Bounding Techniques for the Intrinsic Uncertainty of Channels

Or Ordentlich and Ofer Shayevitz
Dept. EE-Systems, Tel Aviv University, Israel

Abstract—**A channel can generally be defined by a probability distribution on a set of possible actions. These actions determine the output for any possible input, and are independently drawn. The intrinsic uncertainty of a channel is defined as the conditional entropy of the action given the input and output sequences. For many channels, such as the deletion channel, the insertion channel, and various permutation channels, e.g., the trapdoor channel, quantifying the intrinsic uncertainty is the main challenge in determining the capacity. In this paper, we derive an alternative expression for the intrinsic uncertainty via the Laplace variational principle, and utilize it to obtain a general lower bound for the capacity. As an example, we apply our bound to the binary deletion channel and show that for the special case of an i.i.d. input distribution and a range of deletion probabilities, it outperforms the best known lower bound for the mutual information.**

I. Introduction

A channel is traditionally defined via a conditional probability distribution $P_{\mathbf{Y}|\mathbf{X}}$ of the outputs given the inputs. Alternatively, a channel can also be (nonuniquely) defined as a random mapping from an input alphabet to an output alphabet, where the actual mapping applied to the input, namely the channel *action* $\mathbf{A}$, is drawn according to some probability distribution $P_{\mathbf{A}}$ over the set of all possible actions, *independently* of the input. Following this paradigm, the mutual information for a given input distribution $P_{\mathbf{X}}$ can be written as $I(\mathbf{X},\mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{A}) + H(\mathbf{A}|\mathbf{X},\mathbf{Y})$, for any eligible choice of the action distribution $P_{\mathbf{A}}$. A natural quantity to consider is therefore the *intrinsic uncertainty* $H(\mathbf{A}|\mathbf{X},\mathbf{Y})$, that captures the amount of information regarding the channel action revealed by observing its input and output.

For example, the natural choice of action for the Binary Symmetric Channel (BSC) is the modulo-addition of a binary noise sequence $\mathbf{U}$, which is completely revealed given the input and output sequences, resulting in zero intrinsic uncertainty. This is not the case for many other channels: Consider a Z-channel, where the action can again be identified with an i.i.d. binary noise sequence $\mathbf{U}$ such that the channel output is given by $Y_n = X_n(X_n + U_n)$. In this case, the input and output sequences reveal the action only in positions where the input was '1', yielding a generally nonzero intrinsic uncertainty. The same is true for most Discrete Memoryless Channels (DMCs), since in order for the action to be input-independent it needs to separately describe how each possible input would be mapped, yet only a small part of that description (corresponding to what "actually happened") is revealed by observing the input and output.

Generally speaking, channel capacity is clearly obtained by an input distribution that maximizes the (normalized) sum of the output entropy and the intrinsic uncertainty. For a DMC, this interpretation is not particulary useful, as a single letter expression for the capacity exists. In many other cases however, only an infinite-letter expression for the capacity is known. Since the entropy of the action is readily obtained, and the entropy of the output is often relatively easy to compute for many choices of the input process, the main difficulty in determining the capacity is typically encapsulated in evaluating the intrinsic uncertainty. This fact provides impetus for a direct study of the intrinsic uncertainty, which is the main focus of this paper.

The significance of the intrinsic uncertainty as a segue to capacity is perhaps most prominently evident in channels with synchronization errors, originally studied by Gallager [1]. This class of channels includes in particular the deletion channel and the insertion channel, but also finite-state permutation channels such as the trapdoor channel. Loosely speaking, channels in this class act by corrupting the time axis of the input rather than its symbol values. Consider for example the binary deletion channel, where each input bit is independently deleted with probability d. Despite its prima facie simplicity and much effort over the years, the capacity of this channel remains elusive hitherto. The output entropy of the deletion channel is however easy to compute for most reasonable input processes (i.i.d., Markov), and its intrinsic uncertainty is simply the expected logarithm of the number of deletion patterns that could have transformed the input sequence to the output sequence, a quantity that has been extensively yet implicitly tackled in previous work bounding the capacity of the deletion channel.

In this paper we investigate the intrinsic uncertainty for general channels. We employ the Laplace variational principle to derive an exact alternative expression for the intrinsic uncertainty, and then lower bound this expression using the convexity of the relative entropy. We further examine a class of channels whose output sequence, as in the case of the deletion channel, induces a uniform distribution over the associated set of feasible actions, and where our bound admits a simpler more tractable form. Finally, we apply our bound to the deletion channel and show it improves upon the best known lower bound on the mutual information for an i.i.d. input. Applying our techniques to inputs with memory (e.g. Markov), which are known to outperform i.i.d. inputs, is therefore a promising avenue of future research.

II. Bounds for the Intrinsic Uncertainty

In this section we define a channel by its action on its input, and develop general lower bounds on the mutual information between the input and output in terms of the channel action, by bounding the associated intrinsic uncertainty defined below. If the channel is information stable [2], as is the case for most channels of interest, then its capacity is given by the normalized mutual information maximized w.r.t. the input process, in the limit of large block length.

Let $\mathcal{X}, \mathcal{Y}$ be discrete alphabets. Any channel from $\mathcal{X}^n$ to $\mathcal{Y}^*$ can be (nonuniquely) defined by a probability distribution $P(\mathbf{a})$ on a set $\mathcal{A}^{(n)}$ of mappings from $\mathcal{X}^n \mapsto \mathcal{Y}^*$, to which we refer to below as *actions*. Each action $\mathbf{a}(\cdot) \in \mathcal{A}^{(n)}$ is defined for all possible inputs, and the channel action is chosen independently of the input, yielding the output $\mathbf{Y} = \mathbf{A}(\mathbf{X}) \in \mathcal{Y}^*$. The length $*$ of the channel output may itself be a random variable, as different actions in $\mathcal{A}^{(n)}$ may return vectors of different lengths. For any eligible choice of the action distribution $P(\mathbf{a})$, the *intrinsic uncertainty* of the channel with respect to the input distribution $P(\mathbf{x})$ is defined to be $H(\mathbf{A}|\mathbf{X}, \mathbf{Y})$. Note that while the intrinsic uncertainty may depend on the choice of the action distribution, the difference $H(\mathbf{A}) - H(\mathbf{A}|\mathbf{X}, \mathbf{Y})$ does not; we therefore have the freedom to choose the action distribution that is most convenient to work with.

For most channels of interest, the set of possible actions $\mathcal{A}^{(n)}$ extends naturally as n grows. We shall therefore omit the index n and simply denote this set by $\mathcal{A}$.

Example 1 (Deletion Channel): In a deletion channel, each transmitted symbol is either deleted or received uncorrupted. The set $\mathcal{A}$ includes 2^n actions, each corresponding to a different subset of the input indices $[1:n]$ marked for deletion. In an i.i.d. deletion model symbols are independently deleted with probability d, which induces a binomial distribution over the action set. The output's length $*$ equals n minus the number of symbols that were deleted from the transmitted block. Different actions applied to the same input may result in the same output. For example, if $\mathbf{x} = 01100$ we may get the output $\mathbf{y} = 110$ if either the first and fourth symbols or the first and fifth symbols were deleted. Therefore, the intrinsic uncertainty $H(\mathbf{A}|\mathbf{X}, \mathbf{Y})$ is generally positive.

Example 2 (Trapdoor Channel): The trapdoor channel is a simple finite-state channel, defined as follows. Balls labeled "0" or "1" are used to communicate through the channel. The channel starts with a ball already in it, referred to as the initial state. On each channel use, a ball is inserted into the channel by the transmitter, and one of the two balls in the channel is emitted with equal probability. The ball that is not emitted remains inside for the next channel use. In this model, the channel's action consists of choosing the initial state and deciding for each channel use whether to emit the ball that was already inside the channel or the ball that has just entered. Since an input $\mathbf{x}$ can be mapped to an output $\mathbf{y}$ via multiple actions, the intrinsic uncertainty is generally positive.

The mutual information between the input $\mathbf{X}$ and output $\mathbf{Y}$ can be expressed as

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) \\ &= H(\mathbf{Y}) - (H(\mathbf{Y},\mathbf{A}|\mathbf{X}) - H(\mathbf{A}|\mathbf{X},\mathbf{Y})) \\ &= H(\mathbf{Y}) - H(\mathbf{A}|\mathbf{X}) - H(\mathbf{Y}|\mathbf{A},\mathbf{X}) + H(\mathbf{A}|\mathbf{X},\mathbf{Y}) \\ &= H(\mathbf{Y}) - H(\mathbf{A}) + H(\mathbf{A}|\mathbf{X},\mathbf{Y}) \qquad (1) \end{aligned}$$

where (1) follows since the action $\mathbf{A}(\cdot)$ is statistically independent of the input $\mathbf{X}$, and $\mathbf{Y} = \mathbf{A}(\mathbf{X})$. For many channel models, the main challenge in computing (1) for a specific input distribution $P(\mathbf{x})$, is evaluating the intrinsic uncertainty $H(\mathbf{A}|\mathbf{X}, \mathbf{Y})$. The main contribution of this paper is a general technique for lower bounding this term, developed next.

A. General Channels

Denote by $\mathcal{P}(\mathcal{X})$ the set of all probability distributions on $\mathcal{X}$ and by $D(P||Q)$ the relative entropy between the distributions $P, Q \in \mathcal{P}(\mathcal{X})$. The Laplace variational principle [3, Proposition 1.4.2] states (after a straightforward manipulation) that for any distribution $P \in \mathcal{P}(\mathcal{X})$ and any nonnegative function $f(x)$ for which $\mathbb{E}_P \log f(X)$ is finite,

$$\mathbb{E}_P \log f(X) = \min_{Q\in\mathcal{P}(\mathcal{X})} \left(\log \mathbb{E}_Q f(X) + D(P||Q)\right), \qquad (2)$$

and the minimum is uniquely attained by

$$Q^*(x) = \frac{P(x)/f(x)}{\mathbb{E}_P(1/f(x))}, \qquad (3)$$

where by convention we set $1/f(x) = 0$ if $f(x) = 0$.[1]

We would like to obtain an alternative expression for

$$H(\mathbf{A}|\mathbf{X},\mathbf{Y}) = \mathbb{E}\log\frac{1}{P(\mathbf{A}|\mathbf{X},\mathbf{Y})}, \qquad (4)$$

where the expectation is taken with respect to the joint distribution

$$\begin{aligned} P(\mathbf{x},\mathbf{y},\mathbf{a}) &= P(\mathbf{x})P(\mathbf{a}|\mathbf{x})P(\mathbf{y}|\mathbf{x},\mathbf{a}) \\ &= P(\mathbf{x})P(\mathbf{a})\mathbb{1}_{\{\mathbf{y}=\mathbf{a}(\mathbf{x})\}}, \end{aligned}$$

and $\mathbb{1}_{\{B\}}$ is an indicator function for the event B. For brevity, we sometimes refer to this distribution as P.

Define the distribution

$$Q(\mathbf{x},\mathbf{y},\mathbf{a}) \triangleq \frac{P(\mathbf{x},\mathbf{y},\mathbf{a})P(\mathbf{a}|\mathbf{x},\mathbf{y})}{\mathbb{E}_P P(\mathbf{A}|\mathbf{X},\mathbf{Y})}, \qquad (5)$$

which we sometimes refer to as Q. Using the Laplace variational principle, the expectation from (4) can be written as

$$\begin{aligned} \mathbb{E}\log\frac{1}{P(\mathbf{A}|\mathbf{X},\mathbf{Y})} &= \log\mathbb{E}_Q\frac{1}{P(\mathbf{A}|\mathbf{X},\mathbf{Y})} + D(P||Q) \\ &= \log\mathbb{E}_Q\frac{1}{P(\mathbf{A}|\mathbf{X},\mathbf{Y})} + D\left(P(\mathbf{Y})||Q(\mathbf{Y})\right) \\ &\quad + D\left(P(\mathbf{X},\mathbf{A}|\mathbf{Y})||Q(\mathbf{X},\mathbf{A}|\mathbf{Y})\right) \qquad (6) \end{aligned}$$

[1] In the sequel we only use the readily verified fact that (2) is satisfied with equality for $Q = Q^*$. Nevertheless, we note that the general form of (2) can be also used for obtaining upper bounds on the intrinsic certainty.

where (6) follows from the chain rule of relative entropy. The marginal distribution $Q(\mathbf{y})$ is given by

$$\begin{aligned}Q(\mathbf{y}) &= \sum_{\mathbf{x},\mathbf{a}} Q(\mathbf{x},\mathbf{y},\mathbf{a}) \\ &= \frac{1}{\mathbb{E}_P P(\mathbf{A}|\mathbf{X},\mathbf{Y})} \sum_{\mathbf{x},\mathbf{a}} P(\mathbf{x})P(\mathbf{a})\mathbb{1}_{\{\mathbf{y}=\mathbf{a}(\mathbf{x})\}} P(\mathbf{a}|\mathbf{x},\mathbf{y}) \\ &= \frac{\mathbb{E}_{\mathbf{X},\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{y})}{\mathbb{E}_P P(\mathbf{A}|\mathbf{X},\mathbf{Y})}, \qquad (7)\end{aligned}$$

where in (7) we have used the fact that $P(\mathbf{a}|\mathbf{x},\mathbf{y}) = 0$ whenever $\mathbf{y} \neq \mathbf{a}(\mathbf{x})$. Thus,

$$\begin{aligned}&D\left(P(\mathbf{Y})||Q(\mathbf{Y})\right) = \mathbb{E}_{\mathbf{Y}} \log \left(\frac{P(\mathbf{Y})\mathbb{E}_P P(\mathbf{A}|\mathbf{X},\mathbf{Y})}{\mathbb{E}_{\mathbf{X},\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{Y})} \right) \\ &= -H(\mathbf{Y}) + \log \mathbb{E}_P P(\mathbf{A}|\mathbf{X},\mathbf{Y}) - \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{X},\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{Y}). \qquad (8)\end{aligned}$$

In addition,

$$\begin{aligned}\log \mathbb{E}_Q \frac{1}{P(\mathbf{A}|\mathbf{X},\mathbf{Y})} &= \log \sum_{\mathbf{x},\mathbf{y},\mathbf{a}} \frac{Q(\mathbf{x},\mathbf{y},\mathbf{a})}{P(\mathbf{a}|\mathbf{x},\mathbf{y})} \\ &= -\log \mathbb{E}_P P(\mathbf{A}|\mathbf{X},\mathbf{Y}). \qquad (9)\end{aligned}$$

Substituting (8) and (9) into (6) yields

$$\begin{aligned}H(\mathbf{A}|\mathbf{X},\mathbf{Y}) &= -H(\mathbf{Y}) - \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{X},\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{Y}) \\ &\quad + D\left(P(\mathbf{X},\mathbf{A}|\mathbf{Y})||Q(\mathbf{X},\mathbf{A}|\mathbf{Y})\right). \qquad (10)\end{aligned}$$

We are left with the task of evaluating the conditional relative entropy in (10). The conditional distributions that participate in this term are given by

$$P(\mathbf{x},\mathbf{a}|\mathbf{y}) = P(\mathbf{x})P(\mathbf{a}) \frac{\mathbb{1}_{\{\mathbf{y}=\mathbf{a}(\mathbf{x})\}}}{E_{\mathbf{X},\mathbf{A}} \mathbb{1}_{\{\mathbf{y}=\mathbf{A}(\mathbf{X})\}}} \qquad (11)$$

$$Q(\mathbf{x},\mathbf{a}|\mathbf{y}) = P(\mathbf{x})P(\mathbf{a}) \frac{P(\mathbf{a}|\mathbf{x},\mathbf{y})}{E_{\mathbf{X},\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{y})} \qquad (12)$$

and therefore

$$\begin{aligned}&D\left(P(\mathbf{X},\mathbf{A}|\mathbf{Y})||Q(\mathbf{X},\mathbf{A}|\mathbf{Y})\right) \\ &= \mathbb{E} \log \left(\frac{\mathbb{1}_{\{\mathbf{Y}=\mathbf{A}(\mathbf{X})\}}}{E_{\mathbf{X},\mathbf{A}} \mathbb{1}_{\{\mathbf{Y}=\mathbf{A}(\mathbf{X})\}}} \cdot \frac{E_{\mathbf{X},\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{Y})}{P(\mathbf{A}|\mathbf{X},\mathbf{Y})} \right). \qquad (13)\end{aligned}$$

Unfortunately, an exact computation of (13) involves the computation of $\mathbb{E} \log(1/P(\mathbf{A}|\mathbf{X},\mathbf{Y}))$, which is the exact technical difficulty we are trying to avoid. Instead, we lower bound (13) using the convexity of relative entropy, i.e.,

$$D\left(P(\mathbf{X},\mathbf{A}|\mathbf{Y})||Q(\mathbf{X},\mathbf{A}|\mathbf{Y})\right) \geq D\left(P(\mathbf{X},\mathbf{A})||\tilde{Q}(\mathbf{X},\mathbf{A})\right), \qquad (14)$$

where

$$\begin{aligned}\tilde{Q}(\mathbf{x},\mathbf{a}) &= \sum_{\mathbf{y}} P(\mathbf{y})Q(\mathbf{x},\mathbf{a}|\mathbf{y}) \\ &= P(\mathbf{x},\mathbf{a})\mathbb{E}_{\mathbf{Y}} \frac{P(\mathbf{a}|\mathbf{x},\mathbf{Y})}{E_{\mathbf{X},\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{Y})}. \qquad (15)\end{aligned}$$

Note that other properties of relative entropy, such as the data-processing inequality or Pinsker's inequality, could potentially be useful for bounding (13). Combining (14) and (15) gives,

$$\begin{aligned}&D\big(P(\mathbf{X},\mathbf{A}|\mathbf{Y})||Q(\mathbf{X},\mathbf{A}|\mathbf{Y})\big) \\ &\qquad \geq -\mathbb{E}_{\mathbf{X},\mathbf{A}} \log \mathbb{E}_{\mathbf{Y}} \frac{P(\mathbf{A}|\mathbf{X},\mathbf{Y})}{E_{\mathbf{X},\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{Y})}. \qquad (16)\end{aligned}$$

Substituting (16) into (10) and using (1) yields the following.

Theorem 1: For a channel with action $\mathbf{A}$, input $\mathbf{X}$, and output $\mathbf{Y} = \mathbf{A}(\mathbf{X})$,

$$\begin{aligned}I(\mathbf{X};\mathbf{Y}) &\geq -H(\mathbf{A}) - \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{X},\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{Y}) \\ &\quad - \mathbb{E}_{\mathbf{X},\mathbf{A}} \log \mathbb{E}_{\mathbf{Y}} \frac{P(\mathbf{A}|\mathbf{X},\mathbf{Y})}{E_{\mathbf{X},\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{Y})}. \qquad (17)\end{aligned}$$

B. Channels with Uniform Intrinsic Uncertainty

At this point, we restrict our attention to a class of channels we call channels with *uniform intrinsic uncertainty*. For this class of channels, the bound in Theorem 1 takes a particularly simpler form.

For any $\mathbf{x} \in \mathcal{X}^n$ and $\mathbf{y} \in \mathcal{Y}^*$ let

$$\mathcal{A}(\mathbf{x},\mathbf{y}) \triangleq \{\mathbf{a} \ : \ \mathbf{a}(\mathbf{x}) = \mathbf{y}\} \qquad (18)$$

be the set of all possible actions in $\mathcal{A}$ that map the input $\mathbf{x}$ to the output $\mathbf{y}$. Denote the cardinality of this set by $N(\mathbf{x},\mathbf{y}) \triangleq |\mathcal{A}(\mathbf{x},\mathbf{y})|$. A channel is said to have *uniform intrinsic uncertainty*, if all channel actions that can lead to an output $\mathbf{y}$ are equiprobable, i.e. if for any $\mathbf{y} \in \mathcal{Y}^*$

$$P(\mathbf{a}) = \phi(\mathbf{y}), \ \forall \mathbf{a} \in \bigcup_{\mathbf{x}} \mathcal{A}(\mathbf{x},\mathbf{y}) \qquad (19)$$

for some function $\phi : \mathcal{Y}^* \mapsto [0,1]$. This class includes, among many other examples, the trapdoor channel, and the i.i.d. deletion channel; the latter channel deletes each symbol independently with probability d, so that $P(\mathbf{a}) = d^{n-k}(1-d)^k$ for all channel actions that produce an output $\mathbf{y}$ of length k.

Proposition 1: For a channel with uniform intrinsic uncertainty, the action $\mathbf{A}$ is uniformly distributed over the set $\mathcal{A}(\mathbf{X},\mathbf{Y})$, conditioned on $\mathbf{X}$ and $\mathbf{Y}$.[2]

Proof.

$$\begin{aligned}P(\mathbf{a}|\mathbf{x},\mathbf{y}) &= \frac{P(\mathbf{x},\mathbf{y}|\mathbf{a})P(\mathbf{a})}{P(\mathbf{x},\mathbf{y})} \\ &= \frac{P(\mathbf{y}|\mathbf{x},\mathbf{a})P(\mathbf{a})}{P(\mathbf{y}|\mathbf{x})} = \frac{\mathbb{1}_{\{\mathbf{y}=\mathbf{a}(\mathbf{x})\}}P(\mathbf{a})}{\sum_{\mathbf{a}\in\mathcal{A}(\mathbf{x},\mathbf{y})} P(\mathbf{a})} \\ &\overset{(a)}{=} \frac{\phi(\mathbf{y})\mathbb{1}_{\{\mathbf{a}\in\mathcal{A}(\mathbf{x},\mathbf{y})\}}}{\phi(\mathbf{y})N(\mathbf{x},\mathbf{y})} = \frac{\mathbb{1}_{\{\mathbf{a}\in\mathcal{A}(\mathbf{x},\mathbf{y})\}}}{N(\mathbf{x},\mathbf{y})}, \qquad (20)\end{aligned}$$

where (a) follows from $\mathbb{1}_{\{\mathbf{y}=\mathbf{a}(\mathbf{x})\}} = \mathbb{1}_{\{\mathbf{a}\in\mathcal{A}(\mathbf{x},\mathbf{y})\}}$ and since $P(\mathbf{a}) = \phi(\mathbf{y})$ for all $\mathbf{a} \in \mathcal{A}(\mathbf{x},\mathbf{y})$. ■

Lemma 1: For a channel with uniform intrinsic uncertainty

$$\begin{aligned}-\mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{X},\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{Y}) &= H(\mathbf{A}) \\ &\quad - \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{X}} \mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}. \qquad (21)\end{aligned}$$

[2] Note that the converse is not generally true. As a counterexample, consider the BSC.

Proof. Using Proposition 1,

$$\begin{aligned}\mathbb{E}_{\mathbf{X},\mathbf{A}}P(\mathbf{A}|\mathbf{X},\mathbf{y}) &= \sum_{\mathbf{x}} P(\mathbf{x}) \sum_{\mathbf{a}} P(\mathbf{a}) \frac{\mathbb{1}_{\{\mathbf{a}\in\mathcal{A}(\mathbf{x},\mathbf{y})\}}}{N(\mathbf{x},\mathbf{y})}\\ &= \phi(\mathbf{y})\sum_{\mathbf{x}}P(\mathbf{x})\frac{N(\mathbf{x},\mathbf{y})}{N(\mathbf{x},\mathbf{y})}\mathbb{1}_{\{N(\mathbf{x},\mathbf{y})>0\}}\\ &= \phi(\mathbf{y})\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{y})>0\}}. \qquad (22)\end{aligned}$$

Thus,

$$\begin{aligned}-\mathbb{E}_{\mathbf{Y}}\log\mathbb{E}_{\mathbf{X},\mathbf{A}}P(\mathbf{A}|\mathbf{X},\mathbf{Y}) = &-\mathbb{E}_{\mathbf{Y}}\log\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}\\ &-\mathbb{E}_{\mathbf{Y}}\log\phi(\mathbf{Y}).\end{aligned}$$

The lemma follows since

$$\begin{aligned}H(\mathbf{A}) = -\mathbb{E}_{\mathbf{A}}\log P(\mathbf{A}) &= -\mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{A}|\mathbf{Y}}\log P(\mathbf{A})\\ &= -\mathbb{E}_{\mathbf{Y}}\log\phi(\mathbf{Y}), \qquad (23)\end{aligned}$$

where (23) holds since $P(\mathbf{a})$ equals $\phi(\mathbf{y})$ for all channel actions $\mathbf{a}$ that can produce the output $\mathbf{y}$ for some $\mathbf{x}$. ■

The next lemma lower bounds the last term in (17) for channels with uniform intrinsic uncertainty.

Lemma 2: For a channel with uniform intrinsic uncertainty

$$\begin{aligned}-\mathbb{E}_{\mathbf{X},\mathbf{A}}\log\mathbb{E}_{\mathbf{Y}}&\frac{P(\mathbf{A}|\mathbf{X},\mathbf{Y})}{E_{\mathbf{X},\mathbf{A}}P(\mathbf{A}|\mathbf{X},\mathbf{Y})}\\ &\geq -\mathbb{E}_{\mathbf{X}}\log\mathbb{E}_{\mathbf{Y}}\frac{\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}}{\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}} \geq 0 \qquad (24)\end{aligned}$$

Proof. By virtue of Jensen's inequality,

$$\begin{aligned}-\mathbb{E}_{\mathbf{X},\mathbf{A}}\log\mathbb{E}_{\mathbf{Y}}&\frac{P(\mathbf{A}|\mathbf{X},\mathbf{Y})}{E_{\mathbf{X},\mathbf{A}}P(\mathbf{A}|\mathbf{X},\mathbf{Y})}\\ &\geq -\mathbb{E}_{\mathbf{X}}\log\mathbb{E}_{\mathbf{Y}}\frac{\mathbb{E}_{\mathbf{A}}P(\mathbf{A}|\mathbf{X},\mathbf{Y})}{E_{\mathbf{X},\mathbf{A}}P(\mathbf{A}|\mathbf{X},\mathbf{Y})}.\end{aligned}$$

Using (20) and (22), we have

$$\begin{aligned}\frac{\mathbb{E}_{\mathbf{A}}P(\mathbf{A}|\mathbf{x},\mathbf{y})}{E_{\mathbf{X},\mathbf{A}}P(\mathbf{A}|\mathbf{X},\mathbf{y})} &= \frac{\sum_{\mathbf{a}}P(\mathbf{a})\frac{\mathbb{1}_{\{\mathbf{a}\in\mathcal{A}(\mathbf{x},\mathbf{y})\}}}{N(\mathbf{x},\mathbf{y})}}{\phi(\mathbf{y})\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{y})>0\}}}\\ &= \frac{\mathbb{1}_{\{N(\mathbf{x},\mathbf{y})>0\}}}{\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{y})>0\}}},\end{aligned}$$

establishing the first inequality in (24). The second inequality follows by applying Jensen's inequality again, this time w.r.t. $\mathbb{E}_{\mathbf{X}}$. ■

Combining Theorem 1, Lemma 1 and Lemma 2 establishes the following.

Theorem 2: For a channel with uniform intrinsic uncertainty,

$$\begin{aligned}I(\mathbf{X};\mathbf{Y}) \geq &-\mathbb{E}_{\mathbf{Y}}\log\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}\\ &-\mathbb{E}_{\mathbf{X}}\log\mathbb{E}_{\mathbf{Y}}\frac{\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}}{\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}}. \qquad (25)\end{aligned}$$

It is often convenient to consider some *typical* set $\mathcal{T}\subset\mathcal{Y}^*$, such that $P(\mathbf{Y}\in\mathcal{T})$ approaches 1 exponentially fast in n. The exact notion of typicality used can vary with the channel and the input distribution. Using the shorthand notation $\mathbb{E}_{\mathbf{Y}|\mathcal{T}}$ for $\mathbb{E}_{\mathbf{Y}|\mathbf{Y}\in\mathcal{T}}$, we have that

$$\begin{aligned}-\mathbb{E}_{\mathbf{Y}}\log&\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}\\ &= -\mathbb{E}_{\mathbf{Y}|\mathcal{T}}\log\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}} + \mathrm{o}(n)\\ &\geq -\log\mathbb{E}_{\mathbf{X}}\mathbb{E}_{\mathbf{Y}|\mathcal{T}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}} + \mathrm{o}(n), \qquad (26)\end{aligned}$$

where the last transition follows from Jensen's inequality. We note that it is straightforward to show that the capacity of channels with uniform intrinsic uncertainty is greater than (26). To see this consider a codebook whose codewords are drawn independently according to the distribution $P(\mathbf{x})$, in conjunction with a decoder that declares an error if $\mathbf{y}\notin\mathcal{T}$, and otherwise outputs some codeword $\mathbf{x}$ for which $N(\mathbf{x},\mathbf{y})>0$ (there is always at least one). The probability that $\mathbf{Y}\notin\mathcal{T}$ approaches zero as n increases, and the probability that $N(\mathbf{X},\mathbf{Y})>0$ for some codeword $\mathbf{X}$ that is statistically independent of $\mathbf{Y}$, given that $\mathbf{Y}\in\mathcal{T}$, is $\mathbb{E}_{\mathbf{X}}\mathbb{E}_{\mathbf{Y}|\mathcal{T}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}$. Applying the union bound, we see that any rate smaller than (26), normalized by n, is achievable. Our analysis shows that the gap to the normalized mutual information incurred by using this straightforward bound is the normalized divergence $D(P(\mathbf{X},\mathbf{A}|\mathbf{Y})||Q(\mathbf{X},\mathbf{A}|\mathbf{Y}))/n$, which we have bounded from below by $-(1/n)\mathbb{E}_{\mathbf{X}}\log\mathbb{E}_{\mathbf{Y}}\frac{\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}}{\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}}$.

III. Example: The binary i.i.d. deletion channel

The binary i.i.d. deletion channel operates by independently deleting input bits with probability d. As already discussed, this channel admits the uniform intrinsic uncertainty property, and therefore Theorem 2 holds. In this section, we apply Theorem 2 to obtain a lower bound for $I(\mathbf{X};\mathbf{Y})$ under a uniform i.i.d. input distribution, which outperforms the best known bounds for i.i.d inputs [1], [4]. Since the deletion channel is information stable, any rate smaller than the associated $\lim_{n\to\infty}I(\mathbf{X};\mathbf{Y})/n$ is achievable with uniform i.i.d. codebooks. Note that for a uniform i.i.d. input, the output $\mathbf{Y}$ is also uniform i.i.d. given its length $*$, where the latter is binomial with parameters $(n,1-d)$.

For the i.i.d. deletion channel the quantity $N(\mathbf{x},\mathbf{y})$ corresponds to the number of appearances of $\mathbf{y}$ as a subsequence of $\mathbf{x}$, and $\mathbb{1}_{\{N(\mathbf{x},\mathbf{y})>0\}}$ indicates whether or not $\mathbf{y}$ is a subsequence of $\mathbf{x}$. Define the operation $\langle x\rangle \triangleq \max(x,1/2)$. According to [5, Lemma 3.1], for any $\mathbf{y}$ of length αn we have

$$\sum_{\mathbf{x}\in\mathbb{Z}_2^n}\mathbb{1}_{\{N(\mathbf{x},\mathbf{y})>0\}} = \sum_{j=\alpha n}^{n}\binom{n}{j} \doteq 2^{nH_2(\langle\alpha\rangle)}, \qquad (27)$$

where $H_2(\cdot)$ is the binary entropy function, and $\doteq$ denotes exponential equality in the usual sense. This implies that for any $\mathbf{y}$ of length αn we have $\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{y})>0\}} \doteq 2^{n(H_2(\langle\alpha\rangle)-1)}$. We can now define the typical set $\mathcal{T}$ as the set of all $\mathbf{y}$'s whose normalized length α satisfies $|\alpha-(1-d)|<\epsilon$. For any $\epsilon>0$ and n large enough $P(\mathbf{Y}\in\mathcal{T})$ is arbitrarily close to one, and we can therefore apply (26), yielding

$$-\mathbb{E}_{\mathbf{Y}}\log\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}} \geq n\left(1-H_2(\langle 1-d-\epsilon\rangle)\right). \qquad (28)$$

The right hand side of (28) (normalized by n) is a well known lower bound for the deletion channel capacity, obtained with a uniform i.i.d. input [5]. We now evaluate the second term in (25) in order to improve upon this bound. To this end, we first parse each $\mathbf{x} \in \mathbb{Z}_2^n$ into phrases that contain exactly two bit flips and end immediately after the second flip. For example, the string 0001111011001110001 is parsed into the three phrases 00011110, 11001, 110001. We identify each phrase with three parameters: $b \in \{0,1\}$ is the first bit in the phrase, $k_1 \geq 2$ is the index of the first flip in the phrase, and $k_2 \geq 1$ is such that $k_1 + k_2$ is the total number of bits in the phrase. In our example, the three phrases correspond to $\{b = 0, k_1 = 4, k_2 = 4\}$, $\{b = 1, k_1 = 3, k_2 = 2\}$ and $\{b = 1, k_1 = 3, k_2 = 3\}$, respectively. For any pair of integers $2 \leq k_1 < n$, $1 \leq k_2 < n$ let $\Psi^{k_1,k_2}(\mathbf{x})$ be the number of $\{k_1, k_2\}$-phrases in the parsing of $\mathbf{x}$. For $\epsilon > 0$ we define the typical set

$$\mathcal{S} \triangleq \Big\{\mathbf{x} \in \mathbb{Z}_2^n \ : \ \forall\ 2 \leq k_1 < n,\ 1 \leq k_2 < n \\ \left|\frac{1}{n}\Psi^{k_1,k_2}(\mathbf{x}) - \frac{1}{5}2^{-(k_1+k_2-1)}\right| < \epsilon\Big\}.$$

It can verified that for any $\epsilon > 0$ and n large enough $P(\mathbf{X} \in \mathcal{S})$ is indeed arbitrary close to 1. Thus, applying the same reasoning as in (26) we obtain

$$-\mathbb{E}_{\mathbf{X}} \log \mathbb{E}_{\mathbf{Y}} \frac{\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}}{\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}} \geq -\log \mathbb{E}_{\mathbf{Y}} \frac{\mathbb{E}_{\mathbf{X}|\mathcal{S}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}}{\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}}$$

for n large enough. Recalling that α is the (random) length of $\mathbf{Y}$, we take the expectation $\mathbb{E}_{\mathbf{Y}}$ as $\mathbb{E}_\alpha \mathbb{E}_{\mathbf{Y}|\alpha}$ and use (27) to obtain

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Y}} \frac{\mathbb{E}_{\mathbf{X}|\mathcal{S}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}}{\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}} \\
&\doteq \mathbb{E}_\alpha 2^{n(1-H_2(\langle\alpha\rangle))} \mathbb{E}_{\mathbf{Y}|\alpha}\mathbb{E}_{\mathbf{X}|\mathcal{S}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}} \\
&= \mathbb{E}_\alpha 2^{n(1-H_2(\langle\alpha\rangle))} P\left(N(\mathbf{X},\mathbf{Y}) > 0|\alpha, \mathbf{X} \in \mathcal{S}\right). \qquad (29)
\end{aligned}$$

Now, consider a greedy algorithm for determining whether $\mathbf{y}$ is a subsequence of $\mathbf{x}$, defined as follows [6, Section 3.1]: Scanning from left to right, take the first bit in $\mathbf{y}$ and match it with its first appearance in $\mathbf{x}$. Then take the second bit in $\mathbf{y}$ and match it with its subsequent first appearance in $\mathbf{x}$. Continue until either $\mathbf{x}$ or $\mathbf{y}$ are exhausted, where the latter case is termed success. It is easy to see that the greedy algorithm succeeds if and only if $N(\mathbf{x},\mathbf{y}) > 0$. For statistically independent random vectors $\mathbf{X}$ and $\mathbf{Y}$, we enumerate the phrases in the parsing of $\mathbf{X}$ by $i = 1, \ldots, M(\mathbf{X})$ where $M(\mathbf{X})$ is the (random) number of phrases in $\mathbf{X}$. We define the random variables Z_i as the number of bits in $\mathbf{Y}$ that were matched to bits in the ith phrase of $\mathbf{X}$ by the greedy algorithm. The vector $\mathbf{Y}$ consists of αn i.i.d. uniform bits. To simplify computations, we assume that $\mathbf{Y}$ is always padded so that it consists of n i.i.d. bits, and we calculate the probability that $\sum_i Z_i \geq \alpha n$, which is equal to $P(N(\mathbf{X},\mathbf{Y}) > 0|\alpha)$ since the additional random suffix does not affect the event where the first αn bits in $\mathbf{Y}$ are matched. It can be shown that under this assumption the Z_i's are mutually independent, given that the phrase types $\{k_1^{(i)}, k_2^{(i)}\}_{i=1}^{M(\mathbf{X})}$ of $\mathbf{X}$ are known (but assuming that their first bit identifiers $\{b_i\}_{i=1}^{M(\mathbf{X})}$ remain random). Of course, the distribution of Z_i depends on the parameters $k_1^{(i)}, k_2^{(i)}$ that correspond to the ith phrase in $\mathbf{X}$. Given k_1 and k_2, the (base two) moment generating function of Z_i is

$$\begin{aligned}
&\lambda_{Z_i}^{k_1,k_2}(t) \triangleq \mathbb{E}\left(2^{tZ_i}|k_1,k_2\right) = 2^{k_1(t-1)} \\
&\quad + 2^{t-1}\frac{1-2^{k_1(t-1)}}{1-2^{t-1}}\left(2^{t-1}\frac{1-2^{k_2(t-1)}}{1-2^{t-1}} + 2^{k_2(t-1)-t}\right).
\end{aligned}$$

Noting that by definition, for $\mathbf{X} \in \mathcal{S}$ the number and composition of the $\{k_1, k_2\}$-phrases is essentially deterministic, we can use Cramer's theorem [7] to obtain

$$\begin{aligned}
P(N(\mathbf{X},\mathbf{Y}) > 0|\alpha, \mathbf{X} \in \mathcal{S}) &= P(\sum_i^{M(\mathbf{X})} Z_i \geq \alpha n|\alpha, \mathbf{X} \in \mathcal{S}) \\
&\doteq 2^{-n\Lambda^*(\alpha)},
\end{aligned}$$

where

$$\Lambda^*(\alpha) = \max_{t>0}\left(\alpha t - \frac{1}{5}\sum_{k_1=2}\sum_{k_2=1} 2^{-(k_1+k_2-1)} \log \lambda_{Z_{k_1,k_2}}(t)\right).$$

Substituting into (29), and applying standard large deviations arguments, we obtain

$$-\mathbb{E}_{\mathbf{X}} \log \mathbb{E}_{\mathbf{Y}} \frac{\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}}{\mathbb{E}_{\mathbf{X}}\mathbb{1}_{\{N(\mathbf{X},\mathbf{Y})>0\}}} \geq n \cdot g(d) + \mathrm{o}(n)$$

where

$$g(d) \triangleq \min_{0 \leq \alpha \leq 1} D_2(\alpha||1-d) - (1 - H_2(\langle\alpha\rangle)) + \Lambda^*(\alpha)$$

where $D_2(p||q)$ is the binary relative entropy function. It follows that for a uniform i.i.d. distribution,

$$\lim_{n\to\infty} \frac{1}{n} I(\mathbf{X};\mathbf{Y}) \geq 1 - H_2(\min(d, 1/2)) + g(d). \qquad (30)$$

Numerical evaluation of the term $g(d)$ reveals that it is greater than zero for all $d < 1/2$. Thus, (30) improves over Gallager's well know bound $1 - H_2(d)$ [1]. Recently, Rahmati and Duman [4] used a different technique to lower bound the mutual information for uniform i.i.d. inputs. For small values of d their bound is better than (30), but for larger values of d the right hand side of (30) turns out to be greater than their bound. For example, for $d = 0.2$ our bound improves on $1 - H_2(0.2)$ by ≈ 0.0117 bits (roughly 5%), whereas the improvement of [4] is negligible.

References

[1] R. G. Gallager, "Sequential decoding for binary channels with noise and synchronization errors," Tech. Rep., 1961.

[2] R. Dobrushin, "General formulation of Shannons main theorem in information theory," *Amer. Math. Soc. Transl*, 1963.

[3] P. Dupuis and R. S. Ellis, *A weak convergence approach to the theory of large deviations*. John Wiley & Sons, 2011.

[4] M. Rahmati and T. Duman, "Bounds on the capacity of random insertion and deletion-additive noise channels," *IEEE Trans. Inf. Theory*, 2013.

[5] S. N. Diggavi and M. Grossglauser, "On transmission over deletion channels," in *Allerton*, 2001.

[6] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probability Surveys*, 2009.

[7] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*. Berlin: Springer-Verlag, 2010.