M. A. Tope and J. M. Morris, "Near-optimal Sampling to Optimize Communication Over Discrete Memoryless Channels," 2023 57th Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA, 2023, pp. 1-6, doi: 10.1109/CISS56502.2023.10089651.

https://doi.org/10.1109/CISS56502.2023.10089651

Near-optimal Sampling to Optimize Communication Over Discrete Memoryless Channels

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 develops a strategy to minimize the number of channel probes required to recover the components of the channel law and maximize the reliable communication rate across a discrete memoryless channel (DMC). Based on the aggregate set of observed input-output pairs over time, the algorithm sequentially probes subsets of channel input values. We leverage a non-asymptotic probably approximately correct (PAC) bounds to establish the rate of convergence towards channel capacity as $O(\sqrt{\log(\log(N))\log(N)/N})$, where N is the number of channel probes. For a discrete channel with $|\mathcal{X}|$ input values and $|\mathcal{Y}|$ output values, the sampling strategy may reduce the sample complexity by a factor of nearly $\min(|\mathcal{X}|/|\mathcal{Y}|, 1)$ relative to previous methods.**

I. Introduction

When channel knowledge is incomplete or impaired, channel sampling or probing is often employed to gain channel information to aid the coding and modulation processing chain to provide reliable communications. This paper proposes a near-optimal sampling strategy that meets the limiting convergence rate in recovering such information for the discrete memoryless channel (DMC).

Consider the scenario depicted in Fig. 1, where the channel transition probabilities are initially unknown. The sender, Alice, sequentially selects channel input values and the receiver, Bob, records each the observed channel output value. Once Alice and Bob jointly have sufficient channel knowledge, they design and implement a suitable encoder and decoder to provide reliable communication at rate R with high probability. Our goal is to minimize the sample complexity, which is the number of channel probes/samples required to assure a rate R, where R less than but arbitrarily close to the channel capacity C.

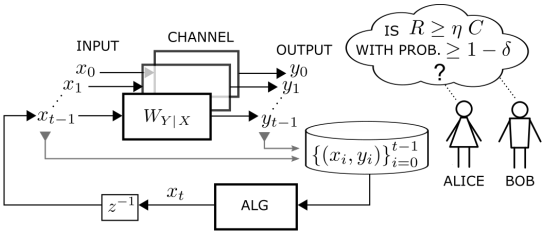

Fig. 1. Maximize a lower bound on channel capacity

During the channel sampling process, suppose Alice and Bob have gathered of 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) $y \in \mathcal{Y} = \{b_0, b_1, \ldots, b_{|\mathcal{Y}|-1}\}$ that corresponds to the instance when input value $x \in \mathcal{X} = \{a_0, a_1, \ldots, a_{|\mathcal{X}|-1}\}$ was applied to the channel. Each sample pair (x_n, y_n) is an instance of a random variable (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 $\mathbf{u} \triangleq [u_{a_0}, u_{a_1}, \ldots, u_{a_{|\mathcal{X}|-1}}] \in \mathcal{P}_{\mathcal{X}}$ and $u_x \triangleq \mathbb{P}\{X = x\}$. Similarly, the marginal *pmf* $\mathbf{v}$ of the channel output RV Y is $\mathbf{v} \triangleq [v_{b_0}, v_{b_1}, \ldots, v_{b_{|\mathcal{Y}|-1}}] \in \mathcal{P}_{\mathcal{Y}}$ where $v_y \triangleq \mathbb{P}\{Y = y\}$.

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

Our approach leverages results for communication over uncertain channels [1], including the compound DMC [2]. We modify the *information spectrum* approach of Han [3] to form a probabilistic compound channel, where constraints are based on probability approximately correct (PAC) bounds [4]. Langford [5] outlined the application of PAC-bounds to machine-learning; however, we require PAC bounds specifically for information theoretic measures. VanderKratts et al. [6] introduced the use of PAC bounds (aka concentration inequalities) on mutual information for datasets with binary input values (and continuous-valued observations). Our sublevel-set (PAC) bound [7] is similar to the PAC-Bayesian bounds of Seldin et al. [8] (also see [9] for a survey and review).

Our channel sampling strategy is based on multi-arm bandit (MAB) algorithms (on-line active learning), where one attempts to 'play' various 'slot-machines' in a manner yielding the highest average payout. Our procedure mimics the upper confidence bound (UCB) methods of Auer et al. [10] and Cesa-Bianchi et al. [11], and we leverage the *law of iterated logarithms* (LIL), which was incorporated into MAB algorithms by Jamieson et al. [12]. In particular, we employ the technique described by Pollard in [13] to create batches of samples in exponentially increasing block sizes to control the number of decisions and ensure all bounds hold with high probability throughout the process.

This paper: (1) extends our results [14] from memoryless binary input-output channels to DMCs and (2) improves upon methods in [15], [16] to provide near-optimal channel input sampling.

The remainder of this paper is as follows: we modify several non-asymptotic sublevel-set bounds (Section II) to reflect the uncertainly about the channel transition probabilities. In Section III, we bound the rate of convergence of a lower bound on channel capacity. In section IV, we describe an online algorithm, whose performance matches this convergence rate. In Section V, we present results from simulated channels. We finish with conclusions and discuss future work (Section VI).

II. Probabilistic Compound Channel

When the channel law $\underline{\mathbf{w}}$ is known, the average mutual information between the RVs X and Y is

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

where the expected output *pmf* $\hat{\mathbf{v}}\,(\mathbf{u}) \triangleq \sum_{x\in\mathcal{X}} u_x \mathbf{w}_x$, and the Kullback Leibler (KL) divergence [2] for discrete RVs P and Q with respective *pmf*s $\mathbf{p} \in \mathcal{P}_\mathcal{Y}$ and $\mathbf{q} \in \mathcal{P}_\mathcal{Y}$ is

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

From $\mathcal{S}_N$, we compute the empirical *pmf*s $\hat{\mathbf{w}}_x = [\hat{w}_{b_0\,|\,x}, \hat{w}_{b_1\,|\,x}, \ldots, \hat{w}_{b_{|\mathcal{Y}|-1}\,|\,x}]$, where

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

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

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

$$\mathbb{P}\left\{\acute{\mathbf{w}} \in \mathit{\Gamma}_\xi^{\text{rev}}\,(\hat{\mathbf{w}})\right\} \geq 1-\delta, \qquad (4)$$

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

$$\mathit{\Gamma}_\xi^{\text{rev}}\,(\hat{\mathbf{w}}) \triangleq \{\acute{\mathbf{w}}\colon D(\hat{\mathbf{w}}\,\|\acute{\mathbf{w}}) \leq \xi,\ \ \forall\acute{\mathbf{w}} \in \mathcal{P}_\mathcal{Y}\}. \qquad (5)$$

Our approach [15] is to use these sub-levelsets to form a compound channel based on the PAC bounds and solve for the maximum assured 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 $\underline{\mathbf{w}} \in \mathit{\Gamma}$. When $\mathit{\Gamma}$ is convex and closed, the channel capacity is given by [1]

$$C = \min_{\mathbf{w}\in\mathit{\Gamma}}\ \max_{\mathbf{u}\in\mathcal{P}_\mathcal{X}} I(\mathbf{u}\ ,\mathbf{w}). \qquad (6)$$

Let R_L be the worst case (minimum capacity) channel over all channel laws within the closed and convex region $\underline{\mathbf{w}} \in \mathit{\Gamma}$, and then there exists a codebook [1] that will support rates up to R_L for any channel law $\underline{\mathbf{w}} \in \mathit{\Gamma}$.

Given the sequential channel sampling process, we repeatedly update the empirical *pmf*s (incorporating newly observed samples) and re-evaluate the probabilistic sub-levelset bounds to guide the selection of inputs values to sample. Eventually, when the sampling process terminates, we want the the final lower bound rate R_L to be valid (to 'hold') with a specified probability of at least $1-\delta_1$. The probability that *any* probabilistic sub-levelset bounds fails to hold over all repeated re-evaluations throughout the entire sampling process must be $< \delta_1$.

A key insight (as noted in [14]) from the law of iterated logarithms (see [13]) is to 'expand' the bound (sub-exponentially) with each re-evaluation according to

$$\delta_\tau = \delta_{\mathit{\Gamma}} \frac{6}{\pi^2\,(\tau+1)^2}\ , \qquad (7)$$

where $\delta_{\mathit{\Gamma}}$ is the probability that one of the sub-levelset bounds holds for all re-evaluations during the entire channel sampling process and τ is the number of *re-evaluations* of the bound thus far. By the union bound,

$$\mathbb{P}\,\{\text{any sub-levelset bound evaluation fails}\} \leq \sum_{\tau=0}^{\infty} \delta_\tau = \delta_{\mathit{\Gamma}}; \qquad (8)$$

therefore, the probability that each sub-levelset bound remains valid (holds) is $\geq 1-\delta_{\mathit{\Gamma}}$.

As the number of observations N increases, the 'size' of the sublevel-set constraints (based on Eq. 5) 'shrink,' according to Sanov's theorem.

Theorem II.1 *Sanov's Theorem (see [17] section 11.4) Let $\hat{\mathbf{w}}$ be the empirical pmf of $\mathcal{S}_N = \{y_0, y_1, \ldots, y_{N-1}\}$, then given **any** region $\mathit{\Gamma} \subset \mathcal{P}_\mathcal{Y}$ and $\mathbf{w}^*$ the 'closest' pmf among all $\acute{\mathbf{w}} \in \mathit{\Gamma}$ to $\mathbf{w}$ in terms of the KL divergence*

$$\mathbf{w}^* = \arg\ \min_{\acute{\mathbf{w}}\in\mathit{\Gamma}} D(\acute{\mathbf{w}}\,\|\mathbf{w}) \qquad (9)$$

then

$$\delta_{\mathit{\Gamma}} \triangleq \mathbb{P}\,\{\hat{\mathbf{w}} \notin \mathit{\Gamma}\} \leq (N+1)^{|\mathcal{Y}|}\exp(-ND(\mathbf{w}^*\,\|\mathbf{w})). \qquad (10)$$

Solving for $\delta_{\mathit{\Gamma}}$, we get a sublevel-set bound, where

$$\xi\,(N, |\mathcal{Y}|\,, \delta_{\mathit{\Gamma}}) = D(\mathbf{w}^*\,\|\mathbf{w}) \leq \frac{|\mathcal{Y}|\ln(N+1) - \ln(\delta_{\mathit{\Gamma}})}{N} \qquad (11)$$

sets the 'size' of the sublevel-set. Inserting the bound 'expansion' to cover repeated re-evaluations (see Eq. 7), we have

$$\begin{aligned}\xi\,(N, |\mathcal{Y}|\,, \delta_{\mathit{\Gamma}}, \tau) &\leq \frac{|\mathcal{Y}|\ln(N+1) - \ln\!\left(\delta_{\mathit{\Gamma}}\frac{6}{\pi^2(\tau+1)^2}\right)}{N} \\ &= \frac{|\mathcal{Y}|\ln(N+1) + 2\ln(\tau+1) + \kappa_0}{N} \qquad (12)\end{aligned}$$

where $\kappa_0 = \ln(\delta_{\mathit{\Gamma}}) + \ln\!\left(\frac{6}{\pi^2}\right)$. If we re-evaluate the sub-levelset bound after each new input-output sample pair is observed, then $\tau = N$ and we have

$$\xi^{\log}\,(N, |\mathcal{Y}|\,, \delta_{\mathit{\Gamma}}) \leq \frac{(|\mathcal{Y}|+2)\ln(N+1) + \kappa_0}{N}, \qquad (13)$$

and the convergence rate of the 'size' ξ of the sub-levelset is $O\,(\ln(N)/N)$.

The following theorem sharpens the 'Sanov' sublevel-set bound to $O\,(\ln(\ln(N))/N)$.

Theorem II.2 *Improved Sub-levelset Bound [7]*
Given the set $\mathcal{S}_N = \{y_0, y_1, \ldots, y_{N-1}\}$ of outcomes from N i.i.d. discrete random variables $Y_n \in \mathcal{Y}$ and $Y_n \sim \mathbf{w}$ for $n = 0, 1, \ldots, N-1$. Let $\hat{\mathbf{w}}$ be the empirical pmf of $\mathcal{S}_N$, and select any $\delta_\Gamma \in (0, 1]$, then $\mathbb{P}\{\hat{\mathbf{w}} \notin \mathit{\Gamma}_\xi(\mathbf{w})\} \leq \delta_\Gamma$ for the sub-levelset $\mathit{\Gamma}_\xi(\mathbf{w}) \triangleq \{\acute{\mathbf{w}} \colon D(\acute{\mathbf{w}} \,\|\mathbf{w}) \leq \xi,\ \ \forall \acute{\mathbf{w}} \in \mathcal{P}_\mathcal{Y}\}$ with 'size'

$$\xi \geq \frac{1}{N}\left(\frac{1}{2}\ln(2\left|\mathcal{Y}\right|) - \frac{3}{2}\ln\left(\frac{\delta_\Gamma}{2}\right) + \left|\mathcal{Y}\right|\ln\left(\log_2(\log_2(N)) \right.\right.$$
$$\left.\left. + \kappa_1\sqrt{\left|\mathcal{Y}\right|} + \log_2(\kappa_2\left|\mathcal{Y}\right|) + 2\right)\right) - \ln\left(\delta_\Gamma \frac{6}{\pi^2\left(\tau+1\right)^2}\right), \tag{14}$$

where $\kappa_1 = 2\sqrt{24}\left(1+\sqrt{2}\right)$, $\kappa_2 = 24$, and τ is the number of re-evaluations of the bound

If we limit the re-evaluation of the iterated log sub-levelset bound to only the moments when the number of samples N is a power of two, then $\tau = \lceil \log_2(N) \rceil$, then the convergence rate remains $O\left(\ln(\ln(N))/N\right)$.

Since the observed empirical *pmf* $\hat{\mathbf{w}}$ is within $\mathit{\Gamma}_\xi(\mathbf{w})$ with probability $> 1-\delta_\Gamma$, then the generator *pmf* $\mathbf{w}$ must be located within the probability space such that $D(\hat{\mathbf{w}} \,\|\mathbf{w}) \leq \xi$ with probability $> 1 - \delta_\Gamma$. We define the sublevel-set $\mathit{\Gamma}^{\text{rev}}_\xi(\hat{\mathbf{w}})$ (see Eq. 5) and the $\mathbb{P}\left\{\mathbf{w} \in \mathit{\Gamma}^{\text{rev}}_\xi(\hat{\mathbf{w}})\right\} \geq 1 - \delta_\Gamma$ to constrain the channel law uncertainty with high probability.

We want the aggregate of $|\mathcal{X}|$ sub-levelsets $\mathit{\Gamma}^{\text{rev}}_\xi(\cdot)$ to 'contain' the 'true' *pmf* $\mathbf{w}$ with high probability (*i.e.* $\geq 1-\delta_1$). So, we set $\delta_\Gamma = \frac{\delta_1}{|\mathcal{X}|}$ to ensure that all sub-levelset bounds simultaneously hold. We 'size' (or 'tune') the sub-levelsets, by setting the ξ parameter of $\mathit{\Gamma}^{\text{rev}}_\xi(\cdot)$ using either: (1) the 'log' sub-levelset bound Eq. 13 or (2) the 'loglog' improved sub-levelset bound $\xi^{\text{loglog}}\left(N_x, |\mathcal{Y}|, \delta_1\right)$ Eq. 14.

The set of $|\mathcal{X}|$ closed convex sub-levelset constraints contain the true channel law with probability $\geq 1 - \delta_1$. Algorithm 1 repeatedly invokes the Blahut-Arimoto algorithm [18] to maximize the lower communication rate bound R_L within these constraints. Algorithm 1 outputs

$$R_L \triangleq \min_{\substack{\mathbf{w}^-_x \in \mathit{\Gamma}^{\text{rev}}_{\xi_w}(\hat{\mathbf{w}}_x) \\ \forall x \in \mathcal{X}}} \ \max_{\mathbf{u}^- \in \mathcal{P}_\mathcal{X}} I\left(\mathbf{u}^-, \underline{\mathbf{w}}^-\right) \tag{15}$$

and the associated unique fixed point solution $\{\mathbf{w}^-_x\}_{x\in\mathcal{X}}$ and $\mathbf{v}^-$. The 'support' of the rate R_L is the set of input values $\mathcal{X}^* \triangleq \{x \colon D(\mathbf{w}^-_x \,\|\mathbf{v}^-) = R_L \ \forall x \in \mathcal{X}\}$. The *pmf*s $\mathbf{w}^-_x$ in the support $\mathcal{X}^*$ will lie on the surface of their respective sub-levelset constraint (see [15] for details).

III. Bounding the channel rate convergence

We want determine how quickly the $\{\mathbf{w}^-_x\}_{x\in\mathcal{X}}$ approaches the true channel law as the number of channel samples increases. Invoking Pinsker's Theorem [19], we have

$$\sqrt{2D(\hat{\mathbf{w}}_x \,\|\mathbf{w}_x)} \geq \|\hat{\mathbf{w}}_x - \mathbf{w}_x\|_1 \tag{16}$$

and

$$\sqrt{2D\left(\hat{\mathbf{w}}_x \,\|\mathbf{w}^-_x\right)} \geq \left\|\hat{\mathbf{w}}_x - \mathbf{w}^-_x\right\|_1 . \tag{17}$$

Algorithm 1 Lower bound on channel capacity given samples

```
Input: $\{\hat{\mathbf{w}}_x, N_x\}_{x\in\mathcal{X}}$, $\delta_1$, $tol \in (0,1)$ $\{\ tol = 10^{-5}\}$
$\xi_x \leftarrow \xi\left(N_x, |\mathcal{Y}|, \delta_1/|\mathcal{X}|\right)\ \forall x \in \mathcal{X}$
$\mathbf{v}^- \in \mathcal{P}_\mathcal{Y}$
repeat
    $\acute{\mathbf{v}} \leftarrow \mathbf{v}^-$
    for $x \in \mathcal{X}$ do
        $\mathbf{w}^-_x \leftarrow \arg \min_{\mathbf{w}'_x \in \mathit{\Gamma}^{\text{rev}}_{\xi_x}(\hat{\mathbf{w}}_x)} D(\mathbf{w}'_x \,\|\acute{\mathbf{v}})$
    end for
    $(\mathbf{u}^-, \mathbf{v}^-, R_L) \leftarrow$ Blahut-Arimoto $(\{\mathbf{w}^-_x\}, tol)$
until $\|\acute{\mathbf{v}} - \mathbf{v}^-\|_2 \leq tol$
Output: $\{\mathbf{w}^-_x\}_{x\in\mathcal{X}}$, $\mathbf{u}^-$, $\mathbf{v}^-$, $R_L$
```

The sublevel-sets $\mathit{\Gamma}^{\text{rev}}_\xi(\cdot)$ are 'sized' such that $D(\hat{\mathbf{w}}_x \,\|\mathbf{w}^-_x) = \xi\left(N_x, |\mathcal{Y}|, \delta_\Gamma\right)$ and according to Thm II.2, $\hat{\mathbf{w}}_x$ falls outside the sublevel-set with probability $\leq \delta_\Gamma$. Likewise $D(\hat{\mathbf{w}}_x \,\|\mathbf{w}_x) \leq \xi\left(N_x, |\mathcal{Y}|, \delta_\Gamma\right)$; therefore, we have an L1-norm bound on how far $\mathbf{w}^-$ can deviate from the true *pmf* $\mathbf{w}$,

$$\begin{aligned}\left\|\mathbf{w}_x - \mathbf{w}^-_x\right\|_1 &= \left\|\hat{\mathbf{w}}_x - \mathbf{w}^-_x\right\|_1 + \left\|\hat{\mathbf{w}}_x - \mathbf{w}_x\right\|_1 \\ &\leq 2\sqrt{2\xi\left(N_x, |\mathcal{Y}|, \delta_\Gamma\right)} \text{ with prob.} \geq 1 - \delta_\Gamma .\end{aligned} \tag{18}$$

Recall that for an RV $X \in \mathcal{X}$ with *pmf* $\mathbf{p}$, the entropy is $H(\mathbf{p}) \triangleq \sum\limits_{x\in\mathcal{X}} -p_x \log_2(p_x)$. 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 [17] Thm 17.3.3),

$$\begin{aligned}|H(\mathbf{p}) - H(\mathbf{p}')| &\leq \Xi\left(\|\mathbf{p}-\mathbf{p}'\|_1, |\mathcal{X}|\right) \\ &= \|\mathbf{p}-\mathbf{p}'\|_1 \left(\log_2(|\mathcal{X}|) - \log_2(\|\mathbf{p}-\mathbf{p}'\|_1)\right) \\ &= O(\|\mathbf{p}-\mathbf{p}'\|_1 \log(1/\|\mathbf{p}-\mathbf{p}'\|_1)) \ ,\end{aligned} \tag{19}$$

where

$$\Xi\left(x, K\right) \triangleq x\log_2(K) - x\log_2(x) \tag{20}$$

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 *pmf*s $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{q}$) may be written as (see [17] Thm 2.4.1)

$$R_{\text{true}} \triangleq I(X;Y) = H(\mathbf{u}) + H(\mathbf{v}) - H(\mathbf{q}) . \tag{21}$$

The *pmf*s $\mathbf{u}^-$, $\mathbf{v}^-$, 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 R_L; therefore, we have

$$R_L = H\left(\mathbf{u}^-\right) + H\left(\mathbf{v}^-\right) - H\left(\mathbf{q}^-\right) . \tag{22}$$

The absolute value of the difference in mutual information is

$$\begin{aligned}\Delta_R &\triangleq |R_{\text{true}} - R_L| \\ &= |H(\mathbf{u}) + H(\mathbf{v}) - H(\mathbf{q}) \\ &\quad - \left(H\left(\mathbf{u}^-\right) + H\left(\mathbf{v}^-\right) - H\left(\mathbf{q}^-\right)\right)| \\ &\leq \left|H(\mathbf{u}) - H\left(\mathbf{u}^-\right)\right| + \left|H(\mathbf{v}) - H\left(\mathbf{v}^-\right)\right| \\ &\quad + \left|H(\mathbf{q}) - H\left(\mathbf{q}^-\right)\right| .\end{aligned} \tag{23}$$

The absolute value of the difference in entropy between each true *pmf* $\mathbf{p}$ and the *pmf* $\mathbf{p}^-$ which lies on the surface of the sub-levelset $\mathit{\Gamma}_{\xi}^{\text{rev}}(\hat{\mathbf{p}})$ is bounded by $O(\|\mathbf{p}-\mathbf{p}^-\|_1 \log(1/\|\mathbf{p}-\mathbf{p}^-\|_1))$ (see Eq. 19) and $\|\mathbf{p}-\mathbf{p}^-\|_1$ converges towards zero with $O\big(\sqrt{\xi}\big)$ (see Eq. 18).

Given some arbitrarily selected input *pmf* $\acute{\mathbf{u}}$, we evaluate $|R_{\text{true}}(\acute{\mathbf{u}}) - R_L(\acute{\mathbf{u}})|$ by bounding the absolute value of each entropy difference (see Eq. 23). Suppose we observed N input-output pairs, and let $\underline{\mathrm{N}} \triangleq [N_x]_{x\in\mathcal{X}}$ 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* $\acute{\mathbf{u}}$. When using this codebook, $\acute{\mathbf{u}}$ would match the *true* input *pmf* $\mathbf{u}$, and so $|H(\mathbf{u}) - H(\acute{\mathbf{u}})| = 0$. For the channel output *pmf* $\mathbf{v}(\acute{\mathbf{u}})$, we have

$$
\begin{aligned}
\left\|\mathbf{v}(\acute{\mathbf{u}}) - \mathbf{v}^-(\acute{\mathbf{u}})\right\|_1 &= \left\|\sum_{x\in\mathcal{X}} \mathbf{w}_x \acute{u}_x - \sum_{x\in\mathcal{X}} \mathbf{w}_x^- \acute{u}_x\right\|_1 \\
&= \left\|\sum_{x\in\mathcal{X}} \left(\mathbf{w}_x - \mathbf{w}_x^-\right) \acute{u}_x\right\|_1 \\
&\le \sum_{x\in\mathcal{X}} \left\|\mathbf{w}_x - \mathbf{w}_x^-\right\|_1 \acute{u}_x \\
&\le \sum_{x\in\mathcal{X}} \acute{u}_x 2\sqrt{2\xi\left(N_x, |\mathcal{Y}|, \delta_{\mathit{\Gamma}}\right)}, \qquad (24)
\end{aligned}
$$

and inserting this result into Eq. 19 and Eq. 20 yields

$$
\left|H(\mathbf{v}(\acute{\mathbf{u}})) - H\big(\mathbf{v}^-(\acute{\mathbf{u}})\big)\right| = \Xi\left(\sum_{x\in\mathcal{X}} \acute{u}_x 2\sqrt{2\xi\left(N_x, |\mathcal{Y}|, \delta_{\mathit{\Gamma}}\right)}\right). \qquad (25)
$$

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

$$
\begin{aligned}
\left|H(\mathbf{q}(\acute{\mathbf{u}})) - H\big(\mathbf{q}^-(\acute{\mathbf{u}})\big)\right| &= \Bigg|\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} w_{y\,|\,x}\acute{u}_x \log\big(w_{y\,|\,x}\acute{u}_x\big) - \\
&\quad \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} w_{y\,|\,x}^-\acute{u}_x \log\Big(w_{y\,|\,x}^- u_x'\Big)\Bigg| \\
&= \Bigg|\sum_{x\in\mathcal{X}} \big(H(\mathbf{w}_x) - H\big(\mathbf{w}_x^-\big)\big)\acute{u}_x - \\
&\quad \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Big(w_{y\,|\,x} - w_{y\,|\,x}^-\Big)\acute{u}_x \log(\acute{u}_x)\Bigg| \\
&\le \sum_{x\in\mathcal{X}} \acute{u}_x \left|H(\mathbf{w}_x) - H\big(\mathbf{w}_x^-\big)\right| \\
&\le \sum_{x\in\mathcal{X}} \acute{u}_x \Xi\left(\left\|\mathbf{w}_x - \mathbf{w}_x^-\right\|_1\right), \qquad (26)
\end{aligned}
$$

and so

$$
\left|H(\mathbf{q}(\acute{\mathbf{u}})) - H\big(\mathbf{q}^-(\acute{\mathbf{u}})\big)\right| \le \sum_{x\in\mathcal{X}} \acute{u}_x \Xi\left(2\sqrt{2\xi\left(N_x, |\mathcal{Y}|, \delta_{\mathit{\Gamma}}\right)}\right). \qquad (27)
$$

Overall we have

$$
\begin{aligned}
\Delta_R\left(\acute{\mathbf{u}}, \underline{\mathrm{N}}, |\mathcal{Y}|, \delta_1\right) &\triangleq \left|R_{\text{true}}(\acute{\mathbf{u}}) - R_L(\acute{\mathbf{u}})\right| \\
&\le \Xi\left(\sum_{x\in\mathcal{X}} \acute{u}_x 2\sqrt{2\xi\left(N_x, |\mathcal{Y}|, \delta_{\mathit{\Gamma}}\right)}\right) + \\
&\quad \sum_{x\in\mathcal{X}} \acute{u}_x \Xi\left(2\sqrt{2\xi\left(N_x, |\mathcal{Y}|, \delta_{\mathit{\Gamma}}\right)}\right) \\
&= O\Big(\sqrt{\xi}\log\Big(1/\sqrt{\xi}\Big)\Big), \qquad (28)
\end{aligned}
$$

where $\xi = \xi\left(N, |\mathcal{Y}|, \delta_1\right)$. Given input-output pairs $\mathcal{S}_N$ and suppose a selected $\acute{\mathbf{u}}$ yields a minimum assured information rate of $R_L(\acute{\mathbf{u}})$), then the convergence rate when using: (1) $\xi^{\log}$ (Eq. 13) is $\Delta_R\left(\acute{\mathbf{u}}, \underline{\mathrm{N}}, |\mathcal{Y}|, \delta_{\mathit{\Gamma}}\right) = O(\log(N)/\sqrt{N})$, and (2) ξ^{loglog} (Eq. 14) the convergence rate is $\Delta_R\left(\acute{\mathbf{u}}, \underline{\mathrm{N}}, |\mathcal{Y}|, \delta_{\mathit{\Gamma}}\right) = O\big(\sqrt{\log(\log(N))\log(N)/N}\big)$.

□

For the remainder of this paper, we shall set $N_x = N_0$ for all input values; therefore, $\Delta_R\left(\acute{\mathbf{u}}, \underline{\mathrm{N}}, |\mathcal{Y}|, \delta_{\mathit{\Gamma}}\right)$ is constant over all input *pmf*s $\acute{\mathbf{u}}$. We will simplify our notation to represent $\Delta_R\left(\acute{\mathbf{u}}, \underline{\mathrm{N}}, |\mathcal{Y}|, \delta_{\mathit{\Gamma}}\right)$ as $\Delta_R(N_0)$.

IV. Near optimal channel sampling

Suppose, $R_L(\acute{\mathbf{u}})$ is a lower bound on the maximum mutual information across the channel under the constraint of a fixed input *pmf* $\acute{\mathbf{u}}$. Let $\mathbf{u}^* = \arg\max_{\acute{\mathbf{u}}\in\mathcal{P}_{\mathcal{X}}} R_L(\acute{\mathbf{u}})$, then $R_L(\mathbf{u}^*) \le C \le R_L(\mathbf{u}^*) + \Delta_R(N_0)$.

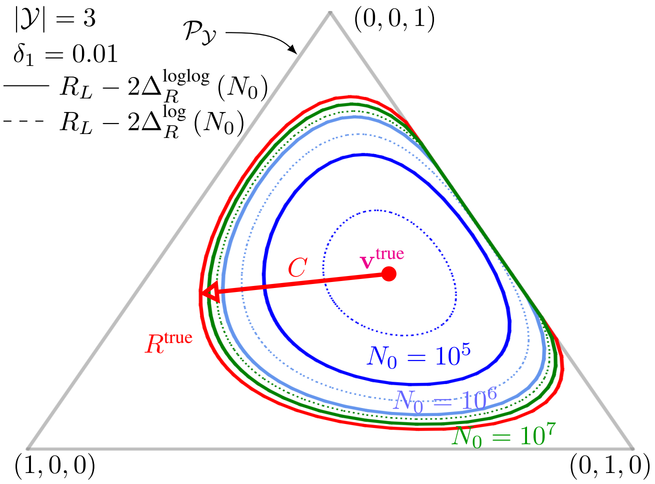

Fig. 2. Input value elimination regions

In Fig. 2, the 'red' contour depicts the surface on which a subset of *pmf*s $\{\mathbf{w}_x : D(\mathbf{w}_x \,\|\mathbf{v}^{\text{true}}) = R^{\text{true}} = C\ \forall\, x \in \mathcal{X}\}$ lie to 'support' achieving channel capacity [15]. If each input value is sampled N_0 times, then the lower bound R_L (output of Algorithm 1) is within $\Delta_R(N_0)$ of R^{true}, and so only the subset of *pmf*s $\{\hat{\mathbf{w}}_x : D(\hat{\mathbf{w}}_x \,\|\mathbf{v}^-) \ge R_L - 2\Delta_R(N_0)\,\forall x \in \mathcal{X}\}$ can potentially improve R_L with additional channel sampling. The $2\Delta_R$ term accounts for: (1) the uncertainty in R_L and (2) the remaining uncertainty in the estimate $\hat{\mathbf{w}}_x$.

Algorithm 2 Near-optimal channel sampling

```
1: Input: $N_{\text{init}}$, $\delta_1$, $\Delta_{\text{STOP}}$, $tol \in (0,1)$
2: $\tau \leftarrow 1$, $\mathcal{S}_\tau \leftarrow \emptyset$, $\mathcal{X}_\tau \leftarrow \mathcal{X}$
3: $N_0 \leftarrow N_{\text{init}}$
4: $\mathbf{v}^- \in \mathcal{P}_\mathcal{Y}$
5: repeat
6:     sample the channel and update $\mathcal{S}_\tau$ such that each input $x \in \mathcal{X}_\tau$ has $N_0$ samples
7:     compute estimates $\{\hat{\mathbf{w}}_x\}_{x\in\mathcal{X}_\tau}$
8:     $(\mathbf{u}^-, \mathbf{v}^-, R_L) \leftarrow \text{ALG1}\left(\{\hat{\mathbf{w}}_x, N_0\}_{x\in\mathcal{X}_\tau}, \delta_1, tol\right)$
9:     compute $\Delta_R(N_0) = \Delta\left(\acute{\mathbf{u}}, N_0, |\mathcal{Y}|, \delta_1/|\mathcal{X}_\tau|\right)$
10:    $\mathcal{X}_{\tau+1} \leftarrow \emptyset$
11:    for $x' \in \mathcal{X}_\tau$ do
12:        if $D(\hat{\mathbf{w}}_{x'} \,\|\mathbf{v}^-) \geq R_L - 2\Delta_R(N_0)$ then
13:            $\mathcal{X}_{\tau+1} \leftarrow \mathcal{X}_{\tau+1} \cup x'$ {retain requisite input values}
14:        end if
15:    end for
16:    $N_0 \leftarrow 2N_0 \quad \forall x \in \mathcal{X}_\tau$ {double number of probes}
17:    $\tau \leftarrow \tau + 1$
18: until $\Delta_R(N_0) \leq \Delta_{\text{STOP}}$ {test stopping criteria}
19: Output: $\{\mathbf{w}_x^-\}_{x\in\mathcal{X}}$, $\mathbf{v}^-$, $\mathbf{u}^-$, $R_L$
```

Suppose $N_0 = 10^5$, then for any input value with an empirical *pmf* $\hat{\mathbf{w}}_x$ located within the region depicted by the solid line 'blue' contour can be eliminated from further sampling (see Fig. 2). The area inside the solid line 'blue' contour, which is based on the iterated log sub-levelset bound, will eliminate more channel estimates than the 'smaller' area enclosed by the dashed line 'blue' contour, which is based on the Sanov sub-levelset bound. As the number of channel probes increases, the contours expand toward the 'red' channel capacity contour.

Algorithm 2 implements this strategy. We start off sampling all input values, then we compute the lower bound on channel capacity R_L and the bound Δ_R on the deviation in mutual information. Next, we discard (stop sampling) any input value x' (and assume that it is not part of the channel) if $D(\hat{\mathbf{w}}_{x'} \,\|\mathbf{v}^-) < R_L - 2\Delta_R$ (i.e. the value x' will never be part of the support that improves the rate R_L). As the the sampling process continues, the algorithm narrows its focus to only the set of input values associated with channel estimates $\{\hat{\mathbf{w}}_x\}_{x\in\mathcal{X}_\tau}$ that are required to approach channel capacity with high probability.

V. Results

Define Algorithm ALL as a simple modification of Algorithm 2 with line 10 replaced with $\mathcal{X}_{\tau+1} \leftarrow \mathcal{X}$ such that all input values are retained for additional channel sampling throughout the entire sampling process. We want to evaluate the reduction in sample complexity for Algorithm 2, which selectively samples input values, against Algorithm ALL, which always samples every input value.

During the channel sampling process, Algorithm 2 whittles down the set of input values as more information about the channel law $\underline{\mathbf{w}}^{\text{true}}$ is observed. We ran Algorithm 2 on a DMC with $|\mathcal{X}| = 100$ and $|\mathcal{Y}| = 4$, each $\mathbf{w}_x^{\text{true}}$ $\forall x \in \mathcal{X}$ was drawn i.i.d. according to an uniform Dirichlet distribution [20] with hyperparameter α i.e. $\mathbf{w}_x \sim Dir(\alpha = 0.75)$, where

$$Dir(\alpha) \propto \prod_{y\in\mathcal{Y}} {w_{y\,|\,x}}^{\alpha-1} \;\forall y \in \mathcal{Y} \tag{29}$$

With parameters $\delta_1 = 0.0001$, $N_{\text{init}} = 100$, and $\Delta_{\text{STOP}} = 0$ (and we interrupt the algorithm at $N_0 = 10^9$ samples), we ran Algorithm 2 twice: (1) first using the Sanov based bound $\xi^{\log}$ (Eq. 13) to 'size' the sub-levelset constraints to contain the true channel law with probability $\geq 1-\delta_1$, and then (2) using the iterated log bound ξ^{loglog} (Eq. 14).

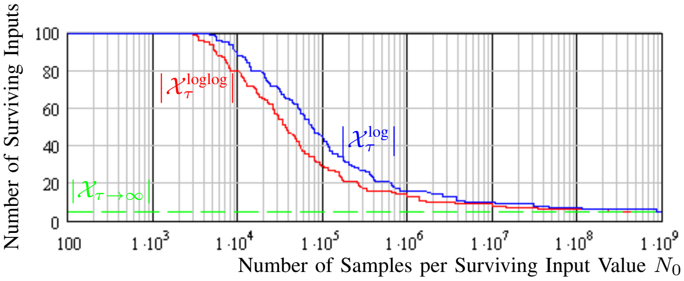

Fig. 3. Number of surviving input value throughout the sampling process

In Fig. 3, the x-axis is the number of samples N_0 observed for each of the $|\mathcal{X}| = 100$ input values. The y-axis shows the number of different input values that were selected and sampled by Algorithm 2 throughout the channel sampling process. At the start all 100 input values are sampled and then at about $N_0 = 3000$ samples, the red curve shows that Algorithm 2 begins to have enough information to eliminate some input values from further consideration. Input x' is eliminated at the point its empirical *pmf* $\hat{\mathbf{w}}_{x'}$ falls within the expanding the contour surfaces depicted in Fig. 2. The green dashed line is the number of input values in the support $\mathcal{X}^*$ at channel capacity (these input values should never be eliminated).

In Fig. 3, the red curve indicates the number of input values remaining when using the iterated log bound ξ^{loglog}. The iterated log bound drops input values quicker than the blue curve, which is based on using the Sanov bound $\xi^{\log}$. The reduction in sample complexity (between the curves) appears modest (approx. factor of 3 or so) for this test case. However, note that for $N_0 > 10^6$, the number of input values sampled is far less than $|\mathcal{X}| = 100$, so Algorithm 2 (with either sub-levelset bound ξ^{loglog} or $\xi^{\log}$) significantly reduces the sample complexity relative to Algorithm ALL.

To evaluate the sample complexity of Algorithm 2 versus Algorithm ALL, we simulated a set of channel laws, where the number of input values was varied $|\mathcal{X}| \in \mathfrak{X} \triangleq \{4, 8, 16, 32, 64, 128, 256, 512\}$ and the number of output values was fixed at $|\mathcal{Y}| = 4$. We used a mixture of Dirichlet distributions to draw the channel law $\underline{\mathbf{w}}^{\text{true}}$ for each trial. We drew $|\mathcal{Y}|$ generator *pmf*s $\mathbf{w}_x$ using $\alpha = 0.6$ and the remaining $|\mathcal{X}| - |\mathcal{Y}|$ generator *pmf*s $\mathbf{w}_x$ using $\alpha = 0.75$. The $|\mathcal{Y}|$ generator *pmf*s $\mathbf{w}_x$ (drawn using $\alpha = 0.6$) are more likely to be in the support $\mathcal{X}^*$ necessary to achieve channel capacity. This test simulation roughly models a scenario where one looks for these specific $|\mathcal{Y}| = 4$ generator *pmf*s among the remaining 'noisier' generator *pmf*s.

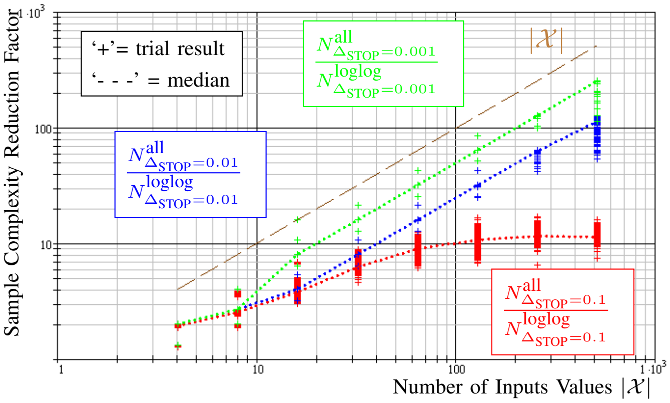

Fig. 4. Sample complexity reduction factor

For each $|\mathcal{X}| \in \mathfrak{X}$, we simulated 200 channel trials, and then we ran both Algorithm 2 and Algorithm ALL with both algorithms using the iterated log bound ξ^{loglog})) and the parameters $\delta_1 = 0.0001$, $N_{\text{init}} = 100$, and $\Delta_{\text{STOP}} \in \{0.1, 0.01, 0.001\}$. Each trial produced a sample complexity N (the total number of channel samples to meet the stopping criteria of $\Delta_R < \Delta_{\text{STOP}}$ (see line 18 in Algorithm 2).

In Fig. 4 each '+' marker is the ratio of the sample complexity $N^{\text{all}}_{\Delta_{\text{STOP}}}$ of Algorithm ALL over the sample complexity $N^{\text{loglog}}_{\Delta_{\text{STOP}}}$ of Algorithm 2 for one trial of the 100 channel trials for each given number of input values $|\mathcal{X}| \in \mathfrak{X}$. The 'dotted' lines are the median for all the trials for a given number of input values $|\mathcal{X}| \in \mathfrak{X}$ and $\Delta_{\text{STOP}} \in \{0.1, 0.01, 0.001\}$.

We see in Fig. 4 that when $\Delta_{\text{STOP}} = 0.001$ (i.e. the green points), the median rises approx. linearly with the number of input values $|\mathcal{X}|$. If we relax the stopping criteria to $\Delta_{\text{STOP}} = 0.1$ (i.e. the red points), then the algorithm terminates quicker with fewer input values eliminated. This lowers the sample complexity reduction factor.

The reduction of the number of samples required to establish a reliable communication rate depends on the specific channel law. Typically, if $|\mathcal{Y}| \leq |\mathcal{X}|$, then $|\mathcal{Y}|$ of the $|\mathcal{X}|$ input values are required to support channel capacity; therefore, our near-optimal channel sampling strategy may reduce the sample complexity by a factor of up to $\min(|\mathcal{X}|/|\mathcal{Y}|, 1)$ relative the sample-all-input-values-equally strategy.

VI. Conclusions and Future Work

We developed and demonstrated an online algorithm (using a near-optimal channel sampling strategy) that establishes a high probability lower bound R_L on channel capacity for a DMC. We proved that the online algorithm matches the same 'big Oh' sample complexity as an off-line algorithm with 'clairvoyance' to request a one-time sufficient batch of input-output samples [16] that meets the stopping criteria.

We did not make any assumptions about the distribution of the channel law $\underline{\mathbf{w}}$. One may leverage additional information about the process that generates the channel law to further reduce sample complexity. For example, if the number of input values is very large, one may initially investigate small subsets of input values to leverage insight on the channel law generating process itself. Still in the worst-case scenario (accommodated here) all input values would eventually need to be evaluated to approach channel capacity.

Future work could perhaps modify the sampling process to adjust with the discovery of new channel output values. While we assumed that the true channel law remain constant, we hope to modify the channel sampling strategy to detect and track channel drift over time.

References

[1] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. on Information Theory*, vol. 44, pp. 2148–2177, 1998.

[2] J. K. I. Csiszár, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[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] 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.

[7] M. A. Tope and J. M. Morris, "Improvements to Sanov and PAC sublevel-set bounds for discrete random variables," *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 2021.

[8] Y. Seldin and N. Tishby, "PAC-Bayesian analysis of co-clustering and beyond," *Journal of Machine Learning Research*, vol. 11, pp. 3595–3636, 2010.

[9] B. Guedj, "A primer on PAC-Bayesian learning," 2019. [Online]. Available: https://arxiv.org/abs/1901.05353

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multi-armed bandit problem," *NeuroCOLT2 Technical Report Series*, 2000.

[11] N. Cesa-Bianchi and L. G, *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[12] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "liluc b: An optimal exploration algorithm for multi-armed bandits," *JMLR: Workshop and Conference Proceedings*, vol. 35, pp. 1–17, 2014.

[13] D. Pollard, *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2002.

[14] 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.

[15] ——, "A PAC-bound on the channel capacity of an observed discrete memoryless channel," *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 2021.

[16] ——, "On the limits of learning a discrete memoryless communication channel," *IEEE Military Communications Conference (MILCOM)*, 2022.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley and Sons, 1991.

[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] 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.

[20] "Dirichlet distribution — Wikipedia, the free encyclopedia," 2021. [Online]. Available: https://en.wikipedia.org/wiki/Dirichlet_distribution