A PAC-bound on the Channel Capacity of an Observed Discrete Memoryless 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 presents a method to compute the channel capacity of an observed (partially known) discrete memoryless channel (DMC) using a probably approximately correct (PAC) bound. Given N independently and identically distributed (i.i.d.) input-output sample pairs, we define a compound DMC with convex sublevel-sets to constrain the channel output uncertainty with high probability. Then we numerically solve an 'K-way' convex optimization to determine an achievable information rate $R_L(N)$ across the channel that holds with a specified high probability. Our approach provides the non-asymptotic 'worst-case' convergence $R_L(N)$ to channel capacity C at the rate of $O(\sqrt{\log(\log(N))/N})$.**

This paper presents a method to compute a communication rate R through a discrete memoryless channel (DMC), where uncertainty remains about the precise channel law (the set of channel transition probabilities).

I. Introduction

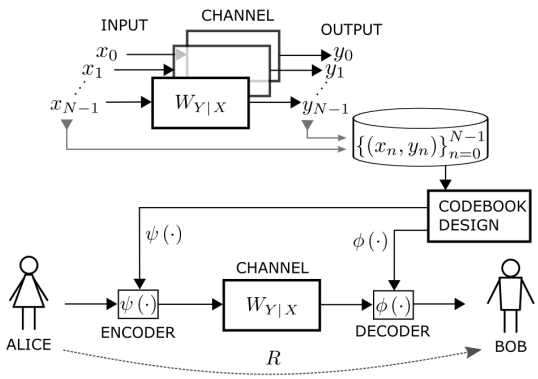

Fig. 1. Given channel samples, determine 'best' R with high probability

We consider the scenario (see Fig. 1) where two engineers (Alice and Bob) wish to establish a one-directional communication link across a discrete memoryless channel (DMC).

The channel is a 'black box,' and at each discrete time instant, Alice can select and apply one of $|\mathcal{X}|$ input symbols (or values) $x \in \mathcal{X}$, where $\mathcal{X}$ is the set of all possible input symbols $\mathcal{X} = \{a_0, a_1, \ldots, a_{|\mathcal{X}|-1}\}$. Each time a input symbol x is applied to the input port an output symbol $y \in \mathcal{Y} = \{b_0, b_1, \ldots, b_{|\mathcal{Y}|-1}\}$ appears on the output port.

The DMC is modeled using the channel law $\underline{\mathbf{w}}$, which is a vector of channel transition probabilities (*ctp*s) indexed by the channel 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 $w_{y\,|\,x} \triangleq \mathbb{P}\{Y \triangleq y \,|\, X = x\}$ $\forall x \in \mathcal{X}$ and $y \in \mathcal{Y}$. We call $\mathbf{w}_x$ the 'generator' *pmf* of the output given input x. Given a fixed channel input x, the channel output RVs $Y \mid X = x$ are independent and identically distributed (i.i.d.). Define the channel input *pmf* ($\mathbf{u} \in \mathcal{P}_{\mathcal{X}}$), as $\mathbf{u} \triangleq [u_{a_0}, u_{a_1}, \ldots, u_{a_{|\mathcal{X}|-1}}]$. Similarly define $\mathbf{v} \in \mathcal{P}_{\mathcal{Y}}$, a *pmf* over the output RV Y, as $\mathbf{v} \triangleq [v_{b_0}, v_{b_1}, \ldots, v_{b_{|\mathcal{Y}|-1}}]$ where $v_y \triangleq \mathbb{P}\{Y = y\}$.

When the channel law $\underline{\mathbf{w}}$ is known, then the channel capacity C is maximum communication rate such that there exist a codebook with an arbitrarily small error rate (see [2] or [1]), and

$$C \triangleq \max_{\acute{\mathbf{u}} \in \mathcal{P}_{\mathcal{X}}} I(\acute{\mathbf{u}}, \underline{\mathbf{w}}), \tag{1}$$

where: (1) $I(\mathbf{u}, \underline{\mathbf{w}}) \triangleq \sum_{x \in \mathcal{X}} u_x D(\mathbf{w}_x \,\|\, \grave{\mathbf{v}}(\mathbf{u}))$ is the mutual information, (2) the expected output is $\grave{\mathbf{v}}(\mathbf{u}) \triangleq \sum_{x \in \mathcal{X}} u_x \mathbf{w}_x$, and (3) 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

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

Suppose, Alice probes the channel by sending various input values $x \in \mathcal{X}$ into the channel N times, and she observes the output $x \in \mathcal{X}$ corresponding to each input to construct the sample set of input-output pairs, i.e. $\mathcal{S}_N = \{(x_0, y_0), (x_1, y_1), \ldots, (x_{N-1}, y_{N-1})\}$. Then, using this set of samples $\mathcal{S}_N$, Alice computes the sample (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

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

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

Both $\mathbf{w}_x$ and $\hat{\mathbf{w}}_x$ lie in the probability space $\mathcal{P}_{\mathcal{Y}}$, and we think of $\hat{\mathbf{w}}_x$ as an estimate of $\mathbf{w}_x$. We define the 'empirical' or 'sample' probability space $\mathcal{P}_{N_x}$ as the set of all possible $\hat{\mathbf{w}}_x$ *pmf*s (i.e. $\mathcal{P}_{N_x} = \{\hat{\mathbf{w}}_x \colon \mathbb{P}\{\hat{\mathbf{w}}_x |\, \mathcal{S}_N\} > 0\}$.

Alice and Bob select δ, and our goal is to find an input *pmf* $\mathbf{u}^*$ (from the samples $\mathcal{S}_N$) such that Alice and Bob then could design and share a codebook (see Fig. 1) allowing communication at or below the rate R_L with an arbitrarily low error rate. Because the observed samples $\mathcal{S}_N$ may not be 'statically typical,' there is a chance that the designed

codebook may fail to provide communication at rate R_L; however, our goal is to ensure that the probability of such a codebook 'failure' is less than $1-\delta$.

For a DMC with known channel law $\underline{\mathbf{w}}$, the channel capacity can be computed using the Blahut Arimoto (BA) algorithm [5]. However for our scenario, the plan is to construct sublevel-sets that jointly contain the 'true' (yet unknown) channel law with probability at least $1-\delta$. These sublevel-sets form a deterministic compound channel that we solve for the best achievable communication rate R_L.

The remainder of this paper is as follows. We review of some relevant previous work, then we describe a novel sketch proof of channel capacity using a multinomial halfspace bound (MHB). Next, we show how to model channel uncertainty using probably approximately correct (PAC) sublevel-set bounds based on the KL Divergence. These sublevel-sets formulate a convex constraint optimization problem that we solve using an iterative algorithm based on the Lambert-W function. Then we describe some simulation results, and finally, we provide conclusions and recommendations for future work.

II. Previous Work

Lapidoth and Narayan provide a summary of results for communication over uncertain channels [3], and recent results for DMCs are covered by Csiszár, J. Körner [2]. Our approach to model channel uncertainty relies on the *compound channel* and the *information spectrum* approach of Han [4].

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 [3]

$$C = \max_{\mathbf{u}\in\mathcal{P}_{\mathcal{X}}} \min_{\mathbf{w}\in\Gamma} I(\mathbf{u},\mathbf{w}). \tag{4}$$

Further, if Γ is convex and closed, then one can swap the order of the $\min$ and $\max$ yielding [3]

$$C = \min_{\mathbf{w}\in\Gamma} \max_{\mathbf{u}\in\mathcal{P}_{\mathcal{X}}} I(\mathbf{u},\mathbf{w}). \tag{5}$$

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 [3] that will support rates up to R_L (simultaneously) for any channel law $\underline{\mathbf{w}} \in \Gamma$.

We rely on the following three probabilistic bounds. The first is the multinomial halfspace bound (MHB).

Theorem II.1 *Multinomial Halfspace Bound [11]*
Given the set $\mathcal{S}_N = \{y_0, y_1, \ldots, y_{N-1}\}$ of outcomes from N i.i.d. 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 sample (empirical) pmf. When given the halfspace Λ (oriented to include the pmf $\mathbf{w}^* \in \mathcal{P}_{\mathcal{Y}}$) defined as

$$\Lambda(\mathbf{w}^*, \mathbf{w}) \triangleq \{\hat{\mathbf{w}} \in \mathcal{P}_{\mathcal{Y}} : \sum_{y\in\mathcal{Y}} \hat{w}_y \ln\left(\frac{w_y^*}{w_y}\right) \le \xi\} \tag{6}$$

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

$$\mathbb{P}\{\hat{\mathbf{w}} \notin \Lambda(\mathbf{w}^*, \mathbf{w})\} \le \exp(-ND(\mathbf{w}^* \,\|\mathbf{w})). \tag{7}$$

Consider the 'forward' sublevel-set $\Gamma_\xi^{\text{fwd}}(\mathbf{w})$ (based on KL divergence), which is 'centered' on $\mathbf{w}$ with a 'size' ξ and defined as

$$\Gamma_\xi^{\text{fwd}}(\mathbf{w}) \triangleq \{\acute{\mathbf{w}} \in \mathcal{P}_{\mathcal{Y}} : D(\acute{\mathbf{w}} \,\|\mathbf{w}) \le \xi\}. \tag{8}$$

From this we define a 'typical set' of empirical *pmf*s computed from N i.i.d.samples generated from *pmf* $\mathbf{w}$ as

$$\mathcal{T}_\delta(\mathbf{w}) \triangleq \left\{\hat{\mathbf{w}} \in \mathcal{P}_{\mathcal{Y}} : \mathbb{P}\left\{\hat{\mathbf{w}} \in \Gamma_\xi^{\text{fwd}}(\mathbf{w})\right\} \ge 1-\delta\right\}, \tag{9}$$

and Sanov's Theorem provide a bound on this typical set.

Theorem II.2 *Sanov's Theorem (see [1] section 11.4)*
*Given the same as Thm. II.1 then given **any** region $\Gamma \subset \mathcal{P}_{\mathcal{Y}}$ and $\mathbf{w}^*$ the 'closest' pmf among all $\acute{\mathbf{w}} \in \Gamma$ to $\mathbf{w}$ in terms of the KL divergence*

$$\mathbf{w}^* = \arg\min_{\acute{\mathbf{w}}\in\Gamma} D(\acute{\mathbf{w}} \,\|\mathbf{w}) \tag{10}$$

then

$$\delta_\Gamma \triangleq \mathbb{P}\{\hat{\mathbf{w}} \notin \Gamma\} \le (N+1)^{|\mathcal{Y}|} \exp(-ND(\mathbf{w}^* \,\|\mathbf{w})). \tag{11}$$

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

$$\xi(N, |\mathcal{Y}|, \delta) = D(\mathbf{w}^* \,\|\mathbf{w}) \le \frac{|\mathcal{Y}| \ln(N+1) - \ln(\delta_\Gamma)}{N} \tag{12}$$

sets the 'size' of the typical set $\mathcal{T}_\delta(\mathbf{w})$.

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

Theorem II.3 *Sublevel-set Bound [11]*
Given the same as Thm. II.1 and select any $\delta_\Gamma \in (0,1]$, then $\mathbb{P}\{\hat{\mathbf{w}} \notin \Gamma_\xi(\mathbf{w})\} \le \delta_\Gamma$ for the sublevel-set $\Gamma_\xi(\mathbf{w})$ (see Eq. 8) with 'size'

$$\xi \ge \frac{1}{N}\left(\frac{1}{2}\ln(2|\mathcal{Y}|) - \frac{3}{2}\ln\left(\frac{\delta_\Gamma}{2}\right) + \kappa_3 + |\mathcal{Y}| \log\left(\log_2(\log_2(N)) + \kappa_1\sqrt{|\mathcal{Y}|} + \log_2(\kappa_2|\mathcal{Y}|) + 2\right)\right) \tag{13}$$

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

Given the sample estimate $\hat{\mathbf{w}}_x$, we define the admissible set to include all *pmf*s that may have generated $\hat{\mathbf{w}}_x$ with high probability as

$$\mathcal{A}_\delta(\hat{\mathbf{w}}) \triangleq \{\acute{\mathbf{w}} \in \mathcal{P}_{\mathcal{Y}} : \hat{\mathbf{w}} \in \mathcal{T}_\delta(\acute{\mathbf{w}})\}, \tag{14}$$

and the admissible region is equivalent to the 'reverse' level-set (also based on KL divergence) defined as

$$\Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}}) \triangleq \{\acute{\mathbf{w}} \in \mathcal{P}_{\mathcal{Y}} : D(\hat{\mathbf{w}} \,\|\acute{\mathbf{w}}) \le \xi\} \tag{15}$$

for $\xi = \xi(N, |\mathcal{Y}|, \delta)$.

III. DMC Capacity Revisited

In the section, we sketch a novel proof of channel capacity of the DMC to help motivate our approach. We start with the 'equal-divergence' property of the channel output *pmf* $\mathbf{v}^*$ that achieves channel capacity.

Consider,

$$C \triangleq \max_{\mathbf{u}\in\mathcal{P}_{\mathcal{X}}} \sum_{x\in\mathcal{X}} u_x D(\mathbf{w}_x \,\|\mathbf{v}\,(\mathbf{u})) \tag{16}$$

$$= \max_{\mathbf{u}\in\mathcal{P}_{\mathcal{X}}} \min_{\mathbf{v}\in\mathcal{P}_{\mathcal{Y}}} \sum_{x\in\mathcal{X}} u_x D(\mathbf{w}_x \,\|\mathbf{v}) \tag{17}$$

$$= \min_{\mathbf{v}\in\mathcal{P}_{\mathcal{Y}}} \max_{\mathbf{u}\in\mathcal{P}_{\mathcal{X}}} \sum_{x\in\mathcal{X}} u_x D(\mathbf{w}_x \,\|\mathbf{v}) \tag{18}$$

$$= \min_{\mathbf{v}\in\mathcal{P}_{\mathcal{Y}}} \max_{x\in\mathcal{X}} D(\mathbf{w}_x \,\|\mathbf{v}). \tag{19}$$

Eq. 17 is a result of

$$\sum_{x\in\mathcal{X}} u_x D(\mathbf{w}_x \,\|\mathbf{v}\,(\mathbf{u})) = \sum_{x\in\mathcal{X}} u_x D(\mathbf{w}_x \,\|\mathbf{v}') - D(\mathbf{v}\,(\mathbf{u}) \,\|\mathbf{v}')$$
$$= \min_{\mathbf{v}'\in\mathcal{P}_{\mathcal{Y}}} \sum_{x\in\mathcal{X}} u_x D(\mathbf{w}_x \,\|\mathbf{v}'), \tag{20}$$

and in Eq. 19, we see that for any $\mathbf{u}^*$ that achieves C, then $D(\mathbf{w}_x \,\|\mathbf{v}^*) = C$ for all $u_x > 0$.

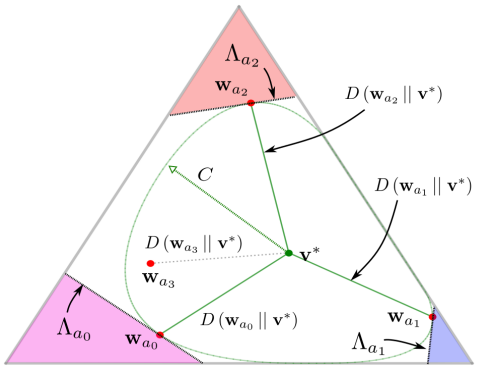

Fig. 2. Channel capacity: known channel law

The Blahut Arimoto (BA) algorithm [5] outputs a non-unique $\mathbf{u}^*$ that achieves channel capacity C, along with the unique average output *pmf* $\mathbf{v}^*$.

In Fig. 2, each *ctp* $\mathbf{w}_x$ (of the channel law $\underline{\mathbf{w}}$) is plotted as a 'red dot.' The 'green curve' marks the level-set $\Gamma^{\text{fwd}}_{\xi=C}(\mathbf{v}^*)$ (see Eq. 8), where the 'size' ξ of the level-set is set to the channel capacity C.

We construct a 'random' codebook Ψ with M L-long codewords, as follows. For each codeword $\psi_m, \forall m \in \{0, 1, \ldots, M-1\}$, we draw L input values, so $\psi_{m,l} \sim \mathbf{u}^*$ for $l = 0, 1, \ldots, L-1$. Suppose that the codeword $\psi_{m'}$ (one row of the codebook) is sent across the channel, and we observe the channel output $\underline{\mathbf{r}} = \{r_l\}_{l=0}^{L-1}$.

At the decoder, we want to test the channel output against every possible codeword to see which one matches. An error occurs whenever either: (1) no codeword matches channel output, (2) more than one codeword matches the channel output, or (3) the wrong codeword matches the channel output. To test whether codeword $\psi_{m''}$ was sent, We gather and sort channel outputs into groups (one group for each $x \in \mathcal{X}$) consisting of all observed channel outputs hypothesized to be generated by *ctp* $\mathbf{w}_x$, and then we compute the 'empirical' *pmf* on each group (indexed by x), i.e.

$$\hat{\mathbf{w}}_x(m'') \triangleq \frac{1}{l_x(m'')} \sum_{l=0}^{L-1} r_l \mathbb{1}_{\{\psi_{m'',l}=x\}}, \tag{21}$$

where each group has $l_x(m'') \triangleq \sum_{l=0}^{L-1} \mathbb{1}_{\{\psi_{m'',l}=x\}}$ channel output samples.

When m'' matches m', then each empirical *pmf* $\hat{\mathbf{w}}_x(m'')$ should be near the 'generator' *ctp* $\mathbf{w}_x$. So we position a halfspace $\Lambda(\mathbf{w}_x, \mathbf{w})$ 'just inside' $\mathbf{w}_x$, and then as $L \to \infty$ each and every $\hat{\mathbf{w}}_x(m'' = m')$ (i.e. matched) should likely fall within the 'shaded' area (outside the halfspace i.e. the decoding region).

When m'' does not match m', then $\hat{\mathbf{w}}_x(m'' \neq m') \sim \mathbf{v}$, $\forall x \in \mathcal{X}$. An error occurs whenever each and every $\hat{\mathbf{w}}_x(m'' \neq m')$ falls into the 'shaded' decoding area. The MHB bounds the probability that $\hat{\mathbf{w}}_x(m'' \neq m')$ incorrectly falls into the decoding area as less than $\exp(-l_x D(\mathbf{w}_x \,\|\mathbf{v}^*))$. So the probability of a decoding error $\psi_{m''=m'}$ (one mismatched codeword) is

$$\epsilon_m \leq \prod_{x\in\mathcal{X}} \exp(-l_x D(\mathbf{w}_x \,\|\mathbf{v}^*)) \tag{22}$$

$$= \exp\left(\sum_{x\in\mathcal{X}} -l_x D(\mathbf{w}_x \,\|\mathbf{v}^*)\right). \tag{23}$$

Recall Eq. 16 and the equal-divergence property, so $D(\mathbf{w}_x \,\|\mathbf{v}^*) = C, \forall x \in \{\mathbf{u}^*_x > 0\}$, so

$$= \exp\left(\sum_{x\in\mathcal{X}} -l_x\, C\right) \tag{24}$$

$$= \exp(-L\, C). \tag{25}$$

As there are $M-1$ incorrect (mismatched) codewords for code rate R ($M = \exp(L\, R)$); therefore, the probability that any of these incorrect codeword decodes as valid (via the Union Bound) is

$$\epsilon(L) \leq \sum_{m=0}^{M-1} \epsilon_m \tag{26}$$

$$= (M-1)\exp(-LC) \tag{27}$$

$$\leq \exp(L\, R)\exp(-L\, C) \tag{28}$$

$$\leq \exp(-L\,(C-R)). \tag{29}$$

so if $R < C$, $\lim_{L\to\infty} \epsilon(L) = 0$.

□

IV. Compound KL sublevel-set channel

Now consider when the channel law $\underline{\mathbf{w}}$ of the DMC is not known precisely, but rather we know that each *ctp* $\mathbf{w}_x$ is within a subset (region) of the probability space, i.e $\mathbf{w}_x \in \Gamma_x$ (for each input x).

Fig. 3 illustrates a compound DMC. We want to determine a *pmf* $\mathbf{u}^*$ that achieves the maximum information rate R. As with the DMC, we place halfspaces (decoding regions) against the 'gray shaded' sublevel-set $\Gamma^{\text{fwd}}_{\xi=R}(\mathbf{v})$. Not all *ctp*s

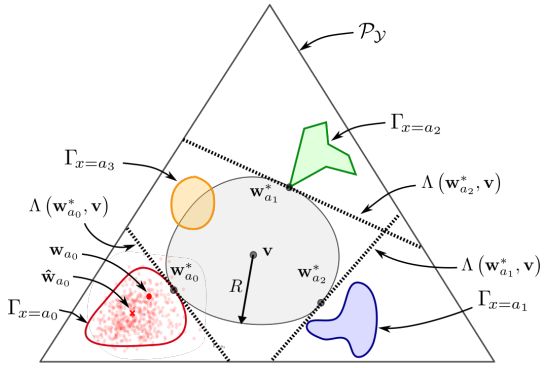

Fig. 3. Compound DMC

in the 'orange' region Γ_{a_3} lie outside of $\Gamma^{\text{fwd}}_{\xi=R}(\mathbf{v})$ and so the input value a_3 will not be used within the codebook to achieve rate R. For the non-convex 'green' Γ_{a_2} and 'blue' Γ_{a_1} the 'decoding' halfspaces are positioned to not overlap their respective region. The shape of the 'gray shaded' sublevel-set controls the 'orientation' of the halfspace; therefore, the *pmf* $\mathbf{w}^*$ may not always lie against the surface of the associated non-convex region. For Γ_{a_2},Γ_{a_2}, and Γ_{a_2}, each 'true' *ctp* is known to be within the respectively region. For our situation (depicted by the 'red' Γ_{a_0}), we define a convex region that contains the *ctp* $\mathbf{w}_{a_0}$ with high probability.

Based on the observed channel samples, we define a compound channel where the *ctp*s are known to be constrained within a set of closed convex admissible regions (see Eq. 5) with high probability. Then we solve for the lower bound on the channel rate $R_L \leq C$, as

$$R_L \triangleq \min_{\substack{\mathbf{w}_x^- \in \mathcal{A}_\delta(\hat{\mathbf{w}}_x) \\ \forall x \in \mathcal{X}}} \max_{\mathbf{u}^- \in \mathcal{P}_{\mathcal{X}}} I\left(\mathbf{u}^-, \underline{\mathbf{w}}^-\right), \tag{30}$$

where

$$\underline{\mathbf{w}}^- = \left[\mathbf{w}_x^-\right]_{x\in\mathcal{X}}. \tag{31}$$

Consider the K-way minimization of an objective function f [7], i.e.

$$f^* = \min_{x_1\in\Gamma_1} \dots \min_{x_K\in\Gamma_K} f(x_1,\dots,x_K). \tag{32}$$

If every Γ_k is a compact convex set, and f is both continuous (with continuous derivatives on $\Gamma_1 \times \cdots \times \Gamma_K$) and bounded from below, then an alternating minimization procedure (a cyclic process where every individual variable is minimized in turn repeatedly) shall converge to f^* (see [7]).

Our objective function Eq. 30 is bounded below by $R = 0$, and it is convex and continuous in $\{\mathbf{w}_x^-\}_{x\in\mathcal{X}}$ with continuous derivatives along the convex constraints (the admissible sets are continuous level-sets); therefore, an alternating minimization procedure will converge to the solution R_L.

For each step of the alternating minimization, we select an input value x, and we want to determine

$$\mathbf{w}_x^- = \arg \min_{\mathbf{w}'_x \in \Gamma^{\text{rev}}_\xi(\hat{\mathbf{w}}_x)} D(\mathbf{w}'_x \| \mathbf{v})$$

Temporarily dropping the index x (to simplify notation), consider solving

$$\mathbf{w}^* = \arg \min_{\mathbf{w}'\in\Gamma_\xi(\hat{\mathbf{w}})} D(\mathbf{w}' \| \mathbf{v})$$

using an Lagrange multiplier equation (with the constraint $\sum_{y\in\mathcal{Y}} w^*_y = 1$ to force the solution to be a *pmf*), i.e.

$$\begin{aligned}\mathcal{L}(\mathbf{w},\omega',\mu') = D(\mathbf{w}\|\mathbf{v}) + \omega'\left(D(\hat{\mathbf{w}}\|\mathbf{w}) - \xi\right) \\ + \mu'\Big(\sum_{y\in\mathcal{Y}} w_y - 1\Big).\end{aligned} \tag{33}$$

To minimize, we take the partial derivative of the Lagrange function, then set to zero (to find a critical point $\mathbf{w}^*$) $\frac{\partial\mathcal{L}(\mathbf{w},\omega',\mu')}{\partial w_y} = 0$. So the a critical point is a solution to the equation

$$\left.\frac{\partial\mathcal{L}(\mathbf{w},\omega',\mu')}{\partial w_y}\right|_{\mathbf{w}=\mathbf{w}^*} = 1 + \ln\left(\frac{w^*_y}{v_y}\right) - \omega'\frac{\hat{w}_y}{w^*_y} + \mu' = 0,$$

and satisfying the constraint $\sum_{y\in\mathcal{Y}} w^*_y = 1$, we get

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

where W_0 is the 'zero-branch' of the Lambert W function [9]. Since the second partial derivative is greater than zero

$$\left.\frac{\partial^2\mathcal{L}(\mathbf{w},\omega',\mu')}{\partial w_y^2}\right|_{\mathbf{w}=\mathbf{w}^*} = \frac{1}{w^*_y} + \omega\frac{\hat{w}_y}{\left(w^*_y\right)^2} \geq 0 \quad \forall y\in\mathcal{Y},$$

whenever $\omega > -\frac{w^*_y}{\hat{w}_y}$, the critical point $\mathbf{w}^*$ is a local minimum for all $\omega \geq 0$ ($\hat{\mathbf{w}}$ and $\mathbf{w}^*$ are *pmf*s; and therefore, $\hat{w}_y \geq 0$ and $w^*_y \geq 0$). We choose ω to satisfy the constraint $D(\hat{\mathbf{w}}\|\mathbf{w}^*) = \xi$, and because the constraint is and convex (with continuous derivatives) over $\mathbf{w}^*$, the local maximum is a global maximum. We using a line search algorithm to find and set $\mathbf{w}^- = \mathbf{w}^*$.

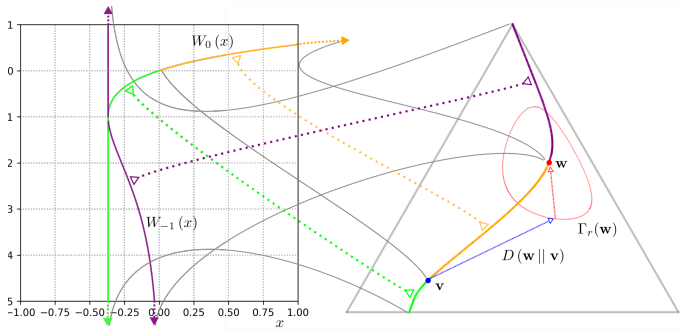

Fig. 4. Connection between critical points and the Lambert W function

Our solution $\mathbf{w}^*$ lies on the 'yellow' portion of the curve in Fig. 4. The Lambert W function is on the left side and the probability space containing the solution is on the right. Various branches of the Lambert W function yield different solutions to points on the admissible region.

Algorithm 1 is a K-way optimization that sequentially 'cycles' over all input values x to update the $\mathbf{w}_x^-$ *pmf*s and the *pmf* $\mathbf{v}^*$ (via the BA algorithm). If there is not significant movement in $\mathbf{v}^*$ over one entire cycle, the algorithm declares convergence.

Algorithm 1: Modified BA algorithm

Input: $\{\hat{\mathbf{w}}_x\}_{x\in\mathcal{X}}$,$N$,$|\mathcal{Y}|$,$\delta$,$tol \in (0,1)$
Output: R_L,$\{\mathbf{w}_x^-\}_{x\in\mathcal{X}}$,$\mathbf{v}_{R_L}^-$,$\mathbf{u}_{R_L}^-$

$\xi \leftarrow \xi(N, |\mathcal{Y}|, \delta)$;
$t \leftarrow 0$, $converged \leftarrow False$;
$(\mathbf{u}^*, \mathbf{v}^*, C') \leftarrow BA(\{\hat{\mathbf{w}}_x\})$;
while $converged == False$ **do**
  $converged \leftarrow True$;
  for $x \in \mathcal{X}$ **do**
    $\mathbf{w}_x^- \leftarrow \arg\min_{\mathbf{w}'_x \in \Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}}_x)} D(\mathbf{w}'_x \| \mathbf{v})$;
    $(\acute{\mathbf{u}}, \acute{\mathbf{v}}, C') \leftarrow BA(\{\mathbf{w}_x^-\})$;
    if $\|\acute{\mathbf{v}} - \mathbf{v}^*\|_2 > tol$ **then**
      $\mathbf{v}^* \leftarrow \acute{\mathbf{v}}$, $\mathbf{u}^* \leftarrow \acute{\mathbf{u}}$;
      $converged \leftarrow False$;
    end
  end
  $t \leftarrow t+1$;
end
$\mathbf{v}_{R_L}^- \leftarrow \mathbf{v}^*$, $\mathbf{u}_{R_L}^- \leftarrow \mathbf{u}^*$, $R_L \leftarrow C'$;

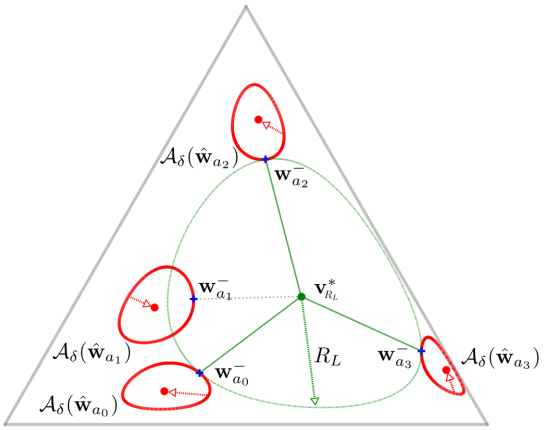

Fig. 5. Lower bound on the channel rate

Figure 5 shows output of Algorithm 1. The'red curves' depict the level-sets of the admissible regions $\{\mathcal{A}_\delta(\hat{\mathbf{w}}_x)\}_{x\in\mathcal{X}}$. The output *pmf* $\mathbf{v}^*$ is plotted as a 'green dot,' and the worst-case channel $\underline{\mathbf{w}}^-$ is plotted as 'blue plus-signs.' The 'green curve' is the surface of $\Gamma_{\xi=R_L}^{\text{fwd}}(\mathbf{v}^*)$.

From this output, we can design a codebook based on R_L and $\mathbf{u}^*$ that will provide reliable communication if the true channel law $\underline{\mathbf{w}}$ lies within the admissible regions.

V. High Probability Bound

We want *all* $|\mathcal{X}|$ sublevel-sets $\Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}}_x)$ (i.e. the admissible regions) to 'contain' the 'true' *ctp* $\mathbf{w}_x$ with high probability ($\geq 1-\delta$). Since there are $|\mathcal{X}|$ inputs, we set $\delta_1 = \frac{\delta}{|\mathcal{X}|}$ and 'tune' the $\mathbb{P}\left\{\mathbf{w}_x \notin \Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}}_x)\right\} \leq \delta_1$

We use a sublevel-set bound to 'size' (i.e. 'tune') the admissible regions, by setting the ξ parameter of $\Gamma_\xi^{\text{rev}}(\hat{\mathbf{w}}_x)$ using either: (1) the 'log' sublevel-set bound (based on Sanov's Theorem) Eq. 12

$$\xi^{\log}(N_x, |Y|, \delta_1) = O\left(\frac{1}{N_x}|\mathcal{Y}|\log(N_x)\right), \tag{34}$$

or (2) the 'loglog' improved sublevel-set bound [11] Eq. 13

$$\xi^{\text{loglog}}(N_x, |Y|, \delta_1) = O\left(\frac{1}{N_x}\left(\sqrt{|\mathcal{Y}|}\log(\log(N_x)) + |\mathcal{Y}|\log(|\mathcal{Y}|)\right)\right). \tag{35}$$

Finally, we consider the rate of convergence as the number of samples increases. Given Pinsker's Theorem [6], we know that

$$\sqrt{2D(\hat{\mathbf{w}}_x \| \mathbf{w}_x)} \geq |\hat{\mathbf{w}}_x - \mathbf{w}_x|_1 \geq \|\hat{\mathbf{w}}_x - \mathbf{w}_x\|_2. \tag{36}$$

and

$$\sqrt{2D(\hat{\mathbf{w}}_x \| \mathbf{w}_x^-)} \geq |\hat{\mathbf{w}}_x - \mathbf{w}_x^-|_1 \geq \|\hat{\mathbf{w}}_x - \mathbf{w}_x^-\|_2. \tag{37}$$

The (deterministic) admissible regions are 'sized' such that

$$D(\hat{\mathbf{w}}_x \| \mathbf{w}_x^-) = \xi(N_x, |Y|, \delta_1) \tag{38}$$

and $\hat{\mathbf{w}}_x$ falls outside the typical set with probability $\leq \delta_1$ if

$$D(\hat{\mathbf{w}}_x \| \mathbf{w}_x) = \xi(N_x, |Y|, \delta_1). \tag{39}$$

So (simultaneously) both $\|\hat{\mathbf{w}}_x - \mathbf{w}_x^-\|_2 \leq \sqrt{\xi(N_x, |Y|, \delta_1)}$ and $\|\hat{\mathbf{w}}_x - \mathbf{w}_x\|_2 \leq \sqrt{\xi(N_x, |Y|, \delta_1)}$ with probability $\geq 1-\delta_1$; therefore,

$$\|\mathbf{w}_x - \mathbf{w}_x^-\|_2 = \|\hat{\mathbf{w}}_x - \mathbf{w}_x^-\|_2 + \|\hat{\mathbf{w}}_x - \mathbf{w}_x\|_2 \tag{40}$$

$$\leq 2\sqrt{\xi(N_x, |Y|, \delta_1)} \text{ with prob. } \geq 1-\delta_1. \tag{41}$$

Recall (Eq. 19) and that $\mathbf{v}^*$ has equal-divergence from all $\mathbf{w}_x^-$ with $u_x > 0$; therefore, $\mathbf{v}^*$ converges to $\mathbf{v}$ at the same rate. And so the overall convergence rate as N increases is $O(\sqrt{\xi(N, |Y|, \delta_1)})$, and specifically the convergence rate when using $\xi^{\log}$ is $O(\frac{1}{N}\log(N))$, and the convergence rate when using ξ^{loglog} is $O(\frac{1}{N}\log(\log(N)))$. □

VI. Results

To test the convergence rate, we generated a set of random channels, where the number of channel inputs was fixed at $|\mathcal{X}| = 3$ and the number of channel outputs was varied as $|\mathcal{Y}| \in \mathcal{Y} = \{5, 7, 10, 15, 20, 25, 30, 35\}$. Each *ctp*s $\mathbf{w}_x$ of the channel law was drawn i.i.d. according to an uniform Dirichlet distribution [10] with hyperparameter $\alpha = 0.8$ or $\mathbf{w}_x \sim Dir(\alpha)$, where

$$Dir(\alpha) \propto \prod_{y\in\mathcal{Y}} {w_{y|x}}^{\alpha-1} \; \forall y \in \mathcal{Y} \tag{42}$$

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

We want to determine the 'tightness' of our level-set bounds against the true probability that a *ctp* $\mathbf{w}_x$ is within an optimally 'sized' admissible region; however, as numerical integration over high dimensional channel output probability spaces is intractable, we used Monte Carlo integration to approximately 'size' each sublevel-sets to the optimal value. Specifically, we generated $M = 8000$ *ctp* samples from a Dirichlet distribution [10] with hyperparameter $\alpha = 1$ or $\mathbf{w}_x \sim Dir(\hat{\mathbf{w}}_x, N_x, \alpha)$,

where

$$Dir\left(\hat{\mathbf{w}}_x, N_x, \alpha\right) \propto \prod_{y \in \mathcal{Y}} {w_{y\,|\,x}}^{N_x \hat{\mathbf{w}}_x + \alpha - 1} \;\; \forall y \in \mathcal{Y} \qquad (43)$$

and then adjusted the ξ parameter of the sublevel-set until $\lfloor M\delta_1 \rfloor$ of the M *ctp* samples fell outside the sublevel-set $\Gamma_{\xi}^{\text{rev}}(\hat{\mathbf{w}}_x)$. Then we used these sublevel-sets as the convex regions in solving for ${R_L}^{\text{mc}}(N, \underline{\mathbf{w}}_{|\mathcal{Y}|})$ (the rate from using Monte Carlo integration).

We compare ${R_L}^{\text{mc}}(N, \underline{\mathbf{w}}_{|\mathcal{Y}|})$ to the R_L sublevel-set rate bounds: ${R_L}^{\log}(N, \underline{\mathbf{w}}_{|\mathcal{Y}|})$ and ${R_L}^{\text{loglog}}(N, \underline{\mathbf{w}}_{|\mathcal{Y}|})$.

We ran algorithm on each of the test channel using parameters $\delta = 0.01$, and $N_0 \in \left\{10 \cdot 2^k\right\}_{k=0}^{19}$, where every input was was sampled N_0 times (so the total number of samples per test channel was $|\mathcal{X}|\, N_0$.

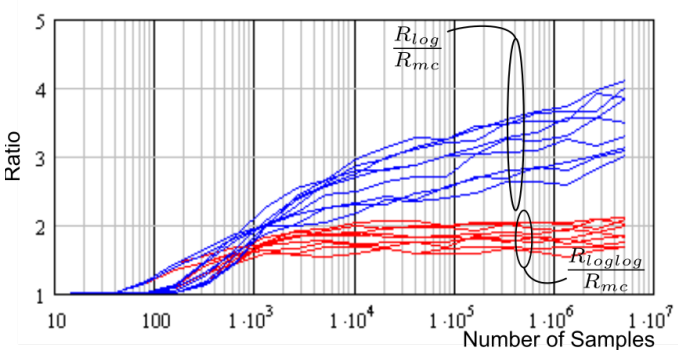

Fig. 6. Convergence as the number of samples increases

Fig. 6 shows one curve for each test channel as the number of samples N is increased. The R_{log}/R_{mc} curves are the ratio of sublevel-set bound ${R_L}^{\log}(N, \underline{\mathbf{w}}_{|\mathcal{Y}|})$ 'over' the Monte Carlo estimate ${R_L}^{\text{mc}}(N, \underline{\mathbf{w}}_{|\mathcal{Y}|})$, and R_{loglog}/R_{mc} is similarly the ratio of sublevel-set bound ${R_L}^{\text{loglog}}(N, \underline{\mathbf{w}}_{|\mathcal{Y}|})$ over the Monte Carlo estimate.

We see is that the R_{loglog}/R_{mc} curves 'flatten out' (for $N > 1000$ samples, and so the 'loglog' sublevel-set bound appears to track (match) the optimal convergence rate within a constant. While 'loglog' sublevel-set bound shows some constant lost (the ratio is not = 1) as N increases, the 'log' (Sanov-based) sublevel-set bound diverges from the optimal estimate as N increases).

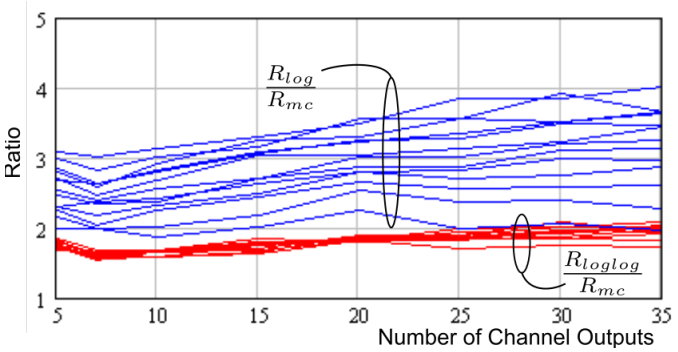

Fig. 7. Convergence as the channel output space increases

Fig. 6 shows one curve for each test channel as the number of channel outputs $|\mathcal{Y}|$ increases (each curve represents N held constant, there is one curve for each $N \in \left\{10 \cdot 2^k\right\}_{k=0}^{19}$). We see a small increasing loss as $|\mathcal{Y}|$ increases. So both the 'log' and 'loglog' levelset bound(s) appear to be 'gradually' diverging from to the optimal Monte Carlo estimate as $|\mathcal{Y}|$ increases. We hope to investigate whether further improvements can remedy this in the future.

VII. Conclusion and Future Work

We developed and demonstrated an algorithm that establishes a high probability lower bound on the maximum information rate through an observed DMC. In addition, we provided a novel transparent proof of the channel capacity of a DMC. Further development of our approach may gain additional insight on the compound DMC, and the finite-block length regime for the DMC.

We expect that this approach could be extended to develop a high probability *upper bound* on the maximum information rate through an observed DMC, and with other modifications, one could establish both high probability lower and upper bounds on the mutual information given the channel samples.

We provided some evidence that our approach is near optimal (it appears to have the same 'bigOh' rate of convergence) as the number of observed samples increases, but more investigation is required to reduce the 'gradually' rising loss when the number of channel output symbols is increased.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, ISBN 0-471-06259-6, John Wiley and Sons Inc., New York, 1991

[2] I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, ISBN 978-0521196819, Cambridge University Press, June 2011

[3] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. on Information Theory*, Vol. IT-44, No. 6, pp. 2148-2177, Oct 1998

[4] T.S. Han, *Information-Spectrum Methods in Information Theory*, ISBN 3-540-43581, Springer-Verlag, New York, 2003

[5] R.E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. on Information Theory*, Vol. IT-18, No. 4, pp. 460 - 473, July 1972

[6] 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

[7] R. W. Yeung and T. Berger, "Multi-way alternating minimization," Proceedings of 1995 IEEE International Symposium on Information Theory, Whistler, BC, Canada, 1995, pp. 74-.

[8] Wikipedia contributors, "Minimax Theorem," *Wikipedia, The Free Encyclopedia*, , (accessed December 1, 2020)

[9] Wikipedia contributors, "Lambert W function," *Wikipedia, The Free Encyclopedia*, , (accessed February 18, 2019)

[10] Wikipedia contributors, "Dirichlet Distribution," *Wikipedia, The Free Encyclopedia*, , (accessed April 10, 2019)

[11] M.A. Tope and J.M. Morris, "Improvements to Sanov and PAC Sublevel-set Bounds for Discrete Random Variables." CISS 2021, submitted, 2021