An Entropy Sumset Inequality and Polynomially Fast Convergence to Shannon Capacity Over All Alphabets

Venkatesan Guruswami[*] and Ameya Velingker[†]

Computer Science Department, Carnegie Mellon University
Pittsburgh, PA, USA
`guruswami@cmu.edu, avelingk@cs.cmu.edu`

─── **Abstract** ───

We prove a lower estimate on the increase in entropy when two copies of a conditional random variable $X|Y$, with X supported on $\mathbb{Z}_q = \{0, 1, \ldots, q-1\}$ for prime q, are summed modulo q. Specifically, given two i.i.d. copies (X_1, Y_1) and (X_2, Y_2) of a pair of random variables (X, Y), with X taking values in $\mathbb{Z}_q$, we show

$$H(X_1 + X_2 \mid Y_1, Y_2) - H(X|Y) \geq \alpha(q) \cdot H(X|Y)(1 - H(X|Y))$$

for some $\alpha(q) > 0$, where $H(\cdot)$ is the normalized (by factor $\log_2 q$) entropy. In particular, if $X|Y$ is not close to being fully random or fully deterministic and $H(X|Y) \in (\gamma, 1-\gamma)$, then the entropy of the sum increases by $\Omega_q(\gamma)$. Our motivation is an effective analysis of the finite-length behavior of polar codes, for which the linear dependence on γ is quantitatively important. The assumption of q being prime is necessary: for X supported uniformly on a proper subgroup of $\mathbb{Z}_q$ we have $H(X + X) = H(X)$. For X supported on infinite groups without a finite subgroup (the torsion-free case) and no conditioning, a sumset inequality for the absolute increase in (unnormalized) entropy was shown by Tao in [20].

We use our sumset inequality to analyze Arıkan's construction of polar codes and prove that for any q-ary source X, where q is any fixed prime, and any $\epsilon > 0$, polar codes allow *efficient* data compression of N i.i.d. copies of X into $(H(X) + \epsilon)N$ q-ary symbols, *as soon as N is polynomially large in $1/\epsilon$*. We can get capacity-achieving source codes with similar guarantees for composite alphabets, by factoring q into primes and combining different polar codes for each prime in factorization.

A consequence of our result for noisy channel coding is that for *all* discrete memoryless channels, there are explicit codes enabling reliable communication within $\epsilon > 0$ of the symmetric Shannon capacity for a block length and decoding complexity bounded by a polynomial in $1/\epsilon$. The result was previously shown for the special case of binary-input channels [7, 9], and this work extends the result to channels over any alphabet.

---

1   Introduction

In a remarkable work, Arıkan [2] introduced the technique of channel polarization, and used it to construct a family of binary linear codes called polar codes that achieve the symmetric Shannon capacity of binary-input discrete memoryless channels in the limit of large block lengths. Polar codes are based on an elegant recursive construction and analysis guided by information-theoretic intuition. Arıkan's work gave a construction of binary codes, and this was subsequently extended to general alphabets in [18]. In addition to being an approach to realize Shannon capacity that is radically different from prior ones, channel polarization turns out to be a powerful and versatile primitive applicable in many other important information-theoretic scenarios. For instance, variants of the polar coding approach give solutions to the lossless and lossy source coding problem [3, 13], capacity of wiretap channels [15], the Slepian-Wolf, Wyner-Ziv, and Gelfand-Pinsker problems [11], coding for broadcast channels [5], multiple access channels [19, 1], interference networks [21], etc. We recommend the well-written survey by Şaşoğlu [17] for a detailed introduction to polar codes.

The advantage of polar codes over previous capacity-achieving methods (such as Forney's concatenated codes that provably achieved capacity) was highlighted in a recent work [7] where *polynomial convergence to capacity* was shown in the *binary* case (this was also shown independently in [9]). Specifically, it was shown that polar codes enable approaching the symmetric capacity of binary-input memoryless channels within an additive gap of ϵ with block length, construction, and encoding/decoding complexity all bounded by a polynomially growing function of $1/\epsilon$. Polar codes are the first and currently only known construction which provably have this property, thus providing a formal complexity-theoretic sense in which they are the first constructive capacity-achieving codes.

The main objective of this paper is to extend this result to the non-binary case, and we manage to do this for *all* alphabets in this work. We stress that the best previously proven complexity bound for communicating at rates within ϵ of capacity of channels with non-binary inputs was *exponential* in $1/\epsilon$. Our work shows the polynomial solvability of the central computational challenge raised by Shannon's non-constructive coding theorems, in the full generality of *all discrete sources* (for compression/noiseless coding) and *all discrete memoryless channels* (for noisy coding).

The high level approach to prove the polynomially fast convergence to capacity is similar to what was done in [7], which is to replace the appeal to general martingale convergence theorems (which lead to ineffective bounds) with a more direct analysis of the convergence rate of a specific martingale of entropies.[1] However, the extension to the non-binary case is far from immediate, and we need to establish a quantitatively strong "entropy increase lemma" (see details in Section 4) over all prime alphabets. The corresponding inequality admits an easier proof in the binary case, but requires more work for general prime alphabets. For alphabets of size m where m is not a prime, we can construct a capacity-achieving code by combining together polar codes for each prime dividing m.

In the next section, we briefly sketch the high level structure of polar codes, and the crucial role played by a certain "entropy sumset inequality" in our effective analysis. Proving this entropic inequality is the main new component in this work, though additional technical work is needed to glue it together with several other ingredients to yield the overall coding result.

---

[1] The approach taken in [9] to analyze the speed of polarization for the binary was different, based on channel Bhattacharyya parameters instead of entropies. This approach does not seem as flexible as the entropic one to generalize to larger alphabets.

2 Overview of the Contribution

In order to illustrate our main contribution, which is an inequality on conditional entropies for inputs from prime alphabets, in a simple setting, we will focus on the source coding (lossless compression) model in this paper. The consequence of our results for channel coding, which is not immediate but follows in a standard manner from compression of sources with side information (see for instance [17, Sec 2.4]), is stated in Theorem 2.3.

Let $\mathbb{Z}_q = \{0, 1, \ldots, q-1\}$ denote the additive group of integers modulo q. Suppose X is a source (random variable) over $\mathbb{Z}_q$ (with q prime), with entropy $H(X)$ (throughout the paper, by entropy we will mean the entropy normalized by a $\lg q$ factor, so that $H(X) \in [0, 1]$). The source coding problem consists of compressing N i.i.d. copies $X_0, X_1, \ldots, X_{N-1}$ of X to $\approx H(X)N$ (say $(H(X)+\epsilon)N$) symbols from $\mathbb{Z}_q$. The approach based on channel polarization is to find an explicit permutation matrix $A \in \mathbb{Z}_q^{N \times N}$, such that if $(U_0, \ldots, U_{N-1})^t = A(X_0, \ldots, X_{N-1})^t$, then in the limit of $N \to \infty$, for most indices i, the conditional entropy $H(U_i|U_0, \ldots, U_{i-1})$ is either ≈ 0 or ≈ 1. Note that the conditional entropies at the source $H(X_i|X_0, \ldots, X_{i-1})$ are all equal to $H(X)$ (as the samples are i.i.d.). However, after the linear transformation by A, the conditional entropies get *polarized* to the boundaries 0 and 1. By the chain rule and conservation of entropy, the fraction of i for which $H(U_i|U_0, \ldots, U_{i-1}) \approx 1$ (resp. ≈ 0) must be $\approx H(X)$ (resp. $\approx 1 - H(X)$).

The polarization phenomenon is used to compress the X_i's as follows: The encoder only outputs U_i for indices $i \in B$ where $B = \{i \mid H(U_i|U_0, \ldots, U_{i-1}) > \zeta\}$ for some tiny $\zeta = \zeta(N) \to 0$. The decoder (decompression algorithm), called a *successive cancellation decoder*, estimates the U_i's in the order $i = 0, 1, \ldots, N-1$. For indices $i \in B$ that are output at the encoder, this is trivial, and for other positions, the decoder computes the maximum likelihood estimate $\hat{u}_i$ of U_i, assuming $U_0, \ldots, U_{i-1}$ equal $\hat{u}_0, \ldots, \hat{u}_{i-1}$, respectively. Finally, the decoder estimates the inputs at the source by applying the inverse transformation A^{-1} to $(\hat{u}_0, \ldots, \hat{u}_{N-1})^t$.

The probability of incorrect decompression (over the randomness of the source) is upper bounded, via a union bound over indices outside B, by $\sum_{i \notin B} H(U_i|U_0, \ldots, U_{i-1}) \leq \zeta N$. Thus, if $\zeta \ll 1/N$, we have a reliable lossless compression scheme. Thus, in order to achieve compression rate $H(X) + \epsilon$, we need a polarizing map A for which $H(U_i|U_0, \ldots, U_{i-1}) \ll 1/N$ for at least $1 - H(X) - \epsilon$ fraction of indices. This in particular means that $H(U_i|U_0, \ldots, U_{i-1}) \approx 0$ or ≈ 1 for all but a vanishing fraction of indices, which can be compactly expressed as $\mathbf{E}_i\big[H(U_i|U_0, \ldots, U_{i-1})\big(1 - H(U_i|U_0, \ldots, U_{i-1})\big)\big] \to 0$ as $n \to \infty$.

Such polarizing maps A are in fact implied by a source coding solution, and exist in abundance (a random invertible map works w.h.p.). The big novelty in Arıkan's work is an explicit recursive construction of polarizing maps, which further, due to their recursive structure, enable efficient maximum likelihood estimation of U_i given knowledge of $U_0, \ldots, U_{i-1}$.

Arıkan's construction is based on recursive application of the basic 2×2 invertible map $K = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.[2] While Arıkan's original analysis was for the binary case, the same construction based on the matrix K also works for any prime alphabet [18]. Let A_n denote the matrix of the polarizing map for $N = 2^n$. In the base case $n = 1$, the outputs are $U_0 = X_0 + X_1$ and $U_1 = X_1$. If $X_0, X_1 \sim X$ are i.i.d., the entropy $H(U_0) = H(X_0 + X_1) > H(X)$ (unless $H(X) \in \{0, 1\}$), and by the chain rule $H(U_1|U_0) < H(X)$, thereby creating a small separation in the entropies. Recursively, if $(V_0, \ldots, V_{2^{n-1}-1})$ and $(T_0, \ldots, T_{2^{n-1}-1})$ are the outputs of

---

[2] Subsequent work established that polarization is a common phenomenon that holds for most choices of the "base" matrix instead of just K [12].

A_{n-1} on the first half and second half of $(X_0, \ldots, X_{2^n-1})$, respectively, then the output $(U_0, \ldots, U_{2^n-1})$ satisfies $U_{2i} = V_i + T_i$ and $U_{2i+1} = T_i$. If H_n denotes the random variable equal to $H(U_i | U_0, \ldots, U_{i-1})$ for a random $i \in \{0, 1, \ldots, 2^n - 1\}$, then the sequence $\{H_n\}$ forms a bounded martingale. The polarization property, namely that $H_n \to \text{Bernoulli}(H(X))$ in the limit of $n \to \infty$, can be shown by appealing to the martingale convergence theorem. However, in order to obtain a finite upper bound on $n(\epsilon)$, the value of n needed for $\mathbf{E}[H_n(1 - H_n)] \leq \epsilon$ (so that most conditional entropies to polarize to $< \epsilon$ or $> 1 - \epsilon$), we need a more quantitative analysis. This was done for the binary case in [7], by quantifying the increase in entropy $H(V_i + T_i | V_0, \ldots, V_{i-1}, T_0, \ldots, T_{i-1}) - H(V_i | V_0, \ldots, V_{i-1})$ at each stage, and proving that the entropies diverge apart at a sufficient pace for H_n to polarize to $0/1$ exponentially fast in n, namely $\mathbf{E}[H_n(1 - H_n)] \leq \rho^n$ for some absolute constant $\rho < 1$.

The main technical challenge in this work is to show an analogous entropy increase lemma for all prime alphabets. The primality assumption is necessary, because a random variable X uniformly supported on a proper subgroup has $H(X) \notin \{0, 1\}$ and yet $H(X + X) = H(X)$. Formally, we prove:

▶ **Theorem 2.1.** *Let (X_i, Y_i), $i = 1, 2$ be i.i.d. copies of a correlated random variable (X, Y) with X supported on $\mathbb{Z}_q$ for a prime q. Then for some $\alpha(q) > 0$,*

$$H(X_1 + X_2 | Y_1, Y_2) - H(X|Y) \geq \alpha(q) \cdot H(X|Y)(1 - H(X|Y)). \tag{1}$$

The *linear* dependence of the entropy increase on the quantity $H(X|Y)(1 - H(X|Y))$ is crucial to establish a speed of polarization adequate for polynomial convergence to capacity. A polynomial dependence is implicit in [16], but obtaining a linear dependence requires lot more care. For the case $q = 2$, Theorem 2.1 is relatively easy to establish, as it is known that the extremal case (with minimal increase) occurs when $H(X|Y = y) = H(X|Y)$ for all y in the support of Y [17, Lem 2.2]. This is based on the so-called "Mrs. Gerber's Lemma" for binary-input channels [23, 22], the analog of which is not known for the non-binary case [10]. This allows us to reduce the binary version of (1) to an inequality about simple Bernoulli random variables with no conditioning, and the inequality then follows, as the sum of two p-biased coins is $2p(1 - p)$-biased and has higher entropy (unless $p \in \{0, \frac{1}{2}, 1\}$). In the q-ary case, no such simple characterization of the extremal cases is known or seems likely [17, Sec 4.1]. Nevertheless, we prove the inequality in the q-ary setting by first proving two inequalities for unconditioned random variables, and then handling the conditioning explicitly based on several cases.

More specifically, the proof technique for Theorem 2.1 involves using an *averaging* argument to write the left-hand side of (1) as the expectation, over $y, z \sim Y$, of $\Delta_{y,z} = H(X_y + X_z) - \frac{H(X_y) + H(X_z)}{2}$, the entropy increase in the sum of random variables X_y and X_z with respect to their average entropy (this increase is called the *Ruzsa distance* between the random variables X_y and X_z, see [20]). We then rely on inequalities for *unconditioned* random variables to obtain a lower bound for this entropy increase. In general, once needs the entropy increase to be at least $c \cdot \min\{H(X_y)(1 - H(X_y)), H(X_z)(1 - H(X_z))\}$, but for some cases, we actually need such an entropy increase with respect to a larger *weighted* average. Hence, we prove the stronger inequality given by Theorem 4.10, which shows such an increase with respect to $\frac{2H(X_y) + H(X_z)}{3}$ for $H(X_y) \geq H(X_z)$[3]. Moreover, for some cases

---

[3] While the weaker inequality $H(A + B) \geq \frac{H(A) + H(B)}{2} + c \cdot \min\{H(A)(1 - H(A)), H(B)(1 - H(B))\}$ seems to be insufficient for our approach, it should be noted that the stronger inequality $H(A + B) \geq \max\{H(A), H(B)\} + c \cdot \min\{H(A)(1 - H(A)), H(B)(1 - H(B))\}$ is generally not true. Thus, Theorem 4.10

of the proof, it suffices to bound $\Delta_{y,z}$ from below by $\frac{|H(X_y)-H(X_z)|}{2}$, which is provided by Lemma 4.9, another inequality for unconditional random variables.

We note a version of Theorem 2.1 (in fact with tight bounds) for the case of unconditioned random variables X taking values in a torsion-free group was established by Tao in his work on entropic analogs of fundamental sumset inequalities in additive combinatorics [20] (results of similar flavor for integer-valued random variables were shown in [8]). Theorem 2.1 is a result in the same spirit for groups with torsion (and which further handles conditional entropy). While we do not focus on optimizing the dependence of $\alpha(q)$ on q, pinning down the optimal dependence, especially for the case without any conditioning, seems like a natural question; see Remark 4.2 for further elaboration.

Given the entropy sumset inequality for conditional random variables, we are able to track the decay of $\sqrt{H_n(1 - H_n)}$ and use Theorem 2.1 to show that for $N = \text{poly}(1/\epsilon)$, at most $H(X) + \epsilon$ of the conditional entropies $H(U_i|U_0, \ldots, U_{i-1})$ exceed ϵ. However, to construct a good source code, we need $H(X) + \epsilon$ fraction of the conditional entropies to be $\ll 1/N$. This is achieved by augmenting a "fine" polarization stage that is analyzed using an appropriate Bhattacharyya parameter. The details of this step are similar to the binary case and appear in the full version of this paper [6].

The efficient construction of the linear source code (i.e., figuring out which entropies polarize very close to 0 so that those symbols can be dropped), and the efficient implementation of the successive cancellation decoder are similar to the binary case [7] and omitted here. Upon combining these ingredients, we get the following result on lossless compression with complexity scaling polynomially in the gap to capacity:

▶ **Theorem 2.2.** *Let X be a q-ary source for q prime with side information Y (which means (X, Y) is a correlated random variable). Let $0 < \epsilon < \frac{1}{2}$. Then there exists $N \leq (1/\epsilon)^{c(q)}$ for a constant $c(q) < \infty$ depending only on q and an explicit (constructible in $\text{poly}(N)$ time) matrix $L \in \{0,1\}^{(H(X|Y)+\epsilon)N \times N}$ such that $\vec{X} = (X_0, X_1, \ldots, X_{N-1})^t$, formed by taking N i.i.d. copies $(X_0, Y_0), (X_1, Y_1), \ldots, (X_{N-1}, Y_{N-1})$ of (X, Y), can, with high probability, be recovered from $L \cdot \vec{X}$ and $\vec{Y} = (Y_0, Y_1, \ldots, Y_{N-1})^t$ in $\text{poly}(N)$ time.*

Moreover, can obtain Theorem 2.2 for *arbitrary* (not necessarily prime) q with the modification that the map $\mathbb{Z}_q^N \to \mathbb{Z}_q^{H(X|Y)+\epsilon)N}$ is no longer linear. This is obtained by factoring q into primes and combining polar codes over prime alphabets for each prime in the factorization.

Channel coding. Using known methods to construct channel codes from polar source codes for compressing sources with side information (see, for instance, [17, Sec 2.4] for a nice discussion of this aspect), we obtain the following result for channel coding, enabling reliable communication at rates within an additive gap ϵ to the *symmetric capacity* for discrete memoryless channels over any fixed alphabet, with overall complexity bounded polynomially in $1/\epsilon$. Recall that a discrete memoryless channel (DMC) W has a finite input alphabet $\mathcal{X}$ and a finite output alphabet $\mathcal{Y}$ with transition probabilities $p(y|x)$ for receiving $y \in \mathcal{Y}$ when $x \in \mathcal{X}$ is transmitted on the channel. The entropy $H(W)$ of the channel is defined to be $H(X|Y)$ where X is uniform in $\mathcal{X}$ and Y is the output of W on input X; the symmetric capacity of W, which is the largest rate at which one can reliably communicate on W when the inputs have a uniform prior, equals $1 - H(W)$. Moreover, it should be noted that if W

---

provides the right middle ground. A limitation of similar spirit for the entropy increase when summing two integer-valued random variables was pointed out in [8].

is a *symmetric* DMC, then the symmetric capacity of W is precisely the Shannon capacity of W.

▶ **Theorem 2.3.** *Let $q \geq 2$, and let W be any discrete memoryless channel capacity with input alphabet $\mathbb{Z}_q$. Then, there exists an $N \leq (1/\epsilon)^{c(q)}$ for a constant $c(q) < \infty$ depending only on q, as well as a deterministic $\mathrm{poly}(N)$ construction of a q-ary code of block length N and rate at least $1 - H(W) - \epsilon$, along with a deterministic $N \cdot \mathrm{poly}(\log N)$ time decoding algorithm for the code such that the block error probability for communication over W is at most $2^{-N^{0.49}}$. Moreover, when q is prime, the constructed codes are linear.*

The structure of our paper will be as follows. Section 3 will introduce notation, describe the construction of polar codes, and define channels as a tool for analyzing entropy increases for a pair of correlated random variables. Section 4 will then present a sketch of our main theorem and describe the "rough" and "fine" polarization results that follow from the main theorem and allow us to achieve Theorem 2.2. Section 5 shows how polar codes for prime alphabets may be combined to obtain a capacity-achieving construction over all alphabets, thereby achieving a variant of Theorem 2.2 over non-prime alphabets, as well its channel-coding counterpart, Theorem 2.3.

3   Construction of Polar Codes

Notation. We begin by setting some of the notation to be used in the rest of the paper. We will let lg denote the base 2 logarithm, while ln will denote the natural logarithm.

For our purposes, unless otherwise stated, q will be a prime integer, and we identify $\mathbb{Z}_q = \{0, 1, 2, \ldots, q-1\}$ with the additive group of integers modulo q. We will generally view $\mathbb{Z}_q$ as a q-ary alphabet.

Given a q-ary random variable X taking values in $\mathbb{Z}_q$, we let $H(X)$ denote the *normalized entropy* of X:

$$H(X) = -\frac{1}{\lg q} \sum_{a \in \mathbb{Z}_q} \Pr[X = a] \lg(\Pr[X = a]).$$

In a slight abuse of notation, we also define $H(p)$ for a probability distribution p. If p is a probability distribution over $\mathbb{Z}_q$, then we shall let $H(p) = H(X)$, where X is a random variable sampled according to p. Also, for nonnegative constants $c_0, c_1, \ldots, c_{q-1}$ summing to 1, we will often write $H(c_0, \ldots, c_{q-1})$ as the entropy of the probability distribution on $\mathbb{Z}_q$ that samples i with probability c_i. Moreover, for a probability distribution p over $\mathbb{Z}_q$, we let $p^{(+j)}$ denote the jth *cyclic shift* of p, namely, the probability distribution $p^{(+j)}$ over $\mathbb{Z}_q$ that satisfies

$$p^{(+j)}(m) = p(m - j)$$

for all $m \in \mathbb{Z}_q$, where $m - j$ is taken modulo q. Note that $H(p) = H(p^{(+j)})$ for all $j \in \mathbb{Z}_q$.

Also, let $\| \cdot \|_1$ denote the ℓ_1 norm on $\mathbb{R}^q$. In particular, for two probability distributions p and p', the quantity $\|p - p'\|_1$ will correspond to twice the total variational distance between p and p'.

Finally, given a row vector (tuple) $\vec{v}$, we let $\vec{v}^t$ denote a column vector given by the transpose of $\vec{v}$.

3.1   Encoding Map

Let us formally define the polarization map that we will use to compress a source X. Given $n \geq 1$, we define an invertible linear transformation $G : \mathbb{Z}_q^{2^n} \to \mathbb{Z}_q^{2^n}$ by $G = G_n$, where $G_t : \mathbb{Z}_q^{2^t} \to \mathbb{Z}_q^{2^t}$, $0 \leq t \leq n$ is a sequence of invertible linear transformations defined as follows: G_0 is the identity map on $\mathbb{Z}_q$, and for any $0 \leq k < n$ and $\vec{X} = (X_0, X_1, \ldots, X_{2^{k+1}-1})^t$, we recursively define $G_{k+1}\vec{X}$ as

$$G_{k+1}\vec{X} = \pi_{k+1}(G_k(X_0, \ldots, X_{2^k-1}) + G_k(X_{2^k}, \ldots, X_{2^{k+1}-1}), G_k(X_{2^k}, \ldots, X_{2^{k+1}-1})),$$

where $\pi_{k+1} : \mathbb{Z}_q^{2^{k+1}} \to \mathbb{Z}_q^{2^{k+1}}$ is a permutation such that $\pi_n(v)_j = v_i$ for $j = 2i$, and $\pi_n(v)_j = v_{i+2^k}$ for $j = 2i + 1$.

G also has an explicit matrix form, namely, $G = B_n K^{\otimes n}$, where $K = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix} \right)$, $\otimes$ is the Kronecker product, and B_n is the $2^n \times 2^n$ bit-reversal permutation matrix for n-bit strings (see [3]).

In our set-up, we have a q-ary source X, and we let $\vec{X} = (X_0, X_1, \ldots, X_{2^n-1})^t$ be a collection of $N = 2^n$ i.i.d. samples from X. Moreover, we encode $\vec{X}$ as $\vec{U} = (U_0, U_1, \ldots, U_{2^n-1})^t$, given by $\vec{U} = G \cdot \vec{X}$. Note that G only has $0, 1$ entries, so each U_i is the sum (modulo q) of some subset of the X_i's.

3.2   Channels

For purposes of our analysis, we define a *channel* $W = (A; B)$ to be a pair of correlated random variables A, B; moreover, we define the *channel entropy* of W to be $H(W) = H(A|B)$, i.e., the entropy of A conditioned on B.[4]

Given a channel W, we can define two channel transformations $-$ and $+$ as follows. Suppose we take two i.i.d. copies $(A_0; B_0)$ and $(A_1; B_1)$ of W. Then, W^- and W^+ are defined by

$$W^- = (A_0 + A_1; B_0, B_1)$$
$$W^+ = (A_1; A_0 + A_1, B_0, B_1).$$

By the chain rule for entropy, we see that

$$H(W^-) + H(W^+) = 2H(W). \tag{2}$$

In other words, splitting two copies of W into W^- and W^+ preserves the total channel entropy. These channels are easily seen to obey $H(W^+) \leq H(W) \leq H(W^-)$, and the key to our analysis will be quantifying the separation in the entropies of the two split channels.

The aforementioned channel transformations will help us abstract each step of the recursive polarization that occurs in the definition of G. Let $W = (X; Y)$, where X is a source taking values in $\mathbb{Z}_q$, and Y can be viewed as side information. Then, $H(W) = H(X|Y)$. One special case occurs when $Y = 0$, which corresponds to an absence of side information.

Note that if start with W, then after n successive applications of either $W \mapsto W^-$ or $W \mapsto W^+$, we can obtain one of $N = 2^n$ possible channels in $\{W^s : s \in \{+, -\}^n\}$. (Here, if

---

[4] It should be noted W can also be interpreted as a communication channel that takes in an input A and outputs B according to some conditional probability distribution. This is quite natural in the noisy channel coding setting in which one wishes to use a polar code for encoding data in order to achieve the channel capacity of a symmetric discrete memoryless channel. However, since we focus on the problem of source coding (data compression) rather than noisy channel coding in this paper, we will simply view W as a pair of correlated random variables.

$s = s_0 s_1 \cdots s_{n-1}$, with each $s_i \in \{+, -\}$, then W^s denotes $(\cdots ((W^{s_0})^{s_1}) \cdots)^{s_{n-2}})^{s_{n-1}})$. By successive applications of (2), we know that

$$\sum_{s \in \{+,-\}^n} W^s = 2^n H(W) = 2^n H(X|Y).$$

Moreover, it can be verified (see [17]) that if $0 \leq i < 2^n$ has binary representation $\overline{b_{n-1} b_{n-2} \cdots b_0}$ (with b_0 being the least significant bit of i), then

$$H(U_i | U_0, U_1, \ldots, U_{i-1}, Y_0, Y_1, \ldots, Y_{N-1}) = H(W^{s_{n-1} s_{n-2} \cdots s_0}),$$

where $s_j = -$ if $b_j = 0$, and $s_j = +$ if $b_j = 1$. As shorthand notation, we will define the channel

$$W_n^{(i)} = W^{s_{n-1} s_{n-2} \cdots s_0},$$

where $s_0, s_1, \ldots, s_{n-1}$ are as above. [18] shows that all but a vanishing fraction of the N channels W^s will be have channel entropy close to 0 or 1:

▶ **Theorem 3.1.** *For any $\delta > 0$, we have that*

$$\lim_{n \to \infty} \frac{|\{s \in \{+, -\}^n : H(W^s) \in (\delta, 1 - \delta)\}|}{2^n} = 0.$$

Hence, one can then argue that as n grows, the fraction of channels with channel entropy close to 1 approaches $H(X|Y)$. In particular, for any $\delta > 0$, if we let

$$\mathsf{High}_{n,\delta} = \{i : H(U_i | U_0, U_1, \ldots, U_{i-1}, Y_0, Y_1, \ldots, Y_{N-1}) > \delta\}, \tag{3}$$

then

$$\frac{|\mathsf{High}_{n,\delta}|}{2^n} \to H(X|Y),$$

as $n \to \infty$. Thus, it can be shown that for any fixed $\epsilon > 0$ and small $\delta > 0$, there exists suitably large n such that $\{U_i\}_{i \in \mathsf{High}_{n,\delta}}$ gives a source coding of $\vec{X} = (X_0, X_1, \ldots, X_{N-1})$ (with side information $\vec{Y} = (Y_0, Y_1, \ldots, Y_{N-1})$ with rate $\leq H(X|Y) + \epsilon$.

Our goal is to show that $N = 2^n$ can be taken to be just polynomial in $1/\epsilon$ in order to obtain a rate $\leq H(X|Y) + \epsilon$.

3.3 Bhattacharyya Parameter

In order to analyze a channel $W = (X; Y)$, where X takes values in $\mathbb{Z}_q$, we will define the *q-ary source Bhattacharyya parameter* $Z_{\max}(W)$ of the channel W as

$$Z_{\max}(W) = \max_{d \neq 0} Z_d(W),$$

where

$$Z_d(W) = \sum_{x \in \mathbb{Z}_q} \sum_{y \in \mathrm{Supp}(Y)} \sqrt{p(x, y) p(x + d, y)}.$$

Here, $p(x, y)$ is the probability that $X = x$ and $Y = y$ under the joint probability distribution (X, Y).

Now, the *maximum likelihood decoder* attempts to decode x given y by choosing the most likely symbol $\hat{x}$:

$$\hat{x} = \arg\max_{x' \in \mathbb{Z}_q} \Pr[X = x' | Y = y].$$

Let $P_e(W)$ be the probability of an error under maximum likelihood decoding, i.e., the probability that $\hat{x} \neq x$ (or the defining arg max for $\hat{x}$ is not unique) for random $(x, y) \sim (X, Y)$. It is known (see Proposition 4.7 in [17]) that $Z_{\max}(W)$ provides an upper bound on $P_e(W)$:

▶ **Lemma 3.2.** *If W is a channel with q-ary input, then the error probability of the maximum-likelihood decoder for a single channel use satisfies $P_e(W) \leq (q - 1)Z_{\max}(W)$.*

Next, the following proposition shows how the $Z_{\max}$ operator behaves on the polarized channels W^- and W^+. For a proof, see Proposition 4.16 in [17].

▶ **Lemma 3.3.** $Z_{\max}(W^+) \leq Z_{\max}(W)^2$, *and* $Z_{\max}(W^-) \leq q^3 Z_{\max}(W)$.

Finally, the following lemma shows that $Z_{\max}(W)$ is small whenever $H(W)$ is small.

▶ **Lemma 3.4.** $Z_{\max}(W)^2 \leq (q - 1)^2 H(W)$.

The proof follows from Proposition 4.8 of [17].

4   Quantification of Polarization

Our goal is to show "rough" polarization of the channel. More precisely, we wish to show that for some $m = O(\lg(1/\epsilon))$ and constant K, we have $\Pr_i[Z(W_m^{(i)}) \leq 2^{-Km}] \geq 1 - H(W) - \epsilon$. The above polarization result will then be used to show the stronger notion of "fine" polarization, which will establish the polynomial gap to capacity.

The main ingredient in showing polarization is the following theorem, which quantifies the splitting that occurs with each polarizing step.

▶ **Theorem 4.1.** *For any channel $W = (A; B)$, where A takes values in $\mathbb{Z}_q$, we have*

$$H(W^-) \geq H(W) + \alpha(q) \cdot H(W)(1 - H(W)),$$

where $\alpha(q)$ is a constant depending only on q.

Theorem 4.1 follows as a direct consequence of Theorem 2.1, whose proof we sketch in Section 4.2. Section 4.1 focuses on proving Theorem 4.10 (tackling the unconditioned case), which will be used in the proof of Theorem 2.1.

4.1   Unconditional Entropy Gain

In this section, we sketch some results (Lemma 4.9 and Theorem 4.10) that provide a lower bound on the normalized (unconditional) entropy $H(A + B)$ of a sum of random variables A, B in terms of the individual entropies.

First, we present some lemmas, whose proofs appear in the full version of the paper [6] (with the exception of Lemma 4.3, which appears, with proof, as Lemma 4.5 in [17]).

Lemmas 4.2 and 4.3 are used to show that the entropy of a linear combination of cyclic shifts of a distribution exhibits an increase over the entropy of the original distribution.

▶ **Lemma 4.2.** *Let p be a distribution over $\mathbb{Z}_q$. Then, if $\lambda_0, \lambda_1, \ldots, \lambda_{q-1}$ are nonnegative numbers adding up to 1, we have*

$$H(\lambda_0 p^{(+0)} + \lambda_1 p^{(+1)} + \cdots + \lambda_{q-1} p^{(+(q-1))}) \geq H(p) + \frac{1}{2 \lg q} \cdot \frac{\lambda_i \lambda_j}{\lambda_i + \lambda_j} \|p^{(+i)} - p^{(+j)}\|_1^2,$$

for any $i \neq j$ such that $\lambda_i + \lambda_j > 0$.

▶ **Lemma 4.3.** *Let p be a distribution over $\mathbb{Z}_q$, where q is prime. Then,*

$$\|p^{(+i)} - p^{(+j)}\|_1 \geq \frac{(1 - H(p)) \lg q}{2q^2(q-1) \lg e}.$$

Lemmas 4.4 and 4.5 show matching lower and upper bounds (up to a constant factor) on the entropy of a low-entropy random variable with high mass on one symbol.

▶ **Lemma 4.4.** *Suppose $0 < \epsilon < 1$. If p is a distribution on $\mathbb{Z}_q$ with mass $1 - \epsilon$ on one symbol, then*

$$H(p) \geq \frac{\epsilon \lg(1/\epsilon)}{\lg q}.$$

▶ **Lemma 4.5.** *Suppose $0 < \epsilon \leq \min\{\epsilon_1, \epsilon_2\}$, where $\epsilon_1 = \frac{1}{500}$ and $\epsilon_2 = \frac{1}{(q-1)^4}$. If p is a distribution on $\mathbb{Z}_q$ with mass $1 - \epsilon$ on one symbol, then*

$$H(p) \leq \frac{17 \epsilon \lg(1/\epsilon)}{12 \lg q}.$$

Lemma 4.6 provides a lower bound on $H(X + Y)$ for low-entropy random variables X, Y that each have most of their mass on a single symbol.

▶ **Lemma 4.6.** *Let X, Y be random variables taking values in $\mathbb{Z}_q$ such that $H(X) \geq H(Y)$, and assume $0 < \epsilon, \epsilon' \leq \min\{\epsilon_1, \epsilon_2\}$, where $\epsilon_1 = \frac{1}{500}$ and $\epsilon_2 = \frac{1}{(q-1)^4}$. Suppose that X has mass $1 - \epsilon$ on one symbol, while Y has mass $1 - \epsilon'$ on a symbol. Then,*

$$H(X + Y) - \frac{2H(X) + H(Y)}{3} \geq \frac{1}{51} \cdot H(Y)(1 - H(Y)). \tag{4}$$

Lemma 4.7 provides a lower bound as well as upper bound on high-entropy random variables that are close to uniform distributions.

▶ **Lemma 4.7.** *Suppose p is a distribution on $\mathbb{Z}_q$ such that for each $0 \leq i \leq q - 1$, we have $p(i) = \frac{1}{q} + \delta_i$ with $\max_{0 \leq i < q} |\delta_i| = \delta$. Then,*

$$1 - \frac{q^2}{\ln q} \delta^2 \leq H(p) \leq 1 - \frac{q^2(q \ln q - (q-1))}{(q-1)^3 \ln q} \delta^2.$$

Lemma 4.8 shows provides a lower bound on $H(X + Y)$ for high-entropy random variables X and Y that are both close to uniform distributions.

▶ **Lemma 4.8.** *Let X and Y be random variables taking values in $\mathbb{Z}_q$ such that $H(X) \geq H(Y)$. Also, assume $0 < \delta, \delta' \leq \frac{1}{2q^2}$. Suppose $\Pr[X = i] = \frac{1}{q} + \delta_i$ and $\Pr[Y = i] = \frac{1}{q} + \delta'_i$ for $0 \leq i \leq q - 1$, such that $\max_{0 \leq i < q} |\delta_i| = \delta$ and $\max_{0 \leq i < q} |\delta'_i| = \delta'$. Then,*

$$H(X + Y) - H(X) \geq \frac{\ln q}{16q^2} \cdot H(X)(1 - H(X)). \tag{5}$$

Now, we present some results showing a lower bound on $H(A + B)$ in terms of $H(A)$ and $H(B)$.

▶ **Lemma 4.9.** *Let A and B be random variables taking values over $\mathbb{Z}_q$. Then,*

$$H(A + B) \geq \max\{H(A), H(B)\}.$$

Proof. Without loss of generality, assume $H(A) \geq H(B)$. Let p be the underlying probability distribution for A. Let $\lambda_i = \Pr[B = i]$. Then, the underlying probability distribution of $A + B$ is $\lambda_0 p^{(+0)} + \lambda_1 p^{(+1)} + \cdots + \lambda_{q-1} p^{(+(q-1))}$. The desired result then follows directly from Lemma 4.2. ◀

The next theorem provides a different lower bound for $H(A + B)$.

▶ **Theorem 4.10.** *Let A and B be random variables taking values over $\mathbb{Z}_q$ such that $H(A) \geq H(B)$. Then,*

$$H(A + B) \geq \frac{2H(A) + H(B)}{3} + c \cdot \min\{H(A)(1 - H(A)), H(B)(1 - H(B))\}$$

for $c = \frac{\gamma_0^3 \lg q}{48q^5(q-1)^3 \lg(6/\gamma_0) \lg^2 e}$, where $\gamma_0 = \frac{1}{500(q-1)^4 \lg q}$.

The proof of this theorem appears in the full version of our paper [6], but we provide a brief overview of the proof below.

Overview of proof. The proof of Theorem 4.10 splits into various cases depending on where $H(A)$ and $H(B)$ lie. Note that some of these cases overlap. The overall idea is as follows. If $H(A)$ and $H(B)$ are both bounded away from 0 and 1 (Case 2), then the desired inequality follows from the concavity of the entropy function, using Lemmas 4.2 and 4.3 (note that this uses primality of q). Another setting in which the inequality can be readily proven is when $H(A) - H(B)$ is bounded away from 0 (which we deal with in Cases 4 and 5).

Thus, the remaining cases occur when $H(A)$ and $H(B)$ are either both small (Case 1) or both large (Case 3). In the former case, one can show that A must have most of its weight on a particular symbol, and similarly for B (note that this is why we must choose $\gamma_0 \ll \frac{1}{\log q}$; otherwise, A could be, for instance, supported uniformly on a set of size 2). Then, one can use the fact that a q-ary random variable having weight $1 - \epsilon$ has entropy $\Theta(\epsilon \log(1/\epsilon))$ (Lemmas 4.4 and 4.5) in order to prove the desired inequality (using Lemma 4.6).

For the latter case, we simply show that each of the q symbols of A must have weight close to $1/q$, and similarly for B. Then, we use the fact that such a random variable whose maximum deviation from $1/q$ is δ has entropy $1 - \Theta(\delta^2)$ (Lemma 4.7) in order to prove the desired result (using Lemma 4.8). ◀

4.2   Conditional Entropy Gain

Theorem 4.1 now follows as a simple consequence of our main theorem, which we restate below. The proof appears in the full version of our paper [6], but we provide an overview of the proof below.

▶ **Theorem 2.1.** *Let (X_i, Y_i), $i = 1, 2$ be i.i.d. copies of a correlated random variable (X, Y) with X supported on $\mathbb{Z}_q$ for a prime q. Then for some $\alpha(q) > 0$,*

$$H(X_1 + X_2 | Y_1, Y_2) - H(X|Y) \geq \alpha(q) \cdot H(X|Y)(1 - H(X|Y)). \tag{1}$$

▶ Remark. We have not attempted to optimize the dependence of $\alpha(q)$ on q, and our proof gets $\alpha(q) \geq \frac{1}{q^{O(1)}}$. It is easy to see that $\alpha(q) \leq O(1/\log q)$ even without conditioning (i.e., when $Y = 0$). Understanding what is the true behavior of $\alpha(q)$ seems like an interesting and basic question about sums of random variables. For random variables X taking values from a torsion-free group G and with sufficiently large $H_2(X)$, it is known that $H_2(X_1 + X_2) - H_2(X) \geq \frac{1}{2} - o(1)$ and that this is best possible [20], where $H_2(\cdot)$ denotes the *unnormalized* entropy (in bits). When G is the group of integers, a lower bound $H_2(X_1 + X_2) - H_2(X) \geq g(H_2(X))$ for an increasing function $g(\cdot)$ was shown for all $\mathbb{Z}$-valued random variables X [8]. For groups G with torsion, we cannot hope for any entropy increase unless G is finite and isomorphic to $\mathbb{Z}_q$ for q prime (as G cannot have non-trivial finite subgroups), and we cannot hope for an absolute entropy increase even for $\mathbb{Z}_q$. So determining the asymptotics of $\alpha(q)$ as a function of q is the analog of the question studied in [20] for finite groups.

Overview of proof. Let X_y denote $X|Y = y$. Then, we use an averaging argument: We reduce the desired inequality to providing a lower bound for $\Delta_{y,z} = H(X_y + X_z) - \frac{H(X_y) + H(X_z)}{2}$, whose expectation over $y, z \sim Y$ is the left-hand side of (1). Then, one splits into three cases for small, large, and medium values of $H(X|Y)$.

Thus, we reduce the problem to aruguing about unconditional entropies. As a first step, one would expect to prove $\Delta_{y,z} \geq \min\{H(X_y)(1 - H(X_y)), H(X_z)(1 - H(X_z))\}$ and use this in the proof of the conditional inequality. However, this inequality turns out to be too weak to deal with the case in which $H(X|Y)$ is tiny (case 2). This is the reason we require Theorem 4.10, which provides an increase for $H(X_y + X_z)$ over a higher *weighted* average instead of the simple average of $H(X_y)$ and $H(X_z)$. Additionally, we use the inequality $H(X_y + X_z) \geq \max\{H(X_y), H(X_z)\}$ to handle certain cases, and this is provided by Lemma 4.9.

In cases 1 and 3 (for $H(X|Y)$ in the middle and high regimes), the proof idea is that either (1) there is a significant mass of $(y, z) \sim Y \times Y$ for which $H(X_y)$ and $H(X_z)$ are separated, in which case one can use Lemma 4.9 to bound $\mathbf{E}[\Delta_{y,z}]$ from below, or (2) there is a significant mass of $y \sim Y$ for which $H(X_y)$ lies away from 0 and 1, in which case $H(X_y)(1 - H(X_y))$ can be bounded from below, enabling us to use Theorem 4.10. ◄

4.3 Rough Polarization

Now that we have established Theorem 4.1, we are ready to show rough polarization of the channels $W_n^{(i)}$, $0 \leq i < 2^n$, for large enough n. The precise theorem showing rough polarization is as follows.

▶ **Theorem 4.11.** *There is a constant $\Lambda < 1$ such that the following holds. For any $\Lambda < \rho < 1$, there exists a constant b_ρ such that for all channels W with q-ary input, all $\epsilon > 0$, and all $n > b_\rho \lg(1/\epsilon)$, there exists a set*

$$\mathcal{W}' \subseteq \{W_n^{(i)} : 0 \leq i \leq 2^n - 1\}$$

such that for all $M \in \mathcal{W}'$, we have $Z_{\max}(M) \leq 2\rho^n$ and $\Pr_i[W_n^{(i)} \in \mathcal{W}'] \geq 1 - H(W) - \epsilon$.

The proof of Theorem 4.11 follows from the following lemma:

▶ **Lemma 4.12.** *Let $T(W) = H(W)(1 - H(W))$ denote the* symmetric entropy *of a channel W. Then, there exists a constant $\Lambda < 1$ (possibly dependent on q) such that*

$$\frac{1}{2}\left(\sqrt{T\left(W_{n+1}^{(2j)}\right)} + \sqrt{T\left(W_{n+1}^{(2j+1)}\right)}\right) \leq \Lambda\sqrt{T\left(W_n^{(j)}\right)} \tag{6}$$

for any $0 \leq j < 2^n$.

The proofs of Theorem 4.11 and 4.12 follow from arguments similar to those found in the proofs of Proposition 5 and Lemma 8 in [7], except that we work with $Z_{\max}$. For completeness, he proofs appear in the full version of this paper [6].

4.4   Fine Polarization

Now, we describe the statement of "fine polarization." This is quantified by the following theorem.

▶ **Theorem 4.13.** *For any $0 < \delta < \frac{1}{2}$, there exists a constant c_δ that satisfies the following statement: For any q-ary input memoryless channel W and $0 < \epsilon < \frac{1}{2}$, if $n_0 > c_\delta \lg(1/\epsilon)$, then*

$$\Pr_i \left[Z_{\max}(W_{n_0}^{(i)}) \le 2^{-2^{\delta n_0}} \right] \ge 1 - H(W) - \epsilon.$$

The proof follows from arguments similar to those in [4, 7] and appears in the full version of the paper [6].

As a corollary, we obtain the following result on lossless compression with complexity scaling polynomially in the gap to capacity:

▶ **Theorem 2.2.** *Let X be a q-ary source for q prime with side information Y (which means (X, Y) is a correlated random variable). Let $0 < \epsilon < \frac{1}{2}$. Then there exists $N \le (1/\epsilon)^{c(q)}$ for a constant $c(q) < \infty$ depending only on q and an explicit (constructible in $\mathrm{poly}(N)$ time) matrix $L \in \{0, 1\}^{(H(X|Y)+\epsilon)N \times N}$ such that $\vec{X} = (X_0, X_1, \ldots, X_{N-1})^t$, formed by taking N i.i.d. copies $(X_0, Y_0), (X_1, Y_1), \ldots, (X_{N-1}, Y_{N-1})$ of (X, Y), can, with high probability, be recovered from $L \cdot \vec{X}$ and $\vec{Y} = (Y_0, Y_1, \ldots, Y_{N-1})^t$ in $\mathrm{poly}(N)$ time.*

Proof. Let $W = (X; Y)$, and fix $\delta = 0.499$. Also, let $N = 2^{n_0}$. Then, by Theorem 4.13, for any $n_0 > c_\delta \lg(1/\epsilon)$, we have that

$$\Pr_i \left[Z_{\max}(W_{n_0}^{(i)}) \le 2^{-2^{\delta n_0}} \right] \ge 1 - H(X) - \epsilon.$$

Moreover, let $N = 2^{n_0}$. Recall the notation in (3). Then, letting $\delta' = 2^{-2^{\delta n_0}}$, we have that $\Pr_i[i \in \mathsf{High}_{n_0, \delta'}] \le H(X|Y) + \epsilon$ and $Z(W_{n_0}^{(i)}) \ge \delta'$ for all $i \in \mathsf{High}_{n_0, \delta'}$. Thus, we can take L to be the linear map G_{n_0} projected onto the coordinates of $\mathsf{High}_{n_0, \delta'}$.

By Lemma 3.2 and the union bound, the probability that attempting to recover $\vec{X}$ from $L \cdot \vec{X}$ and $\vec{Y}$ results in an error is given by

$$\sum_{i \notin \mathsf{High}_{n_0, \delta'}} P_e(W_{n_0}^{(i)}) \le \sum_{i \notin \mathsf{High}_{n_0, \delta'}} (q - 1) Z_{\max}(W_{n_0}^{(i)}) \le (q - 1) N \delta' = (q - 1) 2^{n_0 - 2^{\delta n_0}}, \quad (7)$$

which is $\le 2^{-N^{0.49}}$ for $N \ge (1/\epsilon)^\mu$ for some positive constant μ (possibly depending on q). Hence, it suffices to take $c(q) = 1 + \max\{c_\delta, \mu\}$.

Finally, the fact that both the construction of L and the recovery of $\vec{X}$ from $L \cdot \vec{X}$ and $\vec{Y}$ can be done in $\mathrm{poly}(N)$ time follows in a similar fashion to the binary case (see the binning algorithm and the successive cancellation decoder in [7] for details). Also, the entries of L are all in $\{0, 1\}$ since L can be obtained by taking a submatrix of $B_n K^{\otimes n_0}$, where B_n is a permutation matrix, and $K = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix} \right)$ (see [3]). ◀

5   Extension to Arbitrary Alphabets

In the previous sections, we have shown polarization and polynomial gap to capacity for polar codes over *prime* alphabets. We now describe how to extend this to obtain channel polarization and the explicit construction of a polar code with polynomial gap to capacity over *arbitrary* alphabets.

The idea is to use the multi-level code construction technique sketched in [18] (and also recently in [14] for alphabets of size 2^m). We outline the procedure here. Suppose we have a channel $W = (X; Y)$, where $X \in \mathbb{Z}_q$ and $Y \in \mathcal{Y}$. Moreover, assume that $q = \prod_{i=1}^{s} q_i$ is the prime factorization of q.

Now, we can write $X = (U^{(1)}, U^{(2)}, \ldots, U^{(s)})$, where each $U^{(i)}$ is a random variable distributed over $[q_i]$. We also define the channels $W^{(1)}, W^{(2)}, \ldots, W^{(s)}$ by $W^{(j)} = (U^{(j)}; Y, U^{(1)}, U^{(2)}, \ldots, U^{(j-1)})$. Note that

$$
\begin{aligned}
H(W) = H(X|Y) &= H(U^{(1)}, U^{(2)}, \ldots, U^{(s)}|Y) \\
&= \sum_{j=1}^{s} H(U^{(j)}|Y, U^{(1)}, U^{(2)}, \ldots, U^{(j-1)}) \\
&= \sum_{j=1}^{s} H(W^{(j)}),
\end{aligned}
$$

which means that W splits into $W^{(1)}, W^{(2)}, \ldots, W^{(s)}$. Since each $W^{(j)}$ is a channel whose input is over a prime alphabet, one can polarize each $W^{(j)}$ separately using the procedure of the previous sections. More precisely, the encoding procedure is as follows. For N large enough (as specified by Theorem 2.2), we take N copies $(X_0; Y_0), (X_1; Y_1), \ldots, (X_{N-1}; Y_{N-1})$ of W, where $X_i = (U_i^{(1)}, U_i^{(2)}, \ldots U_i^{(s)})$. Then, sequentially for $j = 1, 2, \ldots, s$, we encode $U_0^{(j)}, U_1^{(j)}, \ldots, U_{N-1}^{(j)}$ using $\left\{ \left(Y_i, U_i^{(1)}, U_i^{(2)}, \ldots, U_i^{(j-1)} \right) \right\}_{i=0,1,\ldots,N-1}$ as side information (which can be done by the procedure in previous sections, since U_j is a source over a prime alphabet).

For decoding, one can simply use s stages of the successive cancellation decoder. In the jth stage, one uses the successive cancellation decoder for $W^{(j)}$ in order to decode $U_0^{(j)}, U_1^{(j)}, \ldots, U_{N-1}^{(j)}$, assuming that $\left\{ U_i^{(k)} \right\}_{k<j}$ has been recovered correctly from the previous stages of successive canellation decoding. Note that the error probability in decoding $X_0, X_1, \ldots, X_{N-1}$ can be obtained by taking a union bound over the error probabilities for each of the s stages of successive cancellation decoding. Since each individual error probability is exponentially small (see (7)), it follows that the overall error probability is also negligible.

As a consequence, we obtain Theorem 2.2 for non-prime q, with the additional modification that the map $\mathbb{Z}_q^N \to \mathbb{Z}_q^{H(X|Y)+\epsilon)N}$ is not linear. Moreover, using the translation from source coding to noisy channel coding (see [17, Sec 2.4]), we obtain the following result for channel coding.

▶ **Theorem 2.3.** *Let $q \geq 2$, and let W be any discrete memoryless channel capacity with input alphabet $\mathbb{Z}_q$. Then, there exists an $N \leq (1/\epsilon)^{c(q)}$ for a constant $c(q) < \infty$ depending only on q, as well as a deterministic $\mathrm{poly}(N)$ construction of a q-ary code of block length N and rate at least $1 - H(W) - \epsilon$, along with a deterministic $N \cdot \mathrm{poly}(\log N)$ time decoding algorithm for the code such that the block error probability for communication over W is at most $2^{-N^{0.49}}$. Moreover, when q is prime, the constructed codes are linear.*

▶ Remark. If q is prime, then the q-ary code of Theorem 2.3 is, in fact, linear.

───── **References** ───

1 Emmanuel Abbe and Emre Telatar. Polar codes for the m–user multiple access channel. *IEEE Transactions on Information Theory*, 58(8):5437–5448, 2012.

2 Erdal Arıkan. Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on Information Theory*, 55(7):3051–3073, 2009.

3 Erdal Arıkan. Source polarization. In *Proceedings of 2010 IEEE International Symposium on Information Theory*, pages 899–903, 2010.

4 Erdal Arıkan and Emre Telatar. On the rate of channel polarization. In *Proceedings of 2009 IEEE International Symposium on Information Theory*, pages 1493–1495, 2009.

5 Naveen Goela, Emmanuel Abbe, and Michael Gastpar. Polar codes for broadcast channels. In *Proceedings of the 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, July 7-12, 2013*, pages 1127–1131, 2013.

6 Venkatesan Guruswami and Ameya Velingker. An entropy sumset inequality and polynomially fast convergence to shannon capacity over all alphabets. *Electronic Colloquium on Computational Complexity (ECCC)*, 21:165, 2014.

7 Venkatesan Guruswami and Patrick Xia. Polar codes: Speed of polarization and polynomial gap to capacity. In *FOCS*, pages 310–319, 2013. Full version to appear in *IEEE Trans. on Info. Theory*, Jan. 2015.

8 Saeid Haghighatshoar, Emmanuel Abbe, and Ì. Emre Telatar. A new entropy power inequality for integer-valued random variables. *IEEE Transactions on Information Theory*, 60(7):3787–3796, 2014.

9 Seyed Hamed Hassani, Kasra Alishahi, and Rüdiger L. Urbanke. Finite-length scaling of polar codes. *CoRR*, abs/1304.4778, 2013.

10 Varun Jog and Venkat Anantharam. The entropy power inequality and mrs. gerber's lemma for groups of order 2^n. *IEEE Transactions on Information Theory*, 60(7):3773–3786, 2014.

11 Satish Babu Korada. Polar codes for Slepian-Wolf, Wyner-Ziv, and Gelfand-Pinsker. In *Proceedings of the 2010 IEEE Information Theory Workshop*, pages 1–5, 2010.

12 Satish Babu Korada, Eren Sasoglu, and Rüdiger L. Urbanke. Polar codes: Characterization of exponent, bounds, and constructions. *IEEE Transactions on Information Theory*, 56(12):6253–6264, 2010.

13 Satish Babu Korada and Rüdiger L. Urbanke. Polar codes are optimal for lossy source coding. *IEEE Transactions on Information Theory*, 56(4):1751–1768, 2010.

14 Jingbo Liu and Emmanuel Abbe. Polynomial complexity of polar codes for non-binary alphabets, key agreement and slepian-wolf coding. In *48th Annual Conference on Information Sciences and Systems, CISS 2014, Princeton, NJ, USA, March 19-21, 2014*, pages 1–6, 2014. Available online at `http://arxiv.org/abs/1405.0776`.

15 Hessam Mahdavifar and Alexander Vardy. Achieving the secrecy capacity of wiretap channels using polar codes. *IEEE Transactions on Information Theory*, 57(10):6428–6443, 2011.

16 Eren Sasoglu. An entropy inequality for q-ary random variables and its application to channel polarization. In *ISIT*, pages 1360–1363. IEEE, 2010.

17 Eren Sasoglu. Polarization and polar codes. *Foundations and Trends in Communications and Information Theory*, 8(4):259–381, 2012.

18 Eren Sasoglu, Emre Telatar, and Erdal Arıkan. Polarization for arbitrary discrete memoryless channels. *CoRR*, abs/0908.0302, 2009.

19 Eren Sasoglu, Emre Telatar, and Edmund M. Yeh. Polar codes for the two-user multiple-access channel. *IEEE Transactions on Information Theory*, 59(10):6583–6592, 2013.

20   Terence Tao. Sumset and inverse sumset theory for Shannon entropy. *Combinatorics, Probability and Computing*, 19(4):603–639, 2010.

21   Lele Wang and Eren Sasoglu. Polar coding for interference networks. *CoRR*, abs/1401.7293, 2014.

22   Hans S. Witsenhausen. Entropy inequalities for discrete channels. *IEEE Transactions on Information Theory*, 20(5):610–616, 1974.

23   Aaron D. Wyner and Jacob Ziv. A theorem on the entropy of certain binary sequences and applications-I. *IEEE Transactions on Information Theory*, 19(6):769–772, 1973.