

Log-Logarithmic Time Pruned Polar Coding

Hsin-Po Wang^{1b} and Iwan M. Duursma^{2b}

Abstract—A pruned variant of polar coding is proposed for binary erasure channel (BEC). Fix any BEC. For sufficiently small $\varepsilon > 0$, we construct a series of capacity achieving codes with block length $N = \varepsilon^{-4.9}$, code rate $R = \text{Capacity} - O(\varepsilon)$, block error probability $P = \varepsilon$, and encoding and decoding time complexity $\text{bC} = O(\log|\log \varepsilon|)$ per information bit. The given per-bit complexity bC is log-logarithmic in N , in $\text{Capacity} - R$, and in P . Beyond BEC, there is a generalization: Fix a prime q and fix a symmetric, q -ary-input, discrete-output memoryless channel. For sufficiently small $\varepsilon > 0$, we construct a series of error correction codes with block length $N = \varepsilon^{-\text{constant}}$, code rate $R = \text{Capacity} - O(\varepsilon)$, block error probability $P = \varepsilon$, and encoding and decoding time complexity $\text{bC} = O(\log|\log \varepsilon|)$ per information bit. Over general channels, this family of codes has the lowest per-bit time complexity among all capacity-achieving codes known to date.

Index Terms—Capacity-achieving codes, low-complexity codes, polar codes, tree pruning.

I. INTRODUCTION

IN THE theory of two-terminal error correcting codes, four of the most essential parameters of block codes are block length N , code rate R , block error probability P , and per-bit time complexity bC . We brief the history below followed by our contribution over existing works.

Shannon [1] proved that for any discrete memoryless channel (DMC), there exists a series of block codes such that R approaches a number denoted by Capacity and P converges to 0. This property is called *capacity achieving*. The price of achieving capacity is that N must approach infinity, i.e., it is not possible to achieve capacity at finite block length. Another price is that bC grows exponentially in N by the nature of random coding. This makes Shannon's (and Fano and Gallager's) construction unsuitable for practical purposes.

Coding theorists characterize how fast does the triple (N, R, P) approach $(\infty, \text{Capacity}, 0)$, extending Shannon's theory. They treat $R(N)$ and $P(N)$ as functions in N and argue about their asymptote. They showed that P alone can be as good as 2^{-N} (*error exponent regime*) [2], [3]. They also showed that R alone can be as good as $\text{Capacity} - N^{-1/2}$ (*scaling exponent regime*) [4], [5]. But together it is impossible to achieve $(R, P) = (\text{Capacity} - N^{-1/2}, 2^{-N})$ at once; the proper asymptote is

$$(R, P) = \left(\text{Capacity} - N^{-\text{constant}}, 2^{-N^{\text{constant}}} \right).$$

Manuscript received May 31, 2019; revised September 29, 2020; accepted November 13, 2020. Date of publication December 1, 2020; date of current version February 17, 2021. (Corresponding author: Hsin-Po Wang.)

The authors are with the Department of Mathematics, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: hpwang2@illinois.edu; duursma@illinois.edu).

Communicated by A. Rudra, Associate Editor for Complexity.

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TIT.2020.3041523>.

Digital Object Identifier 10.1109/TIT.2020.3041523

This later paradigm is called *moderate deviations regime* borrowed from probability theory. All three aforementioned regimes use random coding as the main tool, so their bC are on the order of 2^N . See [6]–[9] for recent progress; see also [10] for an explicit bound that implies a similar result.

Beyond random coding, Reed–Muller code is one of the earliest codes with explicit construction. To decode Reed–Muller codes, various algorithms are proposed, each giving its own trade-off among N, R, P, bC . Among them the most significant one is that Reed–Muller codes achieve capacity under the maximum a posteriori (MAP) decoding over binary erasure channels (BEC) by Kudekar *et al.* published in 2017 [11]. That they achieve capacity is worthwhile by itself so the authors do not continue to write down an explicit parametrization of N, R , and P . That being said, we believe that it is possible to infer a parametrization from their proof. (Remark: bC over BECs is polynomial in N thanks to Gaussian elimination. However, bC is exponential over general channels.)

On a different track, low density parity check (LDPC) codes are invented to generate codes with proper (N, R, P, bC) -quadruples for practical use. The construction of LDPC codes gives the priority to lowering bC . But it is difficult to infer any parametrization of N, R , and P . It was only recently, in 2013, that Kudekar *et al.* proved that LDPC codes achieve capacity [12]. Yet, their proof does not explicitly parametrize N and P . Meanwhile, a variant of LDPC codes called repeat-accumulate (RA) codes puts all efforts on reducing bC . They finally arrived at capacity achieving codes with bounded bC over BEC [13], [14]. Bounded complexity is the best possibility because the encoder should at least read in all inputs. Similar to before, their proofs do not explicitly parametrize N and P .

In 2009, Arıkan observed the phenomenon of *channel polarization* and proposed accordingly polar codes [15]. Using Doob's martingale convergence theorem, Arıkan is able to show that polar codes achieve capacity with $\text{bC} = O(\log N)$. Since then, researchers try to tune polar codes and characterize the corresponding (N, R, P, bC) asymptote. They find that P is on the order of $2^{-N^{\text{constant}}}$ and that $\text{Capacity} - R$ (aka *gap to capacity*) is on the order of $N^{-\text{constant}}$ [16]–[22]. (Just like random codes except that the constants are off.) In particular, the following choice of constants is realizable by a series of polar codes over BECs (see Lemma 3 and below):

$$(N, \text{Capacity} - R, P, \text{bC}) = \left(N, N^{-1/4.9}, 2^{-N^{1/120}}, O(\log N) \right). \quad (1)$$

See Table I for a comparison.

Our main contribution is to construct a pruned variant of polar codes and characterize its (N, R, P, bC) asymptote.

TABLE I
A COMPARISON ABOUT THE R - P -bC ASYMPTOTES OF SOME
WELL-KNOWN CAPACITY-ACHIEVING CODES. THE PROPOSED
CODE IS IN THE LAST ROW. "S." MEANS SYMMETRIC

code	$I - R$	P	bC	channel
random	$N^{-\text{const.}}$	$2^{-N^{\text{const.}}}$	exponential	DMC
Reed-Muller	$\rightarrow 0$	$\rightarrow 0$	$O(N^2)$	BEC
LDPC	$\rightarrow 0$	$\rightarrow 0$	unclear	S. BDMC
RA family	$\rightarrow 0$	$\rightarrow 0$	$O(1)$	BEC
classic polar	$N^{-\text{const.}}$	$2^{-N^{\text{const.}}}$	$O(\log N)$	DMC
old pruned polar	$O(1)$	$2^{-N^{1/2}}$	$\Theta(\log N)$	S. BDMC
this work	$N^{-\text{const.}}$	$N^{-\text{const.}}$	$O(\log \log N)$	S. prime

More precisely, take any arbitrary BEC as an example. Theorem 1 plus Lemma 3 provide a series of pruned polar codes with

$$\begin{aligned}
 (N, \text{Capacity} - R, P, \text{bC}) \\
 &= (N, O(N^{-1/4.9}), N^{-1/4.9}, O(\log \log N)) \\
 &= (\varepsilon^{-4.9}, O(\varepsilon), \varepsilon, O(\log |\log \varepsilon|)).
 \end{aligned}$$

Here $\varepsilon > 0$ is an auxiliary parameter meant to be small. As $\varepsilon \rightarrow 0$ this asymptote is clearly capacity achieving. In contrast to Asymptote (1), our pruned polar codes loosen P from $2^{-N^{1/120}}$ to $N^{-1/4.9}$ but improve bC from $O(\log N)$ to $O(\log \log N)$. The lowered bC is now log-logarithmic in N , in P , and in Capacity $- R$ (gap to capacity). This justifies the title. This is the first time polar codes are tuned and proven to have bC as low as $O(\log \log N)$.

Here is a brief preview of the proof technique: we mentioned above that Arikan observed the channel polarization phenomenon. The phenomenon is caused by the channel transformation T_{Ar} . What T_{Ar} does is to transform a channel into two other channels, one of them has its Bhattacharyya parameter squared. After n rounds of applying T_{Ar} , the majority of good channels has gone through roughly $n/2$ times of squaring. Thus the Bhattacharyya parameters of these good channels are on the order of $2^{-2^{n/2}}$ [23]. We realize that it takes only $O(\log n)$ times of squaring to achieve the order of 2^{-2^n} . An order of 2^{-2^n} suffices for achieving capacity; the remaining applications of T_{Ar} can be pruned. Since on average we prune all but $O(\log n)$ many applications of T_{Ar} , the per-bit time complexity is $\text{bC} = O(\log n) = O(\log \log 2^n) = O(\log \log N)$.

Last but not the least, as polar coding applies to a wide family of channels beyond BEC, our result applies to binary symmetric channels (BSC), symmetric binary-input discrete-output memoryless channels (BDMC), and more non-binary channels. Colloquially speaking, channel polarization is a universal phenomenon that occurs over any DMC—more precisely, all but polynomially many channels polarize, and the pace of polarization is doubly-exponential. In addition to that, pruning T_{Ar} is a versatile technique that harvests channels as early as when they are sufficiently polarized. When pruning is done properly, the introduced log-logarithm asymptote

reappears over a wide variety of channels. As a consequence, the proposed code becomes one of the fastest codes (in terms of the asymptote of the per-bit time complexity) over general channels.

A. Pruning as a Practical Technique

That T_{Ar} can be pruned is not our novel idea. Recent works on the implementation of polar coding develop a toolbox of gadgets (including pruning) that accelerate the performance of polar codes in the real world.

For instance, [24] introduced the so-called *simplified successive cancellation* decoder. It works as follows: During the construction of polar codes, some synthetic channel, for instance $(W^-)^-$, may find that all its descendants are frozen (potentially because $(W^-)^-$ is too bad). In such case, it is unnecessary to establish the part of the circuit that corresponds to $(W^-)^-$'s children. This results in circuits and trees like Fig. 4.

[24] called the synthetic channel $(W^-)^-$ a *rate-zero node*. Similarly, a *rate-one node* is a synthetic channel that is so good that all of its descendants are utilized. In such case, the authors argue that it could save some time by shortcutting the classical successive cancellation decoder. In particular, they replace butterfly devices that do soft-decision (calculation of a posteriori probabilities) by simplified butterfly devices that do hard-decision (exclusive-or of bits). This, however, does not achieve the log-logarithmic asymptote because there are $O(\log N)$ many decisions to do per information bit.

[25] introduced the so called *relaxed polarization* where a channel is further polarized only if a certain criterion is met. For instance, if $(W^-)^+$ is the only channel that meets the criterion among all depth-2 synthetic channels, it will undergo another round of polar transformation that results in circuits and trees like Fig. 5. In contrast to [24], they replace soft-decisions by nothing, so the total number of decisions is reduced. See [25, Section IV.C] for more clarification. This notion of lazy-polarization is dual to our notion of pruning. Similar ideas can be found in [26]–[32].

Alongside their huge success in optimizing practical polar codes, [25] analyzed the mathematical asymptote of how many decisions are reduced for the first time. They showed that one can, and only can, save the number of butterfly devices by a constant fraction while keeping the exponential block error probability. This means that their complexity remains $O(\log N)$ per information bit. See also [33] for the analysis of the latency of [24], [25]'s code.

On top of that, we deliberately tolerate a polynomial error probability in exchange for a log-logarithmic complexity. To sum up, both [25] and we are pruning polar codes; it is the new pruning rule we come up with that results in a different limit behavior of codes.

B. Organization

Section II reviews channel polarization and develops a general tree notation for later use. Section III states the main result, Theorem 1, and demonstrates applications to BEC and to other channels. Section IV proves the main result.

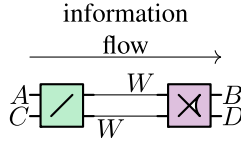




Fig. 1. The starting point of polar code construction. Two horizontal lines marked W are two independent copies of a BEC W . The box on the left is the building block of encoder. The box on the right is the building block of decoder. Pin A to pin B form a BEC which is denoted by W^- . It is a synthetic channel that is more noisy than W . Pin C to pin D form another BEC which is denoted by W^+ . It is a synthetic channel that is more reliable than W . Cf. [15, Fig. 1].

II. PRELIMINARY

A. Channel Polarization and Tree Notation

Channel polarization [15] is a method to synthesize some channels to form some extremely-reliable channels and some extremely-noisy channels. The user then can transmit uncoded messages through extremely-reliable ones while padding predictable symbols through extremely-noisy ones. We summarize channel polarization as follows.

Say we are going to communicate over a BEC W . One of Arıkan's contributions is the abstraction of two *butterfly devices*  and . (Cf. [15, Figs. 9, 10, and 5].) The butterfly devices work in a way that when we wire two independent copies of W like Fig. 1 does, pin A and B form a more noisy synthetic channel W^- while pin C and D form a more reliable synthetic channel W^+ .

Arıkan treated Fig. 1 as a recursive function where nested calls to the function will generate circuits like Fig. 2. In particular, the circuit in Fig. 2 generates eight synthetic channels $((W^-)^-)^-$, $((W^-)^-)^+$, and all the way up to $((W^+)^+)^+$. As the circuit gets larger and larger, we will end up getting $2^{\text{number of calls}}$ channels, from $(\dots (W^-)^- \dots)^-$ to $(\dots (W^+)^+ \dots)^+$. Arıkan observed that synthetic channels generated in this way tend to be either extremely reliable or extremely noisy. That is to say, they *polarize*. He called this phenomenon *channel polarization*.

The relation among $W, W^-, \dots, ((W^+)^+)^+$ is summarized by a channel transformation $T_{\text{Arı}}$ as is discussed in [15, Section II]. We reproduce and improve [15, Fig. 6] in Fig. 3. It is a tree whose vertexes are channels. Each parent-child-child triple represents the fact that the butterfly devices turn two independent copies of the parent channel into an upper child channel and a lower child channel.

We introduce in the next subsection that it is possible to prune circuits and trees to reduce complexity. We will take advantage of the fact that circuits and trees correspond to each other and only argue about trees. Eventually we will show how to prune trees without having to sacrifice R and P too much.

B. Pruning Circuits and Trees

As mentioned in the introduction, the observation that circuits and trees can be pruned to attain a lower complexity has been made several times in the past ([24]–[32]). For instance, Fig. 4 illustrates a circuit-tree pair that saves two butterfly devices, which potentially saves some time comparing to

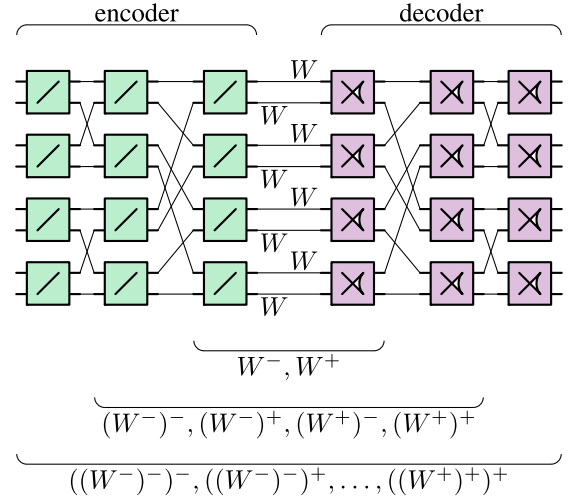


Fig. 2. Fig. 1 works like a recursive function. We can call the function three times to obtain this circuit. At the middle column there are eight independent copies of BEC W . The inner layer of butterfly devices will turn them into four independent copies of W^- and four independent copies of W^+ . The second layer of butterfly devices will turn them into $(W^-)^-$, $(W^-)^+$, $(W^+)^-$, and $(W^+)^+$, each of two independent copies. Finally the outer layer of butterfly devices will turn them into $((W^-)^-)^-$, $((W^-)^-)^+$, $((W^-)^+)^-$, $((W^-)^+)^+$, $((W^+)^-)^-$, $((W^+)^-)^+$, $((W^+)^+)^-$, and $((W^+)^+)^+$. Cf. [15, Figs. 2 and 3].

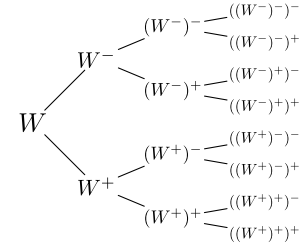


Fig. 3. Fig. 2 explains how the circuit transforms a channel to another. This operation can be encoded by a tree with auxiliary labels. In the tree, each vertex is a channel. A vertex is either a leaf or has two children. When a channel has two children, they form a parent-child-child triangle which represents the fact that the parent channel, say w , is transformed into w^- (upper child) plus w^+ (lower child) by the butterfly devices. Instead of verbosely spamming “butterfly devices,” we put a $T_{\text{Arı}}$ at the center of each such triangle. It represents that butterfly devices serve as a channel transformation and that it is Arıkan who first recognizes/invents this transformation.

Figs. 2 and 3. Fig. 5 illustrates another circuit-tree pair that saves six butterfly devices, which potentially saves more time.

Roughly speaking, we expect that the more the circuit and the tree are pruned, the more butterfly devices are saved. Having less butterfly devices potentially saves more time. However, the saving in time, if any, does not come for free. Since the resulting synthetic channels are different from before, P changes. Thus a user has to recompute/remeasure P and to make sure whether the new P is affordable. An extreme, degenerate case is that one does not discard any butterfly device; use polar coding as it was proposed by Arıkan. Another extreme point is that one drops all butterfly devices; this saves 100% of time but then there is no coding at all. The log-logarithmic behavior is somewhere in between.

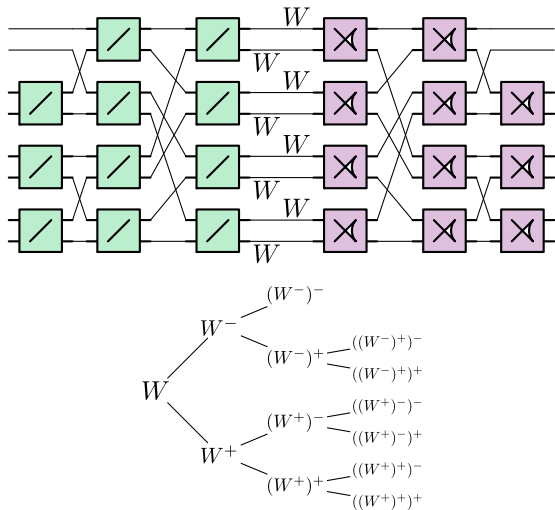


Fig. 4. The top part is a pruned circuit where the butterfly devices applied to $(W^-)^-$ are dropped. As a result, this circuit does not generate $((W^-)^-)^-$ or $((W^-)^-)^+$ and leaves the two copies of $(W^-)^-$ intact. The complete list of generated channels reads: $(W^-)^-$, $(W^-)^-$, $((W^-)^+)^-$, $((W^-)^+)^+$, $((W^+)^-)^-$, $((W^+)^-)^+$, $((W^+)^+)^-$, $((W^+)^+)^+$. The bottom part is a pruned tree that illustrates the fact that $(W^-)^-$ does not undergo the third round of application of T_{An} and has no children. On the other hand, other “depth-2” channels $(W^-)^+$, $(W^+)^-$, $(W^+)^+$ undergo T_{An} and generate what they used to generate in Fig. 3.

In the next subsection, we review channel parameters H , I , P_e , and Z , followed by the stochastic processes $\{K_i\}$, $\{I_i\}$, $\{Z_i\}$, and $\{H_i\}$. Then we generalize the processes. We will show how they relate to trees, especially to pruned trees. Being able to relate trees to processes makes it possible to control the behavior of codes properly.

C. Channel Parameters and Processes

Let $W: \mathcal{X} \rightarrow \mathcal{Y}$ be a symmetric DMC. Let q be the size of the input alphabet \mathcal{X} . We assume the uniform input distribution.

The conditional entropy $H(W)$ of a channel W is the Shannon entropy of the input conditioned on the output (assuming base- q logarithm). More formally,

$$H(W) := - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbb{P}_{XY}(x, y) \log_q \mathbb{P}_{X|Y}(x | y)$$

It measures the noise—unreliability—of a channel. For BEC, $H(W)$ coincides with the erasure probability of W .

The symmetric capacity $I(W)$ of W is defined to be the mutual information between the input and the output (assuming base- q logarithm).

$$I(W) := \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbb{P}_{XY}(x, y) \log_q \frac{\mathbb{P}_{X|Y}(x | y)}{\mathbb{P}_X(x)}$$

Over symmetric channels with uniform inputs, it coincides with the complement $1 - H(W)$.

The bit error probability $P_e(W)$ of W is the error probability of the MAP decoder applied to W . In detail,

$$P_e(W) := \sum_{y \in \mathcal{Y}} \mathbb{P}_Y(y) (1 - \max_{x \in \mathcal{X}} \mathbb{P}_{X|Y}(x | y)).$$

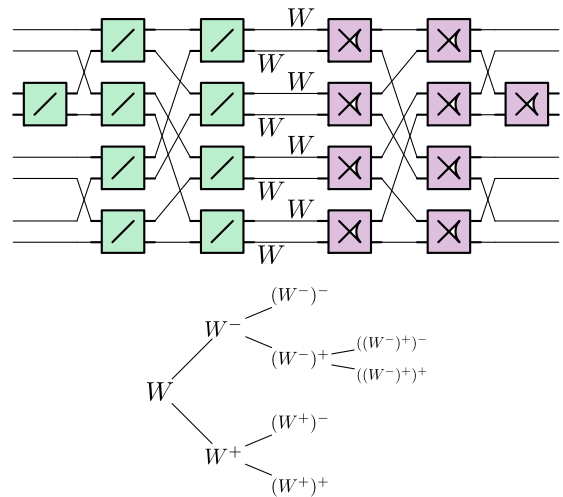


Fig. 5. The top part is a pruned circuit where the butterfly devices applied to $(W^-)^-$, $(W^+)^-$, and $(W^+)^+$ are dropped. They (each of two copies) are left intact. The bottom part is a pruned tree that encodes what happens in the circuit: only $(W^-)^+$ undergoes the third round of application of T_{An} and has children. The complete list of generated channels reads: $(W^-)^-$, $(W^-)^-$, $((W^-)^+)^-$, $((W^-)^+)^+$, $(W^+)^-$, $(W^+)^-$, $(W^+)^+$, $(W^+)^+$.

The Bhattacharyya parameter of W , denoted by $Z(W)$, is designed to be an upper bound on $P_e(W)$. It is defined in Arkan’s work over binary channels [15] and in [34], [35] over general channels. The definitions that will be mentioned in the sequel is the $q = 2$ version

$$Z(W) := \sum_y \sqrt{W(y | 0)W(y | 1)}.$$

Over BECs, $Z(W)$ coincides with $H(W)$. We will use them interchangeably.

Recall the processes $\{K_i\}$, $\{I_i\}$, and $\{Z_i\}$ as defined in [15, Section IV, third paragraph]. Therein, $\{K_i\}$ is the process starting from $K_0 := W$; and K_{i+1} is either K_i^- or K_i^+ , each with probability $1/2$. The process of mutual information $\{I_i\}$ is defined to be $I_i := I(K_i)$. The process of Bhattacharyya parameter $\{Z_i\}$ is defined to be $Z_i := Z(K_i)$. Moreover, we define the process of conditional entropy $\{H_i\}$ to be $H_i := H(K_i)$. Clearly $I_i + H_i = 1$ over symmetric channels with uniform inputs. Here is our generalization.

Denote by \mathcal{T} a finite rooted tree of channels with root channel W . Convention: the root has depth 0; the depth of a tree is the depth of the deepest leaf; and the tree with only one vertex has depth 0. Therefore, the circuit corresponding to \mathcal{T} consists of an array of butterfly devices with $2^{\text{depth}(\mathcal{T})}$ columns and $2^{\text{depth}(\mathcal{T})-1}$ rows. The array contains $2^{\text{depth}(\mathcal{T})}$ independent copies of W . For any leaf channel w , the circuit generates $2^{\text{depth}(\mathcal{T})-\text{depth}(w)}$ copies of w .

Given a finite channel tree \mathcal{T} with root channel W , define three discrete-time stochastic processes $\{K_{i \wedge \tau}\}$, $\{I_{i \wedge \tau}\}$, $\{H_{i \wedge \tau}\}$ and a stopping time τ as follows: Start from the root channel $K_{0 \wedge \tau} := W$. For any $i \geq 0$, if $K_{i \wedge \tau}$ is a leaf, let $K_{i+1 \wedge \tau}$ be $K_{i \wedge \tau}$. If, otherwise, $K_{i \wedge \tau}$ has two children, choose either child with equal probability as $K_{i+1 \wedge \tau}$. Since \mathcal{T} is finite, there is always a smallest index j such that $K_{j \wedge \tau}$

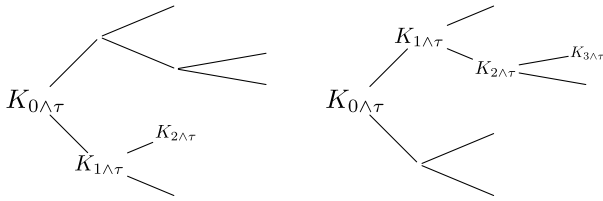


Fig. 6. Recall the tree in Fig. 5. On the left is a possible trajectory of the process $\{K_{i \wedge \tau}\}$. We begin with $K_{0 \wedge \tau}$ being the root channel W . It has children. The first “coin toss” chooses the lower child W^+ as $K_{1 \wedge \tau}$. It has children. The second coin toss chooses the upper child $(W^+)^-$ as $K_{2 \wedge \tau}$. It has no child. The process stabilizes. So $K_{2 \wedge \tau} = K_{3 \wedge \tau} = K_{4 \wedge \tau} = \dots = K_{\tau}$ and $\tau = 2$. The probability measure of this trajectory is $1/8$. On the right is another possible trajectory of the process $\{K_{i \wedge \tau}\}$. We begin with $K_{0 \wedge \tau}$ being the root channel. It has children. The first coin toss chooses the upper child W^- as $K_{1 \wedge \tau}$. It has children. The second coin toss chooses the lower child $(W^-)^+$ as $K_{2 \wedge \tau}$. It has children. The third coin toss chooses the upper child $((W^-)^+)^-$ as $K_{3 \wedge \tau}$. It has no child. The process stabilizes with $K_{3 \wedge \tau} = K_{4 \wedge \tau} = K_{5 \wedge \tau} = \dots = K_{\tau}$ and $\tau = 3$. The probability measure of this trajectory is $1/4$.

reaches a leaf, or equivalently $K_{j \wedge \tau} = K_{j+1 \wedge \tau} = K_{j+2 \wedge \tau} = \dots$ *ad infinitum*. Define a random variable τ to be this smallest index. Then τ is the *stopping time* that records when $K_{i \wedge \tau}$ “stops evolving.” Let K_{τ} be the channel $K_{i \wedge \tau}$ when it stops evolving. That is, $K_{\tau} = \lim_{i \rightarrow \infty} K_{i \wedge \tau}$. Let $I_{i \wedge \tau} := I(K_{i \wedge \tau})$. Let $H_{i \wedge \tau} := H(K_{i \wedge \tau})$. Let $I_{\tau} := I(K_{\tau}) = \lim_{i \rightarrow \infty} I_{i \wedge \tau}$. Let $H_{\tau} := H(K_{\tau}) = \lim_{i \rightarrow \infty} H_{i \wedge \tau}$.

Readers might have noticed that the notations $K_{i \wedge \tau}$, $I_{i \wedge \tau}$, and $H_{i \wedge \tau}$ coincide with what Gallager calls *stopped process* [36, Theorem 9.7.1]. One may as well stick to the operational definition presented above.

Recall the pruned tree in Fig. 5. We give two possible trajectories in Fig. 6. Note that this tree is a nontrivial example where τ is not a constant. As a random variable, τ depends on which child of $K_{i \wedge \tau}$ is chosen at each step. It turns out that $\mathbb{P}\{\tau = 2\} = 3/4$ and $\mathbb{P}\{\tau = 3\} = 1/4$. For the tree in Fig. 4, $\mathbb{P}\{\tau = 2\} = 1/4$ and $\mathbb{P}\{\tau = 3\} = 3/4$. For the tree in Fig. 3, however, $\tau = 3$ with probability 1.

By [15, Proposition 8], $\{I_i\}$ is a martingale. Hence $\{I_{i \wedge \tau}\}$ is a martingale by [37, Theorem 5.2.6]. Since W is symmetric, $H_{i \wedge \tau} = 1 - I_{i \wedge \tau}$ forms a martingale as well. A useful consequence by applying [37, Theorem 5.7.6] to $\{I_i\}$ is

$$I(W) = I_0 = \mathbb{E}[I_{\tau}]. \quad (2)$$

Remark: $\{I_i\}$ being a martingale plays two crucial roles in Arkan’s work. For one: the martingale convergence theorem applies. For two: $I(W) = I_0 = \mathbb{E}[I_n]$ so $\mathbb{P}\{I_{\infty} = 1\} = I(W)$. Equality (2) generalizes this argument in the manner that we can now decide whether to prune a branch or not on a channel-by-channel basis. This creates a new level of flexibility to balance bC and other parameters.

In the next subsection we show how trees and processes relate to codes. Only after we establish the relation between trees and (N, R, P, bC) can we optimize how we are going to prune the tree.

D. From Trees to Codes and Communication

Recall that in a given tree \mathcal{T} , non-leaf vertexes represent channels that are consumed to obtain their children. They are

not available to users. Leafs of \mathcal{T} , on the other hand, represent channels that are available to users. A user who wants to send messages using \mathcal{T} can: 1) choose a subset \mathcal{A} of leafs of \mathcal{T} ; 2) transmit uncoded messages through leaf channels in \mathcal{A} ; and 3) pad predictable symbols through the remaining leaf channels. Leafs in \mathcal{A} and the corresponding channels are said to be *chosen* or *utilized*. Leafs outside \mathcal{A} and the corresponding channels are said to be *frozen*.

This makes the tree-leafs pair $(\mathcal{T}, \mathcal{A})$ a block code. We want to characterize this block code by analyzing these four parameters: block length N , code rate R , block error probability P , and per-bit time complexity bC. Here is how to read-off these parameters from $(\mathcal{T}, \mathcal{A})$.

The *block length* N of $(\mathcal{T}, \mathcal{A})$ is the number of copies of W in the corresponding circuit. In terms of a function of the tree, it is

$$N := 2^{\text{depth}(\mathcal{T})}.$$

N does not depend on \mathcal{A} , so we can talk about “the block length of \mathcal{T} ” without defining \mathcal{A} in advance. The *multiplicity* of a synthetic channel w is the number of occurrences of w in the circuit; this is further equal to $2^{\text{depth}(\mathcal{T}) - \text{depth}(w)} = N \mathbb{P}\{\{K_i\} \text{ passes } w\}$.

The *code rate* R of $(\mathcal{T}, \mathcal{A})$ is the number of synthetic channels w in \mathcal{A} (counting with multiplicity) divided by N . In terms of stochastic processes, it is the probability that K_{τ} is utilized.

$$R := \mathbb{P}\{K_{\tau} \in \mathcal{A}\}.$$

The *block error probability* P of $(\mathcal{T}, \mathcal{A})$ is the probability that any utilized leaf channel in \mathcal{A} fails to transmit the message. For a classical polar code, the block error probability is at most $\sum_{w \in \mathcal{A}} Z(w)$ as stated in [15, Proposition 2]. For a pruned polar code, the block error probability is at most the weighted sum

$$P \leq \sum_{w \in \mathcal{A}} (N \mathbb{P}\{K_{\tau} = w\}) P_e(w).$$

To paraphrase, since $P_e(w)$ is the bit error probability of each individual utilized channel, it suffices to apply the union bound where each $P_e(w)$ is weighted by the multiplicity of w in the circuit.

The *per-block time complexity* of $(\mathcal{T}, \mathcal{A})$ is how long \mathcal{T} ’s circuit takes to execute. It is bounded from above by the number of butterfly devices multiplied by the time each butterfly device spends. (No parallelism allowed in our model.) The design of the butterfly devices suggests that each butterfly device spends constant time. Thus the per-block time complexity is proportional to the number of butterfly devices. As each leaf channel K_{τ} at depth τ must go through τ many $\begin{bmatrix} \text{---} \\ \text{---} \end{bmatrix}$ ’s and τ many $\begin{bmatrix} \text{---} \\ \text{---} \end{bmatrix}$ ’s, the total number of butterfly devices is exactly $2N\mathbb{E}[\tau]$. Hence the per-block time complexity is proportional to $N\mathbb{E}[\tau]$.

The *per-bit time complexity* bC is the amortized time each information bit should pay. Naturally it is proportional to $N\mathbb{E}[\tau]/NR = \mathbb{E}[\tau]/R$. In this work, we are pursuing capacity achieving codes so $R \approx I(W)$ is about a constant. Therefore,

the per-bit time complexity is proportional to

$$\mathbb{E}[\tau]. \quad (3)$$

$\mathbb{E}[\tau]$ does not depend on \mathcal{A} , so we can talk about “the complexity of \mathcal{T} ” without defining \mathcal{A} in advance.

We are almost ready to show readers how to prune trees except that we will phrase pruning in an opposite tone: Instead of starting from a huge, heavy tree and pruning the vast majority of its vertexes, we grow a tree from scratch and decide channel-by-channel whether or not each channel should have children. Doing so fits the stochastic processes paradigm more properly because usually we are not allowed to look into the future (see the descendants) before we make the decision (whether it should have children or not). We assure that this is a matter of wording style and has nothing to do with the actual properties of codes.

In this context, we say *apply T_{An} to w* if we want w to have children. We say *do not apply T_{An} to w* if we want the opposite, that w should be a leaf. Here are two heuristic rules that guide the application of T_{An} : 1) That $N := 2^{\text{depth}(\mathcal{T})}$ suggests that we should set a boundary n and do not apply T_{An} once we reach depth n . This guarantees that the block length N will never exceed 2^n . We assume the worst case scenario $N := 2^n$. 2) A mediocrelly reliable channel increases P too much if we utilize it, but sacrifices R too much if we freeze it. Either way it becomes an obstacle to capacity achieving. To avoid the dilemma, the only chance is applying T_{An} to polarize it further. This suggests that we should make decisions based on a threshold for “mediocre reliability.”

E. Growing Tree and Choosing Leafs as Code Construction

We showed how to estimate the parameters of a block code $(\mathcal{T}, \mathcal{A})$ if \mathcal{T} and \mathcal{A} are explicitly given. Now we state how we are going to grow a good tree of prescribed depth n (instead of pruning the perfect binary tree of depth n). Here n is an integer to be assigned. Let $\varepsilon > 0$ be small. Let $Y(w)$ be $\min\{I(w), H(w)\}$.

Begin with W as the only vertex of a new rooted tree. We announce the following rule:

$$\boxed{\text{Apply } T_{\text{An}} \text{ to } w \text{ if and only if} \\ \text{depth}(w) < n \text{ and } Y(w) > \varepsilon 2^{-n}.} \quad (4)$$

The rule says: for each leaf w , if both $\text{depth}(w) < n$ and $Y(w) > \varepsilon 2^{-n}$ are met, apply T_{An} to w to obtain w^- and w^+ ; and then append w^- and w^+ as the children of w . If, otherwise, either criterion is not met, we do not apply T_{An} and leave w as a leaf. See Appendix A for a possible execution of the rule. We will see later that $Y(w)$ serves as a judgement of whether w is sufficiently polarized or not. Having \mathcal{T} , we declare \mathcal{A} by a criterion

$$\boxed{w \in \mathcal{A} \text{ if and only if} \\ w \text{ is a leaf and } H(w) \leq \varepsilon 2^{-n}.} \quad (5)$$

In Rule 4 and Criterion 5 we implicitly divide channels into four classes: 1) For channels that are mediocrelly reliable, i.e., $\varepsilon 2^{-n} < I(w) < 1 - \varepsilon 2^{-n}$, we apply T_{An} to polarize w further. 2) For channels that are sufficiently reliable,

i.e., $H(w) \leq \varepsilon 2^{-n}$, we stop applying T_{An} and collect them in our pocket \mathcal{A} . Doing so as early as possible maximizes the save on butterfly devices. Nevertheless, every channel we put in \mathcal{A} contributes to the overall block error probability P . We must choose wisely what to and what not to put in \mathcal{A} . 3) For channels that are incredibly noisy, i.e. $I(w) \leq \varepsilon 2^{-n}$, it becomes inefficient to extract any capacity from w . We should just “let go” the noisy channels and save butterfly devices. The earlier we let them go the more butterfly devices we save. Nevertheless, since $\mathbb{E}[I_\tau] = I(W)$ is conserved, letting go a channel means giving up the capacity it carries. We must not give up too much capacity as we want $R \rightarrow I(W)$. 4) For channels that are mediocrelly reliable *at depth n* , there is no chance to polarize them further. We shall let them go.

We now have both \mathcal{T} and \mathcal{A} properly defined in terms of W , ε and n . We will show in the coming section how this $(\mathcal{T}, \mathcal{A})$, as a block code, performs.

III. MAIN RESULT AND APPLICATIONS

Theorem 1 (Main Theorem): Let W be any symmetric channel. Assume that there exist $\beta' > 0$ and $\mu' < \infty$ such that as $i \rightarrow \infty$,

$$\mathbb{P}\{I_i < 2^{-2^{\beta' i}}\} \geq H(W) - O(2^{-i/\mu'}) \quad (6)$$

and

$$\mathbb{P}\{H_i < 2^{-2^{\beta' i}}\} \geq I(W) - O(2^{-i/\mu'}). \quad (7)$$

Then there exists a series of pruned polar codes with block length N , code rate R , block error probability P , and per-bit time complexity bC satisfying

$$(N, I(W) - R, P, \text{bC}) \\ = (N, O(N^{-1/\mu'}), N^{-1/\mu'}, O(\log \log N))$$

as $N \rightarrow \infty$. Let ε be $N^{-1/\mu'}$, then the quadruple can be rewritten as

$$(\varepsilon^{-\mu'}, O(\varepsilon), \varepsilon, O(\log |\log \varepsilon|))$$

as $\varepsilon \rightarrow 0$.

Sketch: The general strategy is to grow a tree according to the framed rule (4) and choose/freeze leafs according to the framed criterion (5). The processes $\{K_i\}$ and the random variable K_τ are thus uniquely determined. In order to control how K_τ behaves, we gain control of how $\{K_i\}$ behaves in terms of Inequalities (6) and (7).

Section IV serves as a formal proof of the theorem. The code will be constructed in Section IV-A. We characterize its block length N , per-bit time complexity bC, and block error probability P in Section IV-B. Section IV-C computes the code rate R . ■

A. On the Precondition and Applications

The preconditions, Inequality (6) and (7), characterize the so-called *moderate deviations behavior* of polar codes. As commented in the introduction, it is expected that all polar codes enjoy some sort of moderate deviations asymptote, meaning that

$$\mathbb{P}\{H_i \leq 2^{-N^{\text{constant}}}\} \geq I(W) - O(N^{-\text{constant}}).$$

On the other hand, the precise constants over various channels are yet to be figured out. On this path, the earliest result dated back to 2013, by Guruswami–Xia [16].

Proposition 2 [16, Theorem 1]: Let W be a symmetric BDMC. There exists $\mu' < \infty$ such that

$$\mathbb{P}\{Z_i \leq 2^{-2^{0.49i}}\} \geq I(W) - O(2^{-i/\mu'}). \quad (8)$$

Intuitively speaking, this lemma shows that Z_i goes to zero doubly-exponentially fast. Recall that in Rule 4, we do not apply T_{An} if $H_i \leq \varepsilon 2^{-n}$. Here $\varepsilon 2^{-n}$ is polynomial in N so H_i will reach this threshold in log-logarithmic steps. All T_{An} afterwards are pruned. This is the main reason why the complexity is log-logarithmic.

The prior result is followed by a generalization with a family of explicit constants.

Lemma 3 [17, Theorem 3 and Inequality (56)]: Let W be a symmetric BDMC. Let μ be the scaling exponent and γ be such that $1/(1+\mu) < \gamma < 1$. Then

$$\mathbb{P}\{Z_i \leq 2^{-2^{i\gamma h_2^{-1}(\frac{2\mu+\gamma-1}{\gamma\mu})}}\} \geq I(W) - O(2^{-\frac{i(1-\gamma)}{\mu}}). \quad (9)$$

Here h_2^{-1} is the inverse function of the binary entropy function. And the scaling exponent μ is a number such that ([38]–[41])

$$\mathbb{P}\{Z_i \text{ is “small”}\} \geq I(W) - O(2^{-i/\mu}).$$

Here “small” is some function in i that examines if a channel is reliable enough. Different works use different functions but it could be proven that a large class of functions all determine the same μ .

Over BECs, we know $Z(W) = H(W) = 1 - I(W)$ and the recursion reads

$$\begin{aligned} Z_{i+1} &= \begin{cases} 1 - (1 - Z_i)^2 & \text{w.p. } 1/2, \\ Z_i^2 & \text{w.p. } 1/2; \end{cases} \\ I_{i+1} &= \begin{cases} I_i^2 & \text{w.p. } 1/2, \\ 1 - (1 - I_i)^2 & \text{w.p. } 1/2. \end{cases} \end{aligned}$$

As a consequence, Inequalities such as (8) and (9) can be “flipped” to obtain their counterparts at the noisy-end over BECs. For instance

$$\mathbb{P}\{I_i \leq 2^{-2^{i\gamma h_2^{-1}(\frac{\gamma\mu+\gamma-1}{\gamma\mu})}}\} \geq H(W) - O(2^{-\frac{i(1-\gamma)}{\mu}}). \quad (10)$$

This and Inequality (9) together provide instances of constants β' and μ' for the precondition of the main theorem. More precisely, we know the scaling exponent μ over BECs is at most 3.639 [17, Theorem 2]. So we let $\mu := 3.639$ and $\gamma := 1261/4900$. Then Inequality (9) becomes

$$\mathbb{P}\{H_i \leq 2^{-2^{i/119.5}}\} \geq I(W) - O(2^{i/4.9}).$$

Its noisy-end counterpart becomes

$$\mathbb{P}\{I_i \leq 2^{-2^{i/119.5}}\} \geq H(W) - O(2^{i/4.9}).$$

So we know Theorem 1 holds for $(\mu', \beta') = (4.9, 1/120)$ over BECs.

Corollary 4: Let W be any BEC. There exists a series of pruned polar codes with block length N , code rate R , block error probability P , and per-bit time complexity bC satisfying

$$(N, I(W) - R, P, \text{bC}) = (\varepsilon^{-4.9}, O(\varepsilon), \varepsilon, O(\log|\log \varepsilon|))$$

as $\varepsilon \rightarrow 0$.

Over other channels, how to fulfill Inequalities (6) and (7) is not so clear. One reason is that known results about moderate deviations behavior (generalizations of Inequalities (8) and (9)) are limited to symmetric prime-ary-input discrete-output memoryless channels. In fact, the following is the best known result.

Lemma 5 [20]: Let W be a symmetric prime-ary-input discrete-output memoryless channel. For some $\beta' > 0$, there exists $\mu' < \infty$ such that

$$\mathbb{P}\{H_i \leq 2^{-2^{\beta' i}}\} \geq I(W) - O(2^{-i/\mu'}). \quad (11)$$

(Appendix B translates their result into our terminology.)

Moreover, even if a moderate deviations behavior is available, it does not immediately imply its noisy-end counterpart (cf. how we derive Inequality (10)). Therefore, we have to provide the flipped version by itself.

Lemma 6: Let W be a symmetric prime-ary-input discrete-output memoryless channel. There exist $\beta' > 0$ and $\mu' < \infty$ such that

$$\mathbb{P}\{I_i \leq 2^{-2^{\beta' i}}\} \geq H(W) - O(2^{-i/\mu'}). \quad (12)$$

The proof of the lemma is heavily inspired by [20], [41]. It will be given in Appendix C. Now Lemmas 5 and 6 and Theorem 1 jointly imply a general version of Corollary 4.

Corollary 7: Let W be any symmetric prime-ary-input discrete-output memoryless channel. For some $\mu' < \infty$, there exists a series of pruned polar codes with block length N , code rate R , block error probability P , and per-bit time complexity bC satisfying

$$(N, I(W) - R, P, \text{bC}) = (\varepsilon^{\mu'}, O(\varepsilon), \varepsilon, O(\log|\log \varepsilon|))$$

as $\varepsilon \rightarrow 0$.

That constitutes the application of the main theorem to the pruned polar code over general channels. We prove the main theorem in the next section.

IV. PROOF OF THEOREM 1 (THE MAIN THEOREM)

In this section, we prove the main theorem. First of all, Inequalities (6) and (7) imply

$$\mathbb{P}\{Y_i > 2^{-2^{\beta' i}}\} \leq O(2^{-i/\mu'}). \quad (13)$$

Here $Y_i := \min(H_i, I_i)$. We are going to use the preconditions in this particular form. In the upcoming subsections, we will construct the code by growing a tree and selecting leafs. We will compute the “average depth” $\mathbb{E}[\tau]$ of the tree. We will then capture the code’s N , bC, P , and R in this order.

A. Tree Construction and Average Depth

We are ready to analyze the stated construction. We will first prove a lemma regarding $\mathbb{E}[\tau]$ and then analyze N, bC, P, R in that order. Once we can control all four parameters we obtain the main theorem, Theorem 1.

Lemma 8: Given W and ε , assign $n := -\mu' \log_2 \varepsilon$. Then Rule (4), i.e.,

$$\boxed{\text{Apply } T_{\text{An}} \text{ to } w \text{ if and only if} \\ \text{depth}(w) < n \text{ and } Y(w) > \varepsilon 2^{-n}},$$

grows a channel tree \mathcal{T} with $\mathbb{E}[\tau] = O(\log|\log \varepsilon|)$.

Proof: Grow the tree and observe the processes $\{K_{i \wedge \tau}\}$ and $\{Y_{i \wedge \tau}\}$. By the rule, the channel $K_{i \wedge \tau}$ has children if and only if $\text{depth}(K_{i \wedge \tau}) < n$ and $Y(K_{i \wedge \tau}) > \varepsilon 2^{-n}$. Conversely, the channel $K_{i \wedge \tau}$ has no child if and only if $\text{depth}(K_{i \wedge \tau}) \geq n$ or $Y(K_{i \wedge \tau}) \leq \varepsilon 2^{-n}$. The stopping time τ , by definition, is the least index j such that $K_{j \wedge \tau}$ has no child. So τ is the least index j such that $\text{depth}(K_j) \geq n$ or $Y(K_j) \leq \varepsilon 2^{-n}$. Equivalently, τ is the least index j such that $j \geq n$ or $Y_j \leq \varepsilon 2^{-n}$. More formally,

$$\tau = \min(\{j : Y_j \leq \varepsilon 2^{-n}\} \cup \{n\}).$$

For stopping times defined in the form “when is the first time *something* happens,” they are usually studied through the events $\{\tau > i\}$ indexed by $i \in \mathbb{N} \cup \{0\}$. To rephrase it, knowing “when does *something* first happen” is equivalent to knowing “whether *something* had happened before time i .” In our case, the predicate $\tau > i$ is equivalent to whether $i \geq n$ or $Y_j \leq \varepsilon 2^{-n}$ for some $j \leq i$. We relax the criteria to whether $Y_i \leq \varepsilon 2^{-n}$. Symbolically,

$$\{\tau > i\} \subseteq \{Y_i > \varepsilon 2^{-n}\} = \{Y_i > \varepsilon^{1+\mu'}\}.$$

The equality is due to our choice of $n := -\mu' \log_2 \varepsilon$.

Whether $Y_i > \varepsilon^{1+\mu'}$ or not can be relaxed to a disjunction $Y_i > 2^{-2^{\beta' i}}$ or $2^{-2^{\beta' i}} > \varepsilon^{1+\mu'}$. We have seen the first disjunct before (in Inequality (13)) and can control how often it happens. The second disjunct is new; we solve for i and deduce that $2^{-2^{\beta' i}} > \varepsilon^{1+\mu'}$ implies $i < O(\log|\log \varepsilon|)$. More formally,

$$\{\tau > i\} \subseteq \{Y_i > 2^{-2^{\beta' i}} \text{ or } i < O(\log|\log \varepsilon|)\}.$$

Now whether $\tau > i$ happens is divided into two cases: 1) If i is small enough— $i < O(\log|\log \varepsilon|)$ —we have little idea about whether $\tau > i$ or not. (It probably is; we do not expect decent polarization this early.) 2) If i is large enough to violate $i < O(\log|\log \varepsilon|)$, then $\{\tau > i\}$ is dominated by the first disjunct $Y_i > 2^{-2^{\beta' i}}$. And Inequality (13) bounds its probability measure from above. Putting 1) and 2) together, we have a joint bound

$$\mathbb{P}\{\tau > i\} \leq \begin{cases} 1 & \text{when } i < O(\log|\log \varepsilon|), \\ O(2^{-i/\mu'}) & \text{otherwise.} \end{cases} \quad (14)$$

Now we recall a useful restatement of Fubini’s theorem in probability theory [37, Lemma 2.2.8]; it reads $\mathbb{E}[\tau] = \sum_{i=0}^{\infty} \mathbb{P}\{\tau > i\}$. This reassures what we claimed above—that when does *something* first happen (LHS) is related to whether

that thing happened before i (RHS). The summation on the RHS is from $i = 0$ to ∞ but we divide them into two cases: 1) When $0 \leq i < O(\log|\log \varepsilon|)$, we have little control over the probability of $Y_i > 2^{-2^{\beta' i}}$. We hence sum $O(\log|\log \varepsilon|)$ many 1’s (the trivial upper bound of probabilities). The sum is $O(\log|\log \varepsilon|)$. 2) When $O(\log|\log \varepsilon|) \leq i \leq \infty$ we have the upper bound $O(2^{-i/\mu'})$. We are summing a geometric series; the sum is $O(1)$. Putting 1) and 2) together, we have a complete estimate

$$\begin{aligned} \mathbb{E}[\tau] &= \sum_{i=0}^{\infty} \mathbb{P}\{\tau > i\} \\ &= \sum_{i=0}^{O(\log|\log \varepsilon|)} \mathbb{P}\{\tau > i\} + \sum_{i=O(\log|\log \varepsilon|)}^{\infty} \mathbb{P}\{\tau > i\} \\ &\leq \sum_{i=0}^{O(\log|\log \varepsilon|)} 1 + \sum_{i=O(\log|\log \varepsilon|)}^{\infty} O(2^{-i/\mu'}) \\ &= O(\log|\log \varepsilon|) + O(1) \\ &= O(\log|\log \varepsilon|). \end{aligned}$$

This finishes the computation of the average depth $\mathbb{E}[\tau]$. ■

B. Code Length, Error Probability, and Complexity

Lemma 8 contains the most technical steps in this work. This is the first time the concept of stopping time is introduced to the field of polar codes, and it plays key roles in the proof. Now we have completed Lemma 8, i.e., the construction of the tree \mathcal{T} and the computation of its average depth $\mathbb{E}[\tau]$, it remains to: 1) read off N and bC from \mathcal{T} ; 2) define \mathcal{A} ; and 3) read off P and R from $(\mathcal{T}, \mathcal{A})$.

For the block length N : Rule 4 stops applying T_{An} at depth n . In other words, the rule grows a tree of depth (at most) n , where n was defined to be $-\mu' \log_2 \varepsilon$ in Lemma 8. No matter what \mathcal{A} will be selected, the code $(\mathcal{T}, \mathcal{A})$ possesses block length $N \leq 2^n = \varepsilon^{-\mu'}$.

For per-bit time complexity bC : As proven in Lemma 8, the stopping time of the tree \mathcal{T} has expectation $\mathbb{E}[\tau] = O(\log|\log \varepsilon|)$. By the discussion that leads to Formula (3), the code $(\mathcal{T}, \mathcal{A})$ possesses per-bit time complexity $\text{bC} = \mathbb{E}[\tau] = O(\log|\log \varepsilon|) = O(\log \log N)$ regardless of how \mathcal{A} will be chosen.

For block error probability P : This quantity depends on \mathcal{A} so we now declare \mathcal{A} by Criterion (5), i.e.,

$$\boxed{w \in \mathcal{A} \text{ if and only if} \\ w \text{ is a leaf and } H(w) \leq \varepsilon 2^{-n}.}$$

Next we attempt to calculate P :

$$\begin{aligned} P &\leq \sum_{w \in \mathcal{A}} N \mathbb{P}\{K_\tau = w\} P_e(w) && \text{(union bound)} \\ &\leq \sum_{w \in \mathcal{A}} N \mathbb{P}\{K_\tau = w\} H(w) && \text{(by [42, (16)])} \\ &\leq \sum_{w \in \mathcal{A}} N \mathbb{P}\{K_\tau = w\} \varepsilon 2^{-n} && \text{(by (5))} \\ &\leq N \varepsilon 2^{-n} = \varepsilon. && \text{(see below)} \end{aligned}$$

Here (see below) uses that $\{K_\tau = w\}$ are disjoint events so their probability measures sum to 1, at most. This proves that the code $(\mathcal{T}, \mathcal{A})$ possesses block error probability $P \leq \varepsilon$.

So far we proved that the code $(\mathcal{T}, \mathcal{A})$ has parameter triple

$$(N, P, \text{bC}) = (N, N^{-1/\mu'}, O(\log \log N)) \\ = (\varepsilon^{\mu'}, \varepsilon, O(\log |\log \varepsilon|)).$$

The code rate of the code $(\mathcal{T}, \mathcal{A})$ is less straightforward to see so we place the calculation in a separate subsection.

C. The Code Rate

We claim and are going to prove that the code $(\mathcal{T}, \mathcal{A})$ possesses code rate $R \geq I(W) - O(\varepsilon)$.

The sample space of the process $\{K_{i \wedge \tau}\}$ is partitioned into the following three events:

$$G := \{1 - \varepsilon 2^{-n} \leq I_\tau \leq 1\}, \\ M := \{\varepsilon 2^{-n} < I_i < 1 - \varepsilon 2^{-n} \text{ for all } i \leq n\}, \\ B := \{0 \leq I_\tau \leq \varepsilon 2^{-n}\}.$$

Compare this to the analysis after Criterion 5. Event G means K_τ is a good channel; corresponding to 2). Event M means $\tau = n$ and K_n is mediocre; corresponding to 4). Event B means K_τ is a bad channel; corresponding to 3).

M is contained in event $\{\tau > n - 1\}$ (that K_i is sufficiently polarized for some i does not happen). By the proof of Inequality (14) we have

$$\mathbb{P}\{\tau > n - 1\} \leq \begin{cases} 1 & \text{if } n - 1 < O(\log |\log \varepsilon|), \\ O(2^{-\frac{n-1}{\mu'}}) & \text{otherwise.} \end{cases}$$

Recall $n := -\mu' \log_2 \varepsilon$, so $n - 1 < O(\log |\log \varepsilon|)$ does not happen as $\varepsilon \rightarrow 0$. The “otherwise” bound $O(2^{-(n-1)/\mu'})$ applies:

$$\mathbb{P}(M) \leq \mathbb{P}\{\tau > n - 1\} = O(2^{-\frac{n-1}{\mu'}}) = O(2^{-n/\mu'}). \quad (15)$$

Use this to rewrite the capacity as follows; here $\mathbb{I}(\bullet)$ is the indicator function of events:

$$\begin{aligned} I(W) &= I_0 = \mathbb{E}[I_\tau] && \text{(by (2))} \\ &= \mathbb{E}[I_\tau \mathbb{I}(G)] + \mathbb{E}[I_\tau \mathbb{I}(M)] + \mathbb{E}[I_\tau \mathbb{I}(B)] && \text{(partition)} \\ &\leq \mathbb{E}[\mathbb{I}(G)] + \mathbb{E}[\mathbb{I}(M)] + \varepsilon 2^{-n} \mathbb{E}[\mathbb{I}(B)] && \text{(see below)} \\ &= \mathbb{P}(G) + \mathbb{P}(M) + \varepsilon 2^{-n} \mathbb{P}(B) && (\mathbb{E} \text{ is } \mathbb{P}) \\ &\leq \mathbb{P}(G) + O(2^{-n/\mu'}) + \varepsilon 2^{-n}. && \text{(by (15))} \end{aligned}$$

Here (see below) is by $I_\tau \leq 1$ for G and M , and by $1 - \varepsilon 2^{-n} \leq I_\tau$ for B . Use the last line to bound the code rate:

$$\begin{aligned} R &= \mathbb{P}\{K_\tau \in \mathcal{A}\} = \mathbb{P}(G) && \text{(by (5))} \\ &\geq I(W) - O(2^{-n/\mu'}) - \varepsilon 2^{-n} && \text{(rewrite } I(W)) \\ &= I(W) - O(\varepsilon) && (n := -\mu' \log_2 \varepsilon) \end{aligned}$$

This proves the claim that $R \geq I(W) - O(\varepsilon)$.

The proof of Theorem 1 ends here.

V. DISCUSSION

We anticipate generalizations of the main theorem to all DMCs. More precisely, let W be any DMC. Let $\{K_i\}$ be the stochastic process of synthetic channels generated by some well-selected ℓ -by- ℓ kernel. Let $\{H_i\}$ be the stochastic process of the conditional entropies of $\{K_i\}$. We ask the following question.

Question 9: Assume that the kernel is polarizing. That is, $\lim_{i \rightarrow \infty} H_i \in \{0, 1\}$. Do there exist $\beta' > 0$ and $\mu' < \infty$ such that as $i \rightarrow \infty$,

$$\mathbb{P}\{1 - H_i < 2^{-\ell^{\beta' i}}\} \geq H(W) - O(\ell^{-i/\mu'})$$

and

$$\mathbb{P}\{H_i < 2^{-\ell^{\beta' i}}\} \geq 1 - H(W) - O(\ell^{-i/\mu'})?$$

Furthermore, if it does, we want to know the answer to the question below.

Question 10: Does there exist a series of pruned polar codes (presumably generated in a similar way) with block length N , code rate R , block error probability P , and per-bit time complexity bC satisfying

$$(N, I(W) - R, P, \text{bC}) = (\varepsilon^{-\mu'}, O(\varepsilon), \varepsilon, O(\log |\log \varepsilon|))$$

as $\varepsilon \rightarrow 0$?

We expect both question can be answered affirmatively as there are promising tools available. For instance, [43] showed how to achieve the true capacity of a non-symmetric channel. The framework in [35] deals with prime-power input alphabet; and [34] deals with arbitrary finite input alphabet. The arguments in [20], [41] seem to be able to show the existence of scaling exponents over more general channels.

APPENDIX

A. Execution of Rule (4)

We present a possible execution of Rule (4), i.e.,

Apply T_{An} to w if and only if $\text{depth}(w) < n$ and $Y(w) > \varepsilon 2^{-n}$.

Let W be a BEC with erasure probability $H(W) = Z(W) = 0.6$; let $\varepsilon = 1.2$. We should have determined n by ε ; but we choose $n = 3$ for simplicity. Note that $\varepsilon 2^{-n} = 0.15$. Moreover, we should have applied T_{An} to *channels*. But for BEC, the Bhattacharyya parameter uniquely determines the channel; so we made a shortcut: by applying T_{An} to a number a to obtain other numbers b and c , we mean to apply T_{An} to the BEC of erasure probability a to obtain two BECs of erasure probabilities b and c , respectively. See Fig. 7 for steps zero to three. See Fig. 8 for steps four to seven.

B. Comments on Lemma 5

[20] proved that there exists polar code with error probability $\exp(-N^\beta)$ and polynomial gap to capacity. We know that the error probability is about the size of H_i when K_i is

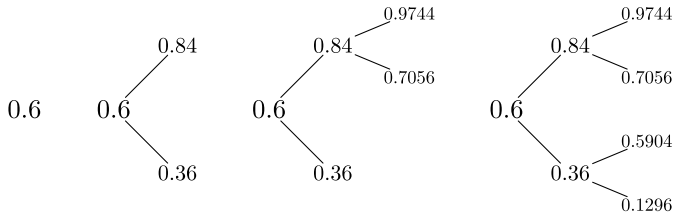


Fig. 7. Step zero: Start with W and write down $Z(W)$, which is 0.6. Step one: Both 0.6 and $1 - 0.6$ are larger than $\varepsilon 2^{-n} = 0.15$. Apply T_{An} to 0.6 to obtain two synthetic channels $1 - (1 - 0.6)^2 = 0.84$ and $0.6^2 = 0.36$. Append them as the children of 0.6. Step two: Both 0.84 and $1 - 0.84$ are larger than $\varepsilon 2^{-n}$. Apply T_{An} to 0.84 to obtain two synthetic channels $1 - (1 - 0.84)^2 = 0.9744$ and $0.84^2 = 0.7056$. Append them as the children of 0.84. Step three: Both 0.36 and $1 - 0.36$ are larger than $\varepsilon 2^{-n}$. Apply T_{An} to 0.36 to obtain two synthetic channels $1 - (1 - 0.36)^2 = 0.5904$ and $0.36^2 = 0.1296$. Append them as the children of 0.36.

a synthetic channel that is used to communicate. Thus their result implies that

$$\begin{aligned} \mathbb{P}\{H_i < e^{-2^{\beta i}}\} &= \text{code rate} \\ &\geq I(W) - \text{polynomial}(\text{block length}) \\ &= I(W) - 2^{-i/O(1)}. \end{aligned}$$

To be even more precise, we now show how [19, Lemmas 2.8 and 3.1 and Theorem 2.5] yield that inequality for Arikan's kernel $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Lemma 3.1 of [19] reads: *Consider $M = \begin{bmatrix} 1 & 0 \\ \alpha & 1 \end{bmatrix}$ for nonzero $\alpha \in \mathbb{F}_q$. For every $\varepsilon > 0$, matrix $M^{\otimes 2}$ satisfies $(1/4, 2 - \varepsilon)$ -exponential polarization. We plug in $\alpha = 1$ and $\varepsilon = 1/4$, then $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes 2}$ satisfies $(1/4, 7/4)$ -exponential polarization.*

Lemma 2.8 of [19] reads: *If mixing matrix M satisfies (η, b) -exponential polarization, then Arikan martingale associated with M is (η, b) -exponentially locally polarizing. We apply this with $(\eta, b) = (1/4, 7/4)$. Realize that the Arikan martingale of $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes 2}$ is the even terms of the usual martingale $\{H_i\}$. Thus, we conclude that $\{H_{2i}\}$ is $(1/4, 7/4)$ -exponentially locally polarizing.*

Theorem 2.5 of [19] reads: *Let $\Lambda < \eta \log_2 b$. Then if a $[0, 1]$ -bounded martingale X_0, X_1, X_2, \dots satisfies (η, b) -exponential[isic] local polarization then it also satisfies Λ -exponentially strong polarization. From here, we know that $\{H_{2i}\}$ satisfies Λ -exponential strong polarization for some $\Lambda < \log_2(7/4)/4$. Take $\Lambda = \log_2(3/2)/4$.*

According to Definition 2.1 of [19], $\{X_t\}$ has Λ -exponentially strong polarization if for every $0 < \gamma < 1$ there exist constants $\alpha < \infty$ and $0 < \rho < 1$ such that for every t ,

$$\mathbb{P}\{2^{-2^{\Lambda t}} < X_t < 1 - \gamma^t\} \leq \alpha \cdot \rho^t.$$

Therefore, from all we got above, we made a choice $\gamma = 1/2$ and obtain

$$\mathbb{P}\{2^{-1.5^{i/4}} < H_{2i} < 1 - 2^{-i}\} \leq \alpha \cdot \rho^i.$$

Similar to Section IV-C, we now partition the sample space into three events

$$\begin{aligned} G &:= \{1 - 2^{-1.5^{i/4}} \leq I_{2i} \leq 1\}, \\ M &:= \{2^{-i} < I_{2i} < 1 - 2^{-1.5^{i/4}}\}, \\ B &:= \{0 \leq I_{2i} \leq 2^{-i}\}. \end{aligned}$$

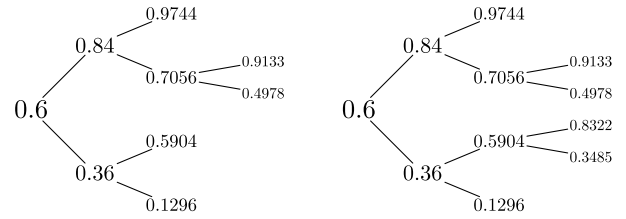


Fig. 8. Steps four and five: $1 - 0.9744$ is smaller than $\varepsilon 2^{-n}$. Do not apply T_{An} ; let 0.9744 be a leaf. Both 0.7056 and $1 - 0.7056$ are larger than $\varepsilon 2^{-n}$. Apply T_{An} to 0.7056 to obtain two synthetic channels $1 - (1 - 0.7056)^2 = 0.9133$ and $0.7056^2 = 0.4978$. Append them as the children of 0.7056. Steps six and seven: Both 0.5904 and $1 - 0.5904$ are larger than $\varepsilon 2^{-n}$. Apply T_{An} to 0.5904 to obtain two synthetic channels $1 - (1 - 0.5904)^2 = 0.8322$ and $0.5904^2 = 0.3485$. Finally 0.1296 is smaller than $\varepsilon 2^{-n}$. Do not apply T_{An} ; let 0.1296 be a leaf. Now we reach depth $n = 3$; terminate.

Note that $\mathbb{P}(M) = \mathbb{P}\{2^{-1.5^{i/4}} < H_{2i} < 1 - 2^{-i}\} = O(\rho^t)$. Then we can rewrite the capacity as follows

$$\begin{aligned} I(W) &= I_0 = \mathbb{E}[I_{2i}] && \text{(martingale)} \\ &= \mathbb{E}[I_{2i}\mathbb{I}(G)] + \mathbb{E}[I_{2i}\mathbb{I}(M)] + \mathbb{E}[I_{2i}\mathbb{I}(B)] && \text{(partition)} \\ &\leq \mathbb{E}[\mathbb{I}(G)] + \mathbb{E}[\mathbb{I}(M)] + 2^{-i}\mathbb{E}[\mathbb{I}(B)] && \text{(by definition)} \\ &= \mathbb{P}(G) + \mathbb{P}(M) + 2^{-i}\mathbb{P}(B) && (\mathbb{E}\mathbb{I} = \mathbb{P}) \\ &\leq \mathbb{P}(G) + O(\rho^i) + 2^{-i}. \end{aligned}$$

Here (by definition) is by $I_{2i} \leq 1$ for G and M , and by $I_i \leq 2^{-i}$ for B .

Keep only $\mathbb{P}(G)$ on the right hand side and move the rest to the left; we finally arrive at $\mathbb{P}(G) \geq I(W) - 2^{-i} - O(\rho^i) = I(W) - O(\rho^i)$ where we assume $1/2 < \rho < 1$ (if not, replace ρ by $1/2$). So

$$\mathbb{P}\{H_{2i} < 2^{-1.5^{i/4}}\} \geq I(W) - O(\rho^i).$$

We are almost there except that we need to bound the odd terms H_{2i+1} . Invoking the fact that $H_{i+1} \leq 2H_i$ (due to martingale), we know that an odd term is at most twice its preceding even term. That is to say,

$$\mathbb{P}\{H_{2i+1} < 2 \cdot 2^{-1.5^{i/4}}\} \geq I(W) - O(\rho^i).$$

Now choose a $\beta' > 0$ such that $2^{2\beta'} < 1.5^{1/4}$ and choose a $\mu' < \infty$ such that $2^{-2/\mu'} > \rho$. Then we conclude Inequality (11) of LEMMA 5.

C. Proof of Lemma 6

Fix a prime q . Fix a q -ary-input discrete symmetric memoryless channel W . We want to find constants $\mu' < \infty$ and $\beta' > 0$ such that the process $\{I_i\}$ satisfies

$$\mathbb{P}\{I_i \leq 2^{-2^{\beta' i}}\} \geq 1 - I(W) - O(2^{-i/\mu'}).$$

We borrow terminologies and lemmas from [41] for a head start.

By [40, Definition 1.8], the matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is mixing. By [ibid., Theorem 1.10], the process $\{I_i\}$ corresponding to $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is locally polarizing. By [ibid., Theorem 1.6], the process $\{I_i\}$ corresponding to $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ is strongly polarizing. By [ibid., Definition 1.4], the process $\{I_i\}$ is such that for all $\gamma > 0$ there

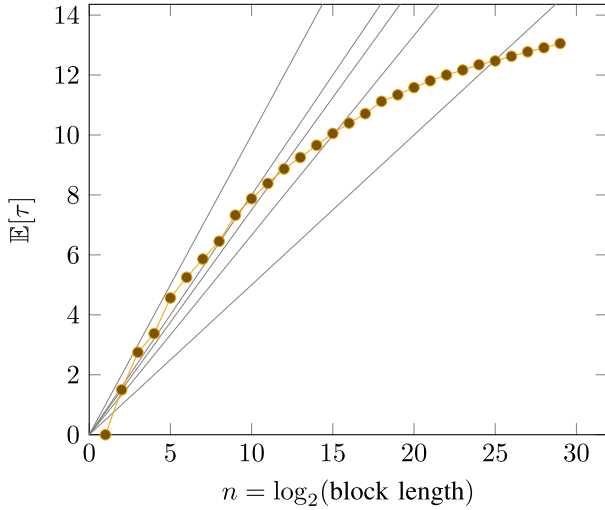


Fig. 9. A script grows trees using Rule 4 and computes the exact $\mathbb{E}[\tau]$ accordingly. The result for various n is shown above. The gray, thin rays are of slopes 1, $4/5$, $3/4$, $2/3$, and $1/2$, respectively.

exist $\eta < 1$ and $\beta < \infty$ such that I_i is $(\gamma^i, \beta\eta^i)$ -polarizing. By [ibid., Definition 1.2], I_i is such that for all $\gamma > 0$ there exist $\eta < 1$ and $\beta < \infty$ such that $\mathbb{P}\{I_i \in (\gamma^i, 1 - \gamma^i)\} < \beta\eta^i$.

Choose $\gamma = 1/2$. We obtain: there exists $\eta < 1$ such that $\mathbb{P}\{I_i \in (2^{-i}, 1 - 2^{-i})\} < O(\eta^i)$. Since $\eta < 1$, the right hand side $O(\eta^i)$ converges to 0 exponentially fast. This means that the majority of I_i are either exponentially small (i.e., $0 \leq I_i \leq 2^{-i}$) or exponentially close to 1 (i.e., $1 - 2^{-i} \leq I_i \leq 1$). What we want to show consists of two parts: 1) The proportion of I_i that is exponentially small is about $1 - I(W)$; the proportion of I_i that is exponentially close to 1 is about $I(W)$. 2) Exponentially small I_i 's are doubly-exponentially small (i.e., $0 \leq I_i \leq 2^{-2^{\beta^i}}$); the close-to-1 counterpart is doubly-exponentially close to 1 (i.e., $1 - 2^{-2^{\beta^i}} \leq I_i \leq 1$). (Remark: part of the statement overlaps Lemma 5; we state for both noisy-end and reliable-end for better comparison.)

Now we go for 1). Observation: the result we want to prove and the tool we have in hand are symmetric in I_i and in $1 - I_i$. It suffices to show, say, that the small- I_i part of the statement holds. The close-to-1 part follows by symmetry (or by what we have done in the previous appendix).

Now we show $\mathbb{P}\{0 \leq I_i \leq 2^{-i}\} \geq I(W) - O(\eta^i)$. Similar to Sections B and IV-C, we partition the sample space into three events

$$\begin{aligned} G &:= \{1 - 2^{-i} \leq I_i \leq 1\}, \\ M &:= \{2^{-i} < I_i < 1 - 2^{-i}\}, \\ B &:= \{0 \leq I_i \leq 2^{-i}\}. \end{aligned}$$

Then $\mathbb{P}(M) = \mathbb{P}\{I_i \in (2^{-i}, 1 - 2^{-i})\} = O(\eta^i)$. Next we rewrite the conditional entropy

$$\begin{aligned} 1 - I(W) &= 1 - I_0 = \mathbb{E}[1 - I_i] && \text{(martingale)} \\ &= \mathbb{E}[(1 - I_i)\mathbb{I}(G)] + \mathbb{E}[(1 - I_i)\mathbb{I}(M)] + \mathbb{E}[(1 - I_i)\mathbb{I}(B)] \\ &\leq 2^i \mathbb{E}[\mathbb{I}(G)] + \mathbb{E}[\mathbb{I}(M)] + \mathbb{E}[\mathbb{I}(B)] && \text{(by definition)} \\ &= 2^i \mathbb{P}(G) + \mathbb{P}(M) + \mathbb{P}(B) && (\mathbb{E}\mathbb{I} = \mathbb{P}) \\ &\geq 2^{-i} + O(\eta^i) + \mathbb{P}(B) \end{aligned}$$

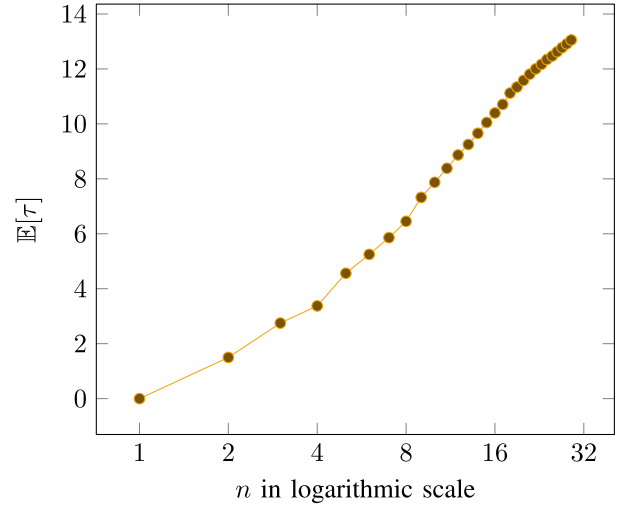


Fig. 10. The same figure as Fig. 9 with log-axis to show the linearity of the data points.

Here (by definition) is by $I_i \geq 0$ for B and M , and by $I_i \geq 1 - 2^{-i}$ for G . Already we have that $\mathbb{P}\{0 \leq I_i \leq 2^{-i}\} = \mathbb{P}(B) \geq 1 - I(W) - O(\eta^i) - O(2^{-i})$. We may assume $\eta > 1/2$. Thus $\mathbb{P}\{0 \leq I_i \leq 2^{-i}\} \geq 1 - I(W) - O(\eta^i)$. The flipped version $\mathbb{P}\{1 - 2^{-i} \leq I_i \leq 1\} \geq I(W) - O(\eta^i)$ also holds by mirroring the argument. This finishes the 1) part.

Now we go for the small- I_i part of 2). We need a lemma; [40, Lemma 6.3] reads: *Let $X_1, X_2 \in \mathbb{F}_q$ be a pair of random variables, and let A_1, A_2 be pair of discrete random variables, such that (X_1, A_1) and (X_2, A_2) are independent. Then*

$$\begin{aligned} 1 - H(X_1 + X_2 | A_1, A_2) \\ \leq (1 - H(X_1 | A_1))(1 - H(X_2 | A_2)) \cdot \text{poly}(q). \end{aligned}$$

Write the constant $\text{poly}(q)$ as Q . Let (X_1, A_1) and (X_2, A_2) be the two copies of K_i , then $(X_1 + X_2 | A_1, A_2)$ is the upper child K_{i+1}^+ . Furthermore, “ $1 - H$ ” is “ I ,” so the lemma means

$$I(K_{i+1}^+) \leq I(K_i)I(K_i)Q,$$

which simplifies to $I_{i+1} \leq QI_i^2$ whenever K_{i+1} is the upper child.

With the lemma in place, we can really start dealing with the small- I_i part of 2). Clearly $I_i < 1/Q^2$ implies $QI_i^2 \leq I_i^{1.5}$. So we deduce that whenever $I_i < 1/Q^2$ and K_{i+1} is the upper child, $I_{i+1} \leq I_i^{1.5}$. Another case is when K_{i+1} is the lower child. We enlarge Q such that $Q \geq 2^5$ (that is, we replace Q with $\max(Q, 2^5)$). Then whenever $I_i < 1/Q^2$ and K_{i+1} is the lower child, $I_{i+1} \leq 2I_i \leq Q^{0.2}I_i < I_i^{-0.1}I_i = I_i^{0.9}$. Combining the two cases of K_{i+1} , we find that if $I_i < 1/Q^2$ then I_{i+1} is (at most) $I_i^{1.5}$ or $I_i^{0.9}$, each with probability $1/2$. We conclude this paragraph by rewriting this formally: when $I_i < 1/Q^2$,

$$I_{i+1} \leq \begin{cases} I_i^{1.5} & \text{w.p. } 1/2 \text{ (upper child case),} \\ I_i^{0.9} & \text{w.p. } 1/2 \text{ (lower child case).} \end{cases}$$

Let n be a large number. We know $\mathbb{P}\{0 \leq I_n \leq 2^{-n}\} \geq 1 - I(W) - O(\eta^n)$. Now we continue the process for $i = n, \dots, 4n$. We want to show that at step $4n$, the bad channels

have doubly-exponentially small capacity. That is, we want $\mathbb{P}\{0 \leq I_{4n} \leq 2^{-2^{4\beta'n}}\} \geq 1 - I(W) - O(\eta^{4n})$ for some β' . There are two obstacles: a) If $I_i \geq 1/Q^2$, we lose control on I_{i+1} . We want to avoid this. b) Even if $I_i < 1/Q^2$, we want I_i to go through the 1.5-th power instead of the 0.9-th power. Let A be the event that $I_n < 2^{-n}$ but $I_i > 1/Q^2$ for some $n < i < 4n$. When that happens, let σ be the lowest i such that $I_i > 1/Q^2$. When that does not happen, we let σ be $4n$. Let B be the event that among $3n$ chances, I_i undergoes the 1.5-th power less than n times. We now control A and B .

For A , we have $\mathbb{P}(A) = \mathbb{P}\{I_\sigma \geq 1/Q^2\} \leq \mathbb{E}[I_\sigma]Q^2 \leq \mathbb{E}[I_n]Q^2 \leq Q^2 2^{-n}$ by [37, Theorem 5.7.6]. For B , by Hoeffding's inequality [44, Theorem 2.8], there exists $\rho < 1$ such that $\mathbb{P}(B) < O(\rho^n)$. Enlarge $\eta < 1$ by replacing it with $\max(\eta, \rho)$. We see that both A and B are rare events in the sense that their probability measures are both in $O(\eta^n)$.

Finally we look at what happens outside $A \cup B$: If $I_n < 2^{-n}$ and neither A nor B happens, then I_i undergoes the 1.5-th power n times, at least; and undergoes the 0.9-th power $2n$ times, at most. Thus I_{4n} is at most I_n to the $(1.5^n \cdot 0.9^{2n})$ -th power. The exponent $1.5^n \cdot 0.9^{2n}$ is at least $2^{0.28n}$, so $I_{4n} \leq (2^{-n})^{2^{0.28n}} \leq 2^{-2^{0.28n}} = 2^{-2^{0.07 \cdot 4n}}$.

We review what we have so far: First the probability that $0 \leq I_n \leq 2^{-n}$ is at least $1 - I(W) - O(\eta^n)$. And then we continue the process for $i = n, \dots, 4n$. We lose some I_i in A ; this costs us $Q^2 2^{-n}$. We lose some I_i in B ; this costs us $O(\eta^n)$. As $n \rightarrow \infty$ the constant Q does not matter; we lose $2O(\eta^n)$. What are left are some I_i such that $I_{4n} \leq 2^{-2^{0.07 \cdot 4n}}$. Therefore, we have just proven that $\mathbb{P}\{I_{4n} \leq 2^{-2^{0.07 \cdot 4n}}\} \leq 1 - I(W) - 3O(\eta^n)$. Now we choose $\mu' > 0$ and $\beta' > 0$ such that $\mathbb{P}\{I_{4n} \leq 2^{-2^{4\beta'n}}\} \geq 1 - I(W) - O(2^{-4n/\mu'})$. This finishes the small- I_i part of 2).

For the close-to-1 part of 2), [20] (i.e., Lemma 5) has that $\mathbb{P}\{1 - I_n \leq 2^{-2^{\beta'n}}\} \geq I(W) - O(2^{-n/\mu'})$ for some constants β', μ' . One can also prove it barehanded by applying the same trick we did for the small- I_i part of 2) to the Bhattacharyya parameters. This is the last piece of the proof. Now 2) is finished. The proof of Lemma 6 is complete.

D. Simulation

We write a script to support Theorem 1. The script: 1) sets $I(W) = 0.6$; 2) loops for $n = 1, \dots, 29$; 3) for each n , evaluates $2^{-n/4.9}$ as ε ; 4) generates the channel tree by Rule 4; and 5) traverses the tree to compute the exact $\mathbb{E}[\tau]$. The $\mathbb{E}[\tau]$ are plotted in Figs. 9 and 10.

Notice that, for one, the curve in Fig. 9 *does not* grow proportionally to n (recall that $\tau = n$ for classical polar codes). For two, it grows linearly in Fig. 10, which reassures $\mathbb{E}[\tau] = O(\log \log N) = O(\log n)$ as Lemma 8 stated.

Starting from $n \geq 10$, our scheme prunes 20% of butterfly devices. After $n \geq 12$ it saves 25% of butterfly devices. At $n = 16$ it reduces by 33% for the first time. Once $n \geq 25$, one-half are left.

REFERENCES

- [1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.
- [2] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.
- [3] A. Barg and G. D. Forney, "Random codes: Minimum distances and error exponents," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2568–2573, Sep. 2002.
- [4] V. Strassen, "Asymptotische abschätzungen in Shannons informationstheorie," in *Trans. 3rd Prague Conf. Inf. Theory*. Prague, Czechia: Publishing House of the Czechoslovak Academy of Sciences, 1962, pp. 689–723. [Online]. Available: <https://www.math.cornell.edu/~pmlut/strassen.pdf>
- [5] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
- [6] Y. Altug and A. B. Wagner, "Moderate deviation analysis of channel coding: Discrete memoryless case," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 265–269.
- [7] Y. Polyanskiy and S. Verdú, "Channel dispersion and moderate deviations limits for memoryless channels," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2010, pp. 1334–1339.
- [8] Y. Altug and A. B. Wagner, "Moderate deviations in channel coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4417–4426, Aug. 2014.
- [9] M. Hayashi and V. Y. F. Tan, "Erasure and undetected error probabilities in the moderate deviations regime," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1821–1825.
- [10] E. Arikan, "A packing lemma for polar codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 2441–2445.
- [11] S. Kudekar, S. Kumar, M. Mondelli, H. D. Pfister, E. Şaşıoğlu, and R. L. Urbanke, "Reed–Muller codes achieve capacity on erasure channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4298–4316, Jul. 2017.
- [12] S. Kudekar, T. Richardson, and R. L. Urbanke, "Spatially coupled ensembles universally achieve capacity under belief propagation," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 7761–7813, Dec. 2013.
- [13] H. D. Pfister, I. Sason, and R. Urbanke, "Capacity-achieving ensembles for the binary erasure channel with bounded complexity," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2352–2379, Jul. 2005.
- [14] H. D. Pfister and I. Sason, "Accumulate-repeat-accumulate codes: Capacity-achieving ensembles of systematic codes for the erasure channel with bounded complexity," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2088–2115, Jun. 2007.
- [15] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, Jul. 2009.
- [16] V. Guruswami and P. Xia, "Polar codes: Speed of polarization and polynomial gap to capacity," in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci.*, Oct. 2013, pp. 310–319.
- [17] M. Mondelli, S. H. Hassani, and R. L. Urbanke, "Unified scaling of polar codes: Error exponent, scaling exponent, moderate deviations, and error floors," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 6698–6712, Dec. 2016.
- [18] S. Fong and V. Tan, "Scaling exponent and moderate deviations asymptotics of polar codes for the AWGN channel," *Entropy*, vol. 19, no. 7, p. 364, Jul. 2017. [Online]. Available: <http://www.mdpi.com/1099-4300/19/7/364>
- [19] J. Błasiok, V. Guruswami, and M. Sudan, "Polar codes with exponentially small error at finite block length," 2018, *arXiv:1810.04298*. [Online]. Available: <http://arxiv.org/abs/1810.04298>
- [20] J. Błasiok, V. Guruswami, and M. Sudan, "Polar codes with exponentially small error at finite block length," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)* (Leibniz International Proceedings in Informatics (LIPIcs)), vol. 116, E. Blais, K. Jansen, J. D. P. Rolim, and D. Steurer, Eds. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018, p. 34. [Online]. Available: <http://drops.dagstuhl.de/opus/volltexte/2018/9438>
- [21] H.-P. Wang and I. Duursma, "Polar-like codes and asymptotic trade-off among block length, code rate, and error probability," 2018, *arXiv:1812.08112*. [Online]. Available: <http://arxiv.org/abs/1812.08112>
- [22] H.-P. Wang and I. Duursma, "Polar codes' simplicity, random codes' durability," 2019, *arXiv:1912.08995*. [Online]. Available: <http://arxiv.org/abs/1912.08995>
- [23] E. Arikan and E. Telatar, "On the rate of channel polarization," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2009, pp. 1493–1495.
- [24] A. Alamdar-Yazdi and F. R. Kschischang, "A simplified successive-cancellation decoder for polar codes," *IEEE Commun. Lett.*, vol. 15, no. 12, pp. 1378–1380, Dec. 2011.

- [25] M. El-Khamy, H. Mahdaviifar, G. Feygin, J. Lee, and I. Kang, "Relaxed polar codes," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 1986–2000, Apr. 2017.
- [26] L. Zhang, C. Zhong, L. Ping, Z. Zhang, and X. Wang, "Simplified successive-cancellation decoding using information set reselection for polar codes with arbitrary blocklength," *IET Commun.*, vol. 9, no. 11, pp. 1380–1387, Jul. 2015.
- [27] Y. Zhang, Q. Zhang, X. Pan, Z. Ye, and C. Gong, "A simplified belief propagation decoder for polar codes," in *Proc. IEEE Int. Wireless Symp. (IWS)*, Mar. 2014, pp. 1–4.
- [28] M. El-Khamy, H. Mahdaviifar, G. Feygin, J. Lee, and I. Kang, "Relaxed channel polarization for reduced complexity polar coding," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2015, pp. 207–212.
- [29] D. Wu, A. Liu, Q. Zhang, and Y. Zhang, "Concatenated polar codes based on selective polarization," in *Proc. 12th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2015, pp. 436–442.
- [30] A. Elkelesh, M. Ebada, S. Cammerer, and S. ten Brink, "Flexible length polar codes through graph based augmentation," in *Proc. 11th Int. ITG Conf. Syst. Commun. Coding*, Feb. 2017, pp. 1–6.
- [31] X. Wu, L. Yang, and J. Yuan, "Information coupled polar codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 861–865.
- [32] X. Wu, L. Yang, Y. Xie, and J. Yuan, "Partially information coupled polar codes," *IEEE Access*, vol. 6, pp. 63689–63702, 2018.
- [33] M. Mondelli, S. A. Hashemi, J. Cioffi, and A. Goldsmith, "Sublinear latency for simplified successive cancellation decoding of polar codes," *IEEE Trans. Wireless Commun.*, early access, Sep. 16, 2020, doi: [10.1109/TWC.2020.3022922](https://doi.org/10.1109/TWC.2020.3022922).
- [34] E. Sasoglu, E. Telatar, and E. Arikan, "Polarization for arbitrary discrete memoryless channels," in *Proc. IEEE Inf. Theory Workshop*, Oct. 2009, pp. 144–148.
- [35] R. Mori and T. Tanaka, "Source and channel polarization over finite fields and Reed–Solomon matrices," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2720–2736, May 2014.
- [36] R. G. Gallager, *Stochastic Processes: Theory for Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2013.
- [37] R. Durrett, *Probability: Theory Examples*, 4th ed. New York, NY, USA: Cambridge Univ. Press, 2010. [Online]. Available: https://services.math.duke.edu/~rtd/PTE/PTE4_1.pdf
- [38] S. B. Korada, A. Montanari, E. Telatar, and R. Urbanke, "An empirical scaling law for polar codes," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 884–888.
- [39] S. H. Hassani, K. Alishahi, and R. L. Urbanke, "Finite-length scaling for polar codes," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 5875–5898, Oct. 2014.
- [40] J. Błasiok, V. Guruswami, P. Nakkiran, A. Rudra, and M. Sudan, "General strong polarization," 2018, *arXiv:1802.02718*. [Online]. Available: <http://arxiv.org/abs/1802.02718>
- [41] J. Błasiok, V. Guruswami, P. Nakkiran, A. Rudra, and M. Sudan, "General strong polarization," in *Proc. 50th Annu. ACM SIGACT Symp. Theory Comput.* New York, NY, USA: Association for Computing Machinery, 2018, pp. 485–492, doi: [10.1145/3188745.3188816](https://doi.org/10.1145/3188745.3188816).
- [42] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 259–266, Jan. 1994.
- [43] J. Honda and H. Yamamoto, "Polar coding without alphabet extension for asymmetric models," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 7829–7838, Dec. 2013.
- [44] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities*. Oxford, U.K.: Oxford Univ. Press, 2013. [Online]. Available: <https://doi.org/10.1093/acprof:oso/9780199535255.001.0001>

Hsin-Po Wang received the bachelor's degree in mathematics from National Taiwan University in 2015. He is currently pursuing the Ph.D. degree with the Department of Mathematics, University of Illinois at Urbana–Champaign.

Iwan M. Duursma received the Ph.D. degree in mathematics from the University of Eindhoven in 1993. After positions with CNRS IML Luminy, University of Puerto Rico, Bell-Labs, AT&T Research, and University of Limoges, he is currently a Professor with the University of Illinois at Urbana–Champaign.