Polar-like Codes and Asymptotic Tradeoff among Block Length, Code Rate, and Error Probability

Hsin-Po Wang and Iwan Duursma
University of Illinois at Urbana–Champaign
{hpwang2, duursma}@illinois.edu

Abstract—A general framework is proposed that includes polar codes over arbitrary channels with arbitrary kernels. The asymptotic tradeoff among block length $N$, code rate $R$, and error probability $P$ is analyzed.

Given a tradeoff between $N, P$ and a tradeoff between $N, R, P$ (Theorem 5). Quantitatively, if $P = \exp(-N^\beta)$ is possible for some $\beta^*$ and if $R = \text{Capacity} - N^{1/\mu^*}$ is possible for some $1/\mu^*$, then $(P, R) = (\exp(-N^{\beta^*}), \text{Capacity} - N^{-1/\mu^*})$ is possible for some pair $(\beta', 1/\mu')$ determined by $\beta^*$, $1/\mu^*$, and auxiliary information. In fancy words, an error exponent regime tradeoff catches up the code rate and error probability of random codes all at once. (Section IX.)

+ Analyzing Arıkan's codes over BEC; [FT17] analyzing Arıkan's codes over general channels. An attempt is made to generalize with complexity slightly larger than $N$. (10)

+ Quantitatively, if $P, R, N$ we return an interpolating tradeoff among block length $N$, code rate $R$, and error probability $P$. (Theorem 5).

+ It has been shown before that the pair of ratios for block codes lies in

\[
(0, 1/2), (0, 0), (1, 0)
\]

and random codes achieve the hypotenuse. This motivates two questions: whether polar codes can achieve the hypotenuse (yes for BEC) and what price we pay in terms of complexity (slightly more than $N \log N$).

See Section IX for big pictures.

A. Channel polarization

Channel polarization [Ari09] is a method to synthesize some channels to form some extremely-unreliable channels and some extremely-reliable channels. The users then can transmit uncoded messages through extremely-reliable ones while transmitting predictable symbols through extremely-unreliable ones.

We summarize channel polarization as follows. Say we are going to communicate over this BEC

\[
\begin{align*}
\text{W} & \\
\end{align*}
\]

We have two magic devices

\[
\begin{align*}
\text{A} & \\
\text{B} & \\
\end{align*}
\]

and

\[
\begin{align*}
\text{W} & \\
\end{align*}
\]

such that if we wire two i.i.d. instances of $W$ as follows

\[
\begin{align*}
\begin{array}{c}
\text{W} \\
\end{array} & \begin{array}{c}
\text{A} \\
\end{array} \\
\begin{array}{c}
\text{B} \\
\end{array} & \begin{array}{c}
\text{D} \\
\end{array}
\end{align*}
\]

then pin $A$ to pin $B$ forms a less reliable synthetic channel $W^\circ$, while pin $C$ to pin $D$ forms a more reliable synthetic channel $W^\otimes$. Graphically, Formula (9) is equivalent to

\[
\begin{align*}
\begin{array}{c}
\text{W} \\
\end{array} & \begin{array}{c}
\text{A} \\
\end{array} \\
\begin{array}{c}
\text{W} \\
\end{array} & \begin{array}{c}
\text{B} \\
\end{array}
\end{align*}
\]

Formula (9) being the base step, the next step is to duplicate Formula (9) and wire them as

\[
\begin{align*}
\begin{array}{c}
\text{W} \\
\end{array} & \begin{array}{c}
\text{W} \\
\end{array} \\
\begin{array}{c}
\text{W} \\
\end{array} & \begin{array}{c}
\text{W} \\
\end{array}
\end{align*}
\]

I. INTRODUCTION

In the theory of two-terminal error correcting codes, three of the most important parameters of block codes are block length $N$, code rate $R$, and error probability $P$. Though we want codes with small $N$, higher $R$, and lower $P$, these goals contradict each other. Thus it becomes essential to quantify the tradeoffs.

Given a memoryless channel $W$ with symmetric capacity $I(W)$, there exists polar codes with

\[
\log(-\log P) \in \Theta(\log N) \quad \text{as } N \to \infty.
\]

It is also shown that there exist polar codes with

\[
-\log(I(W) - R) \in \Theta(\log N) \quad \text{as } N \to \infty.
\]

This work aims to characterize the pairs of ratios

\[
\left( \liminf_{N \to \infty} \frac{\log(-\log P)}{\log N}, \liminf_{N \to \infty} \frac{-\log(I(W) - R)}{\log N} \right)
\]

that are realized by polar codes.
which is equivalent to four synthetic channels as

\[
\begin{array}{c}
W^b \\
W^b \\
W^b \\
\end{array}
\]

or simply

\[
\begin{array}{c}
W^b \\
W^b \\
\end{array}
\]

Further wire Formula (11) as

\[
\begin{array}{c}
W \\
W \\
\end{array}
\]

which is equivalent to

\[
\begin{array}{c}
W^b \\
W^b \\
\end{array}
\]

to

\[
\begin{array}{c}
W^b \\
W^b \\
\end{array}
\]

and to

\[
\begin{array}{c}
(W^b)^b \\
(W^b)^b \\
\end{array}
\]

Here \((W^b)^b\) is a synthetic channel less reliable than \(W^b\); synthetic channel \((W^b)^b\) is more reliable than \(W^b\); synthetic channel \((W^b)^b\) is less reliable than \(W^b\); and synthetic channel \((W^b)^b\) is more reliable than \(W^b\).

After Formula (14), the next, larger construction is two copies of Formula (14) plus four more pairs of magic devices.

It is equivalent to

\[
\begin{array}{c}
W^b \\
W^b \\
\end{array}
\]

to

\[
\begin{array}{c}
(W^b)^b \\
(W^b)^b \\
\end{array}
\]

and finally to

\[
\begin{array}{c}
(W^b)^b \\
(W^b)^b \\
\end{array}
\]

Here \(((W^b)^b)^b\) is a synthetic channel less reliable than \((W^b)^b\); etc.

After Formula (18), the next, larger construction is going to be two copies of Formula (18) plus one extra layer of magic devices.

The game goes on endlessly. Arikan then observes that synthetic channels generated in this way tend to be either extremely reliable or extremely unreliable. That is to say, they polarize.

**B. Channel polarization in Tree Notation**

Draw

\[
W^b \\
T_{An} \\
W^b \\
\]

(24)
to capture the fact that Formula (14)

transforms two instances of $W$ into a $W^\circ$ and a $W^\flat$. We will later call this tree $T_{\text{rect}}^{\text{per}}(W, T_{\text{An}}, 1)$ (guess why). Similarly, draw

$$
\begin{align*}
&\quad \begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
W \\
W \\
W \\
W \\
W \\
W \\
W \\
W \\
\end{array}
\quad
\begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\end{array}
\end{align*}
$$

(25)

to capture the fact that Formula (18)

transforms four instances of $W$ into two pairs of $W^\circ$ and $W^\flat$. Two $W^\circ$ are then transformed into a $(W^\circ)^{\circ}$ and a $(W^\circ)^{\flat}$; two $W^\flat$ are then transformed into a $(W^\flat)^{\circ}$ and a $(W^\flat)^{\flat}$. We will later call this tree $T_{\text{rect}}^{\text{per}}(W, T_{\text{An}}, 2)$ (guess why). Similarly, draw

$$
\begin{align*}
&\quad \begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
W \\
W \\
W \\
W \\
W \\
W \\
W \\
W \\
\end{array}
\quad
\begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\end{array}
\end{align*}
$$

(26)

to capture Formula (18)

That is, eight instances of $W$ are transformed into four pairs of $W^\circ, W^\flat$, into two quadruples of $(W^\circ)^{\circ}, (W^\circ)^{\flat}, (W^\flat)^{\circ}, (W^\flat)^{\flat}$, and finally into $((W^\circ)^{\circ})^{\circ}, ((W^\circ)^{\flat})^{\flat}, ((W^\flat)^{\circ})^{\circ}, ((W^\flat)^{\flat})^{\flat}$. We will later call this tree $T_{\text{rect}}^{\text{per}}(W, T_{\text{An}}, 3)$ (guess why).

It is not hard to imagine that the next construction will transform sixteen instances of $W$ to “some intermediate things”, and finally to $(((W^\circ)^{\circ})^{\circ})^{\circ}$ to $(((W^\flat)^{\flat})^{\flat})^{\flat}$.

C. Generalize the Tree Notation

The tree notation comes with generalizations.

1) Arbitrary Polar Kernels: [KSU10] Given, say, an $\ell$-by-$\ell$ matrix $G_{\text{KSU}}$ as a kernel, it induces a transformation $T_{\text{KSU}}$. We may draw an $\ell$-ary tree, starting from

$$
\begin{align*}
&\quad \begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\end{array}
\quad
\begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\end{array}
\end{align*}
$$

(27)

instead of a binary tree. This, when $\ell = 7$, translates into the circuit setup

$$
\begin{align*}
&\quad \begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\end{array}
\quad
\begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\end{array}
\end{align*}
$$

(28)

Here the top pair of pins forms $W^{(1)}$, and the bottom pair of pins forms $W^{(7)}$. We will later call this tree $T_{\text{rect}}^{\text{per}}(W, T_{\text{KSU}}, 1)$ (guess why).

2) Unbalanced Tree: This is motivated by attempts of optimization of polar codes. The generalization comes in two perspectives.

First perspective [AYK11], [SG13], [SGV+14], [ZZW+15], [ZZP+14]: in a tree like Formula (26) or a larger tree, it could be the case some synthetic channel, say $(W^\circ)^{\circ}$, is so bad that applying further transformations sounds useless. If so, we may remove children of $(W^\circ)^{\circ}$ to get

$$
\begin{align*}
&\quad \begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\end{array}
\quad
\begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\end{array}
\end{align*}
$$

(29)

which translates into the circuit

$$
\begin{align*}
&\quad \begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\end{array}
\quad
\begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\bigcirc \\
\end{array}
\end{align*}
$$

(30)

That is, eight instances of $W$ are transformed into four pairs of $W^\circ, W^\flat$, into two quadruples of $(W^\circ)^{\circ}, (W^\circ)^{\flat}, (W^\flat)^{\circ}, (W^\flat)^{\flat}$, and, notice the difference, while keeping two $(W^\circ)^{\circ}$, the other six are transformed into $((W^\circ)^{\circ})^{\circ}, ((W^\circ)^{\flat})^{\flat}, ((W^\flat)^{\circ})^{\circ}, ((W^\flat)^{\flat})^{\flat}$. Second perspective [EkMf+15], [WLZ15], [递II17], [EkMf+17], [WYY18], [WXY18]: in a tree like Formula (26), it could be that some synthetic channel, say $(W^\circ)^{\circ}$, might not polarize enough, i.e., it is neither extremely good
nor extremely bad. Thus we further polarize it by applying additional $T_{An}$ as follows:

$$ W^{\ell} \xrightarrow{T_{An}} (W^{\ell})^2, \xrightarrow{T_{An}} (W^{\ell})^3, \xrightarrow{T_{An}} (W^{\ell})^4, $$

which translates into the circuit

![Circuit Drawing](image)

That is, eight instances of $W$ are transformed into four pairs of $W^2, W^3, W^4$, into two quadruples of $(W^2)^2, (W^2)^3, (W^2)^4$, and, notice another difference, only the two $(W^2)^4$ are transformed into $((W^2)^2)^2, ((W^2)^3)^2$.

3) Multi-Kernel: \[BBGL17\], \[BGLB17\], \[GBLB17\], \[BCL18\] The Transformation $T_{An}$ generates codes whose block lengths are powers of 2. A transformation $T_{XFM}$ induced by an $\ell$-by-$\ell$ matrix generates codes whose block lengths are powers of $\ell$. For something in between, say length-72, one can apply $T_{An}$ three times and then apply $T_{GGLB}$, a transformation induced by a 3-by-3 matrix, two times.

Here is a small (length-6) example. First $T_{An}$ is applied, and then $T_{GGLB}$ is applied.

![Circuit Drawing](image)

This translates into the circuit drawn below.

![Circuit Drawing](image)

4) Alphabet Extension: \[PSL11\], \[PSL16\] There is a special type of channel transformations corresponding to field extensions $F_q \subset F_p$ for any $q, k$. That is to say, $k$ independent copies of a $q$-ary erasure channel can transmit a $q^k$-bit. We claim an erasure if any of $k$ symbols in the ground field misses. Denote the transformation by $T_{\mathcal{C}}$. Draw

\[ W^{T_{\mathcal{C}}} \xrightarrow{T_{\mathcal{C}}} W^k \] for $W^k$ the $k$-th power of the channel $W$. Here is a $k = 5$

5) Some Convention: Although it is theoretically possible for a tree to have multiple, nested $T_{\mathcal{C}}^k$, each with different parameters $k$, we limit our interest to two small classes of trees. They are

- trees consisting of one transformation $T$ (that is not $T_{\mathcal{C}}^k$);
- trees consisting of three transformations $T_{rat}, T_{\mathcal{C}}^k, T_{err}$, wherein
  - all $T_{rat}^k$ correspond to the same parameter $k$,
  - every root-to-leaf path passes exactly one $T_{\mathcal{C}}^k$,
  - every root-to-$T_{rat}^k$ path passes only $T_{rat}$, and
  - every $T_{rat}^k$-to-leaf path passes only $T_{err}$.

That said, in the second case, $T_{rat}^k$ might not locate at the same depth. It turns out allowing $T_{rat}^k$ to be at different depths bursts the performance, theoretically and practically.

Denote by $\mathcal{T}$ a tree of channels with root channel $W$.

D. Bhattacharyya Parameter and Process

The Bhattacharyya parameter $Z(w)$ of a channel $w$ measures the unreliability, the badness, of the channel. For instance for BDMC

\[ I(w) + Z(w) \geq 1, \]
\[ I(w) + Z(w)^2 \leq 1, \]
\[ I(w) \log 2 + Z(w) \leq 1, \]

by \[JAT18\] Corollary 5. That is to say, this pair of parameters $(I(w), Z(w))$ lies in

\[ (0, 1), (.8, .44), (1, 0). \]

That said, an explicit definition of Bhattacharyya parameters is not presented here since all we need in this work is the following two properties playing as axioms:

- (Regarding transformations) \[MT14\] Lemma 33] For any transformation $T$ we are interested in, it has an operator norm $|T|$ such that for any channel $w$ we are interested in and any outcome $v$ of $T(w)$, the multiple $|T|Z(w)$ bounds $Z(v)$ from above.
- (Regarding error probability) \[MT14\] Lemma 22] For any $q$-ary channel $w$ we are interested in, the multiple $qZ(w)$ bounds from above the probability that a decoder fails to decode a single symbol transmitted through $w$.

Given the nice properties, the general strategy is to fully control $Z(w)$ for as many $w$ as possible in a tree, and then rewrite the resulting inequalities in terms of error probabilities. During this process, it is not important anymore what the
Bhattacharyya parameter is. In theory, it could be replaced by any function that satisfies the aforementioned two axioms. Starting from Section II-C we will call it Z-parameter instead.

Given a channel tree $T$ with root channel $W_r$, define two discrete-time stochastic processes $W_i, Z_i$ and a stopping time $\tau$ as follows: Start from the root channel $W_0 := W$; and let $Z_0 := Z(W_0)$. For any $i \ge 0$, if $W_i$ is a leaf channel, let $\tau := i$. If, otherwise, $W_i$ has children, choose a child uniformly at random as $W_{i+1}$; and let $Z_{i+1} := Z(W_{i+1})$. (c.f. [Ari09 Section IV, third paragraph].)

In case of Ankan’s polar codes, $Z_i$ is a martingale over BEC and is a super-martingale over other BDMC [Ari09 Proposition 9]. For other binary kernels over general BDMC, [KSU10 Remark 5] claims that it is difficult to characterize, but they manage to prove a useful statement [KSU10, Lemma 10]. For larger alphabets, [MTT4 Lemma 33] claims that it is very similar to the binary case. We provide our generalization in Lemma II.

For a tree $T$ as in Formula (31), a possible instance of the process is

$$W_0 \xrightarrow{W_1} (W_2 \xrightarrow{W_3}) \quad \text{(41)}$$

with $\tau = 3$ and $W_\tau = W_3$. The probability measure of this path is $1/8$. For another instance

$$W_0 \xrightarrow{W_1} W_2 \quad \text{(42)}$$

with $\tau = 2$ and $W_\tau = W_2$, the probability measure is $1/4$.

E. Construct Code and Communicate

In a given tree $T$, non-leaf vertices represent channels that are consumed to obtain their children. They are not available to users. Leaves of $T$, however, represent channels that are available to users.

A person who wants to send messages can (a) choose a subset $A$ of leaves, (b) transmit uncoded messages through leaf channels in $A$, and (c) transmit predictable symbols through the remaining leaf channels.

This tree-leaves pair $(T, A)$ determines a block code. A block code has block length $N$, code rate $R$, and error probability $P$. The following is how to read-off these parameters from the pair $(T, A)$.

For every leaf channel $w$ in $T$, the probability $P\{W_\tau = w\}$ is the reciprocal of an integer. This integer is the product of the “cts” of $W_0, W_1, \ldots, W_{\tau-1}$ when $W_\tau = w$.

The block length $N$ of $T$ is the least positive integer such that $N\log P\{W_\tau = w\}$ is an integer for every leaf channel $w$, i.e.,

$$N := \text{lcm}_{w \in \text{leaf}} \frac{1}{P\{W_\tau = w\}}. \quad \text{(43)}$$

when $T^k_\tau$ does not present. When $T^k_\tau$ does present,

$$N := \text{lcm}_{W \in \text{leaf}} \frac{k}{P\{W_\tau = w\}}. \quad \text{(44)}$$

The code rate $R$ of $(T, A)$ is the probability that $W_\tau$ ends up in $A$.

$$R := P\{W_\tau \in A\}. \quad \text{(45)}$$

The error probability $P$ is the probability that any leaf channel in $A$ fails to transmit the message. For Ankan’s polar codes, this quantity is less than the weighted sum

$$\sum_{w \in A} N\log P\{W_\tau = w\} Z(w) \quad \text{(46)}$$

defined by [Ari09 Proposition 2]. For other binary kernels, [KSU10 Formula (12)] claims the same. For larger alphabets, it is still true that the error probability is less than a multiple of the weighted sum [MTT4 Lemma 22]. Later in Section II-C we will define the error probability to be the sum.

F. The Three Regimes

To investigate the tradeoff among block length $N$, code rate $R$, and error probability $P$, researchers have developed three general directions:

- error exponent regime (varying $N, P$);
- scaling exponent regime (varying $N, R$); and
- moderate deviations regime (varying $N, P, R$ at once).

See [MHU16 Abstract and Section 1] for an alternative introduction.

1) Error Exponent Regime: The error exponent regime studies the tradeoff between $N, P$ when $R$ is bounded from below. That is, if we want to communicate at a certain rate $R_{\text{bound}}$ and ask for longer and longer codes, what is the gain of $P$ in exchange for $N$?

For a series of block codes (including random codes), the number

$$\liminf_{N \to \infty} -\frac{\log P}{N} \quad \text{(47)}$$

measures how fast $P$ decays to zero and is called the error exponent [Gal65]. Hence the name error exponent regime. For random codes with $R_{\text{bound}}$ fixed, the error exponent is positive, and it is an interesting $s$ to approximate the error exponent.

However, for other codes such as polar codes or random codes with “fast growing $R$” (will explain soon), $-\log P$ is sub-linear in $N$ so the error exponent vanishes. In such case, the second best thing is the quantity

$$\beta := \liminf_{N \to \infty} \frac{\log (-\log P)}{\log N} \quad \text{(48)}$$

being positive.

The best possible $\beta$ a coding scheme can obtain is denoted by $\beta$ in some literature. For codes with positive error exponent, $\beta = 1$. (And being 1 is optimal.) For Ankan’s polar codes with $R_{\text{bound}}$ fixed, $\beta = 1/2$ [AT09]. For polar codes with arbitrary kernels with $R_{\text{bound}}$ fixed, $\beta$ is the average of logarithms of the partial distances [KSU10]. Chances are that some deliberately selected kernels produce polar codes with $\beta$ arbitrarily close to 1, but not exactly 1.
However, for polar codes with “fast growing $R$” (will explain soon), $\beta'$ is strictly less than $\beta$, and how much $\beta'$ is less than $\beta$ depends largely on how fast $R$ is approaching the capacity. This dependency is the main interest of this work.

In Section II-F we will define the $\partial$-dice which generalizes the usual partial distances. We pretend this extra level of abstraction makes possible application in other paradigms, e.g. LDPC. Readers are invited to read “partial distance” every time they see “\partialDice”. See Appendix A for a note on error exponent regime.

2) Scaling Exponent Regime: The scaling exponent regime studies the tradeoff between $N, R$ when $P$ is bounded from above. That is, if we want to communicate at a certain error on error exponent regime.

The number
\[ \mu' := \liminf_{N \to \infty} \frac{\log N}{-\log(I(W) - R)} \tag{49} \]
measures how fast $R$ approaches the capacity and is sometimes called the scaling exponent. Hence the name scaling exponent regime.

The best possible $\mu'$ a coding scheme can obtain is denoted by $\mu$ in some literature. For random codes with $P_{\text{upper}}$ fixed, $\mu = 2$. (And being 2 is optimal.) For Arıkan’s polar codes with $P_{\text{upper}}$ fixed, $\mu = 3.627$ on BEC [FV14] and $\mu \leq 4.714$ on other channels [MHU16]. For polar codes with arbitrary kernels, it is difficult to approximate but researchers tried to bound [MHU16]. Chances are that some randomly selected kernels produce polar codes with $\mu$ arbitrarily close to 2, but not exactly 2 [FHMV17].

However, for random codes and polar codes with “fast decaying $P$” (will explain soon), $\mu'$ will be strictly more than $\mu$, and how much $\mu'$ is more than $\mu$ depends largely on how fast $P$ is decaying to zero. This dependency is the main interest of this work.

In Section II-C we will define the $\mu^*$-exponent which is a variant of $\mu$. The definition of $\mu^*$ is made so that, say, proving $\mu^* \leq 5$ is much easier than proving $\mu \leq 5$, and then our analysis nonsense (as opposite to abstract nonsense) will complete the rest of proof. See Appendix B for a note on scaling exponent regime.

3) Moderate Deviations Regime: We mentioned above that $\beta' \leq 1$ and 1 can be achieved. We also mentioned that $1/\mu' \leq 1/2$ and 1/2 can be achieved. These poses new questions: Are those all restrictions? Can, in particular, a family of codes achieve $(\beta', 1/\mu') = (1, 1/2)$?

The moderate deviations regime studies $N, R, P$ as a whole to answer these questions. The answer turns out to be NO. There are more fundamental restrictions on the pair $(\beta', 1/\mu')$, i.e., on
\[ \left( \liminf_{N \to \infty} \frac{\log(-\log P)}{\log N}, \liminf_{N \to \infty} \frac{-\log(I(W) - R)}{\log N} \right), \tag{50} \]
that stop a family of codes from achieving $(1, 1/2)$.

The restrictions can be seen in the following way: That $0 \leq 1/\mu' \leq 1/2$ is illustrated by this vertical segment
\[ (0, 1/2) \rightarrow (0, 0) \tag{51} \]
That $0 \leq \beta' \leq 1$ is illustrated by this horizontal segment
\[ (0, 0) \rightarrow (1, 0) \tag{52} \]
The moderate deviations regime then shows that the pair $(\beta', 1/\mu')$ lies in, or on the boundary of, the following right triangle
\[ (0, 1/2) \rightarrow (0, 0) \rightarrow (1, 0) \tag{53} \]
It also shows that every point inside or on the boundary is achievable by random codes
\[ (0, 0) \rightarrow (0, 1/2) \rightarrow (1, 0) \tag{54} \]
So far polar codes achieve
\[ (0, 1/2) \rightarrow (0, 0) \rightarrow (1, 0) \tag{55} \]
on BEC. We will expand it to
\[ (0, 1/2) \rightarrow (0, 0) \rightarrow (1, 0) \tag{56} \]
on BEC.

See Appendix C for a note on moderate deviations regime.

G. Large Deviations Theory

Assume $Y$ is a discrete, bounded random variable. Let $Y_1, Y_2, \ldots$ be i.i.d. copies of $Y$. Let $S_n := Y_1 + Y_2 + \cdots + Y_n$ be the partial sum. Let $y$ be a number that is about, but smaller than, $\mathbb{E}Y$. We want to control the probability
\[ \mathbb{P}\left\{ \frac{S_n}{n} \leq y \right\} \tag{57} \]
in terms of $y$ and the distribution of $Y$.

The canonical argument goes as follows: For every $\lambda < 0$,
\[ \mathbb{P}\left\{ \frac{S_n}{n} \leq y \right\} = \mathbb{P}\{\exp(\lambda S_n) \geq \exp(\lambda ny)\} \tag{58} \]
\[ \leq \mathbb{E}[\exp(\lambda S_n)] \exp(-\lambda ny) \tag{59} \]
\[ = \mathbb{E}[\exp(\lambda Y)]^n \exp(-\lambda y)^n \tag{60} \]
by the Chernoff bound and independency. Take logarithms and divide by $-n$:
\[ -\frac{1}{n} \log \mathbb{P}\left\{ \frac{S_n}{n} \leq y \right\} \geq \lambda y - \log \mathbb{E}[\exp(\lambda Y)]. \tag{61} \]
Since the right hand side of the inequality contains a free parameter $\lambda < 0$, it makes sense to take the supremum and treat it as a function of $y$

$$\frac{-1}{n} \log \mathbb{P}\left\{ \frac{S_n}{n} \leq y \right\} \geq \sup_{\lambda < 0} \lambda y - \log \mathbb{E}[\exp(\lambda Y)].$$

That motivates the definition of the Cramér function

$$\Lambda^*(y) := \sup_{\lambda < 0} \lambda y - \log \mathbb{E}[\exp(\lambda Y)].$$

Two non-obvious comments: (a) Take the supremum over $\lambda \in \mathbb{R}$ still gives the same result for $y < \mathbb{E}Y$. Doing so makes it the Legendre–Fenchel transformation of the cumulant generating function of $Y$. (b) $\Lambda^*$ as defined above is the largest possible function such that Formula (62) holds.

See [DZ10] Theorem 2.1.24 and 2.2.3 for more on this topic.

II. Preliminary

In this section, we consolidate the notations that will be useful to state and prove theorems.

A. Channel Transformation

A communication channel is a triple $(\mathcal{X}, \mathcal{Y}, W)$ of a finite input alphabet $\mathcal{X}$, a finite output alphabet $\mathcal{Y}$, and an one-step Markov process $W : \mathcal{X} \rightarrow \mathcal{Y}$.

To abuse notation, write $W$ to mean the full triple. The cardinality of $\mathcal{X}$ is called the input size of $W$, or the arity of $W$ for short.

Let $\mathcal{C}$ be the set of channels we are interested in. A channel transformation is a triple $(\mathcal{D}, \ell, T)$ of a domain $\mathcal{D} \subset \mathcal{C}$, a length $\ell \in \mathbb{N}$, and a map

$$T : \mathcal{D} \rightarrow \mathcal{C}^{\ell}.$$ (65)

To abuse notation, write $T$ to mean the full triple.

In this work, every $\mathcal{D}$ consists of channels of the same arity. We refer to this number as the arity of $\mathcal{D}$, or the arity of $T$ for short. For instance, $T_{\text{Ar}}$ works on channels of binary input, so $T_{\text{Ar}}$ has arity 2, or it is a binary (2-ary) transformation.

Unless stated otherwise, transformations in this work are such that $\ell \geq 2$ and

$$T : \mathcal{D} \rightarrow \mathcal{D}^{\ell}.$$ (66)

Therefore, it is well-defined when the same transformation is applied iteratively. For instance, Formula (25) begins with $T_{\text{Ar}}(W) = (W_1, W_2)$ and then $T_{\text{Ar}}(W_1) = ((W_1)^2, (W_2)^2)$, and $T_{\text{Ar}}(W_2) = ((W_1)^\ell, (W_2)^\ell)$. They all are of binary input.

We also define an exceptional transformation $(\mathcal{D}, 1, T^1_{\mathcal{C}})$ where

$$T^1_{\mathcal{C}} : \mathcal{D} \rightarrow \mathcal{C}$$ (67)

transforms $q$-ary channels to $q^k$-ary channels, for some integer parameter $k$. This corresponds to the fact that $k$ instances of $q$-ary channels can be seen as a $q^k$-ary channel. Or dually, a $k$-tuple of $q$-bits can be seen as a $q^k$-bit.

B. Channel Tree

A channel tree $T$ is a rooted tree where each vertex is a channel in $\mathcal{C}$, and each non-leaf vertex $w$ corresponds to a transformation $T$ such that $T(w)$ are children of $w$. In this work, channel trees are generated by

- Begin with a channel $W$ as the root of a new tree.
- For each leaf channel $w$, run a deterministic algorithm that observes the current tree and decides whether to apply a certain transformation or not.
- If $T$ is applied, append synthetic channels $T(w)$ as children of $w$.

Most channel trees in this work are finite. In fact, a good algorithm will stop applying transformations once the depth reaches some prescribed number $n$.

For instance, let $T_{\text{fect}}(W, T, n)$ be the channel tree generated as follows:

- Begin with $W$ as the root of a new tree.
- For each leaf channel $w$, apply $T$ if the depth of $w$ is not yet $n$. (The algorithm merely checks the depth.)
- By applying $T$, we mean to append synthetic channels $T(w)$ as children of $w$.

Convention: the root has depth 0; the tree $T_{\text{fect}}(W, T, n)$ has $\ell^n$ leaves, where $\ell$ is the length of $T$. Some examples are Formula (24) being $T_{\text{fect}}(W, T_{\text{Ar}})$; Formula (25) being $T_{\text{fect}}(W, T_{\text{Ar}}, 2)$; Formula (26) being $T_{\text{fect}}(W, T_{\text{Ar}}, 3)$; and Formula (27) being $T_{\text{fect}}(W, T_{\text{GBLB}})$.

More involved example: Formula (33) is $T_{\text{fect}}(W, T_{\text{Ar}}, 1)$ except that a leaf $W^\circ$ is merged with $T_{\text{fect}}(W^\circ, T_{\text{GBLB}})$, and the other leaf $W^\circ$ is merged with $T_{\text{fect}}(W^\circ, T_{\text{GBLB}})$.

Let $T_{\text{fect}}(W, T, \infty)$ be the infinite tree. This is useful when arguing about the process $Z_i$ (defined below) without having to worry about whether $i \leq n$ or not.

C. Z-Parameter and Processes

A Z-parameter will be a function $Z : \mathcal{C} \rightarrow [0, 1]$ measuring the unreliability, the badness, of a given channel. It does not have to be exactly the Bhattacharyya parameter, but could be any function such that a multiple of $Z_i(w)$ bounds, from above, the probability that a decoder fails to decode a single symbol transmitted through $w$.

Given a channel tree with root channel $W$, define a discrete-time stochastic process $W_i$ and a stopping time $\tau$ as follows: Start from the root channel $W_0 := W$. For any $i \geq 0$, if $W_i$ is a leaf channel, let $\tau := i$. If, otherwise, $W_i$ has $\ell$ children, choose an integer $X_{i+1}$ from $1, 2, \ldots, \ell$ uniformly at random, and let $W_{i+1}$ be the $X_{i+1}$-th child of $W_i$.

Be careful that a priori $X_i$ are neither independent nor identical. This is because $X_1$ controls the number of children of $W_1$, which affects the distribution of $X_2$. However, they are i.i.d. in $T_{\text{fect}}(W, T, \infty)$.

Let $Z_i$ be $Z(W_i)$. Let $\underline{Z}_i$ be $\log(\log Z_i / \log Z_{i-1})$; this is the “empirical increment” of $\log(\log Z_i)$. Let $T_{i-1}$ be the transformation applied to $W_{i-1}$. Then the empirical increment can also be written as

$$\underline{Z}_i = \log \frac{\log Z(X_i, \text{th component of } T_{i-1}(W_{i-1}))}{\log Z(W_{i-1})}.$$ (68)
This motivates the definition of the “theoretical increment” (without underline)

\[ Y_i := \liminf_{w \in D} \frac{\log Z(X_i \text{-th component of } T_{i-1}(w))}{\log Z(w)}. \]  

(69)

The purpose of defining two types of “increments” is that \( Y_i \) are i.i.d. in \( T_{\text{fect}}'(W, T, \infty) \) and approximate \( \sum_i \) in a certain context. It is easy to study \( Y_i \) and then predict \( \sum_i \), accordingly.

D. Root-to-Leaf Path as Sample, Vertex as Event

The process \( W_i \) implicitly assumes a sample space: the set of all root-to-leaf paths of \( T \). Each vertex lies on a subset of root-to-leaf paths, which form an event. Thus we can talk about the probability measure of a vertex. It is the probability that the trajectory \( W_0, W_1, \ldots, W_\tau \) passes that vertex.

Furthermore, for any two vertices, their corresponding events are disjoint if and only if neither of them is a descendant of the other. Thus it makes sense to say two vertices are disjoint or not. For a subset of pairwise-disjoint vertices, its probability measure is the sum of probability measures of these vertices. It is also the probability that the trajectory \( W_0, W_1, \ldots, W_\tau \) passes any of these vertices.

Let \( w \) be a synthetic channel at depth \( j \). When \( W_j \) happens to be \( w \), the trajectory \( W_0, W_1, \ldots, W_{j-1} \) is uniquely determined. (In entropy notation, \( H(W_j | W_i) = 0 \) for \( 0 \leq i < j \).)

It also determines \( T_0, X_0, \sum_i Y_0, Z_0 \) and their successors up to \( T_j, X_j, Y_j, Z_j \).

E. Construct Code and Communicate

Let \( \mathcal{T} \) be a channel tree and \( \mathcal{A} \) be a subset of leaves of \( \mathcal{T} \). The pair \( (\mathcal{T}, \mathcal{A}) \) defines a block code.

The block length \( N \) of \( (\mathcal{T}, \mathcal{A}) \) is

\[ N := \frac{1}{\inf_{w \in \mathcal{A}} \mathbb{P}(w)} \]  

(70)

when \( T_{\mathcal{C}}^{k} \) does not present. When \( T_{\mathcal{C}}^{k} \) does present,

\[ N := \frac{k}{\inf_{w \in \mathcal{A}} \mathbb{P}(w)}. \]  

(71)

The code rate \( R \) of \( (\mathcal{T}, \mathcal{A}) \) is

\[ R := \mathbb{P}(\mathcal{A}). \]  

(72)

The error probability \( P \) of \( (\mathcal{T}, \mathcal{A}) \) is defined as

\[ P := \sum_{w \in \mathcal{A}} N \mathbb{P}(w) Z(w). \]  

(73)

F. The \( \delta \)-Dice of a Transformation

Let \( T \), or formally \( (D, \ell, T) \), be a length-\( \ell \) transformation. Let \( X \) be a random integer chosen uniformly from \( 1, 2, \ldots, \ell \). Define the \( \delta \)-dice of \( T \):

\[ Y := \liminf_{w \in D} \frac{\log Z(\text{X-th component of } T(w))}{\log Z(w)}. \]  

(74)

Compare this to Formulae (68) and (69): \( Y \) is the “prototype” of \( Y_i \) in \( T_{\text{fect}}'(W, T, \infty) \), i.e., \( Y_i \) are i.i.d. copies of \( Y \).

Call \( T \) bounded if there exists a number, denoted by \( |T| \), such that, for all \( w \in D \),

\[ \frac{Z(\text{every component of } T(w))}{Z(w)} < |T|. \]  

(75)

Call \( T \) powerful if

\[ \mathbb{P}(Y > 0) > 0. \]  

(76)

The following lemma motivates a necessary condition for our main theorems.

**Lemma 1.** Consider \( T_{\text{fect}}'(W, T, \infty) \) for any \( W \in D \). If \( T \) is bounded, then

\[ Y \geq 0. \]  

(77)

If \( T \) is bounded and powerful, and \( \epsilon > 0 \) is small enough, then there exists \( \delta > 0 \) such that

\[ (Z_i \wedge \delta)^e \text{ is a super-martingale.} \]  

Here \( Z_i \wedge \delta \) is a shorthand for \( \min(Z_i, \delta) \).

**Proof:** For the first statement,

\[ Y \geq \liminf_{w \in D} \frac{\log(Z(w)|T|)}{\log Z(w)} \]  

(79)

\[ = \liminf_{w \in D} \frac{\log(1 + \frac{\log|T|}{\log Z(w)})}{\log Z(w)^\epsilon} \]  

(80)

\[ = \log(1 + 0). \]  

(81)

For the second statement, start from

\[ \mathbb{P}(Y \leq 0) = 1 - \mathbb{P}(Y > 0) < 1. \]  

(82)

Pick a smaller \( \epsilon > 0 \) such that

\[ \mathbb{P}(Y < 2\epsilon) < 1. \]  

(83)

Pick a smaller \( \epsilon > 0 \) such that

\[ |T|^\epsilon \mathbb{P}(Y < 2\epsilon) < 1. \]  

(84)

Pick a number \( \delta > 0 \) such that

\[ |T|^\epsilon \mathbb{P}(Y < 2\epsilon) + \delta^{\epsilon} \mathbb{P}(Y \geq 2\epsilon) \leq 1. \]  

(85)

Pick a smaller \( \delta > 0 \) such that

\[ \inf_{w \in D} \frac{\log Z(\text{X-th component of } T(w))}{\log Z(w)} > Y - \epsilon. \]  

(86)

Note that this is saying

\[ Z_{i-1} < \delta \text{ implies } Y_i > Y_i - \epsilon. \]  

(87)

Now bound \( \mathbb{E}[(Z_i \wedge \delta)^e \mid Z_0, \ldots, Z_{i-1}] \) by considering one plus two cases: (a) If \( Z_{i-1} \geq \delta \), then it is automatically true that

\[ \mathbb{E}[(Z_i \wedge \delta)^e \mid Z_0, \ldots, Z_{i-1}] \leq \mathbb{E}[\delta^e \mid Z_0, \ldots, Z_{i-1}] \leq \delta^e \]  

(88)

\[ = \delta^e \]  

(89)

\[ = (Z_{i-1} \wedge \delta)^e. \]  

(90)

(b-i) If \( Z_{i-1} < \delta \) and \( Y_i < 2\epsilon \), then

\[ (Z_i \wedge \delta)^e \leq Z_i^e \leq Z_{i-1}^e |T|^\epsilon. \]  

(91)
(b-i) If \( Z_{i-1} < \delta \) and \( Y_i \geq 2 \epsilon \), then
\[
Z_i = Z_{i-1}^Y \leq Z_{i-1}^{1+Y_i} < Z_{i-1}^{1+Y_i-\epsilon} \leq Z_{i-1}^{1+\epsilon} < Z_{i-1}^\delta
\]
and hence
\[
(Z_i \land \delta)^\epsilon \leq Z_i^\star \leq Z_{i-1}^\delta - \epsilon, \tag{92}
\]
(b) Combine (b-i) and (b-ii) to get, when \( Z_{i-1} < \delta \),
\[
\mathbb{E}[(Z_i \land \delta)^\epsilon \mid Z_1, \ldots, Z_{i-1}] \leq Z_{i-1}^\delta \mathbb{P}(\{Y_i < 2\epsilon\}) + Z_{i-1}^\delta \mathbb{P}(\{Y_i \geq 2\epsilon\}) \leq \frac{Z_{i-1}^{1+\epsilon}}{\min(Y_i, 2\epsilon)} \leq \frac{Z_{i-1}^{1+\epsilon}}{2\epsilon} \leq \frac{Z_{i-1}^\delta}{2\epsilon}.
\]
where the last inequality is by Formula \( \ast \). Combine (a) and (b) to get
\[
\mathbb{E}(Z_i \land \delta)^\epsilon \mid Z_1, \ldots, Z_{i-1}] \leq (Z_i \land \delta)^\epsilon. \tag{99}
\]
This proves
\[
(Z_i \land \delta)^\epsilon \text{ is a super-martingale.} \tag{100}
\]

For \( T_m \), the lemma does not apply, but it is true that \( Z_i \)

1) The \( \beta^* \)-exponent of \( T \): Define the \( \beta^* \)-exponent:
\[
\beta^* := \mathbb{E}Y \log \ell. \tag{101}
\]

G. The \( \mu^* \)-Exponent of a Transformation

Let \( W \), or formally \((D, \ell, T)\), be a transformation and let \( W \in D \). Let \( A_n \) be the subset of leaves \( w \) in \( T_m \) such that\[
Z(w) < \exp(-n^{2/3}). \tag{102}
\]
Define the \( \mu^* \)-exponent of \( T \):
\[
\mu^* := \sup \limsup_{W \in D, m \to \infty} -\log N_n \log (I(W) - \mathbb{P}(A_n)). \tag{103}
\]
This definition is not perfect because \( I(W) - \mathbb{P}(A_n) \) is not necessary positive. (We can always specify a code whose code rate exceeds the Shannon capacity.) Of course we know that \( I(W) - \mathbb{P}(A_n) \leq 0 \), or even \( I(W) - \mathbb{P}(A_n) \leq O(N_n^{-49}) \), is too good to be true. So we alter the definition a little bit
\[
\mu^* := \sup \limsup_{W \in D, m \to \infty} \frac{-\log N_n}{\log \max(I(W) - \mathbb{P}(A_n), N_n^{-1/2})}. \tag{104}
\]
so that \( \mu^* \) is at least 2.

We will make use of the definition of \( \mu^* \) in the following manner.

Lemma 2. Assume \( E_{m-n\sqrt{n}} \) is an arbitrary subset of disjoint vertices in \( T_m \). Let \( A_n \) be the set of leaves \( w \) that satisfy \( Z(w) < \exp(-m^{2/3}) \) but have no ancestor in \( E_{m-n\sqrt{n}} \). Then
\[
I(W) - \mathbb{P}(E_{m-n\sqrt{n}} \cup A_n) \leq N_m^{-1/\mu^*+o(1)}. \tag{105}
\]

Please be informed that Formula \( \ast \) is not an ad hoc definition. We merely choose a handy instance of quasi-polynomial to avoid being flooded with Big-O notations.

Proof: Every leaf in \( A_{m-n\sqrt{n}} \) has some ancestor in \( E_{m-n\sqrt{n}} \), so \( \mathbb{P}(A_{m-n\sqrt{n}}) \leq \mathbb{P}(E_{m-n\sqrt{n}}) \). This implies \( \mathbb{P}(A_m) \leq \mathbb{P}(E_{m-n\sqrt{n}} \cup A_m) \) and
\[
I(W) - \mathbb{P}(E_{m-n\sqrt{n}} \cup A_m) \leq \mathbb{P}(A_m) \leq N_m^{-1/\mu^*+o(1)}. \tag{106}
\]
The last inequality is a simple consequence of \( \limsup \) in the definition of \( \mu^* \).

In general, we have the following.

Lemma 3. Assume \( A_0 \) is an arbitrary subset of disjoint vertices in \( T_{m,n} \). Let \( \varphi \) be a predicate of channels. Let \( A_m \) be the set of leaves \( w \) that satisfy \( \varphi \) but have no ancestor in \( A_0 \). Then
\[
I(W) - \mathbb{P}(A_0 \cup A_m) \leq I(W) - \mathbb{P}(\varphi \cup A_m). \tag{108}
\]
Proof: Same logic as Lemma 2. By the way, when this lemma is applied, \( \varphi(w) \) will be \( Z(w) < \exp(-\exp(m^{1/3})) \).

H. The Cramér Function

Assume \( Y \) is a discrete, bounded random variable. Let \( Y_1, Y_2, \ldots \) be i.i.d. copies of \( Y \). Define the Cramér function of \( Y \):
\[
\Lambda^*(y) := \sup_{\lambda < 0} \lambda y - \log \mathbb{E}[(\exp)Y]. \tag{109}
\]
It is such that
\[
\mathbb{P}\left\{\frac{Y_1 + Y_2 + \cdots + Y_n}{n} \leq y\right\} \leq \exp(-n\Lambda^*(y)). \tag{110}
\]

III. THE RECRUIT-TRAIN-RETAIN TEMPLATE

The recruit-train-retain template helps us understand the distribution of \( Z_n \) by first understanding the distribution of \( Z_m \) for some \( m < n \).

An over-simplified template is as follows:

Recruit Sometimes \( Z_m \) is quite small. Calculate \( \mathbb{P}(Z_m \text{ is quite small}) \).

Train When \( Z_0 \) is quite small, there is a positive chance that \( Z_{i+1} \) gets smaller. Repeat this for \( i = m, m+1, \ldots, n-1 \); it is very unlikely not to get smaller at all.

Retain By syllogism, most of the \( Z_n \) will be extremely small. Keep these extremely small \( Z_n \), and freeze those \( Z_n \) that are not extremely small enough.

In terms of Sankey diagram:

See Formula \( \ast \) in Appendix \( \ast \) for the big diagram.

This diagram records the fact that synthetic channels at depth \( m \) are classified into two groups based on their \( Z \)-parameters. The upper group consists of bad channels (large
is frozen. The lower group consists of good channels (small $Z_m$) and is recruited and trained in the sense that we want to investigate their children $Z_{m+1}$, grandchildren $Z_{m+2}$, and all the way up to $Z_n$. Then these $Z_n$ are further classified into two groups. Those that are mediocriely small are retained; those that are extremely small are retained — they go to $A_n$.

We will see the template and Sankey diagrams multiple times.

### A. A Brief History

The first appearance dated back to [AT09] with $m := n^{3/4}$.

The argument goes as follows:

Recruit Control $\mathbb{P}\{Z_m < .875^m\}$; this is almost $I(W)$.

Train Condition on the event $\{Z_m < .875^m\}$. For $i = m, m + 1, \ldots, n - 1$,

$$Z_{i+1} \approx \begin{cases} Z_i^4 & \text{with probability } 1/2 \\ Z_i^2 & \text{with probability } 1/2 \end{cases}. \quad (112)$$

That is, it gets squared with probability 1/2.

Retain By syllogism, conditioning on the event $\{Z_m < .875^m\}$, the quantity (how many times it is squared) $\log_2(\log Z_n / \log Z_m)$ is about $(n - m)/2$ with high probability.

That is to say, with probability $I(W) - o(1)$ it holds that

$$\log_2(-\log Z_n) = (n - o(n))/2. \quad (113)$$

Hence the $\beta'$-exponent

$$\frac{\log(-\log P)}{\log N} \approx \frac{\log_2(-\log Z_n)}{\log_2 2^n} \to \frac{1}{2}. \quad (114)$$

The argument can be summarized by the Sankey diagram:

![Sankey Diagram](image)

But part of the proof of [KSU10, Theorem 11] is omitted in the original paper. However, the idea is the same Sankey diagram:

![Sankey Diagram](image)

Another argument appears in [MHU16].

Recruit Control $\mathbb{P}\{Z_m < .5^m\}$, where $m = \gamma n$ for some fixed ratio $0 < \gamma < 1$.

Train Condition on the event $\{Z_m < .5^m\}$. Track the process $Z_m, Z_{m+1}, \ldots, Z_n$.

Retain Control $\log_2(\log Z_n / \log Z_m)$ and $\log(-\log Z_n)$.

The unchanging part of [MHU16] is the Sankey diagram:

![Sankey Diagram](image)

The innovative part of [MHU16] is that $m$ is parameterized by $\gamma$. That is, they are free to choose $\gamma$ before spending $\gamma n$ steps in the recruit phase and $(1 - \gamma)n$ steps in the train phase. A rule of thumb is, a longer recruit phase makes $R$ better; and a longer train phase makes $P$ better. Thus they obtain a tradeoff between $R$ and $P$. The following plot shows pairs of $(\beta', 1/\mu')$ they achieve.

![Plot](image)

### B. Disposing Bad Synthetic Channels

Our contribution in the last work [WD18] is what we now called the disposable recruit-train-retain template. The idea is as follows.

Recruit Control $\mathbb{P}\{Z_m < .5^m\}$ for $m = \sqrt{n}, 2\sqrt{n}, \ldots, n_{\text{rat}}$.

Train Condition on the event $\{Z_m < .5^m \text{ but } Z_i \geq .5 \text{ for } i = 0, \ldots, m - \sqrt{n}\}$.

Retain Control $\log_2(\log Z_n / \log Z_m)$ and $\log(-\log Z_n)$.

In terms of Sankey diagram:

![Sankey Diagram](image)
See Formula (314) in Appendix II for the big diagram.

This approach recruits \( Z_m \) in several rounds to maximize the rate. Plus it trains \( Z_i \) for almost the same depth as [MHU16] does. Thus it should outperform the latter. Our final result, besides what [MHU16] achieved before in gray, is the dark region below.

We personally believe this is the maximum region Arıkan’s polar codes can achieve. We do not see any obvious way to improve our inner bound in [WD18]. Nonetheless, there is a hope that since polar codes generalize to other kernels, they might achieve a larger region.

In fact, Theorem 9 shows any point inside the triangle is achievable. And Corollary 10 extends the conclusion to \( \sqrt{m} \) channels. Therefore, even if some synthetic channel is achievable. We do not see any obvious way to achieve \( \frac{\sqrt{m}}{\sqrt{m}} \) for depth \( \sqrt{m} \).

Based on [MHU16], we do have fine control on \( Z_m \). This approach recruits \( \frac{\sqrt{m}}{\sqrt{m}} \) frozen. However a result like [FV14] is missing for general polar codes. So we came up with a workaround.

In the following version, synthetic channels will be trained in the upcoming rounds. Therefore, even if some synthetic channel is frozen, there is a chance that it’s descendants will be recruited in the upcoming rounds.

This makes it a recyclable recruit-train-retain template.

---

### C. Recycling Bad Synthetic Channels

In Formula (124), a recruited synthetic channel is trained until depth \( n \). As mentioned above, training this much reduces the error probability a lot. But it either requires very fine control on how good \( Z_m \) are to begin with or we will have to freeze a lot of innocent \( Z_n \) (i.e., the \( Z_n \) that we believe are good but are not able to prove), which hurts the rate.

In case of [WD18], which considers only Arıkan’s polar codes, we do have fine control on \( Z_m \) provided by [FV13]. No innocent \( Z_m \) is frozen. However a result like [FV14] is missing for general polar codes. So we came up with a workaround.

In the following version, synthetic channels will be trained for depth \( \sqrt{m} \) and immediately be frozen or retained. And then the next round of recruit-train-retain starts. There will be \( \sqrt{n} \) rounds in total. Therefore, even if some synthetic channel is frozen, there is a chance that its descendants will be recruited in the upcoming rounds.

This makes it a recyclable recruit-train-retain template.

---

### IV. MAIN RESULTS: TO INTERPOLATE \( \beta^* \) AND \( \mu^* \)

We present three statements at once so readers immediately see the similarity. In fact, Theorem 6 can be proven by combining the proofs of Lemma 4 and Theorem 5.

**Lemma 4.** Let \( T \) be a length-\( \ell \), bounded transformation with \( \mu^* \)-exponent \( \mu^* \) and \( \partial \)-dice \( Y \). If

\[
\Pr\{Y = 0\} < e^{-1/\mu^*},
\]

then \( T \) produces block codes \( (T_n, A_n) \) such that

\[
N_n = \ell^n,
\]

\[
R_n > I(W) - N_n^{-1/\mu^*+o(1)}, \text{ and}
\]

\[
P_n < \exp\left(-\exp(n^{1/3})\right).
\]

For \( n \) large enough.

**Proof:** First-time reader may skim Section VI-B. Second-time reader may skim Section VI with white lies in mind: lie number one: \( \sum Y = Y_i \); lie number two: \( a_m \) and \( b_m \) are about \( e^{-(m-\sqrt{m})/\mu} \); lie number three: \( c_m/b_m \) and \( d_m/b_m \) are about \( e^{\sqrt{m}/\mu} \); lie number four: \( g_m-b_m \) is about \( e^{m/\mu} \); lie number five: \( g_m \) is about \( 2m^{1/\mu} \). Third-time reader may realize that the whole proof, Section VI is an attempt to prove those lies but ends up barely proving the lemma by something weaker.

**Theorem 5.** Let \( T \) be a length-\( \ell \), bounded transformation with \( \mu^* \)-exponent \( \mu^* \) and \( \partial \)-dice \( Y \). Let \( \Lambda^* \) be the Cramér function of \( Y \). If

\[
\Pr\{Y = 0\} < e^{-1/\mu^*},
\]

and, for \( \pi \in [0, 1] \),

\[
\frac{(1 - \pi) \log \ell}{\mu^*} < \Lambda^* \left( \frac{\beta^* \mu^* \log \ell}{\mu^* - \pi \mu^*} \right)
\]

then \( T \) produces block codes \( (T_n, A_n) \) such that

\[
N_n = \ell^n,
\]

\[
R_n > I(W) - N_n^{-1/\mu^*}, \text{ and}
\]

\[
P_n < \exp\left(-N_n^{\beta^*}\right)
\]

for \( n \) large enough.

**Proof:** First-time reader may skim Section VI-B. Second-time reader may skim Section VI with white lies in mind: lie number one: \( \sum Y = Y_i \); lie number two: \( a_m \) and \( b_m \) are about \( e^{-(m-\sqrt{m})/\mu} \); lie number three: \( c_m/b_m \) and \( d_m/b_m \) are about \( e^{\sqrt{m}/\mu} \); lie number four: \( f_m/b_m \) is about \( e^{m/\mu} \); lie number five: \( g_m-f_m \) is about \( 2m^{1/\mu} \). Third-time reader may realize that the whole proof, Section VI is an attempt to prove those lies
Theorem 6. Let $T_{rat}$ be a q-ary, length-$\ell$, bounded transformation with $\mu^*$-exponent $\mu_{rat}^*$ and $\partial$-dice $Y_{rat}$. Let $T_k^B$ be transforming $q^k$-ary channels to $q^k$-ary channels. Let $N_{err}^*$ be the Cramér function of $Y_{err}$. If
\[ P\{Y_{rat} = 0\} < \ell^{-1}/\mu_{rat}^* \] (133)
and, for $\pi \in [0,1]$, \[
\frac{(1 - \pi) \log \ell}{\mu' - \pi \mu_{rat}^*} < \frac{\beta \mu' \log \ell}{\mu' - \pi \mu_{rat}^*},
\] (144)
then $T_{rat}, T_{err}^B, T_{err}$ produce block codes $(T_n, A_n)$ such that
\[ N_n = k\ell^n, \] (135)
\[ R_n > I(W) - N_n^{-1/\mu'}, \] (136)
\[ P_n < \exp(-N_n^{\delta'}) \] (137)
for $n$ large enough.

Proof: The first-time reader may believe that this is an easy implication of Lemma 1 and Theorem 5. The second-time reader may realize that it is not an easy implication, but Section VII tries to explain it is an implication.

Readers may notice that in Theorem 6, $(1/\mu', 1/\mu')$ is highly related to $\mu_{rat}^*$ and $Y_{err}$ instead of $\mu_{rat}^*$ or $Y_{rat}$. That is, the code rate is controled by $T_{rat}$ and the error probability is controled by $T_{err}$. This explains why and how we should mix two kernels.

V. PROVE LEMMA 1 BY RECYCLABLE RECRUIT-TRAIN-RETRAIN TEMPLATE

Consider $T_{rat}^B(W, T, n)$. We are going to choose a subset of leaf channels $A_n$.

A. First choose some constants

By Lemma 1 $Y \geq 0$. Start from
\[ \lim_{Y \to +\infty} \mathbb{E}[Y^{-Y}] = \mathbb{P}\{Y = 0\} < \ell^{-1/\mu^*}. \] (138)
Pick a number $Y \gg \exp(1)$ such that
\[ \mathbb{E}[Y^{-Y}] < \ell^{-1/\mu^*}. \] (139)
Pick a number $\epsilon > 0$ such that
\[ \mathbb{E}[Y^{-Y}] \geq 2^\epsilon < \ell^{-1/\mu^*}. \] (140)
Pick a smaller $\epsilon > 0$ and a number $\delta > 0$ such that
\[ (Z_i \wedge \delta)\epsilon \] (141)
as in Lemma 1 Recall from the proof of Lemma 1
\[ \inf_{w \in \mathcal{D}} \log Z(X \text{-th component of } T(w)) \geq Y - \epsilon. \] (142)
Note that this is saying
\[ Z_{i-1} < \delta \implies Y_{i} > Y_i - \epsilon. \] (143)

B. Second fill in the recyclable template

Let $E_0\emptyset$ be the empty set. For $m = \sqrt{n}, 2\sqrt{n}, \ldots, n - \sqrt{n}$, define helically $A_m, B_m, C_m, D_m, E_m, E_0^{m\emptyset}$ as follows:

Recruit Let $A_m$ be the set of synthetic channels $w$ at depth $m$ that satisfy $Z(w) < \exp(-m^{2/3})$ but have no ancestor in $E_0^{m\emptyset}$.

Train Let $B_m$ be the set of synthetic channels at depth $m + \sqrt{n}$ that are descendants of synthetic channels in $A_m$.

Retain Let $C_m$ be the set of synthetic channels $w$ in $B_m$ such that $Z(v) \geq \delta$ for some ancestor $v$ of $w$ at depth $m, m + 1, \ldots, m + \sqrt{n}$. Let $D_m$ be the set of synthetic channels $w$ in $B_m - C_m$ such that
\[ \frac{y_{m+1} + y_{m+2} + \cdots + y_{m+\sqrt{n}}}{\sqrt{n}} \leq 2\epsilon \] (144)
where $y_{m+i}$ are the values that $Y_{m+i}$ take when $W_{m+i}$ happens. Let $E_m$ be
\[ B_m - C_m - D_m. \]
Let $E_m^{\emptyset}$ be $E_0^{m\emptyset} \cup E_m$.

In terms of Sankey diagram:
\[ \sqrt{n} \quad 2\sqrt{n} \quad 3\sqrt{n} \quad n \]

C. Third estimate $c_m/b_m$

It is not hard to see from the definitions that $C_m$ is a subset of $B_m$, so the target quantity
\[ \frac{c_m}{b_m} = \frac{\mathbb{P}(C_m)}{\mathbb{P}(B_m)} = \mathbb{P}(C_m|B_m) \] (146)
is a conditional probability. It is also not hard to see that $B_m$ and $A_m$ refer to the same event, so
\[ \frac{c_m}{b_m} = \mathbb{P}(C_m|B_m) = \mathbb{P}(C_m|A_m). \] (147)
The event defined by $C_m$ is equal to
\[ \{Z_{m+i} \geq \delta \text{ for some } 0 \leq i \leq \sqrt{n}\}. \] (148)
Let $\sigma$ be the stopping time
\[ \min\{0 \leq s \leq \sqrt{n}; Z_{m+s} \geq \delta\} \cup \{\sqrt{n}\}. \] (149)
That is, the first index that makes up the inequality, or the largest index. Then $C_m$ is also equal to
\[ \{Z_{m+\sigma} \geq \delta\} = \{(Z_{m+\sigma} \wedge \delta)\epsilon \geq \delta\}. \] (150)
By how $\delta, \epsilon$ are chosen, $(Z_{m+i} \wedge \delta)^e$ for $i = 0, 1, \ldots, \sqrt{n}$ is a super-martingale. Thus there is a Doob’s inequality-flavor bound (c.f. [Dur10, Theorem 5.4.1])

$$\frac{e_m}{b_m} = \mathbb{P}(\mathcal{C}_m|A_m) \leq \mathbb{E}[(Z_{m+2} \wedge \delta)^e|A_m]\delta^{-\epsilon} \leq \mathbb{E}[(Z_m \wedge \delta)^e|A_m]\delta^{-\epsilon}$$

On the other hand, the event defined by $A_m$ is equal to

$$\{Z_m < \exp(-m^{2/3})\}.$$  

Thus

$$\frac{e_m}{b_m} \leq \mathbb{E}[(Z_m \wedge \delta)^e|A_m]\delta^{-\epsilon} \leq \exp(-m^{2/3}\epsilon)\delta^{-\epsilon} = \exp(-m^{2/3}\epsilon - \epsilon \log \delta).$$

And this is an upper bound on $e_m/b_m$.

**D. Forth estimate $d_m/b_m$**

Since we are in $\mathcal{C}(W, T, n)$, the $\partial$-dices $Y_{m+i}$ are independent of the event $A_m$. As a consequence, the condition imposed by Formula (144) is independent of $A_m$, which we know from the previous subsection refers to the same event as $B_m$ does. Thus $d_m/b_m = \mathbb{P}(\mathcal{D}_m)/\mathbb{P}(A_m)$ is at most the probability that

$$Y_{m+1} + Y_{m+2} + \cdots + Y_{m+\sqrt{n}} \leq 2\epsilon.$$  

To bound the probability measure of this event, it suffices to bound the probability measure of the event

$$\{Y_{m+1} + Y_{m+2} + \cdots + Y_{m+\sqrt{n}} \leq 2\epsilon\sqrt{n}\}.$$  

This is equivalent to the probability measure of

$$\{Y - Y_{m+1} - Y_{m+2} - \cdots - Y_{m+\sqrt{n}} \geq -2\epsilon\sqrt{n}\}.$$  

By the Chernoff bound, it is less than

$$\mathbb{E}[\mathbb{E}[Y - Y_{m+1} - Y_{m+2} - \cdots - Y_{m+\sqrt{n}}]^{2\epsilon\sqrt{n}}]$$

$$\leq \mathbb{E}[\mathbb{E}[Y - Y_{m+1} - Y_{m+2} - \cdots - Y_{m+\sqrt{n}}]^{2\epsilon\sqrt{n}}]$$

$$= (\mathbb{E}[Y - Y_{m+1} - Y_{m+2} - \cdots - Y_{m+\sqrt{n}}]^{2\epsilon\sqrt{n}})$$

$$= (\epsilon^{2\epsilon\mu^e} \sqrt{n})$$

$$= \epsilon^{-2\epsilon\sqrt{n}/\mu^e}$$

where the last inequality is by Lemma [140]. So

$$\frac{1}{\sqrt{n}} (g_m - \sqrt{n} - b_m) \leq N^{-1/\mu^e + o(1)} = \ell^{-m/\mu^e + o(1)},$$

Here $(g_m - \sqrt{n} - b_m)^+$ is max$(g_m - \sqrt{n} - b_m, 0)$. Similarly let $g_m^+$ be max$(g_m, 0)$.

Now we calculate $g_m$

$$g_m = g_m - \sqrt{n} - c_m$$

$$= g_m - \sqrt{n}(1 - \frac{c_m}{b_m}) + (g_m - b_m)\frac{c_m}{b_m}$$

$$\leq g_m^+ (1 - \frac{c_m}{b_m}) + (g_m - b_m)\frac{c_m}{b_m}$$

$$\leq g_m^+ (1 - \frac{c_m}{b_m}) + (g_m - b_m)^+$$

$$= g_m^+ (\frac{c_m}{b_m} + \frac{d_m}{b_m}) + (g_m - b_m)^+$$

$$\leq g_m^+ (\ell^{-\sqrt{n}/\mu^e} + \exp(-m^{2/3}\epsilon - \epsilon \log \delta) + \ell^{-m/\mu^e + o(1)}).$$

Starting from $m \geq O(n^{3/4})$ the term $\ell^{1/\sqrt{n}/\mu^e}$ dominates the term $\exp(-m^{2/3}\epsilon - \epsilon \log \delta)$. Thus it suffices to solve the recurrence relation

$$\left\{ \begin{array}{l}
g_0 = 1; \\
g_m \leq g_m^+ (\frac{c_m}{b_m} + \frac{d_m}{b_m}) + (g_m - b_m)^+ \\
\end{array} \right.$$  

The result is

$$g_{n - \sqrt{n}} \leq \ell^{-n/\mu^e + o(1)} = N^{-1/\mu^e + o(1)}.$$  

By algebra

$$e_0^{-\sqrt{n}} = I(W) - g_{n - \sqrt{n}} \geq I(W) - N^{-1/\mu^e + o(1)}.$$  

**F. Sixth estimate how good synthetic channels in $E_0^{-\sqrt{n}}$ are**

They are synthetic channels such that during the time they are being trained, $Z(W_{m+i})$ is never larger than $\delta$, so $Y_{m+i} > Y_{m+i} - \epsilon$. They are also synthetic channels such that

$$Y_{m+1} + Y_{m+2} + \cdots + Y_{m+\sqrt{n}} > 2\epsilon$$

so

$$Y_{m+1} + Y_{m+2} + \cdots + Y_{m+\sqrt{n}} > \epsilon\sqrt{n}.$$  

Therefore for every $w \in E_m$ and $v$ its ancestor at depth $m$, by telescoping

$$\log(-\log Z(w)) < \log(-\log Z(v)) > \epsilon\sqrt{n}.$$  

But $v \in A_m$ are such that $Z(v) \leq \exp(-m^{2/3})$, so

$$Z(w) < \exp(-\exp(\epsilon\sqrt{n})m^{2/3}).$$

Sum over $E_0^{-\sqrt{n}}$:

$$\sum_{w \in E_0^{-\sqrt{n}}} Z(w) < N_n \exp(-\exp(\epsilon\sqrt{n})m^{2/3}).$$

Let $A_n$ be the set of synthetic channels at depth $n$ that are descendants of synthetic channels in $E_0^{-\sqrt{n}}$. Then the inequality lifts

$$\sum_{w \in A_n} Z(w) < |T|^n \exp(-\exp(\epsilon\sqrt{n})m^{2/3}).$$
Eventually, as \( n \to \infty \), replacing \( \sqrt{n} \) by \( n^{1/3} \) eats up other minor terms:
\[
\sum_{w \in A_{n}} Z(w) < \exp(-\exp(n^{1/3})). \tag{187}
\]

**G. Seventh we announce the code**
\[
(\mathcal{T}^\text{per}_{\text{fect}}(W, T, n), A_n) \tag{188}
\]
has block length
\[
N_n = \ell^n, \tag{189}
\]
and error probability
\[
P_n = \sum_{w \in A_{n}} Z(w) < \exp(-\exp(n^{1/3})). \tag{191}
\]
This finishes the proof of Lemma 4.

**VI. Prove Theorem 5 by Disposable Recruit-Train-Retain Template**

Consider \( T^\text{per}_{\text{fect}}(W, T, n) \). We are going to choose a subset of leaf channels \( A_n \).

**A. First choose some constants**

Pick a number \( \epsilon > 0 \) such that, for all \( \pi \in [0, 1] \),
\[
\frac{(1 - \pi) \log \ell}{\mu' - \pi \mu} < \Lambda' \left( \frac{\beta' \mu' \log \ell}{\mu' - \pi \mu} + \epsilon \right). \tag{192}
\]
Pick a smaller \( \epsilon > 0 \) such that if all \( \mu' \) are replaced by \( \mu' - \epsilon \) in this inequality, then it still holds for all \( \pi \in [0, 1] \). Pick a smaller \( \epsilon > 0 \) and a number \( \delta > 0 \) such that
\[
(Z_i \land \delta)^\epsilon \text{ is a super-martingale} \tag{193}
\]
as in Lemma 1. Recall from the proof of Lemma 1
\[
\inf_{Z(w) < \delta} \frac{\log Z(X \text{-th component of } T(w))}{\log Z(w)} > Y - \epsilon. \tag{194}
\]
Note that this is saying
\[
Z_{i-1} < \delta \implies Y_i > Y_i - \epsilon. \tag{195}
\]

**B. Second fill in the disposable template**

Let \( n_{\text{rat}} = n \mu^* / \mu' \). Let both \( A_0^0 \) and \( E_0^0 \) be the empty set. For \( m = \sqrt{n}, 2 \sqrt{n}, \ldots, n_{\text{rat}} \), define helically \( A_m, A_0^m, B_m, C_m \), \( D_m, E_m, E_0^m \) as follows:

- **Recruit** Let \( A_m \) be the set of synthetic channels \( w \) at depth \( m \) that satisfy \( Z(w) < \exp(-\exp(m^{1/3})) \) but have no ancestor in \( A_{0}^{m-\sqrt{n}} \). Let \( A_0^m \) be \( A_0^{m-\sqrt{n}} \cup A_m \).
- **Train** Let \( B_m \) be the set of synthetic channels at depth \( n \) that are descendants of synthetic channels in \( A_m \).
- **Retain** Let \( C_m \) be the set of synthetic channels \( w \) in \( B_m \)

such that \( Z(v) \geq \delta \) for some ancestor \( v \) of \( w \) at depth \( m, m+1, \ldots, n \). Let \( D_m \) be the set of synthetic channels \( w \) in \( B_m - C_m \) such that
\[
\frac{y_{m+1} + y_{m+2} + \cdots + y_n}{n - m} \leq \frac{\beta' \log \ell}{1 - m/n} + \epsilon. \tag{196}
\]
where \( y_{m+i} \) are the values that \( Y_{m+i} \) take when \( W_n = w \) happens. Let \( E_0^m \) be \( B_m - C_m - D_m \). Let \( E_0^m \) be \( E_0^{m-\sqrt{n}} \cup E_m^m \).

In terms of Sankey diagram:

\[
\text{channels} \quad \sqrt{n} \quad 2\sqrt{n} \quad n_{\text{rat}} \quad n \quad \text{depth}
\]

See Formula 17 in Appendix 2 for the big diagram.

Let \( a_m, b_m, c_m, d_m, e_m, e_0^m \) be the probability measures of the corresponding capital-letter events. Let \( f_m \) be \( I(W) - a_0^m \). Let \( g_m \) be \( I(W) - e_0^m \). Let \( \pi \) be \( m/n_{\text{rat}} \).

 Readers are encouraged to compare this subsection VI-B with Section VI-B and to figure out what the template makes Formula (197) different from Formula (145). More generally, all subsections in this section VI are parallel to those in Section V.

**C. Third estimate \( c_m / b_m \)**

It is not hard to see from the definitions that \( C_m \) is a subset of \( B_m \), so the target quantity
\[
\frac{c_m}{b_m} = \frac{\mathbb{P}(C_m)}{\mathbb{P}(B_m)} = \mathbb{P}(C_m | B_m) \tag{198}
\]
is a conditional probability. It is also not hard to see that \( B_m \) and \( A_m \) refer to the same event, so
\[
\frac{c_m}{b_m} = \mathbb{P}(C_m | B_m) = \mathbb{P}(C_m | A_m). \tag{199}
\]

The event defined by \( C_m \) is equal to
\[
\{Z_{m+i} \geq \delta \text{ for some } 0 \leq i \leq n - m\}. \tag{200}
\]
Let \( \sigma \) be the stopping time
\[
\min\{0 \leq s \leq n - m | Z_{m+s} \geq \delta \} \cup \{n\}. \tag{201}
\]
That is, the first index that makes up the inequality, or the largest index. Then \( C_m \) is also equal to
\[
\{Z_{m+s} \geq \delta \} = \{(Z_{m+s} \land \delta)^\epsilon \geq \delta ^\epsilon \}. \tag{202}
\]
By how \( \delta, \epsilon \) are chosen, \( (Z_{m+i} \land \delta)^\epsilon \) for \( i = 0, 1, \ldots, n - m \) is a super-martingale. Thus there is a Doob’s inequality-flavor bound (c.f. [Dur10], Theorem 5.4.1)
\[
\frac{c_m}{b_m} \leq \mathbb{E}[(Z_{m+s} \land \delta)^\epsilon | A_m] \delta^{-\epsilon} \tag{204}
\]
\[
\leq \mathbb{E}[(Z_{m} \land \delta)^\epsilon | A_m] \delta^{-\epsilon} \tag{205}
\]
On the other hand, the event defined by $A_m$ is equal to
\[ \{Z_m < \exp(-\exp(m^{1/3}))\}. \] (206)

Thus
\[ \frac{c_m}{b_m} \leq \mathbb{E}[ (Z_m \wedge \delta)^\epsilon \mid A_m] \delta^{-\epsilon} < \exp(-\exp(m^{1/3})) \delta^{-\epsilon} \]
\[ = \exp(-\exp(m^{1/3}) \epsilon - \epsilon \log \delta). \] (207) (208) (209)

And this is an upper bound on $c_m/b_m$.

D. Forth estimate $d_m/b_m$

Since we are in $\mathcal{T}(W, T, n)$, the $\partial$-dices $Y_{m+i}$ are independent of the event $A_m$. As a consequence, the condition imposed by Formula (196) is independent of $A_m$, which we know from the previous subsection refers to the same event as $B_m$ does. Thus $d_m/b_m = \mathbb{P}(D_m)/\mathbb{P}(A_m)$ is at most probability that
\[ \frac{Y_{m+1} + Y_{m+2} + \cdots + Y_n}{n-m} \leq \frac{\beta \log \ell}{1-m/n} + \epsilon. \] (210)

Here $m/n = \pi n_{\text{rat}}/n = \pi \mu^*/\mu'$, so the right hand side of the inequality is
\[ \frac{\beta \log \ell}{1-\mu^*/\mu'} + \epsilon = \frac{\beta \mu^* \log \ell}{\mu^* - \pi \mu^*} + \epsilon. \] (211)

By Formula (110) the probability that
\[ \frac{Y_{m+1} + Y_{m+2} + \cdots + Y_n}{n-m} \leq \frac{\beta \mu^* \log \ell}{\mu^* - \pi \mu^*} + \epsilon \]
is bounded from above by
\[ \exp\left(- (n-m) \cdot \Lambda^s \left( \frac{\beta \mu^* \log \ell}{\mu^* - \pi \mu^*} + \epsilon \right) \right). \] (212)

And Formula (192) helps bound this from above by
\[ \exp\left(- (n-m) \frac{(1-\pi) \log \ell}{\mu^* - \pi \mu^*} \right), \] (213)

where the argument of $\exp$ is
\[ -(n-m) \frac{(1-\pi) \log \ell}{\mu^* - \pi \mu^*} = -\left( \frac{n}{\mu^* - \pi \mu^*} \right) \log \ell. \] (214)

Put $\exp$ back; it becomes
\[ \epsilon^{-n/\mu^*+m/\mu^*}. \] (215)

And this is an upper bound on $d_m/b_m$.

E. Fifth estimate $e_{n_{\text{rat}}}^0$

Notice that $E_m$ is a subset of $B_m$, so
\[ 0 \leq \frac{e_m}{b_m} \leq 1. \] (216)

Notice also that
\[ f_m = I(W) - a_m^{m-\sqrt{n}} - a_m = I(W) - \mathbb{P}(A_m^{m-\sqrt{n}} \cup A_m) \leq N_m^{-1/\mu^*+o(1)} \]
\[ = \mathbb{E}[Z_m \wedge \delta] \delta^{-\epsilon} \leq \exp(-\exp(m^{1/3}) \epsilon - \epsilon \log \delta). \] (217) (218) (219) (220)

where the last inequality is by Lemma 4 and 3. So
\[ f_m^+ \leq N_m^{-1/\mu^*+o(1)} = \epsilon^{-m/\mu^*+o(1)}. \] (221)

Here $f_m^+$ is max$(f_m, 0)$. Similarly let $(g_m-\sqrt{n} - b_m)^+$ be max$(g_m-\sqrt{n} - b_m, 0)$.

Now we calculate $g_m - f_m^+$
\[ = g_m-\sqrt{n} - e_m - (f_m-\sqrt{n} - b_m)^+ \]
\[ \leq g_m-\sqrt{n} - e_m - (f_m-\sqrt{n} - b_m)^+ \frac{e_m}{b_m} \]
\[ \leq g_m-\sqrt{n} - e_m - (f_m-\sqrt{n} - b_m)^+ \frac{e_m}{b_m} \]
\[ = g_m-\sqrt{n} - f_m^+ - f_m^+ + f_m^+ \frac{1 - \frac{e_m}{b_m}}{b_m} \]
\[ = g_m-\sqrt{n} - f_m^+ + f_m^+ \frac{e_m + d_m}{b_m} \]
\[ \leq g_m-\sqrt{n} - f_m^+ + \epsilon^{-m/\sqrt{n}+\mu^*} \times \]
\[ \left( \exp(-\exp(m^{1/3}) \epsilon - \epsilon \log \delta) + \epsilon^{-n/\mu^*+m/\mu^*} \right). \] (222) (223) (224) (225) (226) (227)

In the last line, the term $\epsilon^{-n/\mu^*+m/\mu^*}$ dominates the other doubly-exponential term as $n \to \infty$. Thus it suffices to solve the recurrence relation
\[ \begin{align*}
 g_0 - f_0^+ &= 0; \\
 g_m - f_m^+ &\leq g_m-\sqrt{n} - f_m^+ + 2\ell^{-n/\mu^*+\sqrt{n}/\mu^*}. 
\end{align*} \] (228)

The result is
\[ g_{n_{\text{rat}}} - f_{n_{\text{rat}}}^+ \leq \epsilon^{-n/\mu^*+o(1)}. \] (229)

In other words
\[ g_{n_{\text{rat}}} \leq f_{n_{\text{rat}}}^+ + \epsilon^{-n/\mu^*+o(1)} = \epsilon^{-n/\mu^*+o(1)}. \] (230)

Since right after Formula (192) we replaced $\mu^*$ by $\mu' - \epsilon$, this $\epsilon$ cancels $o(1)$ as $n \to \infty$. Hence we can really say that
\[ g_{n_{\text{rat}}} \leq \epsilon^{-n/\mu'} = N_n^{-1/\mu'} \] (231) (232)

and that
\[ e_{n_{\text{rat}}}^0 = I(W) - g_{n_{\text{rat}}} \geq I(W) - N_n^{-1/\mu'}. \] (233)

F. Sixth estimate how good synthetic channels in $E_{n_{\text{rat}}}^0$ are

They are synthetic channels such that during the time they are being trained, $Z(W_{m+i})$ is never larger than $\delta$. Therefore $Y_{m+i} > Y_{m+i} - \epsilon$ holds. They are also synthetic channels such that
\[ \frac{Y_{m+1} + Y_{m+2} + \cdots + Y_n}{n-m} > \beta \log \ell \]
\[ \frac{1-m/n}{1-m/n} + \epsilon \] (234)

so
\[ Y_{m+1} + Y_{m+2} + \cdots + Y_n > \beta' n \log \ell. \] (235)

Therefore for every $w \in E_m$ and $v$ its ancestor at depth $m$, by telescoping
\[ \log(-\log Z(w)) - \log(-\log Z(v)) > \beta' n \log \ell. \] (236)

But $v$ are such that $Z(v) \leq \exp(-\exp(m^{1/3}))$, so
\[ Z(w) < \exp(-\exp(\beta' n \log \ell + m^{1/3})). \] (237)
Sum over $E_{0}^{m}$:
\[
\sum_{w \in E_{0}^{m}} Z(w) < N_{n} \exp\left(-\exp\left(\beta' n \log \ell + m^{1/3}\right)\right). \tag{238}
\]

Let $A_n$ be $E_{0}^{m}$. Eventually, as $n \to \infty$, the term $m^{1/3}$ eats up the term $N_{n}$ in front of exp:
\[
\sum_{w \in A_n} Z(w) < \exp\left(-\exp\left(\beta' n \log \ell\right)\right) = \exp\left(-N_{n}^{\beta'}\right). \tag{239}
\]

G. Seventh we announce the code
\[
(\tau_{\text{per}}^{\text{per}}(W, T, n), A_n) \tag{240}
\]

has block length
\[
N_n = \ell^m, \tag{241}
\]

code rate
\[
R_n = \mathbb{P}(A_n) = c_{0}^{m} \geq I(W) - N_{n}^{1/\mu'}, \tag{242}
\]

and error probability
\[
P_n = \sum_{w \in A_n} Z(w) < \exp(-N_{n}^{\beta'}). \tag{243}
\]

This finishes the proof of Theorem 5.

VII. Prove Theorem 6 by Combining Lemma 4 and Theorem 5

A. First apply Lemma 4 to $T_{\text{rat}}$

The conditions posed in Lemma 4 coincide with conditions posed on $T_{\text{rat}}$. Therefore $T_{\text{rat}}$ produces block codes such that
\[
N_m = \ell^m, \tag{244}
\]
\[
R_m > I(W) - N_m^{-1/\mu'} + o(1), \tag{245}
\]
\[
P_m < \exp\left(-\exp\left(m^{1/3}\right)\right). \tag{246}
\]

B. Second grow a special channel tree
\[
\tau_{\text{per}}^{\text{per}}(W, T_{\text{rat}}, n_{\text{rat}}, T_{\text{err}}, n). \tag{247}
\]

Here $n$ is a positive integer and $n_{\text{rat}} = n \mu' / \mu'$.

\begin{tabular}{|c|c|}
\hline
Stock & Begin with $\tau_{\text{per}}^{\text{per}}(W, T_{\text{rat}}, n)$. \\
Prune & Let $A_m$ and $A_0^{m}$ be defined as in Section VII.B. \\
& For every synthetic channel in $A_0^{m}$, detach its descendants. \\
Graft & To every leaf channel $w$ in the remaining channel tree, append $\tau_{\text{per}}^{\text{per}}(w^k, T_{\text{err}}, n - \text{depth}(w))$. Here $w^k = T_{\ell}^k(w)$ is the $k$-th power of $w$. \\
\hline
\end{tabular}

In terms of Sankey diagram:

\[
\text{channels} \quad \sqrt{n} \quad 2\sqrt{n} \quad n_{\text{rat}} \quad n \quad \text{depth} \tag{248}
\]

See Formula (313) in Appendix B for the big diagram.

Here is a small, but illustrative, example: Stock: choose $T_{\text{An}}$ to be $T_{\text{rat}}$ and prepare $T_{\text{per}}^{\text{per}}(W, T_{\text{An}}, 3)$ to begin with. (Unlike Formula (26), we omit labeling $T_{\text{An}}$.)

Prune: if it happens that $A_0^{m}$ contains $W^A$, $(W^B)^2$, and $(W^C)^4$ (highlighted in yellow background), remove their descendants.

\[
W \quad W^A \quad W^B \quad W^C \tag{250}
\]

Graft: let $k = 1$ (so $T_{\ell}^1$ does nothing) and choose $T_{\text{An}}$ again as $T_{\text{err}}$: attach three trees $\tau_{\text{per}}^{\text{per}}((W^A)^1, T_{\text{An}}, 2)$ and $\tau_{\text{per}}^{\text{per}}((W^B)^2)^1, T_{\text{An}}, 1)$ and $\tau_{\text{per}}^{\text{per}}((W^C)^3)^1, T_{\text{An}}, 0)$ to the corresponding leaves.

\[
W \quad (W^A)^1 \quad (W^B)^2 \quad (W^C)^3 \tag{251}
\]

The depth of the attached subtrees are chosen such that the resulting tree is balanced. It is practically pointless, but legal and coherent, to have $T_{\ell}^k$ at the bottom of a channel tree.

Here is another example. This time $k = 2$ so $T_{\text{rat}}$ and $T_{\text{err}}$ are of different arities. Stock

\[
W \quad (W^1)^{(1)} \quad (W^1)^{(2)} \quad (W^1)^{(3)} \quad (W^1)^{(4)} \quad (W^2)^{(1)} \quad (W^2)^{(2)} \quad (W^2)^{(3)} \quad (W^2)^{(4)} \quad (W^3)^{(1)} \quad (W^3)^{(2)} \quad (W^3)^{(3)} \quad (W^3)^{(4)} \quad (W^4)^{(1)} \quad (W^4)^{(2)} \quad (W^4)^{(3)} \quad (W^4)^{(4)} \quad \text{...
}
\]

(252)
We now verify that these three facts hold in case of \( T_{\text{fct}}(W, T, n) \): The first fact is the definition of \( A_m \), which we inherit in growing the channel tree. The second fact is by Formula \( 253 \) and Lemma \( 3 \). The third fact is by how we grow the channel tree \( T_{\text{fct}}(W, T, n) \). Now we are sure that all three facts hold.

Let \( A_n \) be defined as in Section \( \text{VI} \) the proof of Theorem \( 5 \).

(253)

E. Fifth we announce the code

\[
(\mathcal{T}_{\text{fct}}(W, T, n), A_n)
\]

has block length

\[
N_n = k\ell^n,
\]

code rate

\[
R_n = I(W) - N_n^{-1/\mu'},
\]

and error probability

\[
P_n < \exp(N\beta').
\]

This finishes the proof of Theorem \( 6 \).

VIII. APPLICATION: TO APPROACH THE HYPOTENUSE

In this section, fix the relation \( \ell = 2^k \).

Lemma 7. Assume BEC. There exist binary, length-\( \ell \), bounded transformations \( T_{\text{rat}} \) with \( \mu^* \)-exponents \( \mu_{\text{rat}} \) and \( \partial \)-dices \( Y_{\text{rat}} \) such that

\[
P\{Y_{\text{rat}} = 0\} < \ell^{-1/\mu_{\text{rat}}}
\]

and, as \( \ell \to \infty \),

\[
\mu_{\text{rat}} \to 2.
\]

Proof: That \( \mu_{\text{rat}} \to 2 \) is by \( [FHMV17, \text{Theorem 2 and 3}] \). On BEC, \( Z \)-parameters form a martingale, so transformations are bounded. The condition on \( P\{Y_{\text{rat}} = 0\} \) is a consequence of the fact that an \([n, n - \sqrt{n}]\)-random code has minimal distance at least 2 with high probability or the fact that an \([n, \sqrt{n}]\)-random code has no all-zero column.

Lemma 8. There exist \( \ell \)-ary, length-\( \ell \), bounded transformations \( T_{\text{err}} \) with \( \partial \)-dices \( Y_{\text{err}} \) following the uniform distribution on \( \log 1, \log 2, \ldots, \log \ell \) for all \( \ell : \leq 2^k \).

Proof: \([\text{MT10a}], \text{MT10b}, \text{MT14}\).}

Theorem 9. Assume BEC. For every point \((\beta', 1/\mu')\) inside the right triangle

\[
(0, 1/2), (0, 0), (1, 0)
\]

there exist \( k, \ell \) and transformations \( T_{\text{rat}} \), \( T_{\text{err}} \) that produce block codes \((T_n, A_n)\) such that

\[
N_n = k\ell^n,
\]

\[
R_n > I(W) - N_n^{-1/\mu'}, \text{ and}
\]

\[
P_n < \exp(-N_n\beta').
\]
for \( n \) large enough.

**Proof:** See Section VIII-A right after the corollary below.

**Corollary 10.** Assume BEC. For every point \((\beta', 1/\mu')\) on the hypotenuse of the right triangle

\[
(0, 1/2) \quad (0, 0) \quad (1, 0)
\]

(265)

and every monotonically increasing, unbounded function \( h \), there exist a series of polar-like codes \((T_n, A_n)\) such that

\[
N_n = k \ell^n, \quad R_n > I(W) - N_n^{-1/\mu' + o(1)}, \quad P_n < \exp(-N_n^{\beta' - o(1)}),
\]

(266) (267) (268)

and

\[
\text{complexity} < h(N)N \log N
\]

(269)

for \( n \) large enough.

**Proof:** Approximate the point on the hypotenuse by points inside the right triangle. Apply Theorem 9 to each point and then apply the diagonal argument (as in the proof of Arzelà-Ascoli theorem).

**A. Proof of Theorem 9**

Fix a point \((\beta', 1/\mu')\) inside the right triangle. Since we have Theorem 6 and Lemma 7 and 8, it suffices to find an \( \ell := 2^k \), which determines \( \mu_{\text{rat}} \) (probabilistically) and \( Y_{\text{err}} \), such that, for all \( \pi \in [0, 1] \),

\[
\frac{(1 - \pi) \log \ell}{\mu' - \pi \mu_{\text{rat}}^*} < \Lambda_{\text{err}}^*(\beta' \mu' \log \ell / \mu' - \pi \mu_{\text{rat}}^*).
\]

(270)

Start from the fact that \( Y_{\text{err}} \) follows the uniform distribution on \( \log 1, \log 2, \ldots, \log \ell \). The cumulant generating function satisfies

\[
\log \mathbb{E}[\exp(\lambda Y_{\text{err}})] = \log \mathbb{E}[X_{\text{err}}^\lambda]
\]

(271)

where \( X_{\text{err}} \) follows the uniform distribution on \( 1, 2, \ldots, \ell \). For \( -1 < \lambda < 0 \), the \( \lambda \)-moment is

\[
\mathbb{E}X_{\text{err}}^\lambda = \frac{1}{\ell} \sum_{X=1}^\ell X^\lambda < \frac{1}{\ell} \int_0^\ell X^\lambda \, dX = \frac{\ell^\lambda}{\lambda + 1}.
\]

(272)

This leads to an approximation

\[
\log \mathbb{E}[\exp(\lambda Y_{\text{err}})] < \lambda \log \ell - \log(\lambda + 1).
\]

(273)

The Cramér function is then bounded by

\[
\Lambda_{\text{err}}^*(y) \geq \sup_{\lambda < 0} \lambda y - \lambda \log \ell + \log(\lambda + 1).
\]

(274)

Redeem the supremum at \( \log \ell - y = 1/(\lambda + 1) \) to obtain

\[
\Lambda_{\text{err}}^*(y) > \left( \frac{1}{\log \ell - y - 1} \right) (y - \log \ell) + \log \frac{1}{\log \ell - y}
\]

(275)

\[
= -1 + \log \ell - y - \log(\log \ell - y)
\]

(276)

\[
\geq -1 + \log \ell - y - \log(\log \ell - y) + \frac{y}{\log \ell}
\]

(277)

\[
= (\log \ell - y) \left( 1 - \frac{1}{\log \ell} \right) - \log \log \ell.
\]

(278)

The last line is linear in \( y \). It is \( \log \ell - 1 - \log \log \ell \approx \log \ell \) when \( y = 0 \) and is 0 when

\[
y = y^* := \log \ell - \frac{\log \log \ell}{1 - 1/\log \ell}.
\]

(279)

Back to the fact that \((\beta', 1/\mu')\) is inside the right triangle

\[
(0, 1/2) \quad (\beta', 1/\mu') \quad (0, 0) \quad (1, 0)
\]

(280)

There exist a \( \mu_{\text{rat}}^* > 2 \) (by letting \( \ell \to \infty \))

\[
(0, 1/2) \quad (0, 1/\mu_{\text{rat}}^*) \quad (0, 0) \quad (1, 0)
\]

(281)

and a \( y^*/\log \ell < 1 \) (by letting \( \ell \to \infty \))

\[
(0, 1/2) \quad (y^*/\log \ell, 0) \quad (0, 0) \quad (1, 0)
\]

(282)

such that these three points are collinear

\[
(0, 1/2) \quad (\beta', 1/\mu') \quad (0, 0) \quad (1, 0)
\]

(283)

Fix \( \ell, \mu_{\text{rat}}^*, y^* \) as above. The term

\[
\frac{y^* - y}{y^*/\mu_{\text{rat}}^*} \log \ell
\]

(284)

is also linear in \( y \). It is less than \( \log \ell/2 \) when \( y = 0 \) and is 0 when \( y = y^* \). Thus, for all \( 0 \leq y \leq y^* \),

\[
\frac{y^* - y}{y^*/\mu_{\text{rat}}^*} \log \ell \leq (\log \ell - y) \left( 1 - \frac{1}{\log \ell} \right) - \log \log \ell.
\]

(285)

(286)
because the inequality holds for endpoints and both sides are linear in $y$. Concatenate with Formula (279) to obtain, for all $0 \leq y \leq y^*$,
\[
\frac{y^* - y}{y^* \mu_{rat}^*} \log \ell < \Lambda^*_\text{ent}(y).
\] (287)

On each side of the inequality, replace $y$ with the corresponding side of the equality due to collinearity below
\[
y^*(\mu' - \mu_{rat}^*) = \frac{\beta' \mu' \log \ell}{\mu' - \pi \mu_{rat}^*} \quad (288)
\]
to get
\[
(1 - \pi) \log \ell < \Lambda^*_\text{ent}(\frac{\beta' \mu' \log \ell}{\mu' - \pi \mu_{rat}^*}).
\] (289)

This is exactly what we need to apply Theorem 5.

This proof is very similar to [WD18, Corollary 8].

**IX. FURTHER IMPLICATIONS**

There is another way to state Theorem 5. We put this as a claim since we omit the details of the proof.

**Claim 11.** Let $T$ be a length-$\ell$, bounded transformation with $\mu^*$-exponent $\mu^*$ and $\partial$-dice $Y$. Let $\Lambda^*$ be the Cramér function of $Y$. If $(\beta', 1/\mu')$ does not lie in the convex hull of the point $(0, 1/\mu^*)$, union the epigraph of the function
\[
\beta \mapsto \frac{\Lambda^*(\beta \log \ell)}{\log \ell},
\] (290)
then $(\beta', 1/\mu')$ is possible.

**Sketch:** As a function of $\pi$, consider points
\[
Q(\pi) := \left( \frac{\beta' \mu'}{\mu' - \pi \mu^*}, \frac{1 - \pi}{\mu' - \pi \mu^*} \right).
\] (291)

Here is the trace of $Q(\pi)$ when $\pi = 0, 1, \ldots, 1$: for $\pi = 0$, $Q(0)$ coincides with $(\beta', 1/\mu')$; for $\pi = 1$, $Q(1)$ is on the horizontal axis; for intermediate $\pi$, the $Q(\pi)$ moves along the ray starting at $(0, 1/\mu^*)$ through $(\beta', 1/\mu')$.

From the graph, we learn that $(\beta', 1/\mu')$ does not lie in the convex hull iff $Q(\pi)$ is not in the epigraph for all $\pi \in [0, 1]$; The later happens iff $\mu < \Lambda^*(\beta \log \ell)/\log \ell$ for all $\pi \in [0, 1]$; iff the criteria of Theorem 5 are met. 

Here is a running example: For $T_{\Lambda^*}$, the rescaled Cramér function $\beta \mapsto \Lambda^*(\beta \log 2)/\log 2$ coincides with the relative entropy
\[
\beta \mapsto 1 + \beta \log_2 \beta + (1 - \beta) \log_2 (1 - \beta)
\] (293)
for $0 \leq \beta \leq 1/2$. For $1/2 \leq \beta \leq 1$, the “classical definition” of the Cramér function still coincides with the relative entropy. In our definition, however, we insist that the supremum is taken over negative $\lambda$ so $\Lambda^*$ vanishes. In the following graph, the curve is the relative entropy and the shaded area is the epigraph of $\beta \mapsto \Lambda^*(\beta \log 2)/\log 2$.

Together with $(0, 1/\mu^*)$ they form a convex hull

Back to Claim 11 If $(\beta', 1/\mu')$ is here

then some $Q(\pi)$ is in the epigraph and the criteria of Theorem 5 fail. On the other hand, if $(\beta', 1/\mu')$ is here

then all $Q(\pi)$ are outside the epigraph and Theorem 5 applies. Another interesting case is when $(\beta', 1/\mu')$ is in the tiny tip area at the bottom. Therein all $Q(\pi)$ are outside the epigraph and Theorem 5 applies.
A. Moderate Deviations Regime Recovers Error Exponent Regime as a Special Case

The following is a consequence of the Claim [11] plus the fact that \( A^*(y) \) reaches zero at \( y = EY \).

**Proposition 12.** Let \( T \) be a length-\( \ell \), bounded transformation with \( \mu^*-\)exponent \( \mu^* < \infty \) and \( \beta^*-\)exponent \( \beta^* > 0 \). For any \( \beta' < \beta^* \), there exists \( 1/\mu' > 0 \) such that \((\beta', 1/\mu')\) is possible.

See also [BGST18] Theorem 2.16.

B. Moderate Deviations Regime Recovers Scaling Exponent Regime as a Special Case

The following is another consequence of the Claim 11:

**Proposition 13.** Let \( T \) be a length-\( \ell \), bounded transformation with \( \mu^*-\)exponent \( \mu^* < \infty \) and \( \beta^*-\)exponent \( \beta^* > 0 \). For any \( 1/\mu' < 1/\mu^* \), there exists \( \beta' > 0 \) such that \((\beta', 1/\mu')\) is possible.

This is a generalization of [WD18] Corollary 8.

C. Arıkan’s Polar Codes Attacking on BEC

The three corner dots are \((0, 5), (0, 0), \) and \((1, 0)\). [GX13] proves that it is possible to achieve \((\beta', 1/\mu') = (0.49, O(1))\). It is represented as a point very close to \((5, 0)\). [MHU16] proves an interpolating result. Their curve connects \((0, 1/4.627)\) and \((5, 0)\) and is drawn below. Theorem 5 (and also [WD18]) implies a better curve. This curve connects \((0, 1/3.627)\) and \((5, 0)\). Notice that in this scenario, \( \mu^* = 3.627 \) is given by [FV14].

[Diagram (298)]

D. Arıkan’s Polar Codes Attacking on BDMC

BDMC is not far from BEC in the sense that almost all treatments are the same except \( \mu^* = 4.714 \) instead of 3.627. In particular, the curves are drawn using the same formulae with the new \( \mu^* \). So this time the Theorem 5 curve connects \((0, 1/4.714)\) and \((5, 0)\). And the [MHU16] curve connects \((0, 1/5.714)\) and \((5, 0)\). Notice that in this scenario, \( \mu^* = 4.714 \) is given by [MHU16].

[Diagram (299)]

E. Arıkan’s Polar Codes Attacking on AWGN

[FT17] analyzes the AWGN channel and mimic [MHU16]. They end up with the same curve as the bottom one in the previous plot that connects \((0, 1/5.714)\) and \((5, 0)\). Theorem 5 implies the same curve as the top one in the previous plot that connects \((0, 1/4.714)\) and \((5, 0)\). Notice that in this scenario, \( \mu^* = 4.714 \) is given by [FT17].

[Diagram (300)]

F. Polar Codes with Larger Kernels Attacking on BEC

1) Pessimistic Case: We present two fake curves that illustrates the fact that Theorem 5 can be used to connect \((0, 1/\mu^*)\) and \((\beta^*, 0)\). The left curve with [FHMV17] as an endpoint shows that there are kernels such that \(1/\mu^*\) are arbitrarily close to \(1/2\); while the \( \beta^*\)-exponents of these kernels are unknown. The bottom curve with [KSU10] as an endpoint shows that there are kernels such that \( \beta^*\) are arbitrarily close to \(1\); while the \( \mu^*\)-exponents of these kernels are unknown. Besides the two curves, [BGST18] shows that it is possible to approach where [KSU10] is with positive \(1/\mu'-\)value.

[Diagram (301)]

2) Optimistic Case: Moreover, if there are kernels such that \((\beta^*, 1/\mu^*)\) converges to \((1, 1/2)\), then Theorem 5 will eventually cover the right triangle.

[Diagram (302)]

The existence of such kernels is not clear at this stage. This is one of the reasons why we develop Theorem 6 — which is
basically saying that we can steal the good \( \mu^\ast \)-exponent from a kernel and steal the good \( \beta^\ast \)-exponent from another.

Chances are that random kernels possesses good \( \mu^\ast \) and good \( \beta^\ast \)-exponents. And we can use Hoeffding’s inequality to control the behavior of Cramér functions.

**G. Polar Codes with Larger Kernels Attacking on BDMC**

For binary channels other than BEC, [FHMV17] does not apply anymore. Then [BGN+18] takes place and proves that all kernels, in particular kernels from [KSU10], have positive \( 1/\mu^\ast \). We draw a fake curve to illustrate that Theorem 5 connects the points given by [BGN+18] and by [KSU10].

**H. Polar Codes with Larger Kernels Attacking on General Channels**

For channels that are not binary, [KSU10] does not apply anymore. Then [BGST18] steps in and comments that BCH codes, in general, fill in the blank that there are kernels with \( \beta^\ast \) arbitrarily close to one. We again draw a fake curve to illustrate that Theorem 5 connects the points representing \( 1/\mu^\ast \) and \( \beta^\ast \).

**I. Concatenated Polar Codes Attacking on BEC**

If concatenated polar codes are allowed, then Theorem 9 shows that it is possible to fill the right triangle. We draw a fake to illustrate this.

**J. Concatenated Polar Codes Attacking on General Channels**

For general channels other than BEC, [FHMV17] does not apply. We may apply Theorem 5 or 6 according to whether we want a single kernel or two kernels. We draw a fake to illustrate this.

**K. Arıkan and Reed–Solomon Codes Attacking on BEC**

We consider this a killer application. See [BJE10] for a result similar to [HMTU13]. See [GB14a] for a result similar to [GX13], [BGST18]. See [MEKL13], [MEKL14] for more.

For \( k = 1 \), the transformation \( T_{RS2} \) is \( T_{An} \). There is no concatenation happening.

For \( k = 2 \), transformations \( T_{An}^{\otimes 2}, T_{C}^2, T_{RS4} \) collaboratively beat \( T_{An} \). In particular \( \beta^\ast = (3 + \log_2 3)/8 = .57 \).

For \( k = 3 \), transformations \( T_{An}^{\otimes 3}, T_{C}^3, T_{RS8} \) are even better.
We put $k = 4$ (and $T_{\text{An}}^\otimes 4, T_\mathbb{C}^4, T_{\text{RS16}}$) here just in case the trend is not clear.

It is not hard to see that this series of curves eventually converges to a segment that connects $(0, 1/3.627)$ and $(1, 0)$. As $k \to \infty$, points [GB14a], [BJE10] also converge to $(1, 0)$, at a faster pace.

(L. Arikan and Reed–Solomon Codes Attacking on BDMC and AWGN)

For BDMC and AWGN, curves are connecting $(0, 1/4.714).$

See Appendix D for more types of concatenations.

X. Future Works

What we do not address in this work is whether Theorem 5 and 6 give optimal bounds. For one, it is difficult to imagine that a description as simple as Claim 11 is not the answer. That said, we look forward to a second-order result just like [MT14] extending [AT09].

On the other hand, statements and proofs in this work heavily rely on the magical value $\mu^*$. The problem, as of today, is we can bound or approximate $\mu^*$ but do not know if the limit exists. Should there be distinct $\mu^*$ and $\mu_*$ as limit superior and limit inferior, we expect two curves connecting $(0, 1/\mu^*)$ and $(0, 1/\mu_*)$ to $(\beta^*, 0)$.

Having Theorem 9 and Corollary 10 we like to see if they extend to channels other than BEC. Particularly, does $\mu^*$ achieve 2 for general channels? Furthermore, are there kernels with good $\mu^*$ and $\beta^*$?

XI. Conclusion

We provide a merciful generalization of polar codes and are able to characterize, for a subclass of polar-like codes, the tradeoff among block length, code rate, and error probability asymptotically.

We then show that a grafted variant of polar coding almost catches up the performance of random codes on BEC, if arbitrary kernels are allowed.

If one likes to stick to Reed–Solomon kernels, we characterize the performance as well.

ACKNOWLEDGMENT

H.-P. W. thanks Chi-Hua Wang (Purdue University) for priceless discussions on probability topics.

REFERENCES

[AM14] Sarah E. Anderson and Gretchen L. Matthews. Exponents of polar codes using algebraic geometric code kernels. Designs, Codes and Cryptography, 75(2):699–717, Nov 2014.

[Ari09] E. Arikan. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. IEEE Transactions on Information Theory, 55(7):3051–3073, July 2009.

[Ari15] E. Arikan. A packing lemma for polar codes. In 2015 IEEE International Symposium on Information Theory (ISIT), pages 2441–2445, June 2015.

[AT09] E. Arikan and E. Telatar. On the rate of channel polarization. In 2009 IEEE International Symposium on Information Theory, pages 1493–1495, June 2009.

[AW10] Y. Altuğ and A. B. Wagner. Moderate deviation analysis of channel coding: Discrete memoryless case. In 2010 IEEE International Symposium on Information Theory, pages 265–269, June 2010.

[AW14] Y. Altuğ and A. B. Wagner. Moderate deviations in channel coding. IEEE Transactions on Information Theory, 60(8):4417–4426, Aug 2014.

[AYK11] A. Alamdar-Yazdi and F. R. Kschischang. A simplified successive-cancellation decoder for polar codes. IEEE Communications Letters, 15(12):1378–1380, December 2011.

[BBGL17] Meryem Benammar, Valerio Bioglio, Frederic Gabry, and Ingmar Land. Multi-kernel polar codes: Proof of polarization and error exponents. CoRR, abs/1709.10371, 2017.

[BCL18] V. Bioglio, C. Condo, and I. Land. Memory Management in Successive-Cancellation based Decoders for Multi-Kernel Polar Codes. ArXiv e-prints, September 2018.

[BGLB17] V. Bioglio, F. Gabry, I. Land, and J. Belfiore. Minimum-distance based construction of multi-kernel polar codes. In GLOBECOM 2017 - 2017 IEEE Global Communications Conference, pages 1–6, Dec 2017.

[BGN+18] Jaroslav Blasiok, Venkatesan Guruswami, Preetum Nakkiran, Atri Rudra, and Madhu Sudan. General strong polarization. CoRR, abs/1802.05718, 2018.

[BGS18] J. Blasiok, V. Guruswami, and M. Sudan. Polar Codes with exponentially small error at finite block length. ArXiv e-prints, October 2018.

[BGZ12] Gregory Bonik, Sergei Goreinov, and Nickolai Zamarashkin. Construction and analysis of polar and concatenated polar codes: practical approach. CoRR, abs/1207.4343, 2012.

[BJE10] M. Bakshi, S. Jaggi, and M. Effros. Concatenated polar codes. In 2010 IEEE International Symposium on Information Theory, pages 918–922, June 2010.
APPENDIX

A. Polar Code Error Exponent Regime

Inner and outer bounds for usual polar codes [AT09], [KSU10], [HMTU13], [MT14]. [BBGL17] proposes and solves an interesting question: We have four matrices, say $G_{\text{Ben}}, G_{\text{Bio}}, G_{\text{Gab}}, G_{\text{Lan}}$. They induce four transformations $T_{\text{Ben}}, T_{\text{Bio}}, T_{\text{Gab}}, T_{\text{Lan}}$ and we want to apply them alternately. Even more excitingly, we throw dices to decide which transformation to apply.

In this setup, one may argue that the four transformations actually form a compound transformation $T_{\text{BBGL}}$ with build-in randomness. In particular, the $\delta$-dice $Y_{\text{BBGL}}$ follows a compound distribution derived from $Y_{\text{Ben}}, Y_{\text{Bio}}, Y_{\text{Gab}}, Y_{\text{Lan}}$. Not only their result (an $N$-$P$ tradeoff) follows immediately, but it also automatically upgrades to an $N$-$R$-$P$ tradeoff.

B. Polar Code Scaling Exponent Regime

See [FHMV17] for a good review.

Outer bounds [Dob61], [Str62], [IZ00], [Mon01], [Hay09], [PPV10].

Inner bounds [KMTU10], [HAU14], [GB14b], [MHU16], [FPV14], [PU16], [Has13], [FHMV17].

List decoder [MHU15].

C. Polar Code Moderate Deviations Regime

Outer bound [AW10], [PV10], [AW14], [Art15], [HT15].

Inner bound [GX13], [MHU16], [PT17], [BGN+18], [WD18], [BGS18].

D. Other Types of Concatenations

There are a lot of works trying to concatenate polar codes with Reed–Solomon codes or RS-polar codes. The list includes but is not limited to [BJE10], [KSH11], [MEKL13], [MEKL14], [GB14a], [WZL+17].

Polar with BCH codes [WN14], [WNH16].

Polar with algebraic geometry codes [ED13], [AM14].

Polar with LDPC codes [EPN11], [EPN13], [QQiFS14], [ZLG+14], [MLZ17], [YSL+18], [ZLHC18].

Polar with RA codes [YZ16].

Polar with single parity check code [YM17].

Polar with small, ML-decodable codes [SH10], [BG17].

Polar with arbitrary outer codes [TS11], [GB17].

Polar kernels with various length [BBGL17], [BCL18], [BGL17], [GGL17].
E. Big Sankey Diagram

\[ \sqrt{n} \quad 2\sqrt{n} \quad 3\sqrt{n} \quad n \]

depth

channels

frozen

recruited

trained

retained

(313)

\[ \sqrt{n} \quad 2\sqrt{n} \quad 3\sqrt{n} \quad n \]

depth

channels

recruited

trained

retained

(314)

\[ \sqrt{n} \quad 2\sqrt{n} \quad 3\sqrt{n} \quad n \]

depth

channels

(315)
stop recruiting
recruited trained frozen retained
recruited trained frozen retained
recruited trained frozen retained

(316)

stop recruiting

A
B
C
D
E

A
B
C
D
E

(317)

stop pruning

grafted

(318)