

© 2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

M. A. Tope and J. M. Morris, "On the Limits of Learning a Discrete Memoryless Communication Channel," MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM), Rockville, MD, USA, 2022, pp. 222-228, doi: 10.1109/MILCOM55135.2022.10017597.

<https://doi.org/10.1109/MILCOM55135.2022.10017597>

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

On the Limits of Learning a Discrete Memoryless Communication Channel

Michael A. Tope and Joel M. Morris

Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County, Catonsville, MD 21250
Email: mtope1@umbc.edu, morris@umbc.edu

Abstract—This paper explores bounds on the minimum number of channel probes required to learn sufficient information to establish reliable communication at a given communication rate (below channel capacity) for a discrete memoryless channel. Given a set of discrete channel input-output sample pairs (where for each discrete input value, the associated aggregate set of discrete output values observations is multinomially distributed), we leverage a non-asymptotic probably approximately correct (PAC) bound on the mutual information (channel capacity) between the label (discrete) random variable (RV) and the observation RV to establish a convergence rate for the worst case channel. Previous bounds (such as those based on Sanov’s Theorem) provide high probability (i.e. PAC) bounds on the true mutual information that converge with rate $O(\log(N)/\sqrt{N})$, where N is the number of independent and identically distributed (i.i.d.) samples used to compute the empirical probability mass functions. Using an improved PAC sublevel-set bound, we sharpen the rate of convergence to $O(\sqrt{\log(\log(N))\log(N)/N})$.

In scenarios where channel knowledge may be incomplete or impaired, channel probing is often used to gain channel information to drive the coding and modulation processing chain to provide reliable communications. This paper looks at the limiting convergence rate in gaining such information for the discrete memoryless channel (DMC) with known number of inputs values and known number of output values. Before channel probing process commences, the channel transition probabilities are unknown.

Unlike an additive white Gaussian noise (AWGN) channel (where the variance on the channel output remains constant regardless of the continuous input value), for the discrete channel, the effective ‘variance’ in the channel output may vary. For example, specific channel transition probabilities may rule out certain output values from ever occurring given a specific input value (similar to a Z-channel for the binary input and binary output case where the channel output uncertainty depends significantly given the specific input value). So we would not expect the DMC to behave as the AWGN channel.

While we assume that the channel transition probabilities of the DMC do not change over time, we expect that results presented here are a step towards addressing non-stationary and/or adversarial communication channels including the limiting rate that a system can follow or track channel drift, etc.

Consider the scenario depicted in Fig1. Alice selects channel input values and Bob reports the channel output values that he observed (and associated with each selected input value). Alice and Bob share their channel information and once they

have sufficient channel knowledge to achieve a given communication rate R , they design and implement a suitable encoder and decoder for that rate. From that point on communication at rate R is established with high probability.

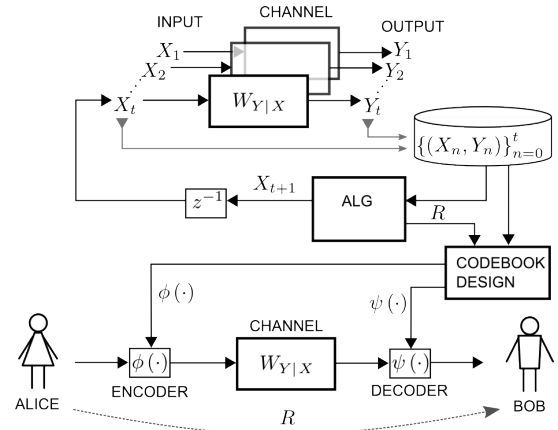


Fig. 1. Given discrete samples pairs, bound $I(X; Y)$

At the completion of the channel probing process (see Fig. 1), suppose Alice and Bob have gathered N input-output sample pairs $\mathcal{S}_N = \{(x_n, y_n)\}_{n=0}^{N-1}$. Each sample pair consists of the observed output value (or symbol) $y \in \mathcal{Y} = \{b_0, b_1, \dots, b_{|\mathcal{Y}|-1}\}$ that corresponds to the instance when input value $x \in \mathcal{X} = \{a_0, a_1, \dots, a_{|\mathcal{X}|-1}\}$ was applied to the channel. We consider each sample pair (x_n, y_n) is an instance of a RV (X, Y) . The marginal probability mass function (pmf) \mathbf{u} of the RV X lies within a discrete probability space $\mathcal{P}_{\mathcal{X}}$ that is $\mathbf{u} \triangleq [u_{a_0}, u_{a_1}, \dots, u_{a_{|\mathcal{X}|-1}}] \in \mathcal{P}_{\mathcal{X}}$ and $u_x \triangleq \mathbb{P}\{X = x\}$. Similarly, the marginal pmf \mathbf{v} of the observation RV Y is $\mathbf{v} \triangleq [v_{b_0}, v_{b_1}, \dots, v_{b_{|\mathcal{Y}|-1}}] \in \mathcal{P}_{\mathcal{Y}}$ where $v_y \triangleq \mathbb{P}\{Y = y\}$.

We seek to recover bounds pertaining to the probabilistic mapping $\underline{\mathbf{w}}$, which is a vector of conditional pmfs indexed by the input symbol x , i.e. $\underline{\mathbf{w}} \triangleq [\mathbf{w}_{a_0}, \mathbf{w}_{a_1}, \dots, \mathbf{w}_{a_{|\mathcal{X}|-1}}]$, where $\mathbf{w}_x \triangleq [w_{b_0|x}, w_{b_1|x}, \dots, w_{b_{|\mathcal{Y}|-1}|x}]$ and $w_{y|x} \triangleq \mathbb{P}\{Y = y | X = x\} \forall x \in \mathcal{X}$ and $y \in \mathcal{Y}$. We call \mathbf{w}_x the ‘generator’ pmf of the observation given the input value x . The observation RV $Y | X = x$ is independent and identically distributed (i.i.d.).

When the channel law (or probabilistic mapping) $\underline{\mathbf{w}}$ is known, the ‘average’ mutual information between the RVs

X and Y is

$$I(X; Y) = I(\mathbf{u}, \underline{\mathbf{w}}) \triangleq \sum_{x \in \mathcal{X}} u_x D(\mathbf{w}_x \| \hat{\mathbf{v}}(\mathbf{u})), \quad (1)$$

where the expected output pmf is $\hat{\mathbf{v}}(\mathbf{u}) \triangleq \sum_{x \in \mathcal{X}} u_x \mathbf{w}_x$, and the Kullback Leibler (KL) divergence [1] for discrete RVs P and Q with respective pmfs $\mathbf{p} \in \mathcal{P}_{\mathcal{Y}}$ and $\mathbf{q} \in \mathcal{P}_{\mathcal{Y}}$ is

$$D(P \| Q) = D(\mathbf{p} \| \mathbf{q}) \triangleq \sum_{y \in \mathcal{Y}} p_y \log_2 \left(\frac{p_y}{q_y} \right). \quad (2)$$

If the input RV and the observed output RV are statistically independent, then the observed output value does not aid in identifying the correct input symbol.

From \mathcal{S}_N , we compute the sample (empirical) pmfs $\hat{\mathbf{w}}_x = [\hat{w}_{b_0 | x}, \hat{w}_{b_1 | x}, \dots, \hat{w}_{b_{|\mathcal{Y}|-1} | x}]$, where

$$\hat{w}_{y | x} \triangleq \frac{1}{N_x} \sum_{n=0}^{N-1} \mathbb{1}_{\{y_n=y \wedge x_n=x\}} \quad \forall y \in \mathcal{Y}, \quad (3)$$

and $N_x = \sum_{n=0}^{N-1} \mathbb{1}_{\{x_n=x\}}$ for each $x \in \mathcal{X}$.

Similarly, we have the input and output empirical pmfs:

$$\hat{u}_x \triangleq \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{1}_{\{x_n=x\}} \quad \forall x \in \mathcal{X} \text{ and} \quad (4)$$

$$\hat{v}_y \triangleq \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{1}_{\{y_n=y\}} \quad \forall y \in \mathcal{Y}. \quad (5)$$

Given an empirical pmf $\hat{\mathbf{w}}$ and N , we want to establish a tight *reverse* probably approximately correct (PAC) bound of the form

$$\mathbb{P} \{ \hat{\mathbf{w}} \in \Gamma_{\xi}^{\text{rev}}(\hat{\mathbf{w}}) \} \geq 1 - \delta, \quad (6)$$

where the pmf $\hat{\mathbf{w}}$ is a possible generator of the empirical pmf $\hat{\mathbf{w}}$. Specifically, we choose a closed convex sub-levelset Γ based on the KL divergence, which is ‘centered’ on $\hat{\mathbf{w}}$ with a ‘size’ ξ . This sub-levelset is defined as

$$\Gamma_{\xi}^{\text{rev}}(\hat{\mathbf{w}}) \triangleq \{ \hat{\mathbf{w}} : D(\hat{\mathbf{w}} \| \hat{\mathbf{w}}) \leq \xi, \quad \forall \hat{\mathbf{w}} \in \mathcal{P}_{\mathcal{Y}} \}. \quad (7)$$

Since

$$\begin{aligned} I(\mathbf{u}, \underline{\mathbf{w}}) &= \sum_{x \in \mathcal{X}} u_x D(\mathbf{w}_x \| \mathbf{v}') - D(\mathbf{v}(\mathbf{u}) \| \mathbf{v}') \\ &= \min_{\mathbf{v}' \in \mathcal{P}_{\mathcal{Y}}} \sum_{x \in \mathcal{X}} u_x D(\mathbf{w}_x \| \mathbf{v}'), \end{aligned} \quad (8)$$

we can solve for a high probability PAC bound on the mutual information $I_L \leq I(X; Y)$ based on the various sublevel-set constraints as

$$I_L \triangleq \min_{\substack{\mathbf{w}_x^- \in \Gamma_{\xi_w}^{\text{rev}}(\hat{\mathbf{w}}_x) \\ \forall x \in \mathcal{X}}} \min_{\mathbf{u}^- \in \Gamma_{\xi_u}^{\text{rev}}(\hat{\mathbf{u}})} \min_{\substack{\mathbf{v}^- \in \Gamma_{\xi_v}^{\text{rev}}(\hat{\mathbf{v}}) \\ x \in \mathcal{X}}} \sum_{x \in \mathcal{X}} u^-_x D(\mathbf{w}_x^- \| \mathbf{v}^-), \quad (9)$$

and parameters ξ_w , ξ_v , and ξ_u are adjusted to meet the overall probability that the bound holds.

This approach is equivalent to forming a compound channel based on the PAC bounds and then solving for the minimum mutual information across the compound channel. For a compound channel, the channel law $\underline{\mathbf{w}}$ is confined to a known

region within the output probability space $\mathcal{P}_{\mathcal{Y}}$, say $\underline{\mathbf{w}} \in \Gamma$, and the channel capacity is given by [2]

$$C = \max_{\mathbf{u} \in \mathcal{P}_{\mathcal{X}}} \min_{\mathbf{w} \in \Gamma} I(\mathbf{u}, \mathbf{w}). \quad (10)$$

Further, if Γ is convex and closed, then one can swap the order of the min and max yielding [2]

$$C = \min_{\mathbf{w} \in \Gamma} \max_{\mathbf{u} \in \mathcal{P}_{\mathcal{X}}} I(\mathbf{u}, \mathbf{w}). \quad (11)$$

So for a convex Γ , one can determine the channel capacity for each channel law $\underline{\mathbf{w}} \in \Gamma$ and then select R_L as the rate of the worst case (minimum capacity) channel, and there exists a codebook [2] that will support rates up to R_L (simultaneously) for any channel law $\underline{\mathbf{w}} \in \Gamma$.

As the number of observations N increases, the ‘size’ of the sublevel-set constraints ‘shrink,’ and we seek to establish and determine the convergence rate of I_L based on a function of the ‘size’ (of the sublevel-set constraints) $\xi(N, |\mathcal{Y}|, \delta)$.

The remainder of this paper is as follows: After briefly describing some previous work (Section I), we review some non-asymptotic sublevel-set bounds (Section II). We leverage those bounds to constrain the uncertainty about the channel transition probabilities (given the input-output sample pairs observed). In Section III, we view these sublevel-set constraints as a compound communication channel, and we develop an algorithm to solve for a lower bound on mutual information using the sub-levelsets as convex constraints on a convex optimization problem. Section IV describes a similar bound on channel capacity. In Section V, we present results from dataset simulations. We finish with conclusions and discuss future work (Section VI).

I. PREVIOUS WORK

Lapidoth and Narayan provide a summary of results for communication over uncertain channels including the compound channel [2], and recent results for compound DMCs are covered by Csiszár, J. Körner [1]. Our approach relies on the *compound channel* and the *information spectrum* approach of Han [3]; however, we use probabilistic constraints to form the compound channel. Each probabilistic constraint is based on a PAC bound.

Valiant [4] developed the probability approximately correct (PAC) concept, where a PAC bound refers to a bound that holds with a prescribed arbitrarily high probability. Langford [5] developed and outlined the application of PAC-bounds to machine-learning. However, we require PAC bounds specifically for information theoretic measures.

Paninski [6], Chattopadhyaya, and Lipson [7] present and discuss a ‘plug-in’ estimator approach of substituting empirical pmfs or ‘flattened’ (or ‘smoothed’) pmfs (such as $\tilde{w}_y \triangleq \frac{w_y + \gamma}{1 + |\mathcal{Y}| \gamma}$ for some $\gamma > 0$) directly into information theoretic measures.

VanderKratts and Banerjee [8] introduced the use of PAC bounds (aka concentration inequalities) on mutual information for datasets with binary labels (and continuous-valued observations). Seldin and Tishby [9] derived PAC-bounds and PAC-Bayesian bounds for discrete RVs, and this paper significantly

leverages that work. Our sublevel-set (PAC) bound is closely related to PAC-Bayesian bounds (see [10] for a survey and review).

This paper extends our results [11] from a memoryless binary input-output channel to a discrete memoryless channel.

II. REVIEW OF SUB-LEVELSET BOUNDS

In this section, we review several probabilistic bounds (see Eq. 6) on the sublevel-set $\Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}})$ (see Eq. 7) by first deriving a bound on the probability $\hat{\mathbf{w}} \in \Gamma_\xi(\mathbf{w}) \triangleq \{\hat{\mathbf{w}}: D(\hat{\mathbf{w}} \parallel \mathbf{w}) \leq \xi, \forall \hat{\mathbf{w}} \in \mathcal{P}_Y\}$. We start by leveraging the multinomial halfspace bound (MHB).

Theorem II.1 Multinomial Halfspace Bound (MHB) [12]

Suppose we have the set $\mathcal{S}_N = \{y_0, y_1, \dots, y_{N-1}\}$ of outcomes from N i.i.d. discrete random variables $Y_n \in \mathcal{Y}$ and $Y_n \sim \mathbf{w}$ for $n = 0, 1, \dots, N-1$, and $\hat{\mathbf{w}}$ is the empirical pmf of \mathcal{S}_N . Given the halfspace Λ (oriented to include the pmf $\mathbf{w}^* \in \mathcal{P}_Y$) defined as

$$\Lambda(\mathbf{w}^*, \mathbf{w}) \triangleq \left\{ \hat{\mathbf{w}} \in \mathcal{P}_Y : \sum_{y \in \mathcal{Y}} \hat{w}_y \ln \left(\frac{w_y^*}{w_y} \right) \leq \xi \right\} \quad (12)$$

where $\xi \triangleq D(\mathbf{w}^* \parallel \mathbf{w})$, then we have

$$\mathbb{P} \{ \hat{\mathbf{w}} \notin \Lambda(\mathbf{w}^*, \mathbf{w}) \} \leq \exp(-ND(\mathbf{w}^* \parallel \mathbf{w})). \quad (13)$$

The Sanov Theorem (a sublevel-set bound) may be constructed from a plurality of MHBs [12].

Theorem II.2 Sanov's Theorem (see [13] section 11.4)

Given the same as Thm. II.1, then given **any** region $\Gamma \subset \mathcal{P}_Y$ and \mathbf{w}^* the 'closest' pmf among all $\hat{\mathbf{w}} \in \Gamma$ to \mathbf{w} in terms of the KL divergence

$$\mathbf{w}^* = \arg \min_{\hat{\mathbf{w}} \in \Gamma} D(\hat{\mathbf{w}} \parallel \mathbf{w}) \quad (14)$$

then

$$\delta_\Gamma \triangleq \mathbb{P} \{ \hat{\mathbf{w}} \notin \Gamma \} \leq (N+1)^{|\mathcal{Y}|} \exp(-ND(\mathbf{w}^* \parallel \mathbf{w})). \quad (15)$$

Solving for δ_Γ , we get a sublevel-set bound, where

$$\xi(N, |\mathcal{Y}|, \delta_\Gamma) = D(\mathbf{w}^* \parallel \mathbf{w}) \leq \frac{|\mathcal{Y}| \ln(N+1) - \ln(\delta_\Gamma)}{N} \quad (16)$$

sets the 'size' of the sublevel-set.

The following theorem sharpens the 'Sanov' sublevel-set bound.

Theorem II.3 Improved Sub-levelset Bound [12]

Given the same as Thm. II.1, then select any $\delta_\Gamma \in (0, 1]$, then $\mathbb{P} \{ \hat{\mathbf{w}} \notin \Gamma_\xi(\mathbf{w}) \} \leq \delta_\Gamma$ for the sub-levelset $\Gamma_\xi(\mathbf{w})$ with 'size'

$$\xi \geq \frac{1}{N} \left(\frac{1}{2} \ln(2|\mathcal{Y}|) - \frac{3}{2} \ln \left(\frac{\delta_\Gamma}{2} \right) + |\mathcal{Y}| \ln \left(\log_2(\log_2(N)) + \kappa_1 \sqrt{|\mathcal{Y}|} + \log_2(\kappa_2 |\mathcal{Y}|) + 2 \right) \right), \quad (17)$$

where $\kappa_1 = 2\sqrt{24}(1 + \sqrt{2})$ and $\kappa_2 = 24$.

We use the sublevel-set bounds to constrain the channel law uncertainty with high probability. Since the observed empirical pmf $\hat{\mathbf{w}}$ is within $\Gamma_\xi(\mathbf{w})$ with probability $> 1 - \delta_\Gamma$, then the generator pmf \mathbf{w} must be located within the probability space such that $D(\hat{\mathbf{w}} \parallel \mathbf{w}) \leq \xi$ with probability $> 1 - \delta_\Gamma$; therefore, we define the sublevel-set $\Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}})$ (see Eq. 7) and the $\mathbb{P} \{ \mathbf{w} \in \Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}}) \} \geq 1 - \delta_\Gamma$. Given the input-output sample pairs \mathcal{S}_N , we compose a sublevel-set bound on every transition probability component \mathbf{w}_x of the channel law $\underline{\mathbf{w}}$.

III. A PAC BOUND ON MUTUAL INFORMATION

We compute a lower bound on mutual information given various sub-levelset constraints according to Eq. 9 by following the approach in [14].

Consider the K -way minimization of an objective function f [15], i.e.

$$f^* = \min_{x_1 \in \Gamma_1} \dots \min_{x_K \in \Gamma_K} f(x_1, \dots, x_K). \quad (18)$$

If every Γ_k is a compact convex set, and f is both continuous (with continuous derivatives on $\Gamma_1 \times \dots \times \Gamma_K$) and bounded from below, then an alternating minimization procedure (where every individual variable is minimized in turn repeatedly in a cyclical fashion) shall converge to f^* (see [15]).

Since the mutual information objective function Eq. 9 is bounded below by $I_L \geq 0$, and it is convex and continuous in the variables $\{\mathbf{w}_x^-\}_{x \in \mathcal{X}}$, \mathbf{u}^- , and \mathbf{v}^- with continuous derivatives along the convex constraints (i.e. the reverse PAC bound sub-levelsets $\Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}})$); therefore, an alternating minimization procedure (over each variable in turn) will converge to the solution I_L .

For each input value x with $\{\mathbf{w}_{x'}^-\}_{x' \in \mathcal{X}/x}$, \mathbf{u}^- , and \mathbf{v}^- held constant, we want to determine

$$\mathbf{w}_x^- = \arg \min_{\mathbf{w}_x' \in \Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}}_x, N_x)} D(\mathbf{w}_x' \parallel \mathbf{v}^-). \quad (19)$$

If $\mathbf{v}^- \in \Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}}_x)$, then set $\mathbf{w}_x^- = \mathbf{v}^-$; otherwise, solve the Lagrange multiplier equation (with the constraint $\sum_{y \in \mathcal{Y}} w_y^* = 1$ to force the solution to be a pmf), i.e.

$$\begin{aligned} \mathcal{L}(\mathbf{w}, \omega', \mu') &= D(\mathbf{w} \parallel \mathbf{v}) + \omega' (D(\hat{\mathbf{w}} \parallel \mathbf{w}) - \xi) \\ &\quad + \mu' \left(\sum_{y \in \mathcal{Y}} w_y - 1 \right). \end{aligned} \quad (20)$$

The Lagrange multipliers ω' and μ' collapse into one 'tuning' parameter ω , and we get

$$w_y^*(\omega) = \frac{1}{Z} \frac{\hat{w}_y}{W_0 \left(\omega \frac{\hat{w}_y}{v_y} \right)}, \text{ where } Z = \sum_{y' \in \mathcal{Y}} \frac{\hat{w}_{y'}}{W_0 \left(\omega \frac{\hat{w}_{y'}}{v_{y'}} \right)}.$$

W_0 is the 'zero-branch' of the Lambert W function [16]. We adjust ω (using line search) such that \mathbf{w}^* touches the 'closest point' on the constraint surface $\Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}}_x)$, and set $\mathbf{w}_x^- = \mathbf{w}^*$.

For the minimization of \mathbf{v}^- and given that $\{\mathbf{w}_{x'}^-\}_{x' \in \mathcal{X}}$, and \mathbf{u}^- are held constant, define $\check{\mathbf{v}} = \sum_{x \in \mathcal{X}} u_x^- \mathbf{w}_x^-$ and consider

$$\mathbf{v}^- = \arg \min_{\mathbf{v}' \in \Gamma_\xi^{\text{rev}}(\check{\mathbf{v}}, N)} D(\mathbf{v}' \parallel \mathbf{v}'). \quad (21)$$

Algorithm 1 Mutual information lower bound

```

1: Input:  $\{\hat{\mathbf{w}}_x, N_x\}_{x \in \mathcal{X}}, \hat{\mathbf{u}}, \hat{\mathbf{v}}, \delta_1, \text{tol} \in (0, 1)$ 
2:  $\xi_{\mathbf{u}} \leftarrow \xi(N, |\mathcal{X}|, \delta_1)$ ,  $\xi_{\mathbf{v}} \leftarrow \xi(N, |\mathcal{Y}|, \delta_1)$ , and  $\xi_x \leftarrow \xi(N_x, |\mathcal{Y}|, \delta_1)$ 
3: pick any  $\mathbf{u}^- \in \mathcal{P}_{\mathcal{X}}$  and  $\mathbf{v}^- \in \mathcal{P}_{\mathcal{Y}}$ 
4: repeat
5:    $\hat{\mathbf{v}} \leftarrow \mathbf{v}^-$ 
6:   for  $x \in \mathcal{X}$  do
7:      $\mathbf{w}_x^- \leftarrow \arg \min_{\mathbf{w}_x^- \in \Gamma_{\xi_x}^{\text{rev}}(\hat{\mathbf{w}}_x)} D(\mathbf{w}_x^- \parallel \mathbf{v}^-)$ 
8:   end for
9:    $\mathbf{v}^- \leftarrow \arg \min_{\mathbf{v}^- \in \Gamma_{\xi_{\mathbf{v}}}^{\text{rev}}(\hat{\mathbf{v}})} D\left(\sum_{x \in \mathcal{X}} u_x^- \mathbf{w}_x^- \parallel \mathbf{v}^-\right)$ 
10:   $\mathbf{u}^- \leftarrow \arg \min_{\mathbf{u}^- \in \Gamma_{\xi_{\mathbf{u}}}^{\text{rev}}(\hat{\mathbf{u}})} \sum_{x \in \mathcal{X}} u_x^- D(\mathbf{w}_x^- \parallel \mathbf{v}^-)$ 
11: until  $\|\hat{\mathbf{v}} - \mathbf{v}^-\|_2 \leq \text{tol}$ 
12: Output:  $\{\mathbf{w}_x^-\}_{x \in \mathcal{X}}, \mathbf{v}^-, \mathbf{u}^-$ 

```

The Lagrange multiplier equation is

$$\mathcal{L}(\mathbf{v}, \lambda', \mu') = D(\hat{\mathbf{v}} \parallel \mathbf{v}) + \lambda' (D(\hat{\mathbf{v}} \parallel \mathbf{v}) - \xi) + \mu' \left(\sum_{y \in \mathcal{Y}} v_y \right). \quad (22)$$

The Lagrange multipliers λ' and μ' collapse into one ‘tuning’ parameter λ , and the solution is

$$v_y^*(\lambda) = \lambda \hat{v}_y + (1 - \lambda) v_y.$$

We adjust λ such that \mathbf{v}^* touches the ‘closest point’ on constraint surface $\Gamma_{\xi}^{\text{rev}}(\hat{\mathbf{v}})$ and set $\mathbf{v}^- = \mathbf{v}^*$.

Finally, for the minimization of \mathbf{u}^- with $\{\mathbf{w}_{x'}^-\}_{x' \in \mathcal{X}}$, and \mathbf{v}^- held constant, we define a vector $\tilde{\mathbf{v}}$ such that $\nu_x = D(\mathbf{w}_x^- \parallel \mathbf{v}^-) \geq 0$ and solve

$$\mathbf{u}^- = \arg \min_{\mathbf{u}^- \in \Gamma_{\xi}^{\text{rev}}(\hat{\mathbf{u}}, N)} \sum_{x \in \mathcal{X}} \nu_x u_x'. \quad (23)$$

The Lagrange multiplier equation is

$$\mathcal{L}(\mathbf{u}, \alpha', \mu') = \sum_{x \in \mathcal{X}} \nu_x u_x + \alpha' (D(\hat{\mathbf{u}} \parallel \mathbf{u}) - \xi) + \mu' \left(\sum_{y \in \mathcal{Y}} u_y \right). \quad (24)$$

The Lagrange multipliers α' and μ' collapse into one ‘tuning’ parameter α , and we get

$$u_x^*(\alpha) = \frac{1}{Z} (\nu_x + \alpha) \hat{u}_x, \text{ where } Z = \sum_{x' \in \mathcal{X}} (\nu_{x'} + \alpha) \hat{u}_{x'}.$$

We adjust α such that \mathbf{u}^* touches the ‘closest point’ on the constraint surface $\Gamma_{\xi}^{\text{rev}}(\hat{\mathbf{u}})$ and set $\mathbf{u}^- = \mathbf{u}^*$.

Suppose that we have the scenario depicted in Fig. 1 except that Alice does not choose the channel inputs, but rather Alice and Bob jointly observe the N input-output sample pairs flowing across the channel. Algorithm 1 computes a lower PAC bound on the mutual information by iterating toward the result of Eq. 9. Algorithm 1 is a K -way optimization that sequentially ‘cycles’ over all variables minimizing each one by one. If there is no significant movement in \mathbf{v}^- over one entire cycle, the algorithm declares convergence.

We want *all* $|\mathcal{X}| + 2$ sub-levelsets $\Gamma_{\xi}^{\text{rev}}(\cdot)$ to ‘contain’ the ‘true’ pmf with high probability (*i.e.* $\geq 1 - \delta$). So, we set $\delta_1 = \frac{\delta}{|\mathcal{X}|+2}$ to ensure that all sub-levelset bounds simultaneously hold. We ‘size’ (or ‘tune’) the sub-levelsets, by setting the ξ parameter of $\Gamma_{\xi}^{\text{rev}}(\cdot)$ using either: (1) the ‘log’ sub-levelset bound, which is based on Sanov’s Theorem see [13] section 11.4,

$$\xi^{\log}(N_x, |\mathcal{Y}|, \delta_1) = \frac{|\mathcal{Y}| \ln(N_x + 1) - \ln(\delta_1)}{N_x}, \quad (25)$$

or (2) the ‘loglog’ improved sub-levelset bound $\xi^{\log\log}(N_x, |\mathcal{Y}|, \delta_1)$ *i.e.* Eq. 17.

To bound the rate of convergence as the number of samples increases, we invoke Pinsker’s Theorem [17]; therefore,

$$\sqrt{2D(\hat{\mathbf{w}}_x \parallel \mathbf{w}_x)} \geq \|\hat{\mathbf{w}}_x - \mathbf{w}_x\|_1 \quad (26)$$

and

$$\sqrt{2D(\hat{\mathbf{w}}_x \parallel \mathbf{w}_x^-)} \geq \|\hat{\mathbf{w}}_x - \mathbf{w}_x^-\|_1. \quad (27)$$

The sublevel-sets $\Gamma_{\xi}^{\text{rev}}(\cdot)$ are ‘sized’ such that $D(\hat{\mathbf{w}}_x \parallel \mathbf{w}_x^-) = \xi(N_x, |\mathcal{Y}|, \delta_1)$ and according to Thm II.3, $\hat{\mathbf{w}}_x$ falls outside the sublevel-set with probability $\leq \delta_1$ if $D(\hat{\mathbf{w}}_x \parallel \mathbf{w}_x) = \xi(N_x, |\mathcal{Y}|, \delta_1)$. Therefore, we have an L1-norm bound on how far \mathbf{w}^- can deviate from the true pmf \mathbf{w} ,

$$\begin{aligned} \|\mathbf{w}_x - \mathbf{w}_x^-\|_1 &= \|\hat{\mathbf{w}}_x - \mathbf{w}_x^-\|_1 + \|\hat{\mathbf{w}}_x - \mathbf{w}_x\|_1 \\ &\leq 2\sqrt{2\xi(N_x, |\mathcal{Y}|, \delta_1)} \text{ with prob. } \geq 1 - \delta_1. \end{aligned} \quad (28)$$

Recall that for a RV $X \in \mathcal{X}$ with pmf \mathbf{p} , the entropy is $H(\mathbf{p}) \triangleq \sum_{x \in \mathcal{X}} -p_x \log_2(p_x)$. Given a second RV $X' \in \mathcal{X}$ with pmf \mathbf{p}' , we have the following bound on entropy that is a function of the L1-norm $\|\mathbf{p} - \mathbf{p}'\|_1$ (see [13] Thm 17.3.3),

$$\begin{aligned} |H(\mathbf{p}) - H(\mathbf{p}')| &\leq \Xi(\|\mathbf{p} - \mathbf{p}'\|_1, |\mathcal{X}|) \\ &= \|\mathbf{p} - \mathbf{p}'\|_1 (\log_2(|\mathcal{X}|) - \log_2(\|\mathbf{p} - \mathbf{p}'\|_1)) \\ &= O(\|\mathbf{p} - \mathbf{p}'\|_1 \log(1/\|\mathbf{p} - \mathbf{p}'\|_1)), \end{aligned} \quad (29)$$

where

$$\Xi(x, K) \triangleq x \log_2(K) - \log_2(x) \quad (30)$$

for $x \in \mathbb{R}^+$ and $K \in \mathbb{Z}^+$.

If we define a joint pmf \mathbf{q} as $q_{x,y} \triangleq w_{y|x} u_x$, the mutual information (for the true pmfs \mathbf{u} , \mathbf{v} , and \mathbf{q}) may be written as (see [13] Thm 2.4.1)

$$I_{\text{true}} \triangleq I(X; Y) = H(\mathbf{u}) + H(\mathbf{v}) - H(\mathbf{q}). \quad (31)$$

The pmfs \mathbf{u}^- , \mathbf{v}^- , and \mathbf{q}^- , (where \mathbf{q}^- is defined as $q_{x,y}^- \triangleq w_{y|x}^- u_x^-$) are all within the various sublevel-set constraints to yield I_L ; therefore, we have

$$I_L = H(\mathbf{u}^-) + H(\mathbf{v}^-) - H(\mathbf{q}^-). \quad (32)$$

The absolute value of the difference in mutual information is

Algorithm 2 Lower bound on channel capacity

```

1: Input:  $\{\hat{\mathbf{w}}_x, N_x\}_{x \in \mathcal{X}}, \delta_1, \text{tol} \in (0, 1)$ 
2:  $\xi_x \leftarrow \xi(N_x, |\mathcal{Y}|, \delta_1)$ 
3:  $\mathbf{v}^- \in \mathcal{P}_Y$ 
4: repeat
5:    $\hat{\mathbf{v}} \leftarrow \mathbf{v}^-$ 
6:   for  $x \in \mathcal{X}$  do
7:      $\mathbf{w}_x^- \leftarrow \arg \min_{\mathbf{w}'_x \in \Gamma_{\xi_x}^{\text{rev}}(\hat{\mathbf{w}}_x)} D(\mathbf{w}'_x \| \hat{\mathbf{v}})$ 
8:   end for
9:    $(\mathbf{u}^-, \mathbf{v}^-, R_L) \leftarrow BA(\{\mathbf{w}_x^-\})$ 
10: until  $\|\hat{\mathbf{v}} - \mathbf{v}^-\|_2 \leq \text{tol}$ 
11: Output:  $\{\mathbf{w}_x^-\}_{x \in \mathcal{X}}, \mathbf{u}^-, \mathbf{v}^-, R_L$ 

```

$$\begin{aligned}
\Delta_I &\triangleq |I_{\text{true}} - I_L| \\
&= |H(\mathbf{u}) + H(\mathbf{v}) - H(\mathbf{q}) \\
&\quad - (H(\mathbf{u}^-) + H(\mathbf{v}^-) - H(\mathbf{q}^-))| \\
&\leq |H(\mathbf{u}) - H(\mathbf{u}^-)| + |H(\mathbf{v}) - H(\mathbf{v}^-)| \\
&\quad + |H(\mathbf{q}) - H(\mathbf{q}^-)|.
\end{aligned} \tag{33}$$

The absolute value of the difference in entropy between each true *pmf* \mathbf{p} and the sublevel-set induced lower bound *pmf* \mathbf{p}^- is bounded by $O(\|\mathbf{p} - \mathbf{p}^-\|_1 \log(1/\|\mathbf{p} - \mathbf{p}^-\|_1))$ (see Eq. 29) and $\|\mathbf{p} - \mathbf{p}^-\|_1$ converges towards zero with $O(\sqrt{\xi})$ (see Eq. 28).

With estimates of the *pdfs* \mathbf{u} , \mathbf{v} , and \mathbf{w} , we can compute a lower (PAC) bound on mutual information I_L (based on \mathcal{S}_N) that converges (as N increases) with $\Delta_I = O(\sqrt{\xi} \log(1/\sqrt{\xi}))$. Specifically, the convergence rate when using ξ^{\log} (Eq. 25) is $\Delta_I = O(\log(N)/\sqrt{N})$, and when using the improved sublevel-set bound $\xi^{\log \log}$ (Eq. 17) the convergence rate is $\Delta_I = O(\sqrt{\log(\log(N)) \log(N)/N})$. \square

We return the scenario (see Fig. 1), where Alice probed the channel without feedback during the sampling process, and now given the N input-output samples pairs she aims to find the input *pmf* \mathbf{u} that yields the largest information rate through the channel (assured with a specified high probability $1 - \delta$).

IV. A PAC BOUND ON CHANNEL CAPACITY

To compute a PAC lower bound on channel capacity via Eq. 11, we leverage the modified Blahut-Arimoto algorithm (see [14]). The Blahut-Arimoto algorithm [18] (denoted by the function $BA(\underline{\mathbf{w}})$) computes the unique output *pmf* \mathbf{v}^* and a non-unique input *pmf* \mathbf{u}^* that maximized the mutual information across the known channel law $\underline{\mathbf{w}}$. Algorithm 2 repeatedly invokes the Blahut-Arimoto find the solution within the channel uncertainty sublevel-set constraints. Our goal is to determine at what rate does Algorithm 2 converge towards the channel capacity as the number of channel probes increases.

Suppose we have observed the set of input-output pairs \mathcal{S}_N , and then we select the specific input *pmf* $\hat{\mathbf{u}}$. Algorithm 3 incorporates the following observation (see [14])

Algorithm 3 Channel rate lower bound given *pmf* $\hat{\mathbf{u}}$

```

1: Input:  $\hat{\mathbf{u}}, \{\hat{\mathbf{w}}_x, N_x\}_{x \in \mathcal{X}}, \delta_1, \text{tol} \in (0, 1)$ 
2:  $\xi_x \leftarrow \xi(N_x, |\mathcal{Y}|, \delta_1)$ 
3:  $\mathbf{v}^- \in \mathcal{P}_Y$ 
4: repeat
5:    $\hat{\mathbf{v}} \leftarrow \mathbf{v}^-$ 
6:   for  $x \in \mathcal{X}$  do
7:      $\mathbf{w}_x^- \leftarrow \arg \min_{\mathbf{w}'_x \in \Gamma_{\xi_x}^{\text{rev}}(\hat{\mathbf{w}}_x)} D(\mathbf{w}'_x \| \hat{\mathbf{v}})$ 
8:   end for
9:    $\mathbf{v}^- \leftarrow \arg \min_{\mathbf{v}' \in \Gamma_{\xi}^{\text{rev}}(\hat{\mathbf{v}})} \max_{x \in \mathcal{X}} D\left(\sum_{x \in \mathcal{X}} \hat{u}_x \mathbf{w}_x^- \| \mathbf{v}'\right)$ 
10: until  $\|\hat{\mathbf{v}} - \mathbf{v}^-\|_2 \leq \text{tol}$ 
11:  $R_L(\hat{\mathbf{u}}) \leftarrow \sum_{x \in \mathcal{X}} \hat{u}_x D(\mathbf{w}_x^- \| \mathbf{v}^-)$ 
12: Output:  $\{\mathbf{w}_x^-\}_{x \in \mathcal{X}}, \mathbf{v}^-, R_L(\hat{\mathbf{u}})$ 

```

$$\begin{aligned}
C &\triangleq \max_{\mathbf{u} \in \mathcal{P}_X} \sum_{x \in \mathcal{X}} u_x D(\mathbf{w}_x \| \mathbf{v}(\mathbf{u})) \\
&= \max_{\mathbf{u} \in \mathcal{P}_X} \min_{\mathbf{v} \in \mathcal{P}_Y} \sum_{x \in \mathcal{X}} u_x D(\mathbf{w}_x \| \mathbf{v}) \\
&= \min_{\mathbf{v} \in \mathcal{P}_Y} \max_{\mathbf{u} \in \mathcal{P}_X} \sum_{x \in \mathcal{X}} u_x D(\mathbf{w}_x \| \mathbf{v}) \\
&= \min_{\mathbf{v} \in \mathcal{P}_Y} \max_{x \in \mathcal{X}} D(\mathbf{w}_x \| \mathbf{v}),
\end{aligned} \tag{34}$$

into line 9 to compute the maximum assured information rate R_L of the channel given the input *pmf* $\hat{\mathbf{u}}$.

Next, we evaluate $|R_{\text{true}}(\hat{\mathbf{u}}) - R_L(\hat{\mathbf{u}})|$ by bounding the absolute value of each entropy difference (see Eq. 33). Suppose we observed N input-output pairs, and let $\underline{\mathbf{N}} \triangleq [N_x]_{x \in \mathcal{X}}$ be the vector of the number of occurrences of each input value x in \mathcal{S}_N . Alice could construct a codebook such that input values are distributed according to any selected *pmf* $\hat{\mathbf{u}}$. When using this codebook, $\hat{\mathbf{u}}$ would match the *true* input *pmf* \mathbf{u} , and so $|H(\mathbf{u}) - H(\hat{\mathbf{u}})| = 0$. For the channel output *pmf* $\mathbf{v}(\hat{\mathbf{u}})$, we have

$$\begin{aligned}
\|\mathbf{v}(\hat{\mathbf{u}}) - \mathbf{v}^-(\hat{\mathbf{u}})\|_1 &= \left\| \sum_{x \in \mathcal{X}} \mathbf{w}_x \hat{u}_x - \sum_{x \in \mathcal{X}} \mathbf{w}_x^- \hat{u}_x \right\|_1 \\
&= \left\| \sum_{x \in \mathcal{X}} (\mathbf{w}_x - \mathbf{w}_x^-) \hat{u}_x \right\|_1 \\
&\leq \sum_{x \in \mathcal{X}} \|\mathbf{w}_x - \mathbf{w}_x^-\|_1 \hat{u}_x \\
&\leq \sum_{x \in \mathcal{X}} \hat{u}_x 2\sqrt{2\xi(N_x, |\mathcal{Y}|, \delta_1)},
\end{aligned} \tag{35}$$

and inserting this result into Eq. 29 and Eq. 30 yields

$$|H(\mathbf{v}(\hat{\mathbf{u}})) - H(\mathbf{v}^-(\hat{\mathbf{u}}))| = \Xi \left(\sum_{x \in \mathcal{X}} \hat{u}_x 2\sqrt{2\xi(N_x, |\mathcal{Y}|, \delta_1)} \right). \tag{36}$$

We see that the bound on $|H(\mathbf{v}(\hat{\mathbf{u}})) - H(\mathbf{v}^-(\hat{\mathbf{u}}))|$ will

vary with the *pmf* $\hat{\mathbf{u}}$ unless the number of channel probes N_x are equal for each input value $x \in \mathcal{X}$.

Finally, for the joint *pmf* \mathbf{q} , we have

$$\begin{aligned}
|H(\mathbf{q}(\hat{\mathbf{u}})) - H(\mathbf{q}^-(\hat{\mathbf{u}}))| &= \left| \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} w_{y|x} \hat{u}_x \log(w_{y|x} \hat{u}_x) - \right. \\
&\quad \left. \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} w_{y|x}^- \hat{u}_x \log(w_{y|x}^- \hat{u}_x) \right| \\
&= \left| \sum_{x \in \mathcal{X}} (H(\mathbf{w}_x) - H(\mathbf{w}_x^-)) \hat{u}_x - \right. \\
&\quad \left. \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (w_{y|x} - w_{y|x}^-) \hat{u}_x \log(\hat{u}_x) \right| \\
&\leq \sum_{x \in \mathcal{X}} \hat{u}_x |H(\mathbf{w}_x) - H(\mathbf{w}_x^-)| \\
&\leq \sum_{x \in \mathcal{X}} \hat{u}_x \Xi(\|\mathbf{w}_x - \mathbf{w}_x^-\|_1),
\end{aligned} \tag{37}$$

and so

$$|H(\mathbf{q}(\hat{\mathbf{u}})) - H(\mathbf{q}^-(\hat{\mathbf{u}}))| \leq \sum_{x \in \mathcal{X}} \hat{u}_x \Xi(2\sqrt{2\xi(N_x, |\mathcal{Y}|, \delta_1)}). \tag{38}$$

Overall we have

$$\begin{aligned}
\Delta_R(\hat{\mathbf{u}}) &\triangleq |R_{\text{true}}(\hat{\mathbf{u}}) - R_L(\hat{\mathbf{u}})| \\
&\leq \Xi \left(\sum_{x \in \mathcal{X}} \hat{u}_x 2\sqrt{2\xi(N_x, |\mathcal{Y}|, \delta_1)} \right) + \\
&\quad \sum_{x \in \mathcal{X}} \hat{u}_x \Xi(2\sqrt{2\xi(N_x, |\mathcal{Y}|, \delta_1)}) \\
&= O(\sqrt{\xi} \log(1/\sqrt{\xi})),
\end{aligned} \tag{39}$$

where $\xi = \xi(N, |\mathcal{Y}|, \delta_1)$. When given the input-output pairs \mathcal{S}_N , then for any selected $\hat{\mathbf{u}}$ the minimum assured information rate is $R_L(\hat{\mathbf{u}})$. The convergence rate when using ξ^{\log} (Eq. 25) is $\Delta_R(\hat{\mathbf{u}}) = O(\log(N)/\sqrt{N})$, and when using the improved sublevel-set bound $\xi^{\log\log}$ (Eq. 17) the convergence rate is $\Delta_R(\hat{\mathbf{u}}) = O(\sqrt{\log(\log(N))} \log(N)/N)$. \square

Define $R_U(\hat{\mathbf{u}}) \triangleq R_L(\hat{\mathbf{u}}) + \Delta(\hat{\mathbf{u}}, \mathbf{N}, |\mathcal{Y}|, \delta_1)$ to be a high probability upper bound on the information rate through the channel given that the input values in the codebook are distributed according to $\hat{\mathbf{u}}$.

Fig. 2 depicts the case where all input values had been probed an equal number of times; therefore, $\Delta_R(\hat{\mathbf{u}})$ is constant in $\hat{\mathbf{u}}$. The x-axis represents the multidimensional probability space $\mathcal{P}_{\mathcal{X}}$. Alice and Bob could vary the proposed codebook (input) *pmf* $\hat{\mathbf{u}}$ to maximize $R_L(\hat{\mathbf{u}})$ by evaluating each proposed $\hat{\mathbf{u}}$ using Algorithm 3. Suppose $\mathbf{u}^* \triangleq \arg \max_{\mathbf{u}' \in \mathcal{P}_{\mathcal{X}}} R_L(\mathbf{u}')$, then we know that the channel capacity C is such that $R_L(\mathbf{u}^*) \leq C \leq R_U(\mathbf{u}^*)$ with high probability.

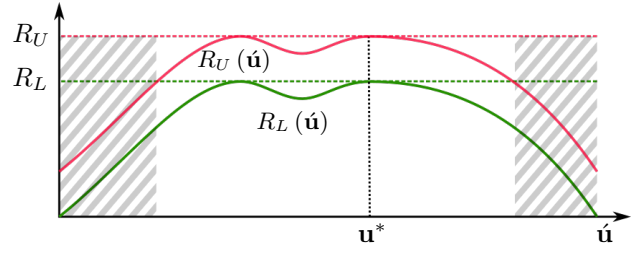


Fig. 2. Equal probing of each input value

If all input values had been sampled more times (before computing $R_L(\hat{\mathbf{u}})$ and $R_U(\hat{\mathbf{u}})$), then the gap between the curves $R_L(\hat{\mathbf{u}})$ and $R_U(\hat{\mathbf{u}})$ decreases such that $R_L(\mathbf{u}^*) = C$ as $N \rightarrow \infty$.

We can remove from consideration any input *pmf* $\hat{\mathbf{u}}$ in the depicted gray-shaded areas because $R_U(\hat{\mathbf{u}}) \leq R_L(\mathbf{u}^*)$, and so these input *pmfs* can not improve the information rate through the channel given any amount of additional channel law information.

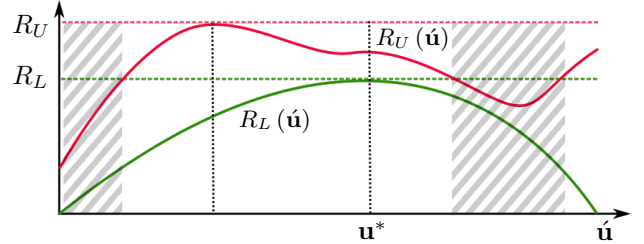


Fig. 3. Unequal probing of each input value

The bounds hold with probability $\geq 1 - \delta$ given a fixed set of input-output pairs \mathcal{S}_N . Fig. 3 depicts the case where inputs are probed an unequal number of times and so the gap between $R_L(\hat{\mathbf{u}})$ and $R_U(\hat{\mathbf{u}})$ varies as a function of $\hat{\mathbf{u}}$. If we can discover an input value x'' such that $\hat{u}_{x''} = 0$ for all $\hat{\mathbf{u}} \in \{\mathbf{u}': R_U(\mathbf{u}') \leq R_L(\mathbf{u}^*)\}$, then additional knowledge about input value x'' would not raise $R_L(\mathbf{u}^*)$. We believe that channel probing could continue and maintain a similar high probability bounds with the same convergence rate by dynamically adjusting δ (i.e. incorporating the results in [11]). We plan to optimize the channel probing strategy in future work.

V. RESULTS

To test the convergence rate of Algorithm 1, we generated a set of channel laws (probabilistic mappings), where the number of input symbol values was fixed at $|\mathcal{X}| = 3$ and the number of discrete observed output values was varied as $|\mathcal{Y}| \in \mathcal{Y} = \{2, 3, 5, 7, 10, 15, 20, 25, 30, 35\}$. Each \mathbf{w}_x of each channel law was drawn i.i.d. according to a uniform Dirichlet distribution [19] with hyperparameter $\alpha = 0.8$ or $\mathbf{w}_x \sim \text{Dir}(\alpha)$, where

$$\text{Dir}(\alpha) \propto \prod_{y \in \mathcal{Y}} w_{y|x}^{\alpha-1} \quad \forall y \in \mathcal{Y} \tag{40}$$

to yield the set of ten test channel laws $W \triangleq \{\mathbf{w}_{|\mathcal{Y}|}\}_{|\mathcal{Y}| \in \mathcal{Y}}$.

We want to determine the ‘tightness’ of our sublevel-set bounds against an optimally sized sub-levelset. We used Monte Carlo integration to approximately ‘size’ each sublevel-set to the optimal value. Specifically, we generated $M = 10000$ samples from a Dirichlet distribution [19] with hyperparameter $\alpha = 1$ or $\mathbf{w}_x \sim \text{Dir}(\hat{\mathbf{w}}_x, N_x, \alpha)$, where

$$\text{Dir}(\hat{\mathbf{w}}_x, N_x, \alpha) \propto \prod_{y \in \mathcal{Y}} w_{y|x}^{N_x \hat{w}_x + \alpha - 1} \quad \forall y \in \mathcal{Y}, \quad (41)$$

and then adjusted the ξ parameter of the sublevel-set until $[M\delta_1]$ of the M samples fell outside the sub-levelset $I_\xi^{\text{rev}}(\cdot)$. These Monte Carlo based sublevel-sets are the convex regions in solving for I_L^{mc} (see Eq. 9) for the bound on mutual information).

We compare I_L^{mc} to the I_L^{log} (based on using sub-levelsets ‘sized’ using ξ^{log} see Eq. 25) and I_L^{loglog} (based on using sub-levelsets ‘sized’ using ξ^{loglog} see Eq. 17).

We ran Algorithm 1 on each of the ten test channel laws using parameters $\delta = 0.01$, and $N_0 \in \{10 \cdot 2^k\}_{k=0}^{19}$. Every input value $x \in \mathcal{X}$ was sampled N_0 times (so the total number of samples N per each tested channel law was $|\mathcal{X}| N_0$).

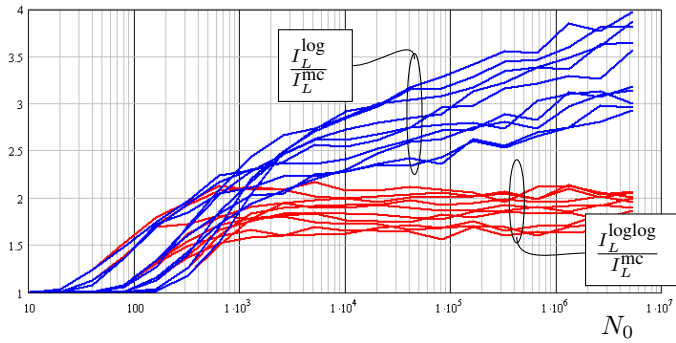


Fig. 4. Convergence as the number of samples per input value increases

Fig. 4 shows one curve for each test channel law as the number of samples N is increased. The $I_L^{\text{log}}/I_L^{\text{mc}}$ curves are the ratio of sublevel-set bound $I_L^{\text{log}}(N)$ ‘over’ the Monte Carlo estimate $I_L^{\text{mc}}(N)$, and $I_L^{\text{loglog}}/I_L^{\text{mc}}$ is similarly the ratio of sublevel-set bound $I_L^{\text{loglog}}(N)$ over the Monte Carlo estimate. We see that the $I_L^{\text{loglog}}/I_L^{\text{mc}}$ curves ‘flatten out’ (for $N_0 > 1000$ samples), and so the ‘loglog’ sublevel-set bound appears to track (match) the optimal convergence rate within a constant factor. While the ‘loglog’ sublevel-set bound tracks the optimal estimate with some fixed constant loss (because the ratio is not identical to one), the ‘log’ (Sanov-based) sublevel-set bound diverges from the optimal estimate as N increases.

VI. CONCLUSIONS AND FUTURE WORK

We developed and demonstrated an algorithm that establishes a high probability lower bound I_L on the mutual information between the input RV and the output RV for a DMC given N input-output sample pairs. We also derived the ‘big Oh’ convergence rate of I_L as the number of samples is increased. We demonstrated via Monte Carlo simulation that

our bound converges near the optimal rate. In addition, we formulated an algorithm that establishes a high probability bound on the maximum assured information rate R_L across a channel that matches the same ‘big Oh’ convergence rate to approach channel capacity.

In future work, we hope that the convergence rate of I_L and R_L (i.e. $O(\sqrt{\log(\log(N)) \log(N)/N})$) can be further reduced by a factor of $\sqrt{\log(N)}$ to match the convergence rate of the L1 norm bound, $\|\mathbf{w}_x - \mathbf{w}_x^-\|_1$ (see Eq. 28); however, this may not be possible.

The ‘big Oh’ convergence rate of I_L and R_L did not require specifying the channel probing strategy; however, a next step is to utilize these bounds to optimize the channel probing process (active online learning).

REFERENCES

- [1] J. K. I. Csiszár, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.
- [2] A. Lapidoth and P. Narayan, “Reliable communication under channel uncertainty,” *IEEE Trans. on Information Theory*, vol. 44, pp. 2148–2177, 1998.
- [3] T. Han, *Information-Spectrum Methods in Information Theory*. Springer-Verlag, 2003.
- [4] L. G. Valiant, “A theory of the learnable,” *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.
- [5] J. Langford, “Tutorial on practical prediction theory for classification,” *Journal of Machine Learning Research*, vol. 6, pp. 273–306, 2005.
- [6] L. Paninski, “Estimation of entropy and mutual information,” *Neural Computation*, vol. 15, pp. 1191–1253, 2003.
- [7] I. Chattopadhyay and H. Lipson, “Computing entropy rate of symbol sources and distribution-free limit theorem,” 2014.
- [8] N. VanderKratts and A. Banerjee, “A finite-sample, distribution-free, probabilistic lower bound on mutual information,” *Journal of Neural Computation*, vol. 23, no. 7, pp. 1862–1898, 2011.
- [9] Y. Seldin and N. Tishby, “PAC-Bayesian analysis of co-clustering and beyond,” *Journal of Machine Learning Research*, vol. 11, pp. 3595–3636, 2010.
- [10] B. Guedj, “A primer on PAC-Bayesian learning,” 2019.
- [11] M. A. Tope and J. M. Morris, “Near optimal channel rate discovery for discrete memoryless binary output channels,” *IEEE Military Communications Conference (MILCOM)*, pp. 483–488, 2017.
- [12] —, “Improvements to Sanov and PAC sublevel-set bounds for discrete random variables,” *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 2021.
- [13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley and Sons, 1991.
- [14] M. A. Tope and J. M. Morris, “A PAC-bound on the channel capacity of an observed discrete memoryless channel,” *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 2021.
- [15] R. W. Yeung and T. Berger, “Multi-way alternating minimization,” *Proceedings of 1995 IEEE International Symposium on Information Theory*, p. 74, 1995.
- [16] “Lambert W function — Wikipedia, the free encyclopedia,” 2021. [Online]. Available: https://en.wikipedia.org/wiki/Lambert_W_function
- [17] T. van Erven and P. Harremoës, “Rényi divergence and Kullback-Leibler divergence,” *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.
- [18] R. Blahut, “Computation of channel capacity and rate-distortion functions,” *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.
- [19] “Dirichlet distribution — Wikipedia, the free encyclopedia,” 2021. [Online]. Available: https://en.wikipedia.org/wiki/Dirichlet_distribution
- [20] D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized*. Cambridge University Press, 2009.
- [21] “Minimax theorem — Wikipedia, the free encyclopedia,” 2021. [Online]. Available: https://en.wikipedia.org/wiki/Minimax_theorem