ISIT'2013 1569720377

Achieving Capacity of Large Alphabet Discrete Memoryless Channels

Yuguang Gao and Aaron B. Wagner
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853
Email: {yg284, wagner}@cornell.edu

***Abstract*—It is observed that some communication situations fall into the *large alphabet* setting, in which the number of channel parameters and the number of channel uses are both large. To model such situations, we consider Discrete Memoryless Channels (DMCs) in which the input and output alphabet sizes increase along with the blocklength n. For known channels, we show that reliable communication at the sequence of channel capacities is possible if and only if the minimum between the square logarithms of the input and the output alphabet sizes grows sublinearly with n. For unknown channels with feedback, we show that universal channel coding can be supported if the input-output product alphabet size grows sublinearly with n.**

I. Introduction

In classical studies of channel capacity, the channel law is fixed, even if it is unknown, and one considers codewords with asymptotically large blocklengths. If the channel is memoryless over a finite alphabet, then this asymptote captures the regime in which the number of channel uses greatly exceeds the alphabet of the channel. If the channel is Markov of fixed order, then it captures the regime in which the number of channel uses greatly exceeds the number of transition probabilities in the channel. In both cases, the number of channel uses greatly exceeds the number of "parameters" in the channel; in fact, they are on different scales.

While this model has enjoyed great success, it is not universally applicable. In high-bandwidth wireless systems in rich fading environments, the number of nonzero "taps" in the channel may not be negligible compared with the blocklength. In packet-switched networks, if one treats the network between two nodes as a "channel," with the packets being the atomic channel symbols, then the alphabet of the channel (the set of all packets) is evidently large compared with the blocklength[1].

We seek to understand how the required blocklength grows as the alphabet size increases. To do this we consider a model in which the channel is discrete memoryless but the alphabet size and blocklength increase together, and we determine which simultaneous growth rates allow for communication at capacity. In practice, of course, the set of channel inputs does not typically vary with the number of channel uses, but this choice of asymptote allows us to focus on the large alphabet regime of interest.

Although little is known about large-alphabet channels, it is worth noting that they are not without precedent in the literature. Reed-Solomon codes exist only if the field size exceeds the blocklength. Moreover, recent work on network coding with adversarial errors involves code constructions with large alphabets and short blocklengths [1] [2], and these works have uncovered new phenomena that have not been observed in the conventional model. Finally, analogous models have received attention in compression and statistics (e.g. [3]-[6]).

For DMCs with known law, we determine the largest possible alphabet growth rate, as a function of the blocklength, such that the sequence of capacities is still achievable. We also study fixed-length universal channel coding for the family of DMCs and derive a sufficient condition on the alphabet growth rate such that universal codes with feedback exist. It is intuitively plausible that the input-output product alphabet size, i.e., the total number of channel parameters in the DMC, should appear in the conditions. This is true of our result for universal channel coding in Section IV, but not for the known channel case, which we shall see in Section III.

II. Notation and Preliminaries

Consider a sequence of DMCs with noiseless feedback $\{W_n : \mathcal{X}_n \to \mathcal{Y}_n\}_{n=1}^{\infty}$, where $\{\mathcal{X}_n\}$ and $\{\mathcal{Y}_n\}$ are sequences of finite input and finite output alphabets, respectively, and $|\mathcal{X}_n||\mathcal{Y}_n|$ is consequently the number of parameters in the channel. Also, for each n, $\{W_n^n : \mathcal{X}_n^{\times n} \to \mathcal{Y}_n^{\times n}\}$ is the n-length operation of the channel, where the set $\mathcal{A}_n^{\times n}$ denotes the n-fold Cartesian product of $\mathcal{A}_n$.

Let $\mathcal{P}(\mathcal{X}_n)$ denote the set of all distributions on $\mathcal{X}_n$, and $\mathcal{P}^n(\mathcal{X}_n)$ the set of all empirical distributions (types) on $\mathcal{X}_n$. For $P_n \in \mathcal{P}(\mathcal{X}_n)$, P_n^N is the distribution of N i.i.d. copies of the random variable $X \sim P_n$. For $Q_n \in \mathcal{P}^n(\mathcal{X}_n)$, $T_{Q_n}^N$ is the typeclass of Q_n, i.e., the set of all N-length strings with type Q_n. For an input distribution $p_n \in \mathcal{P}(\mathcal{X}_n)$ on the channel W_n, we write mutual information between channel input and output as $I(p_n; W_n) = H(p_n W_n) - H(W_n|p_n)$.

Throughout logarithms and exponents are in base e.

A. Known Channels

In the known channel scenario, both the encoder and the decoder are assumed to have complete knowledge of the

[1]If the errors in the packet channel are solely due to physical layer noise, then one could argue that the atomic channel symbols are bits or constellation points that comprise the packet. If the errors are mainly due to jamming or interference, however, then it is natural to model the errors at the packet level.

channel law W_n for each n. We shall adopt the following notion of achievable rate sequence.

Definition 1 (Achievable rate sequence): A sequence of rates $\{R_n\}$ is *achievable with feedback* if for any $\epsilon > 0$, there exists an (n, M_n)-feedback code $\{f_n, g_n\}$, where $f_n = \{f^t\}_{t=1}^n$ with $f^t : \mathcal{M}_n \times \mathcal{Y}_n^{\times(t-1)} \to \mathcal{X}_n$, and $g_n : \mathcal{Y}_n^{\times n} \to \mathcal{M}_n$, where $\mathcal{M}_n = \{1, 2, ..., M_n\}$, such that for all n sufficiently large,

$$\frac{1}{n} \log M_n > R_n - \epsilon$$

and

$$\frac{1}{M_n} \sum_{m=1}^{M_n} \Pr(\hat{M} \neq m | M = m) \leq \epsilon,$$

where M, $\hat{M}$ are encoded and decoded messages, respectively. We define the rate sequence $\{R_n\}$ to be *achievable without feedback* in the natural way.

Define the sequence of information channel capacities $\{C_n\}$ as

$$C_n := \max_{p_n \in \mathcal{P}(\mathcal{X}_n)} I(p_n; W_n). \tag{1}$$

This is the rate sequence that one could achieve if there were no restrictions on the blocklength. We shall be interested in determining when it is possible to achieve this rate sequence if the blocklength is constrained to grow with the alphabet size in a prescribed way. This question is answered in Section III.

B. Unknown Channels

For the family of DMCs $\{W_{n,\theta} : \mathcal{X}_n \to \mathcal{Y}_n, \theta \in \Theta_n\}_{n=1}^\infty$, where Θ_n is the set of all DMCs with the given input and output alphabets, we study fixed-length universal channel coding in the sense of the following definitions (note that $\theta \in \Theta_n$ assumes no prior distribution):

Definition 2 ($\{\mathcal{X}_n, \mathcal{Y}_n, R_n\}$-universal channel code): For the family of DMCs $\{W_{n,\theta}\}$, a sequence of triples $\{\mathcal{X}_n, \mathcal{Y}_n, R_n\}$ admits a deterministic (resp. randomized) universal channel code if for any $\epsilon > 0$, there exists an (n, M_n)-feedback code $\{f_n, g_n\}$ with rate $n^{-1} \log M_n > e^{-\epsilon} R_n$ and for each $W_{n,\theta}$ satisfying

$$\sup_{p_n \in \mathcal{P}(\mathcal{X}_n)} I(p_n; W_{n,\theta}) > R_n + \epsilon, \text{ for all } n, \tag{2}$$

the following holds

$$\frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y^n \in \mathcal{Y}_n^{\times n} : g_n(y^n) \neq m} W_{n,\theta}^n(y^n | f_n(m)) \leq \epsilon,$$

for all n sufficiently large.
Note that we assume a multiplicative back off on the coding rate but an additive back off in (2), which facilitates the analysis.

Definition 3 ($\{\mathcal{X}_n, \mathcal{Y}_n\}$ supports universal channel coding): For the family of DMCs $\{W_{n,\theta}\}$, a sequence of input and output alphabets $\{\mathcal{X}_n, \mathcal{Y}_n\}$ supports deterministic (resp. randomized) universal channel coding if for each rate sequence $\{R_n\}$, the sequence $\{\mathcal{X}_n, \mathcal{Y}_n, R_n\}$ admits a deterministic (resp. randomized) universal channel code.

Note that for unknown channels we only consider codes with feedback. Feedback is an important feature of the problem in that it allows the encoder and the decoder to learn the channel, and in particular the optimal input distribution, via training. Of course, the extent to which this can be done depends on the size of the alphabets relative to the blocklength.

Various models and definitions for universal reliable communication exist in the literature [7]. The definition used here is rather weak. We shall see, however, that establishing the existence of universal channel codes under this weak definition is nontrivial, even for relatively slow alphabet growth rates.

III. Large alphabet channel coding for known channels

When the channel W_n is known for each n, we have the following result regarding the growth rate of $|\mathcal{X}_n|$ and $|\mathcal{Y}_n|$ for which the sequence of capacities (1) is achievable.

Theorem 1: Given a DMC $\{W_n : \mathcal{X}_n \to \mathcal{Y}_n\}_{n=1}^\infty$, if sequences of input alphabets $\{\mathcal{X}_n\}$ and output alphabets $\{\mathcal{Y}_n\}$ satisfy

$$\lim_{n \to \infty} \min \left(\frac{\log^2 |\mathcal{X}_n|}{n}, \frac{\log^2 |\mathcal{Y}_n|}{n} \right) = 0, \tag{3}$$

then $\{C_n\}$ is achievable, even without using feedback codes.

Remark 1: It is interesting to note that if the square logarithm of the input *or* the output alphabet size is $o(n)$, the sequence of capacities is achievable regardless of the growth rate of the other.

Remark 2: While the standard proofs of achievability for fixed-alphabet channel coding theorems can be extended to handle slowly growing alphabets in a straightforward manner, they cannot handle all alphabets satisfying (3). To do so, we use the argument based on Markov's inequality in [9].

We need the following lemma in the proof of the theorem.

Lemma 1 ([6]): For any random variable U and V for which U takes values in $\mathcal{A}$ and any $\epsilon > 0$,

$$p^n \left(\left| -\frac{1}{n} \log p^n(U^n | V^n) - H(U|V) \right| > \epsilon \right) \leq \frac{\max(\log^2 |\mathcal{A}|, 1)}{n\epsilon^2}.$$

Note that the bound on the right-hand side does not depend on the alphabet of V.

Proof of Theorem 1:

Fix $p_n \in \mathcal{P}(\mathcal{X}_n)$. Given an arbitrary $\epsilon > 0$, let $R_n = I(p_n; W_n) - \epsilon$, and let $M_n = e^{nR_n}$. The encoder uses random code selection without feedback and the decoder applies ML decoding rule, i.e., given the received sequence y^n, the decoded message is $\hat{m} = \operatorname{argmax}_{1 \leq m \leq M_n} W_n^n(y^n | x^n(m))$ (ties are resolved arbitrarily).

Assume $x^n(1)$ is transmitted and the channel output is y^n. Let $E_m = \{W_n^n(Y^n | X^n(m)) \geq W_n^n(Y^n | X^n(1))\}$,

$m = 2, ..., M_n$, be the error events. We have

$$\begin{aligned}
&\Pr\{E_m | X^n(1) = x^n(1), Y^n = y^n\} \\
&= \Pr\{W_n^n(Y^n|X^n(m)) \geq W_n^n(Y^n|X^n(1))| \\
&\qquad X^n(1) = x^n(1), Y^n = y^n\} \\
&\leq \frac{E[W_n^n(Y^n|X^n(m))|X^n(1) = x^n(1), Y^n = y^n]}{W_n^n(y^n|x^n(1))} \qquad (4) \\
&= \frac{Q_n^n(y^n)}{W_n^n(y^n|x^n(1))}, \qquad (5)
\end{aligned}$$

where $Q_n^n(y^n) := \sum_{x^n \in \mathcal{X}_n^{\times n}} p_n^n(x^n) W_n^n(y^n|x^n(m) = x^n)$.

Here (4) is obtained by applying Markov's inequality.

Note that

$$\begin{aligned}
&\Pr\left\{\left|-\frac{1}{n}\log\frac{Q_n^n(Y^n)}{W_n^n(Y^n|X^n(1))} - I(p_n; W_n)\right| \leq \frac{\epsilon}{2}\right\} \\
&\geq 1 - \Pr\left\{\left|-\frac{1}{n}\log Q_n^n(Y^n) - H(Q_n)\right| > \frac{\epsilon}{4}\right\} \\
&\quad - \Pr\left\{\left|-\frac{1}{n}\log W_n^n(Y^n|X^n(1)) - H(W_n|p_n)\right| > \frac{\epsilon}{4}\right\} \\
&\geq 1 - \frac{32\max(\log^2|\mathcal{Y}_n|, 1)}{n\epsilon^2},
\end{aligned}$$

where the last inequality follows from Lemma 1.

Now let

$$G_n = \left\{\frac{1}{n}\log\frac{Q_n^n(Y^n)}{W_n^n(Y^n|X^n(1))} \leq -I(p_n; W_n) + \frac{\epsilon}{2}\right\},$$

we have

$$\Pr(G_n) \geq 1 - \frac{32\max(\log^2|\mathcal{Y}_n|, 1)}{n\epsilon^2}. \qquad (6)$$

If we define $\tilde{W}_n$ as the reverse channel from Y^n to $X^n(1)$

$$\tilde{W}_n^n(x^n|y^n) := \frac{p_n^n(x^n(1)) W_n^n(y^n|x^n(1))}{Q_n^n(y^n)},$$

by considering $\frac{p_n^n(X^n(1))}{\tilde{W}_n^n(X^n(1)|Y^n)}$ in place of (5) and via similar arguments we also have

$$\Pr(G_n) \geq 1 - \frac{32\max(\log^2|\mathcal{X}_n|, 1)}{n\epsilon^2}. \qquad (7)$$

Under the assumption (3), and in view of (6) and (7)

$$\begin{aligned}
\Pr(G_n^c) &\leq \min\left(\frac{32\max(\log^2|\mathcal{Y}_n|, 1)}{n\epsilon^2}, \frac{32\max(\log^2|\mathcal{X}_n|, 1)}{n\epsilon^2}\right) \\
&\leq \frac{\epsilon}{2}, \text{ for all } n \text{ sufficiently large.} \qquad (8)
\end{aligned}$$

Therefore, the average probability of error is

$$\begin{aligned}
P_e^{(n)} &\leq \mathbb{E}\left[\Pr\left\{\bigcup_{m=2}^{M_n} E_m | X^n(1), Y^n\right\}\right] \\
&\leq \Pr(G_n^c) + \Pr(G_n)\mathbb{E}\left[\Pr\left\{\bigcup_{m=2}^{M_n} E_m | X^n(1), Y^n\right\} \middle| G_n\right] \\
&\leq \epsilon/2 + M_n \mathbb{E}\left[\Pr(E_2 | X^n(1), Y^n) | G_n\right] \\
&\leq \epsilon/2 + e^{n(I(p_n;W_n)-\epsilon)} e^{-n(I(p_n;W_n)-\epsilon/2)} \\
&= \epsilon/2 + e^{-n\epsilon/2} \leq \epsilon,
\end{aligned}$$

for all n sufficiently large. Since p_n is arbitrary, we can choose $p_n^* = \text{argmax}_{p_n \in \mathcal{P}(\mathcal{X}_n)} I(p_n; W_n)$, for each n, to be the capacity achieving input distribution, and thus $\{C_n\}$ is achievable. ■

Next we show by counterexample that the condition presented in Theorem 1 is also necessary for the sequence of capacities $\{C_n\}$ to be achievable.

Theorem 2: Suppose $\{\mathcal{X}_n\}$, $\{\mathcal{Y}_n\}$ satisfy

$$\limsup_{n\to\infty} \min\left(\frac{\log^2|\mathcal{X}_n|}{n}, \frac{\log^2|\mathcal{Y}_n|}{n}\right) = d > 0, \qquad (9)$$

where d is a real constant. For DMC $\{W_n : \mathcal{X}_n \to \mathcal{Y}_n\}_{n=1}^{\infty}$, where $\mathcal{X}_n = \{1, 2, ..., |\mathcal{X}_n|\}$, $\mathcal{Y}_n = \{1, 2, ..., |\mathcal{Y}_n|\}$, we define for each n,

$$W_n(y|x) = \begin{cases} \frac{1}{2} & \text{if } y = x \\ \frac{1}{|\mathcal{Y}_n|} & \text{if } x > |\mathcal{Y}_n| \\ 0 & \text{if } y > |\mathcal{X}_n| \\ \frac{1}{2(\min(|\mathcal{X}_n|, |\mathcal{Y}_n|)-1)} & \text{otherwise,} \end{cases}$$

for all $x \in \mathcal{X}_n, y \in \mathcal{Y}_n$. Then $\{C_n\}$ is not achievable, even with feedback codes.

Proof of Theorem 2:

From the structure of the channel transition matrix, we see that for each n, only the square matrix of size $\min(|\mathcal{X}_n|, |\mathcal{Y}_n|)$ on the top left corner contributes effectively to communication. The reason is the following. Symbols $y \in \mathcal{Y}_n$ will not occur if $y > |\mathcal{X}_n|$. In the case where $|\mathcal{X}_n| > |\mathcal{Y}_n|$, if some code Ψ_n has codewords with symbols $x \in \mathcal{X}_n$, $x > |\mathcal{Y}_n|$, we can transform it into a random code over the alphabet $\{x : x \leq |\mathcal{Y}_n|\}$. Indeed, transmitting a symbol uniformly selected from this set induces the same distribution over the outputs as transmitting an x with $x > |\mathcal{Y}_n|$. Thus we can restrict the input alphabet to $\{x : x \leq |\mathcal{Y}_n|\}$ if we allow for randomized encoding. Of course, if the channel is known then randomized encoder affords no advantage in terms of average error probability. Thus we may restrict attention to deterministic codes over the input alphabet $\{x : x \leq |\mathcal{Y}_n|\}$.

Now define the effective channel $\{\tilde{W}_n : \tilde{\mathcal{X}}_n \to \tilde{\mathcal{Y}}_n\}$, where $\tilde{\mathcal{X}}_n = \tilde{\mathcal{Y}}_n = \{1, 2, ..., \min(|\mathcal{X}_n|, |\mathcal{Y}_n|)\}$, as

$$\tilde{W}_n(y|x) = W_n(y|x), \text{ for all } x \in \tilde{\mathcal{X}}_n, y \in \tilde{\mathcal{Y}}_n.$$

Fix $p_n \in \mathcal{P}(\tilde{\mathcal{X}}_n)$, and fix an arbitrary $\delta > 0$.

For each n, let $q_n(y) = \sum_{x \in \tilde{\mathcal{X}}_n} p_n(x)\tilde{W}_n(y|x)$ for all $y \in \tilde{\mathcal{Y}}_n$, and $q_n^n(y^n) = \prod_{i=1}^n q_n(y_i)$.

Let m be the message to be transmitted. Define $I(m; y^n) := \log\frac{\tilde{W}_n^n(y^n|m)}{q_n^n(y^n)} = \sum_{i=1}^n I(x_i; y_i)$, where $I(x_i; y_i) = \log\frac{\tilde{W}_n(y_i|m, y^{i-1})}{q_n(y_i)} = \log\frac{\tilde{W}_n(y_i|x_i)}{q_n(y_i)}$.

For any coding scheme, define the probability of correct decoding to be

$$P_c^{(n)} = \frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{y^n \in \mathcal{D}_m^{(n)}} \tilde{W}_n^n(y^n|m),$$

where decoding regions $\{\mathcal{D}_m^{(n)}, m = 1, ..., M_n\}$ are a partition of $\tilde{\mathcal{Y}}_n^{\times n}$, and $x_m^n = (x_{m1}, ..., x_{mn})$ is the codeword of message m.

Suppose $\{C_n\}$ is achievable, then there exists a feedback code with message set $\{M_n\}$ such that $n^{-1} \log M_n > C_n - \delta$. Now let $\gamma = 2\delta$, and let

$$B_m^{(n)} = \{y^n \in \tilde{\mathcal{Y}}_n^{\times n} : I(m; y^n) > n(C_n - \gamma)\}.$$

Then

$$P_c^{(n)} = \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y^n \in \mathcal{D}_m^{(n)} \bigcap (B_m^{(n)})^c} \tilde{W}_n^n(y^n|m) \tag{10}$$

$$+ \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y^n \in \mathcal{D}_m^{(n)} \bigcap B_m^{(n)}} \tilde{W}_n^n(y^n|m), \tag{11}$$

and we upper bound (10) by

$$\frac{1}{M_n} e^{n(C_n-\gamma)} \sum_{m=1}^{M_n} \sum_{y^n \in \mathcal{D}_m^{(n)}} q_n^n(y^n) < e^{-n(C_n-\delta)} e^{n(C_n-\gamma)}$$
$$= e^{-n\delta},$$

thus the first term of $P_c^{(n)}$ tends to 0 as $n \to \infty$.

We then upper bound (11) by

$$\frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y^n \in B_m^{(n)}} \tilde{W}_n^n(y^n|m)$$

and in the following we will show that it is strictly bounded away from 1.

Note that for each $m \in \{1, 2, ..., M_n\}$,

$$\sum_{y^n \in B_m^{(n)}} \tilde{W}_n^n(y^n|m)$$
$$= \Pr\left\{ \sum_{i=1}^{n} \log \frac{\tilde{W}_n(Y_i|X_{mi})}{q_n(Y_i)} > n(C_n - \gamma) | m \right\}. \tag{12}$$

By straightforward calculation we have $C_n = \log \frac{|\tilde{\mathcal{Y}}_n|}{2} - \frac{1}{2} \log(|\tilde{\mathcal{Y}}_n| - 1)$, $n = 1, 2, ...$, achieved by uniform distribution over $\tilde{\mathcal{X}}_n$.

From the channel specification we have

$$\log \frac{\tilde{W}_n(Y_i|X_{mi})}{q_n(Y_i)} = \begin{cases} \log \frac{|\tilde{\mathcal{Y}}_n|}{2}, & \text{w.p. } \frac{1}{2} \\ \log \frac{|\tilde{\mathcal{Y}}_n|}{2} - \log(|\tilde{\mathcal{Y}}_n| - 1), & \text{w.p. } \frac{1}{2} \end{cases}$$

Let $J_n = |\{i : \log \frac{W_n(Y_i|X_{mi})}{q_n(Y_i)} = \log \frac{|\tilde{\mathcal{Y}}_n|}{2} - \log(|\tilde{\mathcal{Y}}_n| - 1)\}|$, and note that J_n is a binomial random variable with parameters $(n, \frac{1}{2})$. Then (12) can be written as

$$\Pr\{J_n(\log \frac{|\tilde{\mathcal{Y}}_n|}{2} - \log(\tilde{\mathcal{Y}}_n| - 1)) + (n - J_n) \log \frac{|\tilde{\mathcal{Y}}_n|}{2}$$
$$> n(\log \frac{|\tilde{\mathcal{Y}}_n|}{2} - \frac{1}{2} \log(|\tilde{\mathcal{Y}}_n| - 1)) - n\gamma\}$$
$$= \Pr\left\{ \frac{J_n - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} < \frac{\sqrt{n} \cdot 2\gamma}{\log(|\tilde{\mathcal{Y}}_n| - 1)} \right\}. \tag{13}$$

Since by assumption $\limsup_{n\to\infty} \frac{\log^2 |\tilde{\mathcal{Y}}_n|}{n} = d > 0$, there exists a real number $d' > 0$ and a subsequence $\{n_k\}$ such that $\lim_{k\to\infty} \frac{\log |\tilde{\mathcal{Y}}_{n_k}|}{\sqrt{n_k}} = d'$. Then for all k sufficiently large, we have $\frac{\sqrt{n_k}}{\log |\tilde{\mathcal{Y}}_{n_k}|} < \frac{2}{d'}$, and therefore

$$p_{n_k}^{n_k} \left\{ \frac{J_{n_k} - \frac{1}{2}n_k}{\frac{1}{2}\sqrt{n_k}} \geq \frac{\sqrt{n_k} \cdot 2\gamma}{\log(|\tilde{\mathcal{Y}}_{n_k}| - 1)} \right\}$$
$$\geq p_{n_k}^{n_k} \left\{ \frac{J_{n_k} - \frac{1}{2}n_k}{\frac{1}{2}\sqrt{n_k}} \geq \frac{4\gamma}{d'} \frac{\log |\tilde{\mathcal{Y}}_{n_k}|}{\log(|\tilde{\mathcal{Y}}_{n_k}| - 1)} \right\},$$

which gives

$$\limsup_{n\to\infty} \Pr\left\{ \frac{J_n - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} \geq \frac{\sqrt{n} \cdot 2\gamma}{\log(|\tilde{\mathcal{Y}}_n| - 1)} \right\} > 0,$$

where the central limit theorem is applied and note that $\frac{4\gamma}{d'} \frac{\log |\tilde{\mathcal{Y}}_n|}{\log(|\tilde{\mathcal{Y}}_n| - 1)} > 0$ for all n. This implies that (12) and thus the second term (11) of $P_c^{(n)}$ is strictly less than 1 for all n. Therefore, the probability of error cannot be made arbitrarily small, which implies that $\{C_n\}$ is not achievable. ■

IV. Large alphabet channel coding of unknown channels

We provide a sufficient condition on the alphabet growth rate for which universal coding is possible. Note that the condition is directly related to the total number of parameters in the channel.

Theorem 3: If $\{\mathcal{X}_n, \mathcal{Y}_n\}$ is a sequence of input and output alphabets satisfying $|\mathcal{X}_n||\mathcal{Y}_n| = O(n^\alpha)$ for some constant $\alpha \in (0, 1)$, then $\{\mathcal{X}_n, \mathcal{Y}_n\}$ supports randomized universal channel coding.

Toward proving the theorem, we construct a fixed-length training-based universal channel code. In the training phase, the encoder sends a training sequence by round robin over the input alphabet, collects the channel feedback, and estimates the channel law using the ML estimator. In the transmission phase, the encoder selects the codewords i.i.d. according to the optimal input distribution for the estimated channel, and the Maximum Mutual Information (MMI) decoder [8] is applied to reconstruct the message. When the hypothesis of the theorem is satisfied, the number of parameters in the channel grows slower than the blocklength, and it is expected that the encoder can learn the channel accurately through training, at which point communication at or near capacity should be possible. The following arguments make this intuition precise.

We analyze the training phase first. Let the length of the training sequence be nD_n, where $D_n = n^{\alpha'-1}$ for some $\alpha < \alpha' < 1$. A sequence of length $d_n = \frac{nD_n}{|\mathcal{X}_n|}$ of each input symbol $x \in \mathcal{X}_n$ is sent, and let $\{Y_k^{(n)}(x)\}_{k=1}^{d_n}$ be the corresponding output sequence from the channel.

The estimated channel is given by

$$\hat{W}_n(y|x) = \frac{1}{d_n} \sum_{k=1}^{d_n} \mathbb{1}_{\{Y_k^{(n)}(x) = y\}}, \text{ for each } x \in \mathcal{X}_n, y \in \mathcal{Y}_n.$$

Proof of Theorem 3:

Given a rate sequence $\{R_n\}$ and an arbitrary $\epsilon > 0$, and given an arbitrary unknown channel $\{W_n\}$ in the family, let $p_n^* = \text{argmax}_{p_n} I(p_n; W_n)$ and $\hat{p}_n = \text{argmax}_{p_n} I(p_n; \hat{W}_n)$ be the optimal input distributions for the true channel W_n and the estimated channel $\hat{W}_n$, respectively.

The length of the transmitted codeword is $N = (1 - D_n)n$. Let the number of codewords be $M_n = e^{NR_n}$, and so the rate of the code $n^{-1} \log M_n = (1 - D_n)R_n$ satisfies the rate requirement in Definition 2.

Let $B_n = \{\sup_{p_n} |I(p_n; W_n) - I(p_n; \hat{W}_n)| \leq \epsilon/12\}$. Using standard concentration results, one can show that $\Pr(B_n)$ tends to one as n tends to infinity, so long as the hypothesis of the theorem holds. If B_n happens, then we have (denote this event by F_n)

$$I(p_n^*; W_n) - \epsilon/6 \leq I(\hat{p}_n; W_n) \leq I(p_n^*; W_n).$$

The encoder uses random code selection according to $\hat{p}_n$, and the decoder reconstructs the message by the rule $\hat{m} = \text{argmax}_{1 \leq m \leq M_n} V_{n,m}^N(y^N | x^N(m))$, where $y^N \in T_{V_{n,m}}(x^N(m))$, which is equivalent to the MMI decoding rule.

Assume $x^N(1)$ is transmitted and the channel output is y^N.

Let $E_m = \{V_{n,m}^N(Y^N|X^N(m)) \geq V_{n,1}^N(Y^N|X^N(1))\}$, $m = 2, ..., M_n$, be the error events. Then

$$\begin{aligned}
&\Pr\{E_m | X^N(1) = x^N(1), Y^N = y^N, \hat{p}_n\} \\
&\leq \frac{E[V_{n,m}^N(Y^N|X^N(m))|X^N(1) = x^N(1), Y^N = y^N, \hat{p}_n]}{V_{n,1}^N(y^N|x^N(1))} \\
&= \frac{\sum_{x^N \in \mathcal{X}_n^{\times N}} \hat{p}_n^N(x^N) V_{n,m}^N(y^N|x^N)}{V_{n,1}^N(y^N|x^N(1))} \\
&= \frac{\sum_{\tilde{Q}_n \in \mathcal{P}^n(\mathcal{X}_n \times \mathcal{Y}_n)} \sum_{x^N : (x^N, y^N) \in T_{\tilde{Q}_n}} \hat{p}_n^N(x^N) \tilde{Q}_n^N(y^N|x^N)}{V_{n,1}^N(y^N|x^N(1))} \\
&\leq \frac{\sum_{\tilde{Q}_n \in \mathcal{P}^n(\mathcal{X}_n \times \mathcal{Y}_n)} \sum_{x^N} \hat{p}_n^N(x^N) \tilde{Q}_n^N(y^N|x^N)}{V_{n,1}^N(y^N|x^N(1))} \\
&\leq \frac{|\mathcal{P}^n(\mathcal{X}_n \times \mathcal{Y}_n)| p_{y^N}^N(y^N)}{V_{n,1}^N(y^N|x^N(1))},
\end{aligned} \tag{14}$$

where the final inequality follows because $p_{y^N}^N$ places the highest probability on y^N among all i.i.d. distributions.

Note that in (14), $N^{-1} \log |\mathcal{P}^n(\mathcal{X}_n \times \mathcal{Y}_n)| \to 0$ when $|\mathcal{X}_n||\mathcal{Y}_n| = O(n^\alpha)$ (cf. [3], Lemma 1).

Let $Q_n(\cdot) = \sum_{x \in \mathcal{X}_n} \hat{p}_n(x) W_n(\cdot|x)$. We have for Y^N distributed according to Q_n^N on $\mathcal{Y}_n^{\times N}$,

$$\frac{1}{N} \log \frac{p_{Y^N}^N(Y^N)}{Q_n^N(Y^N)} = D(p_{Y^N} || Q_n) \xrightarrow{p} 0$$

(see [4], Lemma 7).

Similarly,

$$\frac{1}{N} \log \frac{V_{n,1}^N(Y^N|X^N(1))}{W_n^N(Y^N|X^N(1))} = D(V_{n,1}||W_n|p_{X^N}) \xrightarrow{p} 0.$$

It follows that

$$-\frac{1}{N} \log \frac{Q_n^N(Y^N)}{W_n^N(Y^N|X^N(1))} + \frac{1}{N} \log \frac{p_{Y^N}^N(Y^N)}{V_{n,1}^N(Y^N|X^N(1))}$$

also tends to zero in probability.

Since we have

$$\begin{aligned}
&\Pr\left\{\left|-\frac{1}{N} \log \frac{Q_n^N(Y^N)}{W_n^N(Y^N|X^N(1))} - I(\hat{p}_n; W_n)\right| \leq \frac{\epsilon}{4}\right\} \\
&\geq 1 - \frac{128 \max(\log^2 |\mathcal{Y}_n|, 1)}{N\epsilon^2} \text{ by Lemma 1,}
\end{aligned}$$

if we let G_n denote the set

$$\left\{\frac{1}{N} \log \frac{p_{Y^N}^N(Y^N)}{V_{n,1}^N(Y^N|X^N(1))} \leq -I(\hat{p}_n; W_n) + \frac{\epsilon}{3}\right\},$$

then $\Pr(G_n)$ tends to one.

Thus the average probability of error satisfies

$$\begin{aligned}
P_e^{(n)} &\leq \Pr(F_n^c) + \Pr(G_n^c) + \Pr(\text{error}|F_n, G_n) \\
&\leq \epsilon/2 + M_n \mathbb{E}[\Pr\{E_2|X^N(1), Y^N, \hat{p}_n\}|F_n, G_n] \\
&\leq \epsilon/2 + \mathbb{E}[e^{-N(-I(p_n^*; W_n) + \epsilon + I(\hat{p}_n; W_n) - \epsilon/2)}|F_n, G_n] \\
&\leq \epsilon/2 + e^{-(1-D_n)n\epsilon/3} \leq \epsilon,
\end{aligned}$$

for all n sufficiently large. ∎

Remark 3: It remains to be seen whether the alphabet growth rate in Theorem 3 can support deterministic universal channel coding. On the other hand, we conjecture that when the alphabet sizes satisfy $|\mathcal{X}_n||\mathcal{Y}_n| = \Omega(n)$, universal reliable communication is impossible, even using randomized codes.

ACKNOWLEDGEMENT

This research was supported by the National Science Foundation under grant CCF-1117128.

REFERENCES

[1] O. Kosut, L. Tong and D. Tse, "Polytope codes against adversaries in networks," in *Proceedings IEEE International Symposium on Information Theory (ISIT)*, pp. 2423–2427, Austin, 2010.

[2] E. Ahmed and A. B. Wagner, "Lossy source coding with Byzantine adversaries," in *Proceedings Information Theory Workshop (ITW)*, pp. 462–466, Paraty, Brazil, 2011.

[3] A. Barron, "Uniformly powerful goodness of fit tests," *Ann. Stat.*, vol. 17, no. 1, pp. 107–204, 1989.

[4] B. G. Kelly, A. B. Wagner, T. Tularak and P. Viswanath, "Classification of homogeneous data with large alphabets," *IEEE Trans. Inform. Theory*, vol. 59, no. 2, pp. 782–795, 2013.

[5] A. Orlitsky and N. P. Santhanam, "Speaking of infinity", *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2215-2230, 2004.

[6] B. G. Kelly and A. B. Wagner, "Near-lossless compression of large alphabet sources," in *Proc. Conf. Inf. Sci. and Syst. (CISS)*, Princeton, NJ, 2012.

[7] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2148-2177, 1998.

[8] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, 1981.

[9] Y. Lomnitz and M. Feder, "A simpler derivation of the coding theorem,"[online], Available: http://arxiv.org/abs/1205.1389.