Empirical Properties of Good Channel Codes

Qinghua (Devon) Ding*, Sidharth Jaggi†, Shashank Vatedka‡, and Yihan Zhang†
*Dept. of Computer Science and Engineering, The Chinese University of Hong Kong (qhding@cse.cuhk.edu.hk)
†Dept. of Information Engineering, The Chinese University of Hong Kong ({jaggi, zy417}@ie.cuhk.edu.hk)
‡Dept. of Electrical Engineering, Indian Institute of Technology, Hyderabad (shashankvatedka@iith.ac.in)

***Abstract*—In this article, we revisit the classical problem of channel coding and obtain novel results on properties of capacity-achieving codes. Specifically, we give a linear algebraic characterization of the set of capacity-achieving input distributions for discrete memoryless channels. This allows us to characterize the dimension of the manifold on which the capacity-achieving distributions lie. We then proceed by examining empirical properties of capacity-achieving codebooks by showing that the joint-type of k-tuples of codewords in a good code must be close to the k-fold product of the capacity-achieving input distribution. While this conforms with the intuition that all capacity-achieving codes must behave like random capacity-achieving codes, we also show that some properties of random coding ensembles do not hold for all codes. We prove this by showing that there exist pairs of communication problems such that random code ensembles simultaneously attain capacities of both problems, but certain (superposition ensembles) do not.**

Due to lack of space, several proofs have been omitted but can be found at https://sites.google.com/view/yihan/ [1].

I. Introduction

Shannon's celebrated channel coding theorem [2] in one fell swoop simultaneously derived the fundamental limits of reliable communication over noisy channels, and demonstrated coding strategies that approach these fundamental limits. In this work we revisit this classical problem and derive novel results in a variety of directions.

1) Capacity-achieving input distributions: First, it has long been known [3] in the literature that for some discrete memoryless channels (DMCs) the capacity-achieving distributions are not necessarily unique, and in general may be a convex subset of the probability simplex $\Delta_{\mathcal{X}}$ over the input alphabet $\mathcal{X}$ of the DMC. Prior results [4] in this direction have required solving systems of non-linear equations that do not shed insight into the structure of these optimizing solutions. Others [5], [6], [7] have designed alternating minimization for efficient computation of capacity. Our first result gives a remarkably clean characterization of the set of capacity-achieving input distributions for general DMCs as the intersection of a specific affine space intersected with the probability simplex $\Delta_{\mathcal{X}}$. Consequences of this characterization (and the techniques involved) include:

- A characterization (as the rank of a system of linear equations) of the dimension of the manifold on which capacity-achieving distributions for a given DMC lie (Theorem 1).
- Ancillary characterizations of DMCs with unique optimizing input distributions in terms of linear-algebraic properties of the channel transition law viewed as a matrix (Corollaries 14 and 15).
- The perhaps surprising example of a DMC $W_{Y|X}$ for which the set of optimizing input distributions does not "tensorize" (Claim 3). That is, consider the "two-use channel" corresponding to the DMC with input alphabet $\mathcal{X}\times\mathcal{X}$, output alphabet $\mathcal{Y}\times\mathcal{Y}$, and channel law $W_{Y|X}^{\otimes 2} \coloneqq W_{Y|X}\otimes W_{Y|X}$. We can show that the set of optimizing input distributions for the two-use channel is much larger (indeed, lies in a higher dimensional manifold) than the convex hull of the tensor product of the optimizing input distributions of the underlying DMC. We view this as a somewhat unexpected result, especially in the context of the well-known fact [8] that the capacity of the DMC does indeed tensorize (for any integer k the capacity of the k-use channel is k times the capacity of the one-use channel).

2) Properties all "good" codebooks must satisfy: Next, we move to examining empirical properties of capacity-achieving[1] code ensembles. As a proxy to guide our intuition, we use as benchmarks properties exhibited by the capacity-achieving random code ensembles suggested by Shannon [2], and prove both "positive" and "negative" results in this direction, as described below.

Joint types of codeword k-tuples: One of our results in this direction is that of joint distributions of codewords. As a benchmark, suppose $W_{Y|X}$ has a unique optimizing input distribution P_X^*, then it can be directly verified that for any constant k, an overwhelming fraction of codes in Shannon's capacity-achieving random coding ensemble satisfy the property that "most" k-tuples of codewords have joint type that is "close" (say in total variation distance) to the product distribution $P_X^{*\otimes k}$. Our "positive" result in this context is that indeed such a property must be true for *any* capacity-achieving sequence of codes. That is, given *any* code $\mathcal{C}$ of rate δ-close to the Shannon capacity and with a probability of error of at most ϵ, k codewords sampled uniformly at random from $\mathcal{C}$ will, with probability at least $1-\eta(\epsilon,\delta)$ have joint type at most $\Delta(\epsilon,\delta)$ close to $P_X^{*\otimes k}$, for explicit functions η and Δ that converge to zero as ϵ and δ converge to zero. See Claims 12 and 13, Theorem 4 and Corollaries 6, Lemma 7.

The work of Sidharth Jaggi and Yihan Zhang was supported by the Research Grants Council (RGC) of Hong Kong under Project Numbers 14300617, 14304418 and 14301519.

[1]Given any memoryless channel W, we will interchangeably use "W-good" or "capacity-achieving" to describe codes which have rate arbitrarily close to the capacity of W and have average error probability decaying in blocklength n when transmitted over n channel uses.

Aside from being of intrinsic interest, this fundamental fact about any capacity-achieving code ensemble comes in handy in proving impossibility results for some channel models. In a companion paper [9], we use a corresponding result for AWGN channels to show a novel upper bound on the capacity of a certain two-way adversarially jammed channel. For this outer bound, we critically use the fact that even if the adversary is unable to infer the specific codewords transmitted by the two legitimate users, he is nonetheless able to rely on the fact that with high probability their transmitted codewords are "close to orthogonal", and thereby can tailor his jamming strategy to such pairs of codewords.

Finally, we show some "negative" results – some properties that are satisfied with overwhelming probability (double-exponentially close to one!) by the random coding ensemble are in fact *not* necessarily true for all code ensembles. Specifically, we show:

- *Non-universality:* Random code ensembles are "universal" in the sense that for any two channels $W_{Y|X}$ and $W'_{Y|X}$ with the same capacity and the same optimizing input distribution, the same random code ensemble is capacity-achieving for both channels. However, we demonstrate pairs of channels such that a capacity-achieving ensemble (in particular, superposition code ensembles [10]) for one is provably far from capacity-achieving for the other (Lemma 8).
- *Non-list-decodability:* Random code ensembles are known to *simultaneously* achieve the list-decoding capacity [11] for corresponding "adversarial" channels. That is, even if the noise were "worst case" with the same noise parameters as the underlying DMC for which the random code ensemble was designed, the decoder is able to salvage something by outputting a "small" (constant!) sized list guaranteed to contain the transmitted codeword. We show that again superposition-based codes can violate this correspondence by demonstrating a capacity-achieving code ensemble for a DMC such that it necessarily results in codes which have exponential list sizes (Lemma 9).

A. Prior work

There is a significant body of work formalizing the property that good codes for memoryless channels must induce an output distribution which approximates the capacity-achieving output distribution. Consider a DMC W with finite input alphabet and capacity C. If $\{\mathcal{C}_n\}$ is a sequence of codes of rate $C-\epsilon/2 < R < C$ and achieving $o(1)$ probability of error over W, then the induced output distribution when a random codeword from $\mathcal{C}_n$ is passed through W is close to the capacity-achieving output distribution [12]: $\frac{1}{n}D(P_{\underline{\mathbf{y}}}\|(P_Y^*)^{\otimes n}) \leqslant \epsilon$ for large enough n, where $D(P\|Q)$ denotes the Kullback-Leibler divergence between P and Q and P_Y^* denotes the capacity-achieving output distribution.

This holds even if we are allowed to tolerate a small but nonvanishing probability of error, and even under the total variation distance d_{TV} [13]. These properties hold for all channels that satisfy a strong converse, and in particular the AWGN channel [12].

It is also known that the kth order empirical output distribution of the code approximates the k-fold capacity-achieving output distribution [14]. Let

$$\widehat{Q}^{(k)}_{\underline{\mathbf{x}}}(\underline{a}) \coloneqq \sum_{i=1}^{n-k+1} \frac{1_{\{(x_i,\ldots,x_{i+k-1})=(a_1,\ldots,a_k)\}}}{n-k+1}$$

and define kth order empirical input distribution of the code $\mathcal{C}$ to be

$$\widehat{Q}^{(k)}_{\mathcal{C}}(a_1,\ldots,a_k) = \frac{1}{|\mathcal{C}|}\sum_{\underline{x}\in\mathcal{C}} \widehat{Q}^{(k)}_{\underline{\mathbf{x}}}. \tag{1}$$

It was shown in [14] that

$$d(k) \coloneqq \min_{\bar{X}^k: I(\bar{X}^k;\bar{Y}^k)=kC} D(\mathbb{E}\widehat{Q}^{(k)}_{\underline{\mathbf{x}}}\|P_{\bar{X}^k}),$$

is vanishing in n for $k = O(1)$. It is important to point out a subtlety here. While $\frac{1}{n}D(P_{\underline{\mathbf{y}}}\|(P_Y^*)^{\otimes n}) \to 0$ for a good sequence, $D(P_{\underline{\mathbf{y}}}\|(P_Y^*)^{\otimes n})$ is asymptotically larger than zero. Therefore, the approximability depends on the metric used, and the channel: For the AWGN channel, the asymptotic Wasserstein distance between $P_{\underline{\mathbf{x}}}$ and $(P_X^*)^{\otimes n}$ tends to zero. The convergence of the empirical output distribution to the n-fold capacity-achieving output distribution also holds for certain fading channels [15].

Other necessary conditions for good codes have been studied, including a tight characterization of the peak-to-average-power ratio of good codes for AWGN channels [16], [17]. Properties of the empirical distribution of good codes for multiple access channels were studied in [18], and good quantizers for lossy source coding was studied by [19], [20].

A closely related property of codes is resolvability [12], [21]. Here, the goal is to design codebooks $\mathcal{C}_n$ such that the output distribution when a random codeword is passed through W is close to $P_Y^{*\otimes n}$ in total variation distance, i.e., $d_{\mathrm{TV}}(P_{\underline{\mathbf{y}}}, P_Y^{*\otimes n}) \to 0$ as $n \to \infty$. This is proved to be fundamental in many problems including physical layer security [22], [23], [24], [25] and covert communication [26], [27], [28].

The first systematic study of an ordering of channels was [29] which defined the notion of less noisy and more capable channels. Specifically, a channel V is more capable than W if for every code $\mathcal{C}$ achieving ϵ probability of error over W can be expurgated with negligible loss of rate to achieve ϵ probability of error over V. An equivalent condition is that $I(X;Y_V) \geqslant I(X;Y_W)$ for all input distributions P_X. This has been studied extensively in the context of broadcast channels (see, e.g., [30], [31], [32], [33] for an incomplete list).

A recent paper closely related to ours is [34] which found upper bound on capacity of DMCs with positive invertible channel matrix.

We would like to point out that in contrast to [14] which studied (1), we examine the kth order type of k-tuples of codewords

$$\tau_{\underline{x}_1,\ldots,\underline{x}_k}(a_1,\ldots,a_k) \coloneqq \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_{1i}=a_1,\ldots,x_{ki}=a_k\}},$$

where x_{ij} denotes the jth component of $\underline{x}_i$.

II. Main results

Some of our primary results come from the property that a DMC W is a linear operator from $P_X \in \Delta_{\mathcal{X}}$ to $P_Y \in \Delta_{\mathcal{Y}}$, i.e., representing P_X and P_Y as column vectors[2] $\underline{p}_X$ and $\underline{p}_y$, we can write $W\underline{p}_X = \underline{p}_Y$. We assume that the input and output alphabets are finite.

1) Optimizing input distributions: Our first result is the following linear-algebraic characterization of the space of capacity-achieving input distributions:

Theorem 1 (Characterization of capacity-achieving distributions)**.** *Given $W = (\underline{p}_1, \underline{p}_2, \cdots, \underline{p}_m)$ where $\underline{p}_i$ denotes the ith column, let $\underline{r}^\top \coloneqq (H(\underline{p}_1), H(\underline{p}_2), \cdots, H(\underline{p}_m))$. For any capacity-achieving distribution $\underline{p}^*$ for W, the whole set of capacity-achieving distributions is $\mathcal{P}_X^* = \left\{\underline{p}^* + \ker \binom{W}{\underline{r}}\right\} \cap \mathbb{R}_+^m$.*

This can be generalized to the k-use channel $W^{\otimes k}$.

Theorem 2 (Capacity-achieving distributions for k-use channel)**.** *Given W and $\underline{r}^\top = (r_1, r_2, \cdots, r_m)$ as in Theorem 1. If we have a capacity-achieving distribution P_X^* for W, then the whole set of capacity-achieving distribution for $W^{\otimes k}$ is $\mathcal{P}_{X^k}^* = \left\{P_X^{*\otimes k} + \ker \binom{W^{\otimes k}}{\underline{r}^{(k)}}\right\} \cap \mathbb{R}_+^{m^k}$.*

This yields the (perhaps surprising) result that for $k \geqslant 2$, the space of capacity-achieving input distributions for the k-use channel $\mathcal{P}_{X^k}^*$ can be much larger than the convex hull of $(\mathcal{P}_X^*)^{\otimes k}$, where $\mathcal{P}_X^*$ is th space of optimizing input distributions of W.

Claim 3. *The following noisy typewriter channel*

$$W = \begin{pmatrix} 1/2 & 0 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}. \tag{2}$$

has $\mathcal{P}_X^*$ equal to the convex hull of $\{(1/2, 0, 1/2, 0)^\top$, $(0, 1/2, 0, 1/2)^\top\}$. Specifically, $\dim(\mathcal{P}_X^*) = 1$.

For $k \geqslant 2$, $\dim(\mathcal{P}_{X^k}^*) = 4^k - 3^k > k$.

2) Empirical properties of good codes: Our second result is that the empirical joint distribution of k-tuples of codewords is close to $\mathcal{P}_{X^k}^*$ for $k = O(1)$.

Theorem 4 (Empirical properties of DMC-good codes)**.** *For any $\epsilon > 0$ and $k = O(1)$, any good code for DMC W with $\mathcal{P}_{X^k}^*$ as defined in Theorem 2 satisfies*

$$\frac{1}{|\mathcal{C}|^k} \sum_{(\underline{c}_1, \ldots, \underline{c}_k) \in \mathcal{C}^k} \min_{Q \in \mathcal{P}_{X^k}^*} 1_{\{d_{\mathrm{TV}}(\tau_{\underline{c}_1, \ldots, \underline{c}_k}, Q) > \epsilon\}} = o(1),$$

where $d_{\mathrm{TV}}(\cdot, \cdot)$ denotes total variation distance.

This suggests that all capacity-achieving codes must behave very much like random capacity-achieving ensembles. This behaviour is inherited by good codes for the AWGN channel. Intuition suggests that these must behave like random Gaussian codebooks, in the sense that pairs of codewords are "almost" orthogonal to each other. We can show that this intuition is indeed correct.

Lemma 5. *Given any two deterministic codes $\mathcal{C}_1$ and $\mathcal{C}_2$ that are good for AWGN(P, N) channels, for any constant $\eta \in (0, 1)$, it holds that*

$$\limsup_{n\to\infty} \Pr_{\substack{\underline{\mathbf{x}}_1 \sim \mathcal{C}_1 \\ \underline{\mathbf{x}}_2 \sim \mathcal{C}_2}} [\langle \underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2 \rangle > n\eta] = 0,$$

where the probability is taken over $\underline{\mathbf{x}}_1$ and $\underline{\mathbf{x}}_2$ that are chosen uniformly at random from $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively.

This gives us the following corollaries:

Corollary 6. *Given any deterministic codes $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}$ that are good for AWGN(P, N) channels, for any constant $\eta \in (0, 1)$ and $k \in \mathbb{Z}_{\geqslant 2}$, it holds that*

$$\limsup_{n\to\infty} \Pr_{\substack{\underline{\mathbf{x}}_1 \sim \mathcal{C}_1 \\ \underline{\mathbf{x}}_2 \sim \mathcal{C}_2}} [\langle \underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2 \rangle < -n\eta] = 0, \tag{3}$$

$$\limsup_{n\to\infty} \Pr_{\substack{\underline{\mathbf{x}}_1 \sim \mathcal{C}_1 \\ \underline{\mathbf{x}}_2 \sim \mathcal{C}_2}} [|\langle \underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2 \rangle| > n\eta] = 0, \tag{4}$$

$$\limsup_{n\to\infty} \Pr_{\underline{\mathbf{x}}, \underline{\mathbf{x}}' \overset{i.i.d.}{\sim} \mathcal{C}} \left[\left|\left\langle \underline{\mathbf{x}}, \underline{\mathbf{x}}' \right\rangle\right| > n\eta\right] = 0. \tag{5}$$

$$\limsup_{n\to\infty} \Pr_{\underline{\mathbf{x}}_1, \cdots, \underline{\mathbf{x}}_k \overset{i.i.d.}{\sim} \mathcal{C}} \left[\bigcup_{\substack{i,j \in [k] \\ i \neq j}} \left\{\left|\left\langle \underline{\mathbf{x}}_i, \underline{\mathbf{x}}_j \right\rangle\right| > n\eta\right\} \right] = 0. \tag{6}$$

Using similar ideas, we prove another empirical property that is universal to all AWGN-good codes.

Lemma 7. *Given any deterministic code $\mathcal{C}$ that is good for AWGN(P, N) channels, for any constant $\eta \in (0, 1)$, it holds that*

$$\limsup_{n\to\infty} \Pr_{\underline{\mathbf{x}} \sim \mathcal{C}} \left[\|\underline{\mathbf{x}}\|_2 \leqslant \sqrt{nP(1-\eta)}\right] = 0.$$

3) Impossibility results: While the previous results reinforce the intuition that good codebooks behave like random code ensembles, this does not necessarily hold in all cases. Given a pair of channels V, W having the same capacity-achieving input distribution, a randomly chosen capacity-achieving codebook will, with high probability, achieve vanishingly small probability of error over both U, V. However, there are ensembles of codebooks for which this is not true.

More concretely, consider the ensemble of binary superposition codes [10] of rate R with 2^{nR_1} cloud centers and 2^{nR_2} satellite codewords in each cloud. The cloud centers are chosen uniformly at random from $\mathbb{F}_2^n$ while the satellite codewords are chosen uniformly from a ball of Hamming radius nq around the cloud center. Let us call a typical code from this ensemble $\mathcal{C}_{\sup}(q, R_1, R_2)$. We can prove the following

Lemma 8. *Fix a $p \in (0, 1/2)$ and small $\delta > 0$. A typical code $\mathcal{C}_{\sup}(q, R_1, R_2)$ with $R_1 + R_2 = 1 - H(p) - \delta$ and $R_2 = (1 - H(p))H(q) - \delta$ achieves vanishingly small error probability over the BEC($H(p)$). However, no expurgated subcode of the same rate can achieve vanishingly small error probability over the BSC(p).*

[2]Henceforth, we will use the pmf P_X and its vector form $\underline{p}_X$ interchangeably to denote the same object.

The above result is not surprising, given [32] which showed that the BEC is more capable than the BSC. However, our result gives an entire ensemble of codes where most codes are good for the BEC but cannot have vanishing probability of error over the BSC even after rate-lossless expurgation.

This ensemble of superposition codes is more powerful in giving such counterexamples beyond pairs of discrete memoryless channels. One such question that we can address is whether every capacity-achieving code for the BSC(p) be expurgated without loss of rate to also achieve $O(\mathrm{poly}(n))$ list sizes over the bitflip-p channel. This is true for random binary codes, but not for superposition codes:

Lemma 9. *Fix any $p \in (0, 1/2)$. The code $\mathcal{C}_{\mathrm{sup}}(q, R_1, R_2)$ for any $0 < q < 1/2$ and small $\delta > 0$ with $R_2 = H(q * p) - H(p) - \delta$ and $R_1 + R_2 = 1 - H(p) - \delta$ cannot achieve subexponential list sizes over the bitflip-p channel.*

The other question is whether codes with fractional minimum distance p achieving the GV bound of $1 - H_2(p)$ also achieve the capacity of the BSC(p). This holds, for e.g., for random binary linear codes but once again, not for all ensembles:

Lemma 10. *There exist codes with minimum distance at least np but no expurgated subcodebook of the same asymptotic rate can achieve vanishing probability of error over the BSC(p).*

Let us now proceed to examine each item in more detail.

III. Proof Sketches

A. Linear algebraic characterization of the set of capacity-achieving input distributions

Our goal is to understand the following properties: For $k = 1, 2, \ldots, n$,

1) $\mathbb{P}_k$: the property that $W^{\otimes k}$ is an injective linear operator from $(\mathbb{R}^{|\mathcal{X}|})^{\otimes k}$ to $(\mathbb{R}^{|\mathcal{Y}|})^{\otimes k}$;
2) $\mathbb{S}_k$: the property that $W^{\otimes k}$ as a k-use channel has unique product capacity-achieving input distribution $P_X^{*\otimes k}$;
3) $\mathbb{T}_k$: the property that any code $\mathcal{C}$ achieving $o(1)$ probability of error over W has most k-tuples of codewords "close" to the k-times capacity-achieving input distribution, i.e., for all $\delta > 0$ we have

$$\frac{1}{|\mathcal{C}|^k} \sum_{(\underline{c}_1, \ldots, \underline{c}_k) \in \mathcal{C}^k} 1_{\{d_{\mathrm{TV}}(\tau_{\underline{c}_1, \ldots, \underline{c}_k}, P_X^{*\otimes k}) > \delta\}} = o(1).$$

The following results are elementary:

- $\mathbb{P}_1 \Rightarrow \mathbb{S}_k$
- $\mathbb{S}_1 \not\Rightarrow \mathbb{P}_1$. Indeed, Muroga [3] has the following example

$$W = \begin{pmatrix} 1/2 & 1/4 & 0 \\ 1/2 & 1/4 & 0 \\ 0 & 1/2 & 1 \end{pmatrix}. \tag{7}$$

- $\mathbb{S}_1 \Leftrightarrow \mathbb{S}_k$

As a warmup, we show the following:

Lemma 11 (Linear independence lemma)**.** *If W has unique P_X^*, and $\mathrm{supp}(P_X^*) \coloneqq \mathcal{X}' \subset \mathcal{X}$, then the set of conditional distributions $\{P_{Y|X=i}, i \in \mathcal{X}'\}$ is linearly independent.*

Proof. W.l.o.g., we let $\mathcal{X}' = [m']$ where $m' = |\mathcal{X}'|$ and $m = |\mathcal{X}|$. Denote $\underline{p}(i) = \underline{p}_{Y|X=i}, r_i = H(\underline{p}(i)), \forall i \in \mathcal{X}$. The optimizing input distribution $\underline{p}_X^* = (p_1^*, p_2^*, \cdots, p_{m'}^*, 0, \cdots, 0)^\top$, where $p_i > 0, \forall i \in [m']$.

We now suppose $\{\underline{p}(i), i \in [m']\}$ is not linearly independent. Then we have $\sum_{i=1}^{m'} a_i \underline{p}(i) = 0$ for some $\underline{a} \neq \underline{0}$. Equivalently, $W\underline{a} = \underline{0}$. Note we have $\sum_{i=1}^{m'} a_i = \sum_{i=1}^{m'} a_i \underline{1}^\top \underline{p}(i) = 0$, where $\underline{1}$ is the all-1's vector. Consider a small perturbation of $\underline{p}_X^*$ as $\underline{p}_{X,\epsilon} = (p_1^* + \epsilon a_1, p_2^* + \epsilon a_2, \cdots, p_{m'}^* + \epsilon a_{m'}, 0, \cdots, 0)$ for some $\epsilon > 0$. Clearly $\underline{p}_{X,\epsilon} \neq \underline{p}_X^*$, and for small enough (but non-zero) ϵ (say, for $\epsilon \in [-\alpha, \beta]$ for suitable $\alpha, \beta > 0$), the vector $\underline{p}_{X,\epsilon}$ is a valid pmf.

However, $p_{Y,\epsilon} = W\underline{p}_{X,\epsilon} = W\underline{p}_X^* + \epsilon W\underline{a} = \underline{p}_Y^*$, and $\underline{r}^\top \underline{p}_{X,\epsilon} = \underline{r}^\top \underline{p}_X^* + \epsilon \underline{r}^\top \underline{a} = \underline{r}^\top \underline{p}_X^* + \epsilon(\sum_{i=1}^n a_i r_i)$. If $\sum_{i=1}^n a_i r_i = 0$, then we have $\underline{r}^\top \underline{p}_{X,\epsilon} = \underline{r}^\top \underline{p}_X^*$, and hence $I(X_\epsilon; Y_\epsilon) = H(p_{Y,\epsilon}) - \underline{r}^\top \underline{p}_{X,\epsilon} = H(\underline{p}_Y^*) - \underline{r}^\top \underline{p}_X^* = C$. This contradicts the assumption of uniqueness of $\underline{p}_X^*$. If $\sum_{i=1}^n a_i r_i \neq 0$, then w.l.o.g., assume $\sum_{i=1}^n a_i r_i > 0$, then we have for $\epsilon = -\alpha$, $I(X_\epsilon; Y_\epsilon) = C + \alpha(\sum_{i=1}^n a_i r_i) > C$, leading to a contradiction. □

We now show that if the capacity-achieving input distribution P_X^* is unique, then most codewords of a capacity-achieving code have type close to P_X^*. The proof is via contradiction, where we construct a subcodebook of the same asymptotic rate but a vanishingly small error probability for an input-constrained channel of smaller capacity.

Claim 12. *If W has unique capacity-achieving input distribution P_X^*, then $\mathbb{S}_1 \implies \mathbb{T}_1$.*

Proof. Suppose a W-good code $\mathcal{C}$ satisfies

$$\Pr_{\underline{\mathbf{x}} \sim \mathcal{C}} \left[d_{\mathrm{TV}}(\tau_{\underline{\mathbf{x}}}, P^*) > \epsilon \right] = \eta,$$

for some constant $\eta > 0$. Then $\mathcal{C}' = \{\underline{x} \in \mathcal{C} : d_{\mathrm{TV}}(\tau_x, P^*) > \epsilon\}$ is a large (of size $|\mathcal{C}|\eta$) code which has o(1) probability of error when used on channel W' with the same transition law as W, plus input constraint $d_{\mathrm{TV}}(\tau_{\underline{x}}, P^*) > \epsilon$. Moreover, W' has capacity

$$\max_{P:\, d_{\mathrm{TV}}(P, P^*) > \epsilon} I(X; Y) < C,$$

which contradicts the continuity of mutual information. □

We can strengthen this to show that even the k-th order types of most k-tuples of codewords must be close to $P_X^{*\otimes k}$. The fundamental idea is that if $\mathcal{C}$ is capacity-achieving for W, then $\mathcal{C}^k$ is capacity-achieving for $W^{\otimes k}$. We can then use Claim 12 for the k-use channel.

Claim 13. *If W has unique capacity-achieving input distribution P_X^*, then $\mathbb{S}_1 \Leftrightarrow \mathbb{T}_k$ for $k = O(1)$.*

B. The Space of Capacity-Achieving Distributions

In the following, we will give a characterization of the entire set of capacity-achieving distributions for k-use channels, given one capacity-achieving distribution for W.

1) Proof of Theorem 1: First, we will show that any $\underline{p} \in \mathcal{P}_X^*$ is a valid distribution. Note $\underline{p} = \underline{p}_X^* + \underline{q}$ for some $\underline{q} \in \ker\left(\frac{W}{\underline{r}^\top}\right)$. Since $W\underline{q} = 0$, we have $\underline{1}^\top \underline{q} = \underline{1}^\top W \underline{q} = 0$ (where $\underline{1}$ is the all-ones vector). This then gives $\underline{1}^\top \underline{p} = \underline{1}^\top \underline{p}_X^* = 1$. Since $\underline{p} \in \mathbb{R}_+^m$, it is a valid distribution.

Let us now show that $\underline{p}$ is capacity-achieving. Note that $\underline{p}_Y = W\underline{p} = W(\underline{p} - \delta) = W\underline{p}_X^* = \underline{p}_Y^*$, and $\underline{r}^\top \underline{p} = \underline{r}^\top \underline{p}_X^*$ for the same reason. Thus, $I(X;Y) = H(Y) - H(Y|X) = H(Y^*) - \underline{r}^\top \underline{p} = H(Y^*) - \underline{r}^\top \underline{p}^* = I(X^*;Y^*)$. Thus $\underline{p}$ is capacity-achieving.

Now we prove that any capacity-achieving distribution $\underline{p}$ is in $\mathcal{P}_X^*$. Note $W(\underline{p} - \underline{p}^*) = 0$ by uniqueness of $\underline{p}_Y^*$. And we have $0 = I(X;Y) - I(X^*;Y^*) = \underline{r}^\top(\underline{p} - \underline{p}^*)$. Thus, $\underline{p} - \underline{p}^* \in \ker\left(\frac{W}{\underline{r}^\top}\right)$. Since $\underline{p} \in \mathbb{R}_+^m$, the claim is proved. □

Corollary 14 (Unique P_X^* for weakly symmetric channel)**.** *A weakly symmetric channel*[3] *has unique P_X^* iff W is injective.*

Proof. Since $P_X^* = \frac{1}{m}\underline{1}$ is always a capacity-achieving distribution, any direction in $\ker(W)$ will be feasible in Theorem 1. □

Corollary 15 (Uniqueness of P_X^*)**.** *If $\ker\left(\frac{W}{\underline{r}^\top}\right) = \{0\}$, then $\underline{p}^*$ is unique.*

Proof. Since $\ker\left(\frac{W}{\underline{r}^\top}\right) = \{0\}$, by Theorem 1, $\mathcal{P}_X^* = \{P_X^*\} \cap \mathbb{R}_+^m = \{P_X^*\}$. Thus P_X^* is unique. □

Example 1 (Muroga's channel). The channel W in (7) is not injective, but note $\underline{r}^\top = (1, 3/2, 0)$, and $\ker\left(\frac{W}{\underline{r}^\top}\right) = \{\underline{0}\}$, thus it still has unique P_X^*.

Remark 2. This characterization of uniqueness is not tight in general. We can easily construct counter-examples when $m \geqslant n$.

An interesting consequence of our characterization is that the set of optimizing input distributions for the k-use channel does not necessarily tensorize. Indeed, from Theorem 2, $\mathcal{P}_{X^k}^* = \left\{P_X^{*\otimes k} + \ker\left(\frac{W^{\otimes k}}{\underline{r}^{(k)}}\right)\right\} \cap \mathbb{R}_+^{m^k}$.

2) Proof of Theorem 2: Note $W^{\otimes k}$ is essentially a discrete memoryless channel over $\mathcal{X}^k$ and $P_X^{*\otimes k}$ is one of its capacity-achieving distributions. The only thing left is to find $\underline{r}^{(k)}$. Since $r^{(k)}_{i_1 i_2 \cdots i_k} = H(Y^k | X^k = i_1 i_2 \cdots i_k) = \sum_{k'=1}^k H(Y_{k'} | X_{k'} = i_{k'}) = \sum_{k'=1}^k r_{i_{k'}}$, thus $(\underline{r}^{(k)})^\top$ is the row vector with components $\{r^{(k)}_{i_1 i_2 \cdots i_k}, i_1, i_2, \cdots, i_k \in \mathcal{X}\}$, where the indices are in lexicographical order. For example, we have $(\underline{r}^{(2)})^\top = (r_1 + r_1, r_1 + r_2, \cdots, r_1 + r_m \mid r_2 + r_1, r_2 + r_2, \cdots, r_2 + r_m \mid \cdots \mid r_m + r_1, r_m + r_2, \cdots, r_m + r_m)$. □

As a corollary of the above, we can show the following.

Corollary 16 (Weakly symmetric channel)**.** *For a weakly symmetric channel W (in fact for any channel with constant $H(Y|X)$), if we have a P_X^* for it, then the set of capacity-achieving distribution for $W^{\otimes k}$ is $\mathcal{P}_{X^k}^* = \left\{P_X^{*\otimes k} + \sum_{i=1}^k (\mathbb{R}^m)^{\otimes i-1} \otimes \ker(W) \otimes (\mathbb{R}^m)^{\otimes k-i}\right\} \cap \mathbb{R}_+^{m^k}$.*

[3]A channel is said to be weakly symmetric if every column is a permutation of every other column and all row sums are the same.

3) Proof of Claim 3: Consider the noisy typewriter channel in (2). It can be verified that $\mathcal{P}_X^* = \mathrm{cl}\{(1/2, 0, 1/2, 0)^\top, (0, 1/2, 0, 1/2)^\top\}$, where cl denotes convex hull. The choice $P_X^* = (1/4, 1/4, 1/4, 1/4)^\top$ is capacity-achieving, and the null space $\ker(W) = \mathrm{span}\{(-1, 1, -1, 1)^\top\}$. Since W is symmetric, by Corollary 14, we have the null space of $W \otimes W$ as $\ker(W \otimes W) = \mathrm{span}\{v\} \otimes \mathbb{R}^4 + \mathbb{R}^4 \otimes \mathrm{span}\{v\}$, here $v := (-1, 1, -1, 1)^\top$. Note $\mathbb{R}^4 = \mathrm{span}\{v, u_1, u_2, u_3\}$ for $u_1 = (1, 0, 0, 0)^\top$, $u_2 = (0, 1, 0, 0)^\top$, and $u_3 = (0, 0, 1, 0)^\top$. Then we have $\ker(W \otimes W) = \mathrm{span}(\{v \otimes v, v \otimes u_i, u_i \otimes v, \forall i \in [3]\})$. Since $P_X^* \otimes P_X^* = \frac{1}{16} I$, any direction in $\ker(W \otimes W)$ is admissible. Thus $\dim(\mathcal{P}_{X^2}^*) = \dim(\ker(W \otimes W)) = 7$. In particular, we have $\dim(\mathcal{P}_{X^k}^*) = 4^k - 3^k$. □

C. Counterexamples using superposition codes

Lemma 17. *$\mathcal{C}_{\mathrm{sup}}(q, R_1, R_2)$ with bounded distance decoding achieves vanishingly small probability of error over the BSC(p) as long as*

$$R_1 + R_2 < 1 - H(p) \quad \textit{and} \quad R_2 < H(q * p) - H(p) \tag{8}$$

where $q * p := q(1-p) + p(1-q)$.

Lemma 18. *$\mathcal{C}_{\mathrm{sup}}(q, R_1, R_2)$ with MAP decoding achieves vanishingly small probability of error over the BEC(p) as long as*

$$R_1 + R_2 < 1 - p \quad \textit{and} \quad R_2 < (1-p)H(q). \tag{9}$$

1) Proof of Lemma 8: The fact that $\mathcal{C}_{\mathrm{sup}}(q, R_1, R_2)$ achieves a vanishingly small probability of error follows from Lemma 18. To show that the probability of error over the BSC is large, consider the more powerful decoder where Bob has access to an oracle who reveals the cloud center of the transmitted codeword. Conditioned on the cloud center (and hence the cloud subcode), the corresponding subcodebook can achieve vanishingly small probability of error only if R_2 is less than the input-constrained capacity of the BSC(p) with input Hamming weight constraint of nq, or (see, e.g., [10, Theorem 3.2]), equal to $H(q*p) - H(p)$. However, $(1 - H(p))H(q) > H(q*p) - H(p)$ for p and q in $(0, 1/2)$. This completes the proof. □

The proofs of Lemmas 9 and 10 are similar, and we skip the details.

D. Closing remarks

We have studied general properties of codes and make a variety of novel observations. While this is only a first step, we believe that the tools used in this paper would complement [14], [13] in obtaining a clearer picture of capacity-achieving codes. In particular, we feel that the linear algebraic characterization as well as the superposition code ensemble could be very useful in deriving more general results.

REFERENCES

[1] Q. Ding, S. Jaggi, S. Vatedka, and Y. Zhang, "Empirical properties of good channel codes," *Preprint*, 2020.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[3] S. Muroga, "On the capacity of a discrete channel. i mathematical expression of capacity of a channel which is disturbed by noise in its every one symbol and expressible in one state diagram," *Journal of the Physical Society of Japan*, vol. 8, no. 4, pp. 484–494, 1953.
[4] S. Muroga, "On the capacity of a discrete channel, ii. mathematical expression of the capacity of a noisy channel which is expressible by corresponding two multi-state diagrams," *Journal of the Physical Society of Japan*, vol. 11, no. 10, pp. 1109–1120, 1956.
[5] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.
[6] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.
[7] I. Csiszár and G. Tusnady, "Information geonetry and alternating minimization procedures," *Statistics and decisions*, vol. 1, pp. 205–237, 1984.
[8] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, New York, 1991.
[9] Y. Zhang, S. Vatedka, and S. Jaggi, "Quadratically constrained two-way adversarial channels," *ArXiv preprint, arXiv:2001.02575*, 2020.
[10] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.
[11] V. Guruswami, J. Hastad, and S. Kopparty, "On the List-Decodability of Random Linear Codes," in *Proc. ACM Symp. on Theory of Comp.*, 2010.
[12] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 752–772, 1993.
[13] Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with nonvanishing error probability," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 5–21, 2013.
[14] S. Shamai and S. Verdú, "The empirical distribution of good codes," *IEEE Transactions on Information Theory*, vol. 43, no. 3, pp. 836–846, 1997.
[15] S. L. Fong and V. Y. Tan, "Empirical output distribution of good delay-limited codes for quasi-static fading channels," *arXiv preprint arXiv:1510.08544*, 2015.
[16] Y. Polyanskiy and Y. Wu, "Peak-to-average power ratio of good codes for gaussian channel," *IEEE Transactions on Information Theory*, vol. 60, no. 12, pp. 7655–7660, 2014.
[17] Y. Polyanskiy, "ℓ_p-norms of codewords from capacity-and dispersion-achieveing gaussian codes," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 294–301, IEEE, 2012.
[18] S. L. Fong, "Output distributions of capacity-achieving codes for gaussian multiple access channels," *IEEE Communications Letters*, vol. 20, no. 5, pp. 938–941, 2016.
[19] T. Weissman and E. Ordentlich, "The empirical distribution of rate-constrained source codes," *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 3718–3733, 2005.
[20] V. Kostina and S. Verdú, "The output distribution of good lossy source codes," in *2015 Information Theory and Applications Workshop (ITA)*, pp. 308–312, IEEE, 2015.
[21] J. Liu, P. Cuff, and S. Verdu, "e_γ-resolvability," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 2629–2658, 2016.
[22] M. R. Bloch and J. N. Laneman, "Strong secrecy from channel resolvability," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8077–8098, 2013.
[23] A. Thangaraj, "Coding for wiretap channels: Channel resolvability and semantic security," in *2014 IEEE Information Theory Workshop (ITW 2014)*, pp. 232–236, IEEE, 2014.
[24] W. Yang, R. F. Schaefer, and H. V. Poor, "Wiretap channels: Nonasymptotic fundamental limits," *IEEE Transactions on Information Theory*, 2019.
[25] L. Yu and V. Y. Tan, "Renyi resolvability and its applications to the wiretap channel," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1862–1897, 2018.
[26] M. R. Bloch, "Covert communication over noisy channels: A resolvability perspective," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2334–2354, 2016.
[27] P. H. Che, M. Bakshi, and S. Jaggi, "Reliable deniable communication: Hiding messages in noise," in *2013 IEEE International Symposium on Information Theory*, pp. 2945–2949, IEEE, 2013.
[28] L. Wang, G. W. Wornell, and L. Zheng, "Fundamental limits of communication with low probability of detection," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3493–3503, 2016.
[29] J. Korner and K. Marton, "Comparison of two noisy channels," *Topics in Information Theory, I. Csiszr and P. Elias, Eds., Amsterdam, The Netherlans*, pp. 411–423, 1977.
[30] A. Gamal, "The capacity of a class of broadcast channels," *IEEE Transactions on Information Theory*, vol. 25, no. 2, pp. 166–169, 1979.
[31] C. Nair and A. El Gamal, "The capacity region of a class of three-receiver broadcast channels with degraded message sets," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4479–4493, 2009.
[32] Y. Geng, C. Nair, S. S. Shitz, and Z. V. Wang, "On broadcast channels with binary inputs and symmetric outputs," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 6980–6989, 2013.
[33] D. Sutter and J. M. Renes, "Universal polar codes for more capable and less noisy channels and sources," in *2014 IEEE International Symposium on Information Theory*, pp. 1461–1465, IEEE, 2014.
[34] T. Nguyen and T. Nguyen, "On bounds and closed form expressions for capacities of discrete memoryless channels with invertible positive matrices," *arXiv preprint, arXiv:2001.01847v1*, 2020.