m) \leq h'} P_X^m(a^m) W_{Y|X}^m(\beta_m(a^m, h)|a^m) < \nu \right\} \\ \forall m \in J_\nu. \quad (20)$$

Now, (19) implies that for  $m$  large enough the first sum in (20) vanishes. Setting  $h$  equal to  $h'$ , (15) implies that the second sum in (20) is upper bounded by  $\epsilon$ . Thus if we choose  $\epsilon < \nu$  and set  $h$  equal to  $h'$ , the inequality in (20) will be satisfied for  $m$  large enough. In turn, this implies that  $h_{m,\nu} \leq h'$  for large  $m \in J_\nu$ , contradicting (18). Therefore inequality (17) does not hold.

Assume that the reverse inequality

$$\lim_{\epsilon \rightarrow 0} g_\epsilon > \bar{H}(Y|X)$$

holds. Then there exists  $\epsilon_0 > 0$  such that for every  $\nu < \epsilon_0$

$$g_\nu > h_{n,\epsilon} \quad \text{for } n \text{ large enough,}$$

and the inequality is strict, uniformly in  $\epsilon$ . Choose  $h'$  satisfying

$$\limsup_{n \rightarrow \infty} h_{n,\epsilon} \leq \bar{H}(Y|X) < h' < g_\nu. \quad (21)$$

[The first inequality in (21) holds due to part (a).] By (21) and the definition of  $g_\nu$ , there exists  $\delta > 0$  such that

$$P_X^n(a^n: g_{n,\nu}(a^n) > h') > \delta \quad (22)$$

infinitely often in  $n$ . Again, by definition

$$h_{n,\epsilon} = \inf \left\{ h: \sum_{a^n: g_{n,\nu}(a^n) > h'} P_X^n(a^n) W_{Y|X}^n(\beta_n(a^n, h)|a^n) \right. \\ \left. + \sum_{a^n: g_{n,\nu}(a^n) \leq h'} P_X^n(a^n) W_{Y|X}^n(\beta_n(a^n, h)|a^n) < \epsilon \right\}. \quad (23)$$

By (15) and (22), for  $h = h'$  and some subsequence of the indices  $n$ , the first sum in (23) is larger than  $\delta\nu$ . Choose  $\epsilon < \delta\nu$ . Then according to (23) we must have  $h_{n,\epsilon} > h'$  infinitely often in  $n$ , contradicting (21). Thus part (c) follows. This completes the proof of the lemma.  $\square$

Observe that since  $g_\epsilon$  increases as  $\epsilon \rightarrow 0$ , part (c) of Lemma 3 implies

$$\lim_{n \rightarrow \infty} P_X^n(a^n: g_{n,\epsilon}(a^n) > \bar{H}(Y|X)) = 0 \quad \forall \epsilon > 0. \quad (24)$$

We are now ready to state and prove the achievability part.

*Theorem 1:* For every  $\epsilon > 0$

$$\sigma_\epsilon(\mathbf{XY}|X) \leq \bar{H}(Y|X).$$

*Proof:* Fix  $\nu > 0$ ,  $\delta > 0$ . Define the sets

$$\mathcal{A}_k(\delta) = \{a^k: g_{k,\nu}(a^k) < \bar{H}(Y|X) + \delta\}, \\ \mathcal{A}(\delta) = \{\alpha = \{a^k\}_{k \geq 1}: a^k \in \mathcal{A}_k(\delta)\}. \quad (25)$$

For every  $a^k \in \mathcal{A}_k(\delta)$ , define

$$\beta_k(a^k) = \left\{ b^k: -\frac{1}{n} \log W_{Y|X}^k(b^k|a^k) > g_{k,\nu}(a^k) \right\}. \quad (26)$$

Let  $\hat{b}^k$  be some default word in  $\beta_k(a^k)$ . For every  $a^k \in \mathcal{A}_k(\delta)$  define an approximate distribution

$$V_{Y|X}^k(b^k|a^k) = \begin{cases} W_{Y|X}^k(b^k|a^k), & b^k \notin \beta_k(a^k), \\ 0, & b^k \in \beta_k(a^k) \text{ and } b^k \neq \hat{b}^k, \\ W_{Y|X}^k(\beta_k(a^k)|a^k), & b^k = \hat{b}^k \end{cases}$$

and note that by definition  $W_{Y|X}^k(\beta_k(a^k)|a^k) < \nu$  and therefore the variational distance between  $W_{Y|X}^k(\cdot|a^k)$  and  $V_{Y|X}^k(\cdot|a^k)$  is at most  $\nu$ . Now,  $\{V_{Y|X}^k(\cdot|\alpha)\}_{\alpha \in \mathcal{A}(\delta)}$  is a collection of sources for which  $\bar{H}(Y|X) + \delta$  is an achievable resolution rate [1, Theorem 3]. That is, for every  $\alpha \in \mathcal{A}(\delta)$  and every  $\gamma > 0$ ,  $\epsilon > 0$ , there exists  $\tilde{V}_{Y|X}^n$  such that

$$\frac{1}{n} R(\tilde{V}_{Y|X}^n(\cdot|\alpha_n)) < \bar{H}(Y|X) + \delta + \gamma, \quad (27)$$

$$d(V_{Y|X}^n(\cdot|\alpha_n), \tilde{V}_{Y|X}^n(\cdot|\alpha_n)) < \epsilon \quad (28)$$

for all sufficiently large  $n$ . We claim that the achievability in (27), (28) is uniform over  $\mathcal{A}(\delta)$ . To see this, let  $n_0(\epsilon, \delta, \gamma, \alpha)$  be the minimum  $n_0$  such that (27) and (28), hold for every  $n > n_0$ . The objective is to show that

$$\sup_{\alpha \in \mathcal{A}(\delta)} n_0(\epsilon, \delta, \gamma, \alpha) < \infty.$$

Assume otherwise, Then there exists a sequence of elements of  $\mathcal{A}(\delta)$ ,  $\{\alpha(k)\}_{k \geq 1}$ , such that

$$\bar{n}_k = n_0(\epsilon, \delta, \gamma, \alpha(k))$$

is an increasing sequence. We construct a new element  $\tilde{\alpha}$  as

$$\tilde{\alpha}_n = \alpha_n(k+1) \quad \bar{n}_k \leq n < \bar{n}_{k+1}.$$

By construction, there is no  $n_0$  such that  $V_{Y|X}(\cdot|\tilde{\alpha})$  satisfies (27) and (28) for every  $n > n_0$ . On the other hand,  $\tilde{\alpha} \in \mathcal{A}(\delta)$ , which implies a contradiction.

Define

$$n_\delta = \sup_{\alpha \in \mathcal{A}(\delta)} n_0(\epsilon, \delta, \gamma, \alpha).$$

For every  $n > n_\delta$  and every  $a^n \in \mathcal{A}_n(\delta)$ , we approximate  $V_{Y|X}^n(\cdot|a^n)$  by  $\bar{H}(Y|X) + \delta + \gamma$  bits with error (in the variational sense) less than  $\epsilon$ , and leave all other conditional distributions unimplemented (or use a default distribution instead). We take  $P_X^n \tilde{V}_{Y|X}^n$  as an approximation to  $P_X^n W_{Y|X}^n$ . The resulting variational distance satisfies

$$\begin{aligned} & d(P_X^n W_{Y|X}^n, P_X^n \tilde{V}_{Y|X}^n) \\ & \leq d(P_X^n W_{Y|X}^n, P_X^n V_{Y|X}^n) + d(P_X^n V_{Y|X}^n, P_X^n \tilde{V}_{Y|X}^n) \\ & = \sum_{a^n} P_X^n(a^n) \sum_{b^n} |W_{Y|X}^n(b^n|a^n) - V_{Y|X}^n(b^n|a^n)| \\ & \quad + \sum_{a^n} P_X^n(a^n) \sum_{b^n} |V_{Y|X}^n(b^n|a^n) - \tilde{V}_{Y|X}^n(b^n|a^n)| \\ & \leq \sum_{a^n \in \mathcal{A}_n(\delta)} P_X^n(a^n) \sum_{b^n} |W_{Y|X}^n(b^n|a^n) - V_{Y|X}^n(b^n|a^n)| \\ & \quad + \sum_{a^n \in \mathcal{A}_n(\delta)} P_X^n(a^n) \sum_{b^n} |V_{Y|X}^n(b^n|a^n) - \tilde{V}_{Y|X}^n(b^n|a^n)| \\ & \quad + 4P_X^n(\mathcal{A}_n^c(\delta)) \\ & \leq \nu + \epsilon + 4P_X^n(\mathcal{A}_n^c(\delta)). \end{aligned}$$

By (24), the contribution of the last term vanishes as  $n \rightarrow \infty$ . Since  $\nu, \delta$  are arbitrary, the theorem is proved.  $\square$

We proceed now to show that a converse statement is also true: one cannot hope to approximate the joint distribution arbitrarily well by simulating the channel with less than  $\bar{H}(Y|X)$  bits per input sample. The proof of a converse part with respect to the  $\bar{d}$  distance makes use of continuity properties of the sup-entropy and conditional sup-entropy functions, stated in the next two lemmas. Lemma 4, whose proof can be found in [10], states that the sup-entropy function is uniformly lower semicontinuous (l.s.c.) with respect to  $\bar{d}$ . We use this property in Lemma 5 to show the lower semicontinuity of the conditional sup-entropy function.

*Lemma 4:* For every  $\gamma > 0$  there exists  $\epsilon > 0$  such that if  $P_Y, \tilde{P}_Y$  satisfy

$$\bar{d}(P_Y^n, \tilde{P}_Y^n) < \epsilon$$

for all sufficiently large  $n$ , then

$$\bar{H}(\tilde{Y}) > \bar{H}(Y) - \gamma.$$

*Proof:* Analogous to the proof of upper semicontinuity of the inf-entropy rate  $\underline{H}(Y)$  (cf. [10]).  $\square$

*Lemma 5:* Let  $P_{XY}$  be given. For every  $\gamma > 0$  there exists  $\epsilon > 0$  such that if  $\tilde{P}_{XY}$  satisfies

$$\limsup_{n \rightarrow \infty} \bar{d}(P_{XY}^n, \tilde{P}_{XY}^n) < \epsilon,$$

then

$$\bar{H}(\tilde{Y}|\tilde{X}) > \bar{H}(Y|X) - \gamma.$$

*Remark:* this is more than we need. In channel simulation the input process is fixed. Hence it is enough to show the lower semicontinuity of  $\bar{H}(\tilde{Y}|X)$  in  $\tilde{Y}$ , where  $X$  is held fixed and equal to the  $X$ -marginal of  $P_{XY}$ . However, with almost the same effort we can prove the stronger result in Lemma 5.

*Proof:* Assume the contrary, that exists  $\gamma > 0$  such that for every  $\epsilon > 0$

$$\bar{H}(Y|X) > \bar{H}(\tilde{Y}|\tilde{X}) + \gamma.$$

Then there exists  $\nu' > 0$  such that for every  $\nu < \nu'$

$$g_\nu > \bar{H}(\tilde{Y}|\tilde{X}) + \gamma. \quad (29)$$

Now, pick some  $\epsilon > 0$  and let

$$\bar{d}(P_{XY}^n, \tilde{P}_{XY}^n) < \epsilon^2.$$

Thus, for some  $\mu \in \mathcal{P}(P_{XY}^n, \tilde{P}_{XY}^n)$  we have

$$E_\mu d_n(a^n b^n, \tilde{a}^n \tilde{b}^n) < \epsilon^2.$$

Define the set

$$S_n \triangleq \{a^n \tilde{a}^n : E_\mu [d_n(a^n b^n, \tilde{a}^n \tilde{b}^n) | a^n \tilde{a}^n] < \epsilon\}.$$

The  $\bar{d}$  distance can be lower bounded as

$$\begin{aligned} \epsilon^2 & > E_\mu d_n(a^n b^n, \tilde{a}^n \tilde{b}^n) \\ & \geq \sum_{a^n \tilde{a}^n \in S_n^c} \sum_{b^n \tilde{b}^n} d_n(a^n b^n, \tilde{a}^n \tilde{b}^n) \mu(a^n b^n, \tilde{a}^n \tilde{b}^n) \\ & = \sum_{S_n^c} E_\mu [d_n(a^n b^n, \tilde{a}^n \tilde{b}^n) | a^n \tilde{a}^n] \mu(a^n \tilde{a}^n) \geq \epsilon \mu(S_n^c), \end{aligned}$$

which implies

$$\mu(S_n) \geq 1 - \epsilon. \quad (30)$$

By definition of  $S_n$ , for every sequence of pairs  $a^n \tilde{a}^n \in S_n$ , the sequence of  $\tilde{Y}^n$  distributions  $\{\tilde{W}_{Y|X}^n(\cdot|\tilde{a}^n)\}_{n \geq 1}$  is an  $\epsilon$ -approximation (in  $\bar{d}$  distance) of the sequence of  $\tilde{Y}^n$  distributions  $\{W_{Y|X}^n(\cdot|a^n)\}_{n \geq 1}$ .

Define the sets

$$\hat{D}_n^\nu \triangleq \{a^n \tilde{a}^n : g_{n,\nu}(a^n) > \bar{H}(\tilde{Y}|\tilde{X}) + \gamma\},$$

$$\hat{G}_n^\nu \triangleq \{a^n \tilde{a}^n : \bar{g}_{n,\nu}(\tilde{a}^n) < \bar{H}(\tilde{Y}|\tilde{X}) + \delta\}.$$

By definition of  $g_\nu$  and by (29), we are assured that  $\mu(\hat{D}_n^\nu) \geq \rho$  infinitely often, for some fixed  $\rho > 0$  indepen-

dent of  $\epsilon$ . Also, by the characterization of  $\bar{H}(\tilde{Y}|\tilde{X})$  in terms of  $\tilde{g}_\nu$ , we are assured that  $\mu(\hat{G}_n^\nu) \rightarrow 1$  as  $n$  increases. Define

$$G_n^\nu = S_n \cap \hat{G}_n^\nu \cap \hat{D}_n^\nu.$$

Clearly

$$\mu(G_n^\nu) \geq \rho - \epsilon \text{ infinitely often } n. \quad (31)$$

Next, for every  $a^n \tilde{a}^n \in G_n^\nu$  define

$$\beta_n(\tilde{a}^n) = \left\{ b^n: -\frac{1}{n} \log \tilde{W}_{Y|X}^n(b^n|\tilde{a}^n) > \tilde{g}_{n,\nu}(\tilde{a}^n) \right\}$$

and observe that

$$\tilde{W}_{Y|X}^n(\beta_n(\tilde{a}^n)|\tilde{a}^n) \leq \nu.$$

Let  $\hat{b}^n$  be some default word in  $\beta_n(\tilde{a}^n)$ . For every  $a^n \tilde{a}^n \in G_n^\nu$ , let  $V_{Y|X}^n(\cdot|\tilde{a}^n)$  be an approximation for  $\tilde{W}_{Y|X}^n(\cdot|\tilde{a}^n)$

$$V_{Y|X}^n(b^n|\tilde{a}^n) = \begin{cases} \tilde{W}_{Y|X}^n(b^n|a^n), & \text{if } b^n \notin \beta_n(\tilde{a}^n), \\ \tilde{W}_{Y|X}^n(\beta_n(\tilde{a}^n)|\tilde{a}^n), & b^n = \hat{b}^n, \\ 0, & \text{otherwise,} \end{cases}$$

and note that the  $\bar{d}$  distance between  $V_{Y|X}^n(\cdot|\tilde{a}^n)$  and  $\tilde{W}_{Y|X}^n(\cdot|\tilde{a}^n)$  is at most  $\nu$ . We denote by  $Y$  the output process of the channel  $V_{Y|X}$  with input  $\tilde{X}$ . Finally, let  $J$  be the sequence of indices indicated by (31), and let  $\alpha \tilde{\alpha} = \{a^m \tilde{a}^m\}_{m \in J}$  be a sequence of pairs,  $a^m \tilde{a}^m \in G_m^\nu$ ,  $m \in J$ . For this sequence,  $\{V_{Y|X}^m(\cdot|\tilde{a}^m)\}_{m \in J}$  is a source for which  $\bar{H}(\tilde{Y}|\tilde{X}) + \delta$  is an achievable resolution rate. Thus

$$\bar{H}(\tilde{Y}|\tilde{\alpha}) \leq \bar{H}(\tilde{Y}|\tilde{X}) + \delta$$

and, since  $a^m \tilde{a}^m \in \hat{D}_m^\nu$ ,

$$\bar{H}(Y|X) \geq \limsup_{m \rightarrow \infty} g_{m,\nu}(a^m) > \bar{H}(\tilde{Y}|\tilde{X}) + \gamma,$$

which implies that

$$\bar{H}(Y|X) \geq \bar{H}(\tilde{Y}|\tilde{\alpha}) + \gamma - \delta. \quad (32)$$

On the other hand, the distance between  $V_{Y|X}^m(\cdot|\tilde{a}^m)$  and  $W_{Y|X}^m(\cdot|a^m)$  is at most  $\epsilon^2 + \nu$ , which by Lemma 4 implies

$$\bar{H}(\tilde{Y}|\tilde{\alpha}) \geq \bar{H}(Y|\alpha) - f(\epsilon^2 + \nu),$$

where  $f(u) \rightarrow 0$  as  $u \rightarrow 0$  and is independent of  $\alpha$ . Since  $\delta$  and  $\nu$  are arbitrary, this contradicts inequality (32).  $\square$

We are now in a position to state and prove a converse part with respect to the  $\bar{d}$  distance.

*Theorem 2:*

$$\lim_{\epsilon \rightarrow 0} \bar{\sigma}_\epsilon(XY|X) \geq \bar{H}(Y|X).$$

*Proof:* Assume that

$$\bar{\sigma}(XY|X) < \bar{H}(Y|X).$$

Then there exist  $\gamma > 0$  and a sequence  $\{\tilde{W}_{Y|X}^n\}$  with resolution  $\bar{H}(Y|X) - \gamma$ , such that

$$\lim_{n \rightarrow \infty} \bar{d}(P_X^n W_{Y|X}^n, P_X^n \tilde{W}_{Y|X}^n) = 0.$$

But if  $\tilde{W}_{Y|X}^n$  has resolution  $\bar{H}(Y|X) - \gamma$ , then

$$\bar{H}(\tilde{Y}|X) \leq \bar{H}(Y|X) - \gamma$$

contradicting Lemma 5.  $\square$

Our main result is the following.

*Theorem 3:*

$$\sigma(XY|X) = \bar{\sigma}(XY|X) = \bar{H}(Y|X).$$

*Proof:* The theorem is a direct consequence of (9), Theorem 1, and Theorem 2.  $\square$

In contrast to entropy rates, sup-entropy rates do not obey a simple chain rule; in general

$$\bar{H}(X, Y) \leq \bar{H}(X) + \bar{H}(Y|X) \quad (33)$$

and it is possible to construct sequences of joint distributions for which strict inequality holds in (33). Thus, if we aim at approximating the joint distribution, the number of bits per sample we save by the fact that the  $X$ -marginal is already implemented can be strictly less than  $\bar{H}(X)$ . (Recall Example 3 in Section 1.)

The accuracy measure used in [3], [4] is the  $\bar{d}$  distance between channels. In our setting it is given by

$$\begin{aligned} \bar{d}(W_{Y|X}, \tilde{W}_{Y|X}) \\ = \limsup_{n \rightarrow \infty} \max_{a^n \in A^n} \bar{d}(W_{Y|X}^n(\cdot|a^n), \tilde{W}_{Y|X}^n(\cdot|a^n)). \end{aligned} \quad (34)$$

In a completely analogous way one can define the variational distance between channels as

$$\begin{aligned} d(W_{Y|X}, \tilde{W}_{Y|X}) \\ = \limsup_{n \rightarrow \infty} \max_{a^n \in A^n} d(W_{Y|X}^n(\cdot|a^n), \tilde{W}_{Y|X}^n(\cdot|a^n)). \end{aligned} \quad (35)$$

[Note that both (34) and (35) are pseudometrics rather than metrics.] Now the minimum number of random bits required to approximate a channel in the sense of (34), (35) is actually the worst-case complexity over all input sequences, and is equal to the supremum of  $\bar{H}(Y|\alpha)$ , defined in (10), over  $\mathcal{S}_A$ . When simulating the channel with a fixed input distribution, this turns out to be a pessimistic bound since it brings into account all input sequences, including those that have vanishing probabilities according to  $P_X$  and hence do not affect the simulation accuracy when measured as distance between joint input-output distributions. The following example illustrates this. Let  $P_X^n$  be a uniform distribution on  $\{0, 1\}^n$ , pick an element  $\alpha \in \mathcal{S}_A$ , and let  $W_{Y|X}^n$  be the sequence of conditional distributions defined as

$$W_{Y|X}^n(\cdot|a^n) = \begin{cases} \text{uniform distribution on } \{0, 1\}^n, \\ \text{if } a^n = \alpha_n, \\ \text{identity channel,} \\ \text{otherwise.} \end{cases} \quad (36)$$

Then,  $\bar{H}(Y|X) = 0$  whereas  $\sup_{\alpha \in \mathcal{S}_A} \bar{H}(Y|\alpha) = 1$ .

We can define now the *channel sup-entropy* (cf. [4]) as

$$\bar{H}_c(W_{Y|X}) \triangleq \sup_X \bar{H}(Y|X) \quad (37)$$

where the supremum is taken over all sequences of finite dimensional input distributions.  $\bar{H}_c(W_{Y|X})$  is the minimal rate of pure random bits required by the most efficient channel simulation scheme in order to approximate arbitrarily well any joint input-output distribution. The channel sup-entropy serves as a measure of channel worst-case complexity, and it is actually equal to the worst-case complexity over all input sequences; i.e.,

$$\bar{H}_c(W_{Y|X}) = \sup_{\alpha \in \mathcal{S}_A} \bar{H}(Y|\alpha).$$

To see this observe that, on the one hand, the maximization in (37) is over all input processes, including deterministic inputs, and hence  $\bar{H}_c(W_{Y|X}) \geq \sup_{\alpha \in \mathcal{S}_A} \bar{H}(Y|\alpha)$ . On the other hand,  $\bar{H}_c(W_{Y|X}) \leq \sup_{\alpha \in \mathcal{S}_A} \bar{H}(Y|\alpha)$  since otherwise parts (b) and (c) of Lemma 3 would imply that there exists  $\alpha' \in \mathcal{S}_A$  such that  $\bar{H}(Y|\alpha') > \sup_{\alpha \in \mathcal{S}_A} \bar{H}(Y|\alpha)$ .

If the channel  $W_{Y|X}$  is stationary and ergodic and we change the definition in (37) so that the supremum is taken over all stationary input processes, then  $\bar{H}_c(W_{Y|X})$  coincides with the channel entropy as defined by Neuhoff and Shields in [4]. Since we do not restrict the approximation to be of a primitive channel type, our converse result together with the results in [4] imply that for  $\bar{d}$ -continuous CABI channels one can restrict the simulator to be a primitive channel scheme without losing efficiency (in the sense of required randomness). This is not the case for the general stationary channel; as shown in [3, Theorem 1], primitive channels cannot approximate in the  $\bar{d}$  distance (and hence also in the variational distance) channels that are not  $\bar{d}$ -continuous and CABI. Our results indicate that although the sense of approximation and the method of simulation change, the channel entropy is still the number of pure random bits per sample required for accurate simulation.

#### IV. CONDITIONAL RESOLVABILITY AND CODING WITH SIDE INFORMATION

In this section we provide a connection between conditional resolvability and source coding. In [1] it is shown that the resolvability of a source is equal to its minimal achievable block coding rate. A natural extension that fits nicely in the present setup is the connection between the conditional resolvability and coding with side information. Coding with side information is a special case of the canonical general problem of encoding of correlated sources. The first results on encoding of correlated sources are due to Slepian and Wolf [11], who treated the case where  $XY$  is an iid sequence of pairs  $\{X_i Y_i\}$ . In [12], Cover generalized the results of Slepian and Wolf to the case where  $X, Y$  are jointly ergodic processes. The analysis presented here provides a framework for further generalization of those results to nonergodic sources. We start with several definitions.

*Definition 9:* A  $(n, \exp(nR), \lambda)$  source code for  $Y^n$  with

*side information*  $Y^n$  is an encoder map

$$f(\cdot): B^n \rightarrow \{1, 2, \dots, \exp(nR)\}$$

and a decoder map

$$\phi(\cdot, \cdot): A^n \times \{1, 2, \dots, \exp(nR)\} \rightarrow A^n \times B^n$$

with probability of error less than or equal to  $\lambda$ ; i.e.,

$$P_e^n \triangleq P_{XY}^n(\phi(X^n, f(Y^n)) \neq (X^n, Y^n)) \leq \lambda.$$

*Definition 10:* Let  $P_{XY} = \{P_{XY}^n\}_{n \geq 1}$  be given.  $R$  is an  $\epsilon$ -achievable source coding rate for  $Y$  with side information  $X$  if for every  $\gamma > 0$  and sufficiently large  $n$  there exists a  $(n, \exp\{n(R + \gamma)\}, \epsilon)$  source code for  $Y^n$  with side information  $X^n$ .  $R$  is an achievable source coding rate for  $Y$  with side information  $X$  if it is  $\epsilon$ -achievable for every  $\epsilon > 0$ .

*Definition 11:*  $T(Y|X)$  denotes the minimal achievable source coding rate for  $Y$  with side information  $X$ .

We turn now to prove that the minimal achievable source coding rate for  $Y$  with side information  $X$  is equal to the conditional sup-entropy rate  $\bar{H}(Y|X)$ . We will state and prove first the direct (achievability) part and then the converse.

*Theorem 4:*

$$T(Y|X) \leq \bar{H}(Y|X).$$

*Proof:* We will use a random binning argument, similar to that of Slepian and Wolf in [11]. Thus, independently assign every  $b^n \in B^n$  to one of the  $\exp[n\bar{H}(Y|X) + n\gamma]$  indices according to a uniform distribution  $U$  on  $\{1, 2, \dots, \exp[n\bar{H}(Y|X) + n\gamma]\}$ . Note that every index can have—and usually will have—more than one data block assigned to it. These assignments form the (random) encoder map  $f(\cdot)$ . Assume  $f(\cdot)$  is known to the encoder and decoder.

We turn to the construction of the decoder map. Fix  $\nu > 0$ ,  $\delta > 0$ , and let  $\{\mathcal{A}_n(\delta)\}_{n \geq 1}$  be a sequence of sets as defined in (25). For every  $a^n \in \mathcal{A}_n(\delta)$ , define

$$Z_n(a^n) = \left\{ b^n: -\frac{1}{n} \log W_{Y|X}^n(b^n|a^n) < g_{n,\nu}(a^n) + \delta \right\}.$$

By definition, the following holds

$$W_{Y|X}^n(Z_n(a^n)|a^n) \geq 1 - \nu,$$

$$\frac{1}{n} \log |Z_n(a^n)| < g_{n,\nu}(a^n) + \delta. \quad (38)$$

The sets  $Z_n(a^n)$  will play the role of the jointly typical fans in the Slepian–Wolf coding.

For every realization  $a^n b^n$  emitted from the source  $P_{XY}^n$ , the receiver gets  $a^n$  and the index  $j = f(b^n)$  assigned to the specific  $b^n$  string by the encoder map  $f(\cdot)$ . The decoder assigns  $\phi(a^n, j) = (a^n, b^n)$  only if  $a^n \in \mathcal{A}_n(\delta)$  and there is only one  $b^n \in Z_n(a^n)$  such that  $f(b^n) = j$ . Otherwise, it declares an error.

Define the three error events

$$E_0 = \{X^n \notin \mathcal{A}_n(\delta)\},$$

$$E_1 = \{X^n \in \mathcal{A}_n(\delta) \text{ and } Y^n \notin Z_n(X^n)\},$$

$$E_2 = \{\exists \hat{b}^n \neq Y^n: f(\hat{b}^n) = f(Y^n) \text{ and } \hat{b}^n \in Z_n(X^n)\}.$$

The probability of error associated with a given realization of  $f(\cdot)$  satisfies

$$P_e^n \leq P_X^n(E_0) + P_{XY}^n(E_1) + P_{XY}^n(E_2), \quad (39)$$

where only  $E_2$  depends on the encoder map  $f(\cdot)$ . By (24) we know that  $P_X^n(E_0) \rightarrow 0$  as  $n \rightarrow \infty$ . By (38)

$$P_{XY}^n(E_1) = \sum_{a^n \in \mathcal{A}_n(\delta)} P_X^n(a^n) W_{Y|X}^n(Z_n^c(a^n)|a^n) \leq \nu. \quad (40)$$

It remains to bound the average of the last term in (39) over all encoder realizations. Thus

$$\begin{aligned} & E_U P_{XY}^n(E_2) \\ &= \sum_{a^n \in \mathcal{A}_n(\delta)} \sum_{b^n \in B^n} P_{XY}^n(a^n, b^n) P(\exists \hat{b}^n \neq b^n \\ & \quad f(\hat{b}^n) = f(b^n) \text{ and } \hat{b}^n \in Z_n(a^n)) \\ &= \sum_{a^n \in \mathcal{A}_n(\delta)} \sum_{b^n \in B^n} P_{XY}^n(a^n, b^n) \\ & \quad \cdot \sum_{\hat{b}^n \neq b^n, \hat{b}^n \in Z_n(a^n)} U(f(\hat{b}^n) = f(b^n)) \\ &= \sum_{a^n \in \mathcal{A}_n(\delta)} \sum_{b^n \in B^n} P_{XY}^n(a^n, b^n) \\ & \quad \cdot \exp[-n\bar{H}(Y|X) + n\gamma] |Z_n(a^n)| \\ &\leq \sum_{a^n \in \mathcal{A}_n(\delta)} P_X^n(a^n) \exp\left\{-n[\bar{H}(Y|X) + \gamma\right. \\ & \quad \left.- g_{n,\nu}(a^n) - \delta]\right\} \\ &\leq \exp(n2\delta - n\gamma), \end{aligned} \quad (41)$$

where the last inequality follows from the definition of  $\mathcal{A}_n(\delta)$ . Since  $\delta, \nu$  are arbitrary, the overall probability of error averaged over all realizations of  $f$  can be made arbitrarily small. Thus for every  $\gamma > 0, \epsilon > 0$  there must exist at least one sequence of encoder maps with probability of error not exceeding  $\epsilon$  and rate not exceeding  $\bar{H}(Y|X) + \gamma$ . This proves the theorem.  $\square$

*Theorem 5:*

$$T(Y|X) \geq \bar{H}(Y|X)$$

*Proof:* Assume that  $R$  is an  $\epsilon^2$ -achievable source coding rate for  $Y$  with side information  $X$ . We show that it is also a  $4\epsilon$ -achievable resolution rate of  $XY$  given  $X$ .

Fix  $\gamma > 0$ . Let  $G_n$  be the set of all  $a^n$  sequences such that  $W_{Y|X}^n(D_n|a^n) \leq 1 - \epsilon$  for every set  $D_n \subset B^n$  with cardinality  $|D_n| \leq \exp(nR + n\gamma)$ . The probability of error of any  $(n, \exp(nR + n\gamma), \epsilon^2)$  source code with side information satisfies

$$\epsilon^2 > P_e^n \geq P_X^n(G_n)\epsilon$$

(such a code exists for  $n$  large enough, by assumption). Hence

$$P_X^n(G_n^c) \geq 1 - \epsilon. \quad (42)$$

With every  $a^n \in G_n^c$  we can associate a set  $D_n(a^n) \subset B^n$  which satisfies

$$|D_n(a^n)| \leq \exp(nR + n\gamma),$$

$$W_{Y|X}^n(D_n(a^n)|a^n) \geq 1 - \epsilon.$$

Thus, repeating the arguments in [1, Theorem 1, converse part], for  $n$  large enough (depending only on  $\epsilon, \gamma$ ) and for every  $a^n \in G_n^c$ , we can construct a distribution  $\tilde{W}_{Y|X}^n(\cdot|a^n)$  which satisfies the conditions (i) the resolution of  $\tilde{W}_{Y|X}^n(\cdot|a^n)$  is at most  $n(R + 3\gamma)$ ; (ii)  $d(\tilde{W}_{Y|X}^n(\cdot|a^n), W_{Y|X}^n(\cdot|a^n)) \leq 3\epsilon$ . Since the probability of  $G_n$  is upper bounded by  $\epsilon$ , this implies that we can construct a channel approximation  $\tilde{W}$  such that (i) the resolution of  $\tilde{W}_{Y|X}$  is at most  $n(R + 3\gamma)$ ; (ii)  $d(P_X^n \tilde{W}_{Y|X}^n, P_X^n W_{Y|X}^n) \leq 4\epsilon$ . Since  $\gamma$  is arbitrary, the proof is complete.  $\square$

## V. EXAMPLES

In this section we solve Examples 1–3 given in the Introduction.

*Example 1:* Memoryless binary symmetric channel (BSC). Here we have

$$\begin{aligned} -\frac{1}{n} \log W_{Y|X}^n(Y^n|X^n) &= -\frac{1}{n} \log \prod_{i=1}^n W(Y_i|X_i) \\ &= -\frac{1}{n} \sum_{i=1}^n \log P_Z(Z_i). \end{aligned}$$

Since  $Z_i$  are iid, as  $n \rightarrow \infty$  the last term converges to  $h(\alpha) = -\alpha \log \alpha - (1 - \alpha) \log(1 - \alpha)$ , with probability 1  $P_{XY}$ . Thus  $\bar{H}(Y|X) = \underline{H}(Y|X) = h(\alpha)$  independently of the input statistics. Clearly, this channel is encompassed by the class of channels treated in [4].

*Example 2:* Binary channel which is not  $\bar{d}$ -continuous. Here we consider a binary channel where the distribution of the noise sequence depends on the input sequence via the relation

$$P_{Z_n}(1) = \frac{1}{n} \sum_{i=1}^n X_i.$$

This channel is not  $\bar{d}$ -continuous since it has an infinite input memory and for every  $N < n$  all the  $N$ -blocks of the past share the same weight in determining  $P_{Z_n}(1)$ . Clearly

$$-\frac{1}{n} \log W_{Y|X}^n(Y^n|X^n) = -\frac{1}{n} \sum_{i=1}^n \log W_{Y|X}(Y_i|X^i).$$

We now consider two cases:

(a)  $\{X_i\}$  is iid Bernoulli with  $P_X(1) = \theta$ . We claim that

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \sum_{i=1}^n \log W_{Y|X}(Y_i|X^i) = h(\theta)$$

in probability  $P_{XY}$ , (43)

and thus,  $\bar{H}(Y|X) = \underline{H}(Y|X) = h(\theta)$ . The proof of (43) is given in the Appendix.

(b)  $\{X_i\}$  is a deterministic process. Define

$$\theta_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Then, conditioned on the input,  $\{Z_i\}$  is a sequence of independent random variables with parameter  $P_{Z_i}(1) = \theta_i$ . Application of the law of large numbers yields

$$\begin{aligned} \bar{H}(Y|X) &= \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n h(\theta_i), \\ \underline{H}(Y|X) &= \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n h(\theta_i). \end{aligned}$$

*Example 3:* Here we solve an example of a channel and an input process for which (33) holds with strict inequality. The idea is to construct a nonergodic input process  $X$  and an input-dependent noise process  $Z$  for which  $\bar{H}(X)$  and  $\bar{H}(Y|X)$  correspond to separate ergodic modes of  $X$  and hence "cannot appear together." For convenience, we repeat the details of the example here:  $X$  is an iid Bernoulli process with parameter  $\theta$ , where  $\theta$  is a random variable taking values in  $\{1/4, 1/2\}$  with *a priori* probabilities  $1/2, 1/2$ . The distribution of the channel noise satisfies

$$P_{Z_n}(1) = f\left(\frac{1}{n} \sum_{i=1}^n X_i\right),$$

where

$$f(u) = 2u \bmod 1 \triangleq \begin{cases} 2u, & u \in \left[0, \frac{1}{2}\right], \\ 2u - 1, & u \in \left(\frac{1}{2}, 1\right]. \end{cases}$$

To evaluate  $\bar{H}(X, Y)$ , we first evaluate the sup-entropy of the joint process  $XY$  conditioned on  $\theta$ . Clearly,  $\bar{H}(X|\theta = 1/2) = h(1/2)$ , and

$$\begin{aligned} \bar{H}(X|\theta = 1/2) &\leq \bar{H}(X, Y|\theta = 1/2) \\ &\leq \bar{H}(Y|X, \theta = 1/2) + \bar{H}(X|\theta = 1/2). \end{aligned} \quad (44)$$

Conditioned on  $\theta = 1/2$ ,  $(1/n)\sum_{i=1}^n X_i$  converges with probability 1 to  $1/2$ . By definition of  $f(\cdot)$ , this implies that  $f(\sum_{i=1}^n X_i/n)$  converges with probability 1 to the set  $\{0, 1\}$ , i.e.,

$$P_X\left(\lim_{n \rightarrow \infty} \min_{i \in \{0,1\}} (P_{Z_n}(1) - i) = 0\right) = 1.$$

Using the fact that  $h(1) = h(0) = 0$  and following exactly the lines of Example 2(a), we arrive to the conclusion that  $\bar{H}(Y|X, \theta = 1/2) = 0$ . Hence by (44)

$$\bar{H}(X, Y|\theta = 1/2) = h(1/2) = 1. \quad (45)$$

As for  $\theta = 1/4$ ,

$$\begin{aligned} &-\frac{1}{n} \log P_{XY}^n(X^n Y^n|\theta = 1/4) \\ &= -\frac{1}{n} \sum_{i=1}^n \log P_X(X_i|\theta = 1/4) \\ &\quad -\frac{1}{n} \sum_{i=1}^n \log W_{Y|X}(Y_i|X^i, \theta = 1/4). \end{aligned} \quad (46)$$

Conditioned on  $\theta = 1/4$ ,  $X$  is iid with parameter  $1/4$ . Thus, as  $n \rightarrow \infty$ , the first term of the right-hand side of (46) converges to  $h(1/4)$  with probability 1  $P_X$ . Again, following exactly the lines of Example 2(a) one can verify that

$$\begin{aligned} \lim_{n \rightarrow \infty} -\frac{1}{n} \sum_{i=1}^n \log W_{Y|X}(Y_i|X^i, \theta = 1/4) \\ = h(f(1/4)) = h(1/2) = 1 \quad \text{in probability } P_{XY}. \end{aligned} \quad (47)$$

Combining (46) and (47) we have

$$\bar{H}(X, Y|\theta = 1/4) = \underline{H}(X, Y|\theta = 1/4) = h(1/4) + 1. \quad (48)$$

Therefore, accurate simulation of the joint process  $XY$  at mode  $\theta = 1/2$  requires asymptotically 1 random bit per sample, whereas at mode  $\theta = 1/4$  it requires  $h(1/4) + 1$  random bits per sample. Since both modes have positive probabilities, the worst-case complexity of the joint process  $XY$  is equal to the maximum of the worst-case complexities  $\bar{H}(X, Y|\theta)$  over  $\theta$ . [Note that this is the essence of Lemma 3; see the discussion following (10).] Thus

$$\bar{H}(X, Y) = h(1/4) + 1. \quad (49)$$

We turn now to evaluate  $\bar{H}(X)$  and  $\bar{H}(Y|X)$ . By Example 1,

$$\begin{aligned} \bar{H}(X) &= \max\{\bar{H}(X|\theta = 1/4), \bar{H}(X|\theta = 1/2)\} \\ &= 1 \end{aligned}$$

and by Example 2(a)

$$\bar{H}(Y|X) = \max\{\bar{H}(Y|X, \theta = 1/4), \bar{H}(Y|X, \theta = 1/2)\} = 1.$$

Therefore, with (49) we obtain

$$h(1/4) + 1 = \bar{H}(X, Y) < \bar{H}(Y|X) + \bar{H}(X) = 2.$$

#### APPENDIX

Here we prove (43). To this end, we show that for every  $\delta > 0$ ,  $\gamma > 0$ ,

$$P_{XY}^n\left(\left|-\frac{1}{n} \sum_{i=1}^n \log W_{Y|X}(Y_i|X^i) - h(\theta)\right| > 2\delta\right) < \gamma \quad (50)$$

for sufficiently large  $n$ . Indeed, pick  $\delta > 0$ ,  $\epsilon > 0$ . A simple union bound yields

$$\begin{aligned} & P_{XY}^n \left( \left| -\frac{1}{n} \sum_{i=1}^n \log W_{Y|X}(Y_i|X^i) - h(\theta) \right| > 2\delta \right) \\ & \leq P_{XY}^n \left( -\frac{1}{n} \sum_{i=1}^{n_0-1} \log W_{Y|X}(Y_i|X^i) \right. \\ & \quad \left. + \left| -\frac{1}{n} \sum_{i=n_0}^n \log W_{Y|X}(Y_i|X^i) - h(\theta) \right| > 2\delta \right) \\ & \leq P_{XY}^n \left( -\frac{1}{n} \sum_{i=1}^{n_0-1} \log W_{Y|X}(Y_i|X^i) \geq \delta \right) \\ & \quad + P_{XY}^n \left( \left| -\frac{1}{n} \sum_{i=n_0}^n \log W_{Y|X}(Y_i|X^i) - h(\theta) \right| > \delta \right). \end{aligned} \quad (51)$$

For fixed  $\delta$  and  $n_0$ , the first term in the far right-hand side of (51) can be made arbitrarily small by letting  $n \rightarrow \infty$ . We will choose  $n_0$  later. Define the set

$$G(n_0, n) = \left\{ x^n : \frac{1}{j} \sum_{i=1}^j x_i \in (\theta - \epsilon, \theta + \epsilon) \quad n_0 \leq j \leq n \right\},$$

that is,  $G(n_0, n)$  is the set of all sequences  $x^n \in \{0, 1\}^n$  that exhibit “good” behavior in the interval  $[n_0, n]$  in the sense that their empirical mean up to any  $j \in [n_0, n]$  is close to  $\theta$ . Since  $(1/n)\sum_{i=1}^n X_i$  converges (as  $n \rightarrow \infty$ ) to  $\theta$  with probability 1 ( $P_X$ ), there exists  $n_0 = n_0(\epsilon)$  such that

$$P_X^n(G(n_0, n)) > 1 - \epsilon \quad \forall n > n_0. \quad (52)$$

To see this, define

$$G = \left\{ x \in \{0, 1\}^{\mathbb{N}} : \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i \in (\theta - \epsilon, \theta + \epsilon) \right\}.$$

Clearly,

$$P_X(G) = 1. \quad (53)$$

We now decompose  $G$  into a sequence of increasing sets as follows

$$G(n_0) = \left\{ x \in \{0, 1\}^{\mathbb{N}} : \frac{1}{n} \sum_{i=1}^n x_i \in (\theta - \epsilon, \theta + \epsilon) \quad \forall n \geq n_0 \right\}.$$

Then, by the definition of these sets

$$G(n_0) \subseteq G(n_0 + 1) \subseteq \dots \quad (54)$$

and

$$G = \bigcup_{n_0=1}^{\infty} G(n_0) = \lim_{n_0 \rightarrow \infty} G(n_0).$$

Thus (53) and the continuity of probability measures imply

$$\lim_{n_0 \rightarrow \infty} P_X(G(n_0)) = 1 \quad (55)$$

and due to (54), the convergence in (55) is monotonic. The sequence of sets  $G(n_0)$  decomposes  $G$  from below. Now for any

$n_0$ , we decompose  $G(n_0)$  from above with the sets  $G(n_0, n)$  “lifted” to  $\{0, 1\}^{\mathbb{N}}$ . Thus define

$$G'(n_0, n) = \left\{ x \in \{0, 1\}^{\mathbb{N}} : \frac{1}{j} \sum_{i=1}^j x_i \in (\theta - \epsilon, \theta + \epsilon) \quad n_0 \leq j \leq n \right\}.$$

For these sets we have

$$G'(n_0, n) \supseteq G'(n_0, n+1) \supseteq \dots$$

and

$$G(n_0) = \bigcap_{n > n_0} G'(n_0, n) = \lim_{n \rightarrow \infty} G'(n_0, n)$$

and thus continuity of probability measures again implies

$$\lim_{n \rightarrow \infty} P_X(G'(n_0, n)) = \lim_{n \rightarrow \infty} P_X^n(G(n_0, n)) = P_X(G(n_0)) \quad (56)$$

where the convergence is monotonically decreasing [i.e.,  $P_X^n(G(n_0, n))$  decreases as  $n$  increases]. Equation (52) is implied by (55) and (56).

We choose  $n_0$  so that (52) is satisfied. The second term in the right-hand side of (51) can be bounded as follows

$$\begin{aligned} & P_{XY}^n \left( \left| -\frac{1}{n} \sum_{i=n_0}^n \log W_{Y|X}(Y_i|X^i) - h(\theta) \right| > \delta \right) \\ & \leq \sum_{a^n \in G(n_0, n)} P_{XY}^n \left( \left| -\frac{1}{n} \sum_{i=n_0}^n \log W_{Y|X}(Y_i|X^i) \right. \right. \\ & \quad \left. \left. - h(\theta) \right| > \delta \mid X^n = a^n \right) P_X^n(X^n = a^n) \\ & \quad + P_X^n(G^c(n_0, n)) \\ & = \sum_{a^n \in G(n_0, n)} P_{XY}^n \left( \left| \frac{1}{n} \sum_{i=n_0}^n (-\log W_{Y|X}(Y_i|X^i)) \right. \right. \\ & \quad \left. \left. - \frac{nh(\theta)}{n - n_0 + 1} \right| > \delta^2 \mid X^n = a^n \right) \\ & \quad \cdot P_X^n(X^n = a^n) + P_X^n(G^c(n_0, n)). \end{aligned} \quad (57)$$

Note that for  $a^n \in G(n_0, n)$  and  $n_0 \leq i \leq n$ ,

$$\begin{aligned} \min \{h(\theta - \epsilon), h(\theta + \epsilon)\} & \leq -E[\log W_{Y|X}(Y_i|X^i) \mid X^n = a^n] \\ & \leq \max \{h(\theta - \epsilon), h(\theta + \epsilon)\} \end{aligned}$$

and therefore there exists a function  $q(\epsilon)$ ,  $q(\epsilon) \rightarrow 0$  as  $\epsilon \rightarrow 0$ , such that for  $n$  large enough

$$\begin{aligned} & \left| E \left[ -\log W_{Y|X}(Y_i|X^i) - \frac{nh(\theta)}{n - n_0 + 1} \mid X^n = a^n \right] \right| \\ & < q(\epsilon) \quad \forall a^n \in G(n_0, n), n_0 \leq i \leq n. \end{aligned} \quad (58)$$

Observe that

$$\begin{aligned} & E\left[\left(-\log W_{Y|X}(Y_i|X^i)\right)^2 | X^n = a^n\right] \\ &= \left[\log\left(\frac{1}{j} \sum_{i=1}^j a_i\right)\right]^2 \frac{1}{j} \sum_{i=1}^j a_i \\ &+ \left[\log\left(1 - \frac{1}{j} \sum_{i=1}^j a_i\right)\right]^2 \left(1 - \frac{1}{j} \sum_{i=1}^j a_i\right). \end{aligned}$$

Thus

$$\begin{aligned} & E\left[\left(-\log W_{Y|X}(Y_i|X^i)\right)^2 | X^n = a^n\right] \\ &\leq \max_{\theta - \epsilon \leq \theta' \leq \theta + \epsilon} \{[\log \theta']^2 \theta'\} \\ &+ [\log(1 - \theta')]^2 (1 - \theta') \\ &\triangleq c_1 \quad \forall a^n \in G(n_0, n), n_0 \leq i \leq n \end{aligned}$$

from which it follows that there exists a constant  $c$  such that for  $n$  large enough

$$\begin{aligned} & E\left[\left(-\log W_{Y|X}(Y_i|X^i) - \frac{nh(\theta)}{n - n_0 + 1}\right)^2 | X^n = a^n\right] \\ &\leq c_1 + 2 \frac{nh(\theta)}{n - n_0 + 1} E\left[\log W_{Y|X}(Y_i|X^i) | X^n = a^n\right] \\ &+ \frac{n^2 h(\theta)^2}{(n - n_0 + 1)^2} \\ &< c \quad \forall a^n \in G(n_0, n), n_0 \leq i \leq n. \end{aligned}$$

[We have used here also (58).] Therefore, applying Markov inequality and the fact that conditioned on  $X^n$ ,  $\{Y_i\}_{i=n_0}^n$  are independent we get

$$\begin{aligned} & P_{XY}^n \left( \left| \frac{1}{n} \sum_{i=n_0}^n \left( -\log W_{Y|X}(Y_i|X^i) - \frac{nh(\theta)}{n - n_0 + 1} \right) \right| \right. \\ & \left. > \delta^2 | X^n = a^n \right) \\ & \leq \frac{1}{\delta^2} \left( q^2(\epsilon) + \frac{c}{n} \right) \end{aligned} \quad (59)$$

for sufficiently large  $n$ . Hence, using (59) and (57) in (51) we conclude that

$$\begin{aligned} & P_{XY}^n \left( \left| -\frac{1}{n} \sum_{i=1}^n \log W_{Y|X}(Y_i|X^i) - h(\theta) \right| > 2\delta \right) \\ & \leq \epsilon + \frac{1}{\delta^2} \left( q^2(\epsilon) + \frac{c}{n} \right) + P_X^n(G^c(n_0, n)) \end{aligned} \quad (60)$$

for sufficiently large  $n$  [we have used the fact that the first term in the far right-hand side of (51) can be made arbitrarily small by increasing  $n$ ]. Since  $\epsilon$  is arbitrary and  $q(\epsilon) \rightarrow 0$  as  $\epsilon \rightarrow 0$ , (50) follows from (60) and (52).

#### REFERENCES

- [1] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. IT-39, pp. 752-772, May 1993.
- [2] R. Ahlswede and G. Dueck, "Identification via channels," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 15-29, Jan. 1989.
- [3] D. L. Neuhoff and P. C. Shields, "Channels with almost finite memory," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 4, pp. 440-447, July 1979.
- [4] D. L. Neuhoff and P. C. Shields, "Channel Entropy and Primitive Approximation," *Ann. Probab.*, vol. 10, no. 1, pp. 188-198, 1982.
- [5] T. S. Han and S. Verdú, "Spectrum invariance under output approximation for full-rank discrete memoryless channels," (in Russian), *Problemi Peredachi Informatsii*, 1993, vol. 29, no. 2, pp. 9-27, April-June 1993.
- [6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
- [7] D. S. Ornstein, "An Application of Ergodic Theory to Probability Theory," *Ann. Probab.*, vol. 1, pp. 43-58, 1973.
- [8] D. S. Ornstein, *Ergodic Theory, Randomness, and Dynamical Systems*. New Haven, CT: Yale University Press, 1974.
- [9] R. M. Gray and D. S. Ornstein, "Block coding for discrete stationary  $\bar{d}$ -continuous noisy channels," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 3, pp. 292-306, May 1979.
- [10] S. Vembu and S. Verdú, "On generating random bits from an arbitrary distribution," *Proc. 1992 Allerton Conference on Communication and Control*, Allerton, IL., Oct. 1992.
- [11] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471-480, 1973.
- [12] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 226-228, 1975.