Polar Codes' Simplicity, Random Codes' Durability

Hsin-Po Wang and Iwan M. Duursma

***Abstract*— Over any discrete memoryless channel, we offer error correction codes such that: for one, their block error probabilities and code rates scale like random codes'; and for two, their encoding and decoding complexities scale like polar codes'. Quantitatively, for any constants $\pi, \rho > 0$ such that $\pi + 2\rho < 1$, we construct a sequence of block codes with block length N approaching infinity, block error probability $\exp(-N^{\pi})$, code rate $N^{-\rho}$ less than the Shannon capacity, and encoding and decoding complexity $O(N \log N)$ per code block. The core theme is to incorporate polar coding (which limits the complexity to polar's realm) with large, random, dynamic kernels (which boosts the performance to random's realm). The putative codes are optimal in the following manner: Should $\pi + 2\rho > 1$, no such codes exist over generic channels regardless of complexity.**

***Index Terms*— Capacity-achieving codes, low-complexity codes, polar codes, random codes.**

I. Introduction

RICHARD W. HAMMING is one of the first few people who had the idea that by grouping information in blocks with redundancies, a calculating machine can correct errors by its own and proceed to the next command instead of halting. Their solution, now called Hamming codes, is found in [1]. Claude E. Shannon, a colleague of Hamming at Bell Labs, theorized the *communication channels* and showed that a channel associates to a number called *capacity*, which represents the ultimate limit of the efficiency of communications over that channel.

To brief the rest of the history, we follow the analogy [21] used. Shannon's eternal result, the noisy-channel coding theorem [2], is considered the analog of the law of large numbers (LLN). The theorem implies that there exists a sequence of longer and longer block codes whose block error probabilities approach 0 and code rates approach the capacity, which is analogous to that the empirical average of random variables is close to the mean with high probability. Robert G. Gallager, Shannon, Robert M. Fano, and followers extended the LLN result by looking at how the block error probability P_e scales when the code rate R is fixed. They showed that the error probability P_e scales like $\exp(-E_\mathrm{r}(R)N)$. Here N is the block length, and $E_\mathrm{r}(R)$ is a constant depending on R. This paradigm is considered the analog of the large deviations principle (LDP). See [3]–[11]. Meanwhile, a series of works fix the error probability P_e and looked at how the code rate R scales [12]–[18]. They showed that the code rate R scales like $I - Q^{-1}(P_\mathrm{e})\sqrt{V/N}$ for I the capacity, Q^{-1} the inverse of the standard Q-function, and V an intrinsic parameter of the channel. The parameter V is called the *dispersion* or *varentropy* by different authors. It is the "variance" of the channel while I is the "mean" of the channel. This turns out to be more than an analog—the random variable $\log(W(Y \mid X)/W_\mathrm{out}(Y))$ called *information density* or *information spectrum* has mean I and variance V. This paradigm is considered the analog of the central limit theorem (CLT). Later, Altuğ–Wagner, Polyanskiy–Verdú, and followers considered the joint behavior when both P_e and R vary [19]–[23]. They showed that the quantity $N(I-R)^2/|\log P_\mathrm{e}|$ converges to $2V$, twice the very dispersion appearing in the CLT paradigm. This paradigm is considered the analog of the moderate deviations principle (MDP).

TABLE I

An Analogy Between Probability Theory and Coding Theory

Paradigm	R.v.s' behavior	Codes' behavior
LLN	$\bar{X} \to \mu$	$R \to I$ and $P_\mathrm{e} \to 0$
LDP	$P\{\bar{X} - \mu > x\} \approx \exp(-nI(x))$	$P_\mathrm{e} \approx \exp(-E_\mathrm{r}(R)N)$
CLT	$\dfrac{\sum X - \mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0,1)$	$R \approx I - \dfrac{Q^{-1}(P_\mathrm{e})}{\sqrt{VN}}$
MDP	$\dfrac{\lvert\log P\{\bar{X} - \mu > \varepsilon_n x\}\rvert}{n\varepsilon_n^2} \approx I(x)$	$\dfrac{\lvert\log P_\mathrm{e}\rvert}{N(I-R)^2} \approx \dfrac{1}{2V}$

On a parallel track, the engineering aspects of the communication theory thrive. Codes with excellent practicality are proposed. To name a few, Reed–Muller (1954), convolutional (1955), Bose–Chaudhuri–Hocquenghem (1959), Reed–Solomon (1960), trellis modulation (1970s), turbo (1990s), low-density parity-check (1963 and 1996), repeat-accumulate (1998), fountain (1998), and polar (2009).

Among the long list of inventions, only trellis modulation, low-density parity-check, and polar achieve the LLN paradigm over nontrivial channels—they are *capacity-achieving*. Among these three, polar stands out as the only code that achieves

Manuscript received February 15, 2020; revised August 21, 2020; accepted November 15, 2020. Date of publication December 1, 2020; date of current version February 17, 2021. This work was supported in part by an award from the University of Illinois at Urbana–Champaign Campus Research Board. *(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 X. Kliewer, Associate Editor for Coding Techniques.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2020.3041570.

Digital Object Identifier 10.1109/TIT.2020.3041570

the CLT paradigm (optimally), the only code that achieves the LDP paradigm (optimally), and the only code that achieves the MDP paradigm (suboptimally). If only polar code achieves the optimal MDP paradigm. We brief the history of polar codes below. Unless stated otherwise, I means the *symmetric capacity* in the next three paragraphs.

Erdal Arıkan's original works on channel polarization [24], [25] established the foundation of polar codes, placing polar codes in the LLN paradigm on day one. Arıkan and Telatar [26] characterized the LDP behavior of polar codes, showing that P_{e} scales like $\exp(-\sqrt{N})$ when an $R < I$ is fixed. Later, Korada–Şaşoğlu–Urbanke [27] generalized polar codes from Arıkan's kernel $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ to any invertible ℓ-by-ℓ matrix G, granted that $\ell \geqslant 2$ and G is not column-equivalent to a lower triangular matrix. And then they showed that the LDP behavior is $P_{\mathrm{e}} \approx \exp(-N^{E_{\mathrm{c}}(G)})$ where $E_{\mathrm{c}}(G)$ is a constant depending on the kernel matrix G. The notation $E_{\mathrm{c}}(G)$ is meant to resemble Gallager's error exponent $E_{\mathrm{r}}(R)$ but the former is inserted at $\exp(-N^{\text{this}})$ place while the latter is inserted at $\exp(-\text{this}N)$ place. The LDP behavior of polar codes is then refined in [28]. Therein, P_{e} is approximated by $\exp(-\ell^{\mathcal{E}})$ where $\mathcal{E} = E_{\mathrm{c}}(G)n - \sqrt{V_{\mathrm{c}}(G)n}Q^{-1}(R/I) + o(\sqrt{n})$ is a more accurate exponent, ℓ is the matrix dimension, n is the depth of the code, and $V_{\mathrm{c}}(G)$ is another constant depending on G. The notation $V_{\mathrm{c}}(G)$ is meant to resemble the channel dispersion V. Appearing to be a CLT behavior, this result lies in the corner of the LDP paradigm that touches the MDP paradigm. Finally, Mori–Tanaka [29] generalized everything above to channels of prime power input size. Over arbitrary input alphabets, [30], [31] showed the equivalence of [25], [26]. Over binary but asymmetric channels, [32], [33] showed the counterpart of [25], [26] with I being the Shannon capacity. No further result on the LDP side, e.g. over non-binary asymmetric channels, is known. The present work fills the gap.

The CLT behavior of polar codes turns out to be difficult to characterize. It was Korada–Montanari–Telatar–Urbanke [34] who came up with the idea that approximating an *eigenfunction* tightly bounds the *eigenvalue* $\ell^{-\rho}$. Here $\rho > 0$ is a number such that R scales like $I - N^{-\rho}$ with a fixed P_{e}. They had $0.2669 \leqslant \rho \leqslant 0.2841$ over binary erasure channels (BECs). The upper bound was brought down to $3.553\rho \leqslant 1$ over binary-input discrete-output memoryless channels (BDMCs) [35]. Hassini–Alishahi–Urbanke [36] moved down the upper bound to $3.627\rho \leqslant 1$ over BECs and $3.579\rho \leqslant 1$ over BDMCs. They also proved an lower bound $1 \leqslant 6\rho$ over BDMCs. The latter is suboptimal and [37], [38] improved the bound to $1 \leqslant 5.702\rho$ and to $1 \leqslant 4.714\rho$. Additive white Gaussian noise channles (AWGNCs) have continuous output alphabet, but [39] show that they have $1 \leqslant 4.714\rho$ too. Over BECs particularly, [40], [41] examined a series of larger kernels; the current record is a 64-by-64 kernel believed to have $1 \leqslant 2.9\rho$. Near the end of the road to $2\rho < 1$, [42] showed that by allowing $q \to \infty$, Reed–Solomon kernels achieve $2\rho < 1$ over q-ary channels. This does not really prove that polar codes achieve $2\rho < 1$ over any specific channel, but gave hopes. Fazeli *et al.* [43]–[45], eventually, showed that large random kernels achieve $2\rho < 1$ over BECs, breaking the barrier. Guruswami *et al.* [46], [47] extended their result to all BDMCs utilizing the dynamic kernel technique. Over the remaining channels, the present work fills the gap.

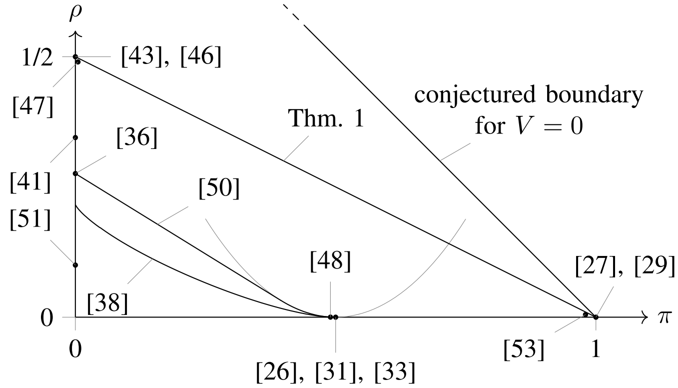

Fig. 1. Recent works on polar coding arranged on a ρ-π plot. Note that results utilizing different kernels over various channels are mixed. The higher ρ, π, the better performance. The curve part of [50] is $\rho = 1 - h_2(\pi)$.

Between LDP and CLT is polar codes' MDP behavior. Guruswami and Xia [48], [49] showed that there exists $\rho > 0$ such that P_{e} scales like $\exp(-N^{0.49})$ while R scales like $I - N^{-\rho}$ over BDMCs. This raised a question about what are the possible pairs (π, ρ) such that (P_{e}, R) scales like $(\exp(-N^{\pi}), I - N^{-\rho})$. Mondelli–Hassani–Urbanke [38] answered this, partially, in the same paper they bounded ρ. They showed that under a certain curve connecting $(0, 1/5.714)$ and $(1/2, 0)$ all (π, ρ) are achievable over BDMCs (see Fig. 1). For BECs the upper left corner is $(0, 1/4.627)$. A straightforward generalization to AWGNCs was also given in [39]. We in [50] improved their result, suggesting that via a combinatorial trick the upper left corner of the curve is $(0, \rho)$ for any ρ that is valid in the CLT regime. The same trick also implicated that over BECs all (π, ρ) such that $\pi + 2\rho < 1$ are achievable, which is mainly owing to [43]'s result that $2\rho < 1$ over BECs is achievable. Meanwhile, [51], [52] made the first step to investigate the general kernel matrices over general prime-ary channels. They showed that it is possible to achieve $\rho > 0$ with $P_{\mathrm{e}} \approx N^{-\Omega(1)}$. This is, strictly speaking, "only" a CLT behavior as the desired error probability in the MDP world is $\exp(-N^{\pi})$. Later, Błasiok *et al.* [53], [54] were able to show that for all $\pi < E_{\mathrm{c}}(G)$ there exists $\rho > 0$ such that (π, ρ) is achievable. This makes it a direct generalization of [48] to all polarizing kernel matrices G over all prime-ary channels. The preprint [47] contains a section that pushes the conference abstract [46] to positive π while maintaining $\rho \approx 1/2$. Over the remaining channels, the present work fills the gap.

The following works, though not counting as predecessors of ours, have impact on us through their insights on the essence of the channel polarization: [55]–[70].

I resumes to be the Shannon capacity. Readers are now prepared to be presented the main theorem.

Theorem 1 (Main Theorem—Polar Codes' Simplicity, Random Codes' Durability): Let W be any discrete memoryless channel. Fix a prime $\varsigma \geqslant 2$. Fix constants $\pi, \rho > 0$ such that $\pi + 2\rho < 1$. There exists a sequence of block codes with encoding and decoding algorithms such that: (cs) the codes accept uniform ς-ary messages; (cn) the block length N approaches infinity; (cp) the block error probability falls below $\exp(-N^\pi)$; (cr) the code rate exceeds $I - N^{-\rho}$; and (cc) the encoding and decoding complexity is $O(N \log N)$ per code block.

The proof of the main theorem spans over Sections II to VIII, lemmas continuing in Appendices A to C. The entry points are Sections II-A for (cs), II-B for (cn), III-B for (cc), IV-B for (cp), and VI-C for (cr). For the special case that W is a BDMC with uniform input, Appendix D lists some simplifications of the general treatments. The main theorem is optimal in the following manner.

Proposition 2 (Optimality): Fix constants $\pi, \rho > 0$ such that $\pi + 2\rho > 1$. Assume $V > 0$. Conditions (cn), (cp), and (cr) cannot hold simultaneously.

Proof: If so, $N(I - R)^2/|\log P_\mathrm{e}| \leqslant NN^{-2\rho}/N^\pi = N^{1-2\rho-\pi} \to 0$ as $N \to \infty$. This contradicts the known result $\liminf_{N\to\infty} N(I - R)^2/|\log P_\mathrm{e}| \geqslant 2V > 0$ [21, Theorem 2], [20, Theorem 6]. Remark: For $V = 0$ channels, the *correct* threshold seems to be $\pi + \rho = 1$ [71, Inequality (3.354)], [21, Remark 1].

Bibliographic remarks: The two citations [21] and [20] point to the same result. The former uses a Gallager flavor technique that is similar to what will be seen in Section VIII-C. The latter uses the information spectrum mentioned in the introduction. This technique was developed by Hayashi but is not the same as what will appear in Section VIII-D. ■

For the rest of the section, we outline the ideas to prove Theorem 1. The proof is a straightforward remix of polar coding techniques and random coding techniques if it were not for a few hurdles.

Hurdle of input alphabet size: The majority of the polar coding theory assumes that the input alphabet of the underlying channel is binary, of prime size, or, less likely, of prime power size. But the main theorem aims for arbitrary finite alphabets. Finite alphabets do possess polarization behavior but the speed of polarization has room for improvement [31, Theorem 3.5]. We will overcome this by adding "dummy symbols" into the input alphabet to make it a prime power.

Hurdle of asymmetric channel: Although asymmetric channels do polarize, the input distributions do not automatically become the uniform distribution. Pre-composing a source coding machinery helps shape the desired distribution and has been proposed before [30, Section III.D]. On the other hand, Honda and Yamanoto [33] showed that *one* polar code can do both source coding and noisy-channel coding at once. We borrow their idea.

Hurdle of kernel selection: Judging and identifying the best-behaved kernel gets harder as we need finer descriptions of the performance of the code. The good result for the BEC case depends heavily on the erasure nature of the channels (that they are *totally ordered* by their capacities). Other general results are not strong enough to meet our goal. To overcome, we borrow a technique called *dynamic kernels* from [72]. The idea is to prepare more than one polarizing kernel and apply a proper one on a channel-by-channel basis. This makes a paradigm shift from *one kernel fits all channels* to *every channel deserves a tailor-made kernel*. We will, once per channel, apply the random coding theory to show the existence of a proper kernel.

Hurdle of output alphabet size: Even with the great freedom to choose one kernel for each and every channel, there lies the difficulty that some performance bounds are proven with one fixed channel in mind to favor the big-O notations. Those bounds are prone to depend on the size of the output alphabet, which grows to infinity as the channel transformations take place. Meanwhile, some *universal bounds* are proven that depend only on the size of the input alphabet, which is invariant under channel transformations. We will borrow a bound derived in [73], [74].

A. Organization

Section II reviews channels and entropy notations; Section II-A explains how to overcome the hurdle of arbitrary input alphabet size. Section III reviews the channel transformations; Section III-A designs the decoder; Section III-B analyzes its complexity; Section III-C designs the encoder, overcoming the hurdle of asymmetric channel. Section IV reviews the channel parameters such as the Bhattacharyya parameter; Section IV-B shows how to control the block error probability. Section V reviews the channel processes; Section V-A argues that the global MDP behaviors of $H(\mathsf{W}_n)$ and $H(\mathsf{V}_n)$ imply the main theorem. The main theorem is thus reduced to the behavior of certain channel processes. Section VI proves that the global MDP behavior we want holds granted that the local LDP and CLT behaviors hold, effectively boiling the main theorem down to the local behaviors. Section VI-B introduces the random kernel trick and Section VI-C introduces the dynamic kernel trick to overcome the hurdle of kernel selection. Section VII confirms the local LDP behavior. The proof distills properties of the weight distributions of random codes. Sections VII-A and VII-B prove the two fundamental theorems of polar coding. Section VIII confirms the local CLT behavior. Contributions from Gallager and Hayashi are utilized. Section VIII-B invokes Chang–Sahai's universal bound, overcoming the hurdle of output alphabet size.

B. Three Families of Randomnesses

The randomnesses from the sender's message, the channel, and the randomized rounding constitute the first family. Typeset in Roman font are random variables $(U, X, Y, \dots)$, probability measures (P, Q, W), entropies (H, I), and other parameters $(P_\mathrm{e}, Z, T, S \dots)$ in this family. The randomness from the channel process, one main technique in the polar coding literature, is the second family. Typeset in sans serif font are stochastic processes $(\mathsf{K}_n, \mathsf{W}_n, \mathsf{H}_n, \mathsf{Z}_n, \dots)$, probability measure (P), and expectation (E) in this family. The randomness from random kernel ensembles, the main technique in the random coding literature, is the third family. Typeset in blackboard bold font are random variables $(\mathbb{G}, \mathbb{X}, \mathbb{K})$, probability measure ($\mathbb{P}$, with exceptions), expectation $(\mathbb{E})$, and Kullback–Leibler divergence ($\mathbb{D}$, with exceptions) in this family.

II. Channel and Entropy Preliminaries

A *discrete memoryless channel* is a finite-state Markov chain

$$W\colon \mathcal{X} \to \mathcal{Y}.$$

Here $\mathcal{X}$ is a finite set of input alphabet; $\mathcal{Y}$ is a finite set of output alphabet; and W is an array of transition probabilities $W(y \mid x) \in [0,1]$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The numbers satisfy $\sum_{y\in\mathcal{Y}} W(y \mid x) = 1$ for all $x \in \mathcal{X}$, which represents the fact that each x must be transitioned to some unique y. When $\mathcal{X}$ and $\mathcal{Y}$ are clear from the context, we call W a *channel*. Although the input distribution is not part of the channel data, we write $W_{\mathrm{in}}(x)$ to denote the input distribution. When $W_{\mathrm{in}}(x)$ is understood from the context, we write $W(x, y)$ to denote the joint distribution $W(y \mid x)W_{\mathrm{in}}(x)$, write $W_{\mathrm{out}}(y)$ to denote the output distribution, and write $W(x \mid y)$ to denote the a posteriori probability $W(x, y)/W_{\mathrm{out}}(y)$. (Thus the interpretation of $W(\bullet \mid \bullet)$ depends on the arguments and the context.) A tuple of inputs $(x_i, x_{i+1}, \ldots, x_j)$ is abbreviated as x_i^j. Same for y_i^j for tuple of outputs, and for u_i^j for general variables. We assume memoryless channels, and write $W^\ell(y_1^\ell \mid x_1^\ell)$ to denote the product measure $\prod_{i=1}^{\ell} W(y_i \mid x_i)$ for consecutive usages. We write $W_{\mathrm{in}}^\ell(x_1^\ell)$, $W_{\mathrm{out}}^\ell(y_1^\ell)$, $W^\ell(x_1^\ell, y_1^\ell)$, and $W^\ell(x_1^\ell \mid y_1^\ell)$ to denote the input, output, joint, and a posteriori probabilities.

Let X and Y be two r.v.s (random variables). Let $H(X)$, $H(X \mid Y)$, and $I(X\,;Y)$ be the standard entropy, conditional entropy, and mutual information. The base of the logarithm will be assigned later (in Section IV). When X is the input fed into some channel $W\colon \mathcal{X} \to \mathcal{Y}$ and Y is the corresponding output, we say $H(W)$ and $I(W)$ to mean $H(X \mid Y)$ and $I(X\,;Y)$. When the distribution of X (the input distribution) is chosen to maximize $I(W)$, it is called a *capacity-achieving input distribution* and $I(W)$ is called the *(Shannon) capacity* of the channel $W\colon \mathcal{X} \to \mathcal{Y}$. Unless stated otherwise, the input distributions will be capacity-achieving.

A. Reduce Input Size to Prime Power

Immediately after we declared what channels are concerned (those with finite input and output alphabets), we show that it suffices to consider input alphabets of prime power size.

Let $W\colon \mathcal{X} \to \mathcal{Y}$ be a channel. Let the input alphabet $\mathcal{X}$ be of size s. Let q be any prime power greater than or equal to s. Degrade the channel W as follows: Let symbols in $\mathcal{X}$ be $\xi_1, \xi_2, \ldots, \xi_s$. Let $\xi_{s+1}, \xi_{s+2}, \ldots, \xi_q$ be $q - s$ extra symbols. Let $\mathcal{X}^\sharp$ be $\mathcal{X} \cup \{\xi_{s+1}, \xi_{s+2}, \ldots, \xi_q\}$; this is the extended alphabet. Define a dummy channel $\natural\colon \mathcal{X}^\sharp \to \mathcal{X}$ by letting $\natural(\xi_{\min(i,s)} \mid \xi_i)$ be 1 for all $i = 1, 2, \ldots, , q$. That is, all extra symbols collapse to ξ_s while the old symbols remain. The composition of the two channels

$$W \circ \natural\colon \mathcal{X}^\sharp \xrightarrow{\natural} \mathcal{X} \xrightarrow{W} \mathcal{Y}$$

forms a degraded channel with prime power input size. By the data processing inequality, the Shannon capacity of the degraded channel $W \circ \natural$ is no greater than W's Shannon capacity. Meanwhile, it is clear that the degraded channel $W \circ \natural$ achieves W's capacity by the same input distribution, ignoring extra symbols. In other words, $I(W \circ \natural) = I(W)$. This constitutes the input size reduction.

Hereafter, we assume the size of the input alphabet $\mathcal{X}$ is q, where q is a prime power. If the sender wants to send uniform binary messages, let q be a power of 2. If the sender wants to send uniform ternary messages, let q be a power of 3. In case the sender wishes to send uniform quaternary messages but does not want to split an information bit over two channel symbols, let q be a power of 4. Bonus: should the sender want to send uniform senary messages, choose q_2 a power of 2 and q_3 a power of 3 such that q_2q_3 is a power of 6; then alternate (aka time-share) between $q = q_2$ and $q = q_3$. That is, the sender breaks every senary bit into a binary component and a ternary component, sends the binary component through the $q = q_2$ code blocks, and send the ternary component through the $q = q_3$ code blocks. For other message alphabets, apply the fundamental theorem of arithmetic.

Fix a q. Let $\mathbb{F}_q$ be the finite field of order q (with the addition and multiplication structure). Identifying $\mathcal{X}$ with $\mathbb{F}_q$, we will use them interchangeably. We say $W : \mathcal{X} \to \mathcal{Y}$ *is a q-ary channel* or, more concisely, *W is a q-ary channel*, depending on whether or not we need to refer to the variables $\mathcal{X}$ and $\mathcal{Y}$. It is worth keeping in mind that for inequalities in this work, $q = 2$ is the most difficult case and $q \geqslant 2$ will be used silently.

We clarified (cs), there are (cn), (cc), (cp), and (cr) to go.

B. On the Message Alphabet and the Block Length

The fact that we have some freedom to choose q blurs the meaning of the block length N since, say, a q^2-bit bears twice as much message as a q-bit does. Notwithstanding, we would like to remind readers that multiplication and division of N by any constant do not alter the semantics of the main theorem. This is because $O(N \log N)$ can absorb any constant; $\exp(-N^\pi)$ and $N^{-\rho}$ can, too, by fluctuating π and ρ a bit.

A more series aftereffect is caused by mixing code blocks with distinct q. When the sender attempts to send uniform 30-ary messages, they choose $q_2, q_3, q_5 \geqslant s$ and switch among the three block codes. The $q = q_2$ blocks have their own block length N_2 just like the other blocks have N_3 or N_5 as block lengths. The de facto block length N, the minimal number of the channel usages before the receiver can decode everything sent so far, is thus three times the least common multiple of N_2, N_3, and N_5. We claim without a proof (but it will be clear once we prove the rest of the main theorem) that it is possible to make $N_2 = N_3 = N_5$ and consequently $N = 3N_2$. Again, increasing N by three-fold does not make any semantic difference. For numbers with more prime factors, a similar reasoning applies.

We recommend readers not to worry about the message alphabet as there exists a powerful solution—to pre-compose another code that re-encodes an arbitrary finite message distribution (not necessarily uniform) to a uniform prime-ary input distribution. The existence of such codes, by duality, is tightly bonded to the existence of error correction codes that carry uniform prime-ary messages over channels of arbitrary arity. (Cf. the philosophy of [58], [59].) The latter is exactly what the main theorem concerns.

We clarified (cs) and (cn) in this section; there are (cc), (cp), and (cr) to go. We continue proving the main theorem in the next section.

III. Channel Transformation

Let $\ell \geqslant 2$ be an integer. This will be the dimension of the kernel matrices. But for now, let us introduce a flexible framework. Fix a q-ary channel $W: \mathcal{X} \to \mathcal{Y}$. Let $U_1, U_2, \ldots, U_\ell$ be r.v.s taking values in $\mathcal{X}$. For $1 \leqslant i \leqslant j \leqslant \ell$, let U_i^j be the joint r.v. $U_i U_{i+1} \cdots U_j$. Let $g^W : \mathcal{X}^\ell \to \mathcal{X}^\ell$ be a bijective map; that is, $H(U_1^\ell \mid g^W(U_1^\ell)) = 0$. We now feed ℓ i.i.d. (independent and identically distributed) copies of the channel W with $X_1^\ell \coloneqq g^W(U_1^\ell)$. Let $Y_1^\ell \in \mathcal{Y}^\ell$ be the corresponding output. The chain rule of conditional entropy reads

$$\begin{aligned} H(U_1^\ell \mid Y_1^\ell) = H(U_\ell \mid U_1^{\ell-1} Y_1^\ell) + H(U_{\ell-1} \mid U_1^{\ell-2} Y_1^\ell) \\ + \cdots + H(U_2 \mid U_1 Y_1^\ell) + H(U_1 \mid Y_1^\ell). \quad (1) \end{aligned}$$

Interpretation: to estimate U_1^ℓ given Y_1^ℓ, we first estimate U_1 given Y_1^ℓ; and then use the estimate $\hat{U}_1$ to further estimate U_2; afterward, we estimate U_3 given $\hat{U}_1$, $\hat{U}_2$, and Y_1^ℓ; and so on. To achieve W's capacity, $g^W(U_1^\ell)$ must follow a certain capacity-achieving distribution. Since g^W is bijective, this induces a distribution of U_1^ℓ. (Remark: we imply nothing about whether $U_1, U_2, \ldots, U_\ell$ are i.i.d or not.) Fix this distribution, then

$$\begin{aligned} I(U_1^\ell \,;\, Y_1^\ell) = I(U_\ell \,;\, Y_1^\ell \mid U_1^{\ell-1}) + I(U_{\ell-1} \,;\, Y_1^\ell \mid U_1^{\ell-2}) \\ + \cdots + I(U_2 \,;\, Y_1^\ell \mid U_1) + I(U_1 \,;\, Y_1^\ell). \end{aligned}$$

These two chain rules motivate the *channel transformation*: Let $[\ell]$ be the set of integers $\{1, 2, \ldots, \ell\}$. For each $i \in [\ell]$, let $W^{(i)}: \mathcal{X} \to \mathcal{X}^{i-1} \times \mathcal{Y}^\ell$ be a channel where $W^{(i)}(u_1^{i-1} y_1^\ell \mid u_i)$ is the probability that $U_1^{i-1} Y_1^\ell = u_1^{i-1} y_1^\ell$ conditioned on $U_i = u_i$. A more lengthy but exact form reads

$$W^{(i)}(u_1^{i-1} y_1^\ell \mid u_i) \coloneqq \frac{\sum_{u_{i+1}^\ell} W^\ell(g^W(u_1^\ell), y_1^\ell)}{\sum_{u_1^{i-1} u_{i+1}^\ell} W_{\text{in}}^\ell(g^W(u_1^\ell))}.$$

Its input distribution $W_{\text{in}}^{(i)}(u_i)$ is determined by that of U_i. It may sound weird that $W^{(i)}$ will tell the receiver the input of $W^{(1)}, W^{(2)}, \ldots, W^{(i-1)}$ for free. But in reality, $W^{(i)}$ acts as an interactive device where the receiver (not the sender) needs to input what U_1^{i-1} is and the device will output something that looks like $U_1^{i-1} Y_1^\ell$; only when the receiver inputs the correct U_1^{i-1} does the device return the correct $U_1^{i-1} Y_1^\ell$. Under this interpretation, the de facto capability of $W^{(i)}$ is thus $I(U_i \,;\, Y_1^\ell \mid U_1^{i-1})$ instead of $I(U_i \,;\, U_1^{i-1} Y_1^\ell)$, which justifies the chain rule of the mutual information. To avoid confusion, we prefer $H(W^{(i)})$ over $I(W^{(i)})$ in calculations.

What makes the idea of channel transformation powerful is that the transformations apply recursively. The precise formulation is as below: Fix any $i \in [\ell]$. Let $(X^{(i)})_1, (X^{(i)})_2, \ldots, (X^{(i)})_\ell \in \mathcal{X}$ be ℓ i.i.d. copies of the capacity-achieving input of $W^{(i)}$; let $(Y^{(i)})_1, (Y^{(i)})_2, \ldots, (Y^{(i)})_\ell \in \mathcal{X}^{i-1} \times \mathcal{Y}^\ell$ be the corresponding outputs. Let $g^{W^{(i)}}: \mathcal{X}^\ell \to \mathcal{X}^\ell$ be a bijection. Define a tuple of r.v.s $(U^{(i)})_1^\ell \coloneqq (g^{W^{(i)}})^{-1}((X^{(i)})_1^\ell)$; that is to say, $g^{W^{(i)}}((U^{(i)})_1^\ell) = (X^{(i)})_1^\ell$. For each $j \in [\ell]$, we define a depth-2 channel $(W^{(i)})^{(j)}: \mathcal{X} \to \mathcal{X}^{j-1} \times (\mathcal{X}^{i-1} \times \mathcal{Y}^\ell)^\ell$ where $(W^{(i)})^{(j)}((u^{(i)})_1^{i-1} (y^{(i)})_1^\ell \mid (u^{(i)})_j)$ is the probability that $(U^{(i)})_1^{j-1} (Y^{(i)})_1^\ell = (u^{(i)})_1^{j-1} (y^{(i)})_1^\ell$ conditioned on $(U^{(i)})_j = (u^{(i)})_j$. To sum up, we can define $(W^{(i)})^{(1)}, (W^{(i)})^{(2)}, \ldots, (W^{(i)})^{(\ell)}$ out of $W^{(i)}$ for any $i \in [\ell]$ in the same way we define $W^{(1)}, W^{(2)}, \ldots, W^{(\ell)}$ out of W. For $i, j \in [\ell]$, each $(W^{(i)})^{(j)}$ is again a channel, so the transformations apply to generate depth-3 channels. In the setup of the classical polar coding, a fixed bijection g is used to define $W^{(i)}$, $(W^{(i)})^{(j)}$, $((W^{(i)})^{(j)})^{(k)}$, et seq. To reach the optimal MDP paradigm, we allow g^W to depend on the channel W. That is to say, we need ℓ (presumably distinct) bijections $g^{W^{(i)}}: \mathcal{X}^\ell \to \mathcal{X}^\ell$ for every $i \in [\ell]$ when we want to define $(W^{(i)})^{(j)}$ out of $W^{(i)}$. Similarly, we need yet another ℓ^2 bijections $g^{(W^{(i)})^{(j)}}: \mathcal{X}^\ell \to \mathcal{X}^\ell$ for every $i, j \in [\ell]$ in defining depth-3 channels. And the recursion goes on ad infinitum.

Prudent readers are invited to check [30, the paragraph before Section III], [46], [50], [72], [75], [76] for a list of inhomogeneous configurations of kernels. See [30], [31], [77] for how nonlinear bijections are similar to (or different from) linear bijections.

A. Design of the Decoder

To implement channel transformations, we define a *DU* (decoding unit) to be an automata as follows: It is a box with ℓ pins on the left and ℓ pins on the right. Each pin is connected to another DU, a CH, an FH, or an IH (to be defined later). Each pin may take input or produce output but not at the same moment. A DU works as follows: Let $W: \mathcal{X} \to \mathcal{Y}$ be the channel it is to transform. (Step 0) For all $i \in [\ell]$, the i-th pin on the left takes the input y_i. The input is passed in the form of the a posteriori distribution $(W(x_i \mid y_i) : x_i \in \mathcal{X})$. This is what Arıkan calls α-representation [78, Section II.A]. (Step 1-a) It computes the a posteriori distribution of U_1 given y_1^ℓ; that is, $(W^{(1)}(u_1 \mid y_1^\ell) : u_1 \in \mathcal{X})$. And then it outputs this tuple of probabilities to the first pin on the right. (Step 1-b) At a later moment, it will receive an estimate $\hat{u}_1$ of U_1 from the first pin on the right. Note that $\hat{u}_1$ is a hard symbol in $\mathcal{X}$, not a soft tuple of probabilities. (Step 2-a) It computes the a posteriori distribution of U_2 given $\hat{u}_1 y_1^\ell$; that is to say, it pretends that U_1 happens to be $\hat{u}_1$ and computes $(W^{(2)}(u_2 \mid \hat{u}_1 y_1^\ell) : u_2 \in \mathcal{X})$ accordingly. And then it outputs this tuple of probabilities to the second pin on the right. (Step 2-b) At a later moment, it will receive an estimate $\hat{u}_2$ of U_2 from the second pin on the right. (Step i-a) In general, it computes $W^{(i)}(u_i \mid \hat{u}_1^{i-1} y_1^\ell)$ for all $u_i \in \mathcal{X}$ and then output the tuple to the i-th pin on the right. (Step i-b) After a while, it will receive $\hat{u}_i$. (Step $\ell+1$) Once it receives $\hat{u}_\ell$ from the last pin on the right, it computes $\hat{y}_1^\ell \coloneqq g^W(\hat{u}_1^\ell)$, and then output $\hat{y}_i$ to the i-th pin on the left for all $i \in [\ell]$. See Figs. 2 to 4 for illustrations.

The general rule to arrange the DUs is as follows: For a depth-n construction, put DUs in an ℓ^{n-1}-by-n array. Each DU is indexed by $(k_1, k_2, \ldots, k_{n-1}; m)$ where $k_1, k_2, \ldots, k_{n-1} \in [\ell]$ and $m \in [n]$. For all $m \in [n-1]$ and all

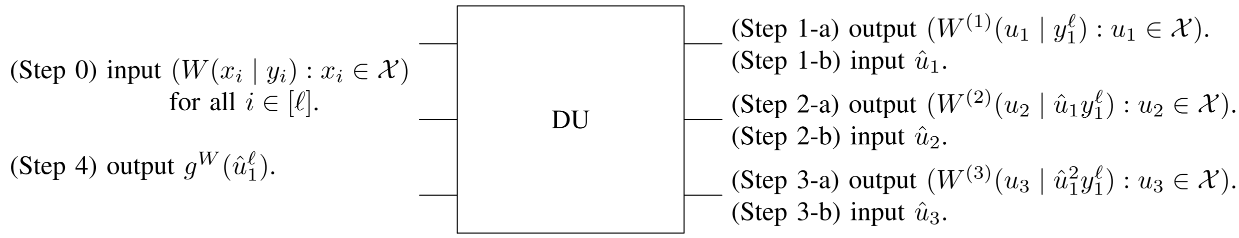

Fig. 2. A DU with $\ell = 3$ and its I/Os.

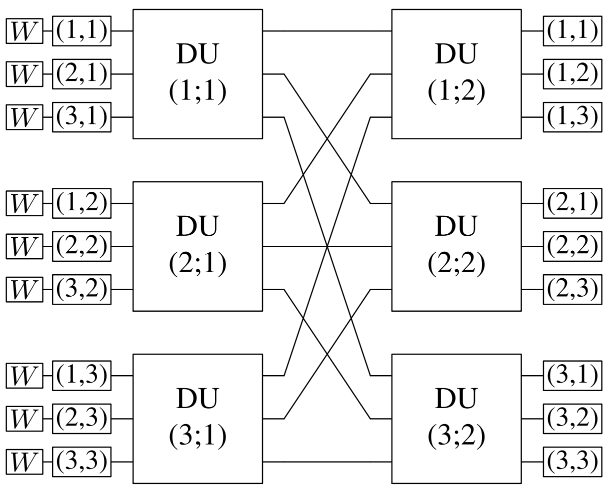

Fig. 3. 6 DUs (with $\ell = 3$ and $n = 2$) are chained together to implement $(W^{(1)})^{(1)}, \dots, (W^{(3)})^{(3)}$. Boxes marked "$W$" are channels. Boxes next to channels are CHs; the labels are their indices. Boxes at the rightmost column are either FHs or IHs; the labels are their indices. Note that DUs in the first column use the same g^W. DUs in the second column use $g^{W^{(1)}}$, $g^{W^{(2)}}$, and $g^{W^{(3)}}$, respectively.

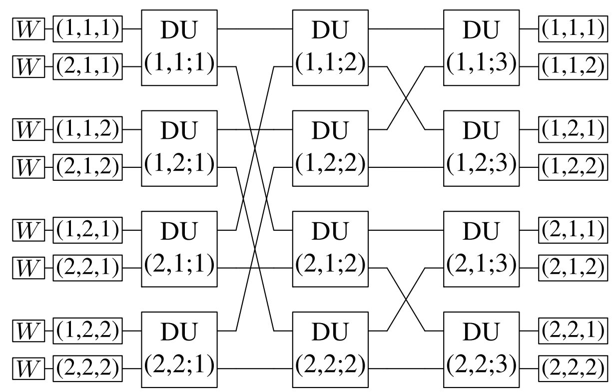

Fig. 4. 12 DUs (with $\ell = 2$ and $n = 3$) are chained together to implement $((W^{(1)})^{(1)})^{(1)}, \dots, ((W^{(2)})^{(2)})^{(2)}$. DUs in the first column use g^W; DUs in the second column use $g^{W^{(1)}}$ and $g^{W^{(2)}}$; DUs in the third column use $g^{(W^{(1)})^{(1)}}$, $g^{(W^{(1)})^{(2)}}$, $g^{(W^{(2)})^{(1)}}$, and $g^{(W^{(2)})^{(2)}}$.

$k_1, k_2, \dots, k_n \in [\ell]$, connect the k_m-th pin on the right of the $(k_1, \dots, k_{m-1}, k_{m+1}, \dots, k_n; m)$-th DU to the k_{m+1}-th pin on the left of the $(k_1, \dots, k_m, k_{m+2}, \dots, k_n; m+1)$-th DU. Here, the $(k_1, k_2, \dots, k_{n-1}; m)$-th DU is to transform the channel $(\cdots(W^{(k_1)})\cdots)^{(k_m)}$ into $(\cdots(W^{(k_1)})\cdots)^{(k_{m+1})}$. The k_1-th pin on the left of the $(k_2, \dots, k_n; 1)$-th DU connects to a *CH* (channel helper) indexed by $(k_1, k_2, \dots, k_n)$. Each CH then connects to the output of a copy of the channel W. The k_n-th pin on the right of the $(k_1, \dots, k_{n-1}; n)$-th DU connects to either an *FH* (frozen bit helper) or an *IH* (information bit helper); in either case, the connected helper is indexed by $(k_1, k_2, \dots, k_n)$. Let $\mathcal{I} \subset [\ell]^n$ be the set of indices $(k_1, k_2, \dots, k_n)$ such that the k_n-th pin on the right of the $(k_1, \dots, k_{n-1}; n)$-th DU connects to an IH. Then $[\ell]^n \setminus \mathcal{I}$ is the set of indices where the pin connects to an FH.

On the left-hand side of the DU array, the task of the $(k_1, k_2, \dots, k_n)$-th CH is to receive the output $Y_{(k_1,k_2,\dots,k_n)} \in \mathcal{Y}$ from the channel and then forward the a posteriori distribution tuple $(W(x_{(k_1,k_2,\dots,k_n)} \mid Y_{(k_1,k_2,\dots,k_n)}) : x_{(k_1,k_2,\dots,k_n)} \in \mathcal{X})$ to the DU array. On the right-hand side, FHs correspond to what Arıkan called *frozen bits*—bits that do not carry information and the receiver knows their values as part of the communication protocol. The task of the $(k_1, k_2, \dots, k_n)$-th FH is to receive the a posteriori distribution of the $(k_1, k_2, \dots, k_n)$-th frozen bit and then return the correct symbol $U_{(k_1,k_2,\dots,k_n)} \in \mathcal{X}$ back to the DU array. IHs correspond to information bits that carry the sender's messages. The task of the $(k_1, k_2, \dots, k_n)$-th IH is to receive the a posteriori distribution of the $(k_1, k_2, \dots, k_n)$-th information bit and then return the most probable symbol $\hat{U}_{(k_1,k_2,\dots,k_n)} \in \mathcal{X}$ back to the DU array. When all IHs are activated once, a code block completes. The most probable symbols they returned to the DU array form the decoded message $\hat{U}_{\mathcal{I}}$, meaning the tuple $(\hat{U}_{(k_1,k_2,\dots,k_n)} : (k_1, k_2, \dots, k_n) \in \mathcal{I})$.

What we just established is the *successive cancellation decoder* of polar codes that could be found in most works that implement polar codes. For instance, [25, Section VIII], [55, Section 3.2], [33, Section III], and [76, Section VI.B]. See especially [46, Section 9] for an almost identical construction albeit they had $q = 2$ in mind. We replicate the whole story to demonstrate that each DU may use a unique bijection "g" without changing the overall structure too much. Whether or not this construction can transmit information reliably is discussed in Section IV-B. There, we will also clarify how to arrange FHs and IHs. The complexity can be estimated prior to further specification.

B. Complexity of the Decoder

There are various models that measure the complexity of a structure. The polar coding community uses a variant of the time complexity where the arithmetic of real numbers costs $O(1)$ and passing probabilities between DUs costs $O(1)$. The complexity of the DU array is thus the number of the DUs multiplied by the complexity of a single DU. The number of

DUs is $\ell^{n-1}n$. The complexity of a DU depends on how a DU computes the a posteriori probabilities $W^{(j)}(u_i \mid u_1^{i-1}y_1^\ell)$ out of $W(x_i \mid y_i)$. The naïve approach is to exhaust every single possible input $u_1^\ell \in \mathcal{X}^\ell$ and calculate the a posteriori probabilities via Bayesian formulas. This costs $O(\ell^{10}q^{\ell+10})$ (here 10 is an overestimate). Hence the overall complexity is $O(\ell^{n-1}n\ell^{10}q^{\ell+10})$. In our setup, however, q is fixed, ℓ will be chosen upon knowing π, ρ, and n goes to infinity afterwards. So we advertise that the complexity is $O(\ell^n n)$, or $O(N \log N)$. Here $N \coloneqq \ell^n$ is the block length, equal to the number of copies of the channel W attached to the DU array. The complexities of the CHs, FHs, and IHs can be computed similarly. They are all bounded by $O(\ell^{n+10}q^{10})$. Thus the decoder as a whole costs $O(N \log N)$.

We claim that the encoder has the same complexity, namely $O(N \log N)$, although we have not defined the encoder yet. The encoder is essentially a special decoder and is the subject of the next subsection.

C. Design of the Encoder

The encoder will be an exact copy of the decoder except that CHs and IHs will behave differently. In greater detail: Let there be an ℓ^{n-1}-by-n array of DUs indexed and connected in the same way described in Section III-A. Each DU executes the exact same task described in Section III-A. The left pins of the DUs in the first column each connect to a CH. The right pins of the DUs in the last column each connect to the same type of device (an IH or an FH) as its twin-DU in the decoder does. Here, as part of the encoder, a CH will output the capacity-achieving input distribution ($W_{\text{in}}(x)$ for all $x \in \mathcal{X}$) into the DU array. For each $(k_1, k_2, \dots, k_n) \in \mathcal{I}$, the $(k_1, k_2, \dots, k_n)$-th IH will receive a recommended distribution of the $(k_1, k_2, \dots, k_n)$-th information bit and then return the message symbol $U_{(k_1,k_2,\dots,k_n)} \in \mathcal{X}$ the sender wants to send back to the DU array. For each $(k_1, k_2, \dots, k_n) \in [\ell]^n \setminus \mathcal{I}$, the $(k_1, k_2, \dots, k_n)$-th FH will receive a recommended distribution of the $(k_1, k_2, \dots, k_n)$-th frozen bit and then return a r.v. $U_{(k_1,k_2,\dots,k_n)} \in \mathcal{X}$ that follows that distribution back to the DU array. This r.v. is simulated by a pseudo random number generator shared between the encoder and the decoder. The twin-FH in the decoder, regardless what distribution it receives, will return the exact same symbol $U_{(k_1,k_2,\dots,k_n)}$ back to the DU array. This step is called *randomized rounding* and is found in [55, Section 3.3], [57, Section III], [79, Section II], and [33, Section III.A]. After all IHs return the sender's messages and all FHs returns randomly rounded bits to the DU array, the CHs will each get a codeword symbol $X_{(k_1,k_2,\dots,k_n)} \in \mathcal{X}$ from the DU array. And then each CH will forward that symbol to an i.i.d. copy of the channel W.

This design is a copy of [33]'s encoder explained in our terminology. It is clear that the encoding complexity will be $O(N \log N)$, too. Alongside the decoder, the encoder creates its own channel transformations. Let $W\colon \mathcal{X} \to \mathcal{Y}$ be a q-ary channel and X be a capacity-achieving input. Define a flattening channel $W_\flat\colon \mathcal{X} \to \{\eta\}$ that erases all information. Then the encoder is effectively synthesizing depth-1 channels $W_\flat^{(i)}\colon \mathcal{X} \to \mathcal{X}^{i-1} \times \{\eta\}^\ell$ for each $i \in [\ell]$, depth-2 channels $(W_\flat^{(i)})^{(j)}\colon \mathcal{X} \to \mathcal{X}^{j-1} \times (\mathcal{X}^{i-1} \times \{\eta\}^\ell)^\ell$ for each $j \in [\ell]$, depth-3 channels $((W_\flat^{(i)})^{(j)})^{(k)}\colon \mathcal{X} \to \mathcal{X}^{k-1} \times (\mathcal{X}^{j-1} \times (\mathcal{X}^{i-1} \times \{\eta\}^\ell)^\ell)^\ell$ for each $k \in [\ell]$, et seq. utilizing the same input distributions and series of bijections. For instance, $W_\flat^{(i)}(u_1^{i-1}y_1^\ell \mid u_i)$ is the probability that $U_1^{i-1} = u_1^{i-1}$ conditioned on $U_i = u_i$, or equally

$$\Big(\sum_{u_{i+1}^\ell} W_{\text{in}}^\ell(g^W(u_1^\ell))\Big) \div \Big(\sum_{u_1^{i-1}u_{i+1}^\ell} W_{\text{in}}^\ell(g^W(u_1^\ell))\Big).$$

Moreover, $H(W_\flat) = H(X)$ and $H(W_\flat^{(i)}) = H(U_i \mid U_1^{i-1})$. No "$Y$" plays any role here since they are constant. The fact that a channel as boring as $W_\flat$ is helpful to our main theorem will be covered later, in Section IV-B.

We clarified (cs), (cn), and (cc) up to this section; there are (cp) and (cr) to go.

IV. Channel Parameters

Let $W\colon \mathcal{X} \to \mathcal{Y}$ be a q-ary channel. Let X be a capacity-achieving input and Y be the corresponding output. Besides H and I, there are several channel parameters that capture the qualities of channels. Here is a list of parameters extracted from the work [29] of Mori and Tanaka.

Both $H(X \mid Y)$ and $H(W)$ are the base-q conditional entropy, the base chosen such that $0 \leqslant H(X \mid Y) \leqslant H(X) \leqslant 1$. Both $I(X \mid Y)$ and $I(W)$ are the base-q mutual information, and hence $0 \leqslant I(X \mid Y) \leqslant H(X) \leqslant 1$.

$P_{\text{e}}(X \mid Y)$ is the error probability of the optimal decoder, the maximum a posteriori (MAP) decoder. The MAP decoder looks at an output $y \in \mathcal{Y}$ and chooses a symbol $\hat{x} \in \mathcal{X}$ that maximizes $W(\hat{x} \mid y)$. When the output is $Y = y$, the probability that the MAP decoder does not choose X as $\hat{x}$ is $1 - \max_{x \in \mathcal{X}} W(x \mid y)$. Therefore, $P_{\text{e}}(X \mid Y) = \sum_{y \in \mathcal{Y}} W_{\text{out}}(y)(1 - \max_{x \in \mathcal{X}} W(x \mid y))$. Within a channel-centric narrative, we also write $P_{\text{e}}(W)$ for $P_{\text{e}}(X \mid Y)$.

$Z(X \mid Y)$ is the rescaled sum of Bhattacharyya coefficients of the transition distribution $W(y \mid x)$ for the uniform input. For non-uniform inputs, a modification is made to generalize the definition and the properties that used to hold. Intuitively speaking, a MAP decoder seeing y is "confident" if $W(x \mid y)$ is small for all but one x, or equivalently, if the product $W(x, y)W(x', y)$ is small for all distinct $x, x' \in \mathcal{X}$. The *Bhattacharyya parameter* measures the "confidence" by

$$Z(X \mid Y) \coloneqq \frac{1}{q-1} \sum_{\substack{x,x' \in \mathbb{F}_q \\ x \neq x'}} \sum_{y \in \mathcal{Y}} \sqrt{W(x, y)W(x', y)}.$$

In addition, define

$$Z_{\text{mad}}(X \mid Y) \coloneqq \max_{0 \neq d \in \mathbb{F}_q} \sum_{x \in \mathbb{F}_q} \sum_{y \in \mathcal{Y}} \sqrt{W(x, y)W(x + d, y)}.$$

We also write $Z(W)$ and $Z_{\text{mad}}(W)$ for these quantities. Remarks: The rescaling is such that $0 \leqslant Z \leqslant Z_{\text{mad}} \leqslant (q-1)Z \leqslant q-1$. Our definition of Z_{mad} is different from the $Z_{\max}$ in [29], but rather a mixture of $Z_{\max}$ and Z_d therein. That said, the definitions of other parameters—H, I, P_{e}, Z, T,

S, and $S_{\max}$—match [29]'s. Cf. [31, Section 3.C] has defined Z_{mad} for prime input size.

$T(X \mid Y)$ is the weighted average of the total variation distances from the a posteriori distributions $(W(x \mid y) : x \in \mathcal{X})$ to the uniform noise $(1/q, 1/q, \ldots, 1/q)$. More formally, it is defined to be $\sum_{y\in\mathcal{Y}} W_{\mathrm{out}}(y) \sum_{x\in\mathcal{X}} |W(x \mid y) - 1/q|$. We also write $T(W)$ for this quantity.

$S(X \mid Y)$ is the weighted average of the L^1-norms of the Fourier coefficients of the a posteriori distributions. The formal definition is as follows: Let $\mathrm{tr}\colon \mathbb{F}_q \to \mathbb{F}_p$ be the field trace, where $\mathbb{F}_q = \mathcal{X}$ and $\mathbb{F}_p$ is the prime subfield. Let $\chi\colon \mathbb{F}_q \to \mathbb{C}$ be an additive character defined as $\chi(x) := \exp(2\pi i \,\mathrm{tr}(x)/p)$, where $2\pi i$ is temporarily the period of $\exp$. Define the Fourier coefficient

$$M(w \mid y) := \sum_{z\in\mathbb{F}_q} W(z \mid y)\chi(wz).$$

Define the S-parameters

$$S(X \mid Y) := \frac{1}{q-1} \sum_{0\neq w\in\mathbb{F}_q} \sum_{y\in\mathcal{Y}} W_{\mathrm{out}}(y) \cdot \Big|M(w \mid y)\Big|,$$
$$S_{\max}(X \mid Y) := \max_{0\neq w\in\mathbb{F}_q} \sum_{y\in\mathcal{Y}} W_{\mathrm{out}}(y) \cdot \Big|M(w \mid y)\Big|.$$

We also write $S(W)$ and $S_{\max}(W)$ for these quantities. Remarks: The rescaling is such that $0 \leqslant S \leqslant S_{\max} \leqslant (q-1)S \leqslant q-1$. An interpretation is as follows: Fix a y. When $W(x \mid y)$ is roughly equal to $1/q$ for all $x \in \mathbb{F}_q$, the Fourier coefficient $M(w \mid y) = \sum_{z\in\mathbb{F}_q} W(z \mid y)\chi(wz)$ should be roughly $\sum_{z\in\mathbb{F}_q} \chi(wz)/q = 0$. The S-parameter measures how far those coefficients are from zero.

A. Relations Among Channel Parameters

The following is a series of lemmas we extract from existing works. They characterize the relations among H, I, P_{e}, Z, T, and S.

Lemma 3 [29, Lemma 22 with $k = 1$]: For any q-ary channel W,

$$\frac{q-1}{q^2}\Big(\sqrt{1+(q-1)Z(W)} - \sqrt{1-Z(W)}\Big)^2 \leqslant P_{\mathrm{e}}(W) \leqslant \frac{q-1}{2} Z(W).$$

Lemma 4 [29, Lemma 23 with $k = q-1$]: For any q-ary channel W,

$$\frac{q-1}{q} - P_{\mathrm{e}}(W) \leqslant \frac{T(W)}{2} \leqslant \frac{q-1}{q} - \frac{1}{q}\Big((q-1)qP_{\mathrm{e}}(W) - (q-1)(q-2)\Big).$$

Lemma 5 [29, Lemma 26 with $k = q-1$]: For any q-ary channel W,

$$1 - \frac{q}{q-1}P_{\mathrm{e}}(W) \leqslant S(W) \leqslant (q-1)q\Big(\frac{q-1}{q} - P_{\mathrm{e}}(W)\Big)\sqrt{1 - \frac{q}{q-1}\frac{q-2}{q-1}}.$$

Lemma 6 [80, Theorem 1]: For any q-ary channel W,

$$h_2(P_{\mathrm{e}}(W)) + P_{\mathrm{e}}(W)\log_2(q-1) \geqslant H(W)\log_2 q \geqslant \begin{cases} 2P_{\mathrm{e}}(W), \\ (q-1)q\log_2 \frac{q}{q-1}(P_{\mathrm{e}}(W) - \frac{q-2}{q-1}) + \log_2(q-1). \end{cases}$$

Here, h_2 is the binary entropy function; $h_2(1/2) = 1$. The upper bound is Fano's inequality. The first lower bound fits when $H(W)$ and $P_{\mathrm{e}}(W)$ are small; the second lower bound fits when $H(W)$ and $P_{\mathrm{e}}(W)$ are close to 1.

The above lemmas inspire the following characterization: Let A and B be two channel parameters, we say A, B are *bi-Hölder at* (a, b) if there exist $c, d > 0$ such that $|A(W) - a| < c|B(W) - b|^d$ and $|B(W) - b| < c|A(W) - a|^d$ for all q-ary channels W. The notion of bi-Hölder is an equivalence relation. In particular, if A, B are bi-Hölder at (a, b) and B, C are bi-Hölder at (b, c), then A, C are bi-Hölder at (a, c). In this case, it makes sense to say A, B, C are bi-Hölder at (a, b, c). This notion generalizes to tuples of more parameters. Now we can summarize Lemmas 3 to 6 in a more concise statement.

Lemma 7 (Implicit Bi-Hölder Tolls): Channel parameters $H, P_{\mathrm{e}}, Z, Z_{\mathrm{mad}}$ are bi-Hölder at $(0, 0, 0, 0)$. Channel parameters $H, P_{\mathrm{e}}, T, S, S_{\max}$ are bi-Hölder at $(1, 1 - 1/q, 0, 0, 0)$.

Proof: Z, Z_{mad} are bi-Hölder at $(0, 0)$ since $Z \leqslant Z_{\mathrm{mad}} \leqslant (q-1)Z$. Lemma 3 implies that P_{e}, Z are bi-Hölder at $(0, 0)$. Lemma 6 (with the first lower bound) implies that P_{e}, H are bi-Hölder at $(0, 0)$. Now apply the transitivity to conclude the first statement. For the second statement, $S, S_{\max}$ are bi-Hölder at $(0, 0)$ since $S \leqslant S_{\max} \leqslant (q-1)S$. Lemma 5 implies that P_{e}, S are bi-Hölder at $(1-1/q, 0)$. Lemma 4 implies that P_{e}, T are bi-Hölder at $(1-1/q, 0)$. Lemma 6 (with the second lower bound) implies that P_{e}, H are bi-Hölder at $(1-1/q, 1)$. Now apply the transitivity to conclude. ∎

See [29, Corollary 28] for what inspired us. They use notation $A \overset{\mathrm{e}}{\sim} B$ to mean A, B are bi-Hölder at $(0, 0)$ and at $(1, 1)$. See also [81] for the relation between Z and symmetric capacity over BDMCs. For some very technical details on the way toward the main theorem, we need explicit Hölder relations among H, Z_{mad}, and $S_{\max}$. We claim them here. The proof is nothing but looking closer into Lemmas 3, 5, and 6. A written-out proof is in Appendix A.

Lemma 8 (Explicit Hölder Tolls): log is natural. For all q-ary channels W, the following hold:

$$Z_{\mathrm{mad}}(W) \leqslant q\sqrt{H(W)\log_4 q}, \tag{2}$$
$$H(W) \leqslant \sqrt{e(q-1)Z_{\mathrm{mad}}(W)/2}, \tag{3}$$
$$S_{\max}(W) \leqslant (q-1)q\sqrt{(1-H(W))\log(q)/2}, \tag{4}$$
$$1 - H(W) \leqslant (q-1)S_{\max}(W)/\log q. \tag{5}$$

B. Control of the Block Error Probability

Let W be the channel we want to communicate over; and let X be any input. In the classical theory of polar coding, the second last step of the construction of the block code is to determine a subset $\mathcal{I} \subset [\ell]^n$ of indices that points to the depth-n channels that transmit information bits. When decoding this code, a block error happens if the successive cancellation decoder fails to decode any information bit. Let $E_{(k_1,k_2,\dots,k_n)}$ be the event that the first error occurs when the decoder is solving for the input to $(\cdots(W^{(k_1)})\cdots)^{(k_n)}$, i.e., when $\hat{U}_{(k_1,k_2,\dots,k_n)} \neq U_{(k_1,k_2,\dots,k_n)}$ and the equality holds for lexicographically earlier indices. Then the event's probability measure $P(E_{(k_1,k_2,\dots,k_n)})$ is no more than the bit error probability $P_{\mathrm{e}}\Big((\cdots(W^{(k_1)})\cdots)^{(k_n)}\Big)$. By the union bound, the block error probability of the decoder is bounded from above by a sum

$$P\{\hat{U}_{\mathcal{I}} \neq U_{\mathcal{I}}\} \leqslant \sum_{(k_1,k_2,\dots,k_n)\in\mathcal{I}} P_{\mathrm{e}}\Big((\cdots(W^{(k_1)})\cdots)^{(k_n)}\Big).$$

Here $U_{\mathcal{I}}$ is the tuple $(U_{(k_1,k_2,\dots,k_n)} : (k_1,k_2,\dots,k_n) \in \mathcal{I})$.

With this observation, we define $\mathcal{I}$ to be the set of indices $(k_1,k_2,\dots,k_n) \in [\ell]^n$ such that $H\Big((\cdots(W^{(k_1)})\cdots)^{(k_n)}\Big) < \theta_n$ for some clever choice of the threshold $\theta_n > 0$. This immediately implies $P_{\mathrm{e}}\Big((\cdots(W^{(k_1)})\cdots)^{(k_n)}\Big) < c\theta_n^d$ for some $c, d > 0$ by Lemma 7. Let θ_n be $\exp(-\ell^{\pi n}n)$. The sum of P_{e} is less than $\ell^n c\theta_n^d < \exp(-\ell^{\pi n})$ for sufficiently large n, which is the block error probability we claimed. Remark: Arıkan used a different criterion $Z < \theta_n$. It still implies $P_{\mathrm{e}} < c\theta_n^d$ and that the sum of P_{e} is less than $\ell^n c\theta_n^d < \exp(\ell^{\pi n})$ for large n. The benefit of controlling P_{e} using other parameters is that some parameters are easier to control (because Theorems 9 and 10 exist).

For the main theorem where the channel W is asymmetric, we want to control both the decoder block error and the encoder block error. Here, the encoder block error is not the encoder's failure to encode a message, but rather its failure to generate the capacity-achieving input distribution of W. To penalize, imagine that we employ an oracle that *claims* an encoder block error whenever the generated codeword should have been another word to fit the ideal distribution. That way, the actual block error probability will not exceed the sum of the encoder and decoder block error probabilities. More rigorously, let P be the probability measure assuming the ideal distribution of $U_{\mathcal{I}}$ and Q be the probability measure assuming the actual $U_{\mathcal{I}}$ generated by the encoder. Then the overall block error probability can be bounded by

$$Q\{\hat{U}_{\mathcal{I}} \neq U_{\mathcal{I}}\} \leqslant P\{\hat{U}_{\mathcal{I}} \neq U_{\mathcal{I}}\} + \|P - Q\|.$$

$P\{\hat{U}_{\mathcal{I}} \neq U_{\mathcal{I}}\}$ as the decoder block error probability is bounded before. The encoder block error probability is represented by $\|P - Q\|$, the total variation distance from P to Q. There is a telescoping argument similar to how we control the decoder error—classifying events by the first input bit where the oracle disagrees with the encoder [55, Lemma 3.5], [57, Lemma 4], [79, Lemma 2], [33, Lemma 1]. It yields that the encoder block error probability is bounded from above by the sum

$$\|P - Q\| \leqslant \sum_{(k_1,k_2,\dots,k_n)\in\mathcal{I}} T\Big((\cdots(W_\flat^{(k_1)})\cdots)^{(k_n)}\Big).$$

In controlling the encoder bit error probability, we strengthen the policy of collecting indices $(k_1,k_2,\dots,k_n) \in [\ell]^n$ for $\mathcal{I}$ by asking for $H\Big((\cdots(W_\flat^{(k_1)})\cdots)^{(k_n)}\Big) > 1 - \theta_n$. The latter criterion immediately implies $T\Big((\cdots(W_\flat^{(k_1)})\cdots)^{(k_n)}\Big) < c\theta_n^d$ by Lemma 7. As a consequence, the overall block error probability is controlled by $Q\{\hat{U}_{\mathcal{I}} \neq U_{\mathcal{I}}\} \leqslant P\{\hat{U}_{\mathcal{I}} \neq U_{\mathcal{I}}\} + \|P - Q\| < 2\ell^n c\theta_n^d < \exp(-\ell^{\pi n})$ for n large.

The preceding argument is a paraphrase of the proof of [33, Theorem 13]; Inequalities (59) and (57) therein are the key ideas. So far the block length, the complexity, and the error aspects of the main theorem are covered, it remains to control the code rate $|\mathcal{I}|/\ell^n$. In other words, we are to compute the cardinality of $\mathcal{I}$ given that $\mathcal{I} \subset [\ell]^n$ is the set of indices such that $H\Big((\cdots(W^{(k_1)})\cdots)^{(k_n)}\Big) < \theta_n$ and $1 - H\Big((\cdots(W_\flat^{(k_1)})\cdots)^{(k_n)}\Big) < \theta_n$, where $\theta_n := \exp(-\ell^{\pi n}n)$.

C. Before and After Channel Transformations

Alongside the relations among different parameters applied to the same channel, there are also relations between the same parameter applied to the original and the transformed channels. That $\sum_{i=1}^{\ell} H(W^{(i)}) = \ell H(W)$ is one. There are two more that are pivotal in the theory of polar coding but require more prerequisites. Assume that $g^W : \mathcal{X}^\ell \to \mathcal{X}^\ell$ is a linear isomorphism given by the multiplication of an invertible matrix G from the right—$g^W(u_1^\ell) := u_1^\ell G$. The following framework extends to nonlinear bijections but we do not need that much. (There is also the paradigm that random linear codes perform better than random codebooks for that a bad linear code tends to hoard a lot of light codewords at once, effectively removing them from the ensemble pool. So there is a good reason to stick to the linear case.) Let $0_1^{i-1}1_i u_{i+1}^\ell \in \mathbb{F}_q^\ell$ be a tuple of $i-1$ many 0 followed by a 1 and $\ell - i$ arbitrary symbols. A *coset code* is a subset of codewords of the form $\{0_1^{i-1}1_i u_{i+1}^\ell G : u_{i+1}^\ell \in \mathbb{F}_q^{\ell-i}\} \subset \mathbb{F}_q^\ell$. The coset codes have weight distributions just like every other code does. Let $\mathrm{wt}(x_1^\ell)$ be the Hamming weight of x_1^ℓ. The weight enumerator of the i-th coset code is defined to be a one-variable polynomial over the integers

$$f_{GZ}^{(i)}(z) := \sum_{u_{i+1}^\ell} z^{\mathrm{wt}(0_1^{i-1}1_i u_{i+1}^\ell G)} \in \mathbb{Z}[z].$$

We can now state the second relation. This is considered the main cause of why polar coding ever exists/works.

Theorem 9 (Fundamental Theorem of Polar Coding—Z-end, FTPCZ): [25, Proposition 5], [27, Lemma 10], [31, Lemma 3.5], [43, Section 4.1], [29, Lemma 33]

$$Z_{\mathrm{mad}}(W^{(i)}) \leqslant f_{GZ}^{(i)}(Z_{\mathrm{mad}}(W)).$$

The proof is postponed until Section VII-A. The fundamental theorems come as a pair. Let $u_1^{i-1}1_i 0_{i+1}^\ell \in \mathbb{F}_q^\ell$ be a tuple of $i-1$ arbitrary symbols followed by a 1 and $\ell - i$ many 0. Let $G^{-\top}$ be the inverse transpose of G. The weight enumerator of the i-th dual coset code is defined to be this one-variable polynomial over the integers

$$f_{GS}^{(i)}(s) := \sum_{u_1^{i-1}} s^{\mathrm{wt}(u_1^{i-1}1_i 0_{i+1}^\ell G^{-\top})} \in \mathbb{Z}[s].$$

We can now state the third relation, the dual of the second. The proof is postponed until Section VII-B.

Theorem 10 (Fundamental Theorem of Polar Coding—S-End, FTPCS): [55, Lemma 5.7], [57, Theorem 19], [79, Lemma 6], [29, Lemma 34], [46, Inequalities (74) and (75)]

$$S_{\max}(W^{(i)}) \leqslant f_{GS}^{(i)}(S_{\max}(W)).$$

Remark: These two bounds are not tight—the equality does not hold for BECs. In detail, Arıkan's original bound reads $Z_{\mathrm{mad}}(W^{(1)}) \leqslant 2Z_{\mathrm{mad}}(W) - Z_{\mathrm{mad}}(W)^2$ while our bound turns into $Z_{\mathrm{mad}}(W^{(1)}) \leqslant 2Z_{\mathrm{mad}}(W)$, the subtraction term missing. We are simply not able to prove a version that degenerates to an equality over erasure channels, nor does any prior work seem to. This causes a serious aftermath that $Z_{\mathrm{mad}}(\mathsf{W}_n)$ (to be defined later) is no longer a supermartingale. Nonetheless, this bound is strong enough to collaborate with the random coding theory. See, for example, how we compensate in Appendix C-A.

We clarified (cs), (cn), (cc), and (cp) up to this section; there is (cr) to go.

V. Channel Processes

Let $\mathsf{K}_1, \mathsf{K}_2, \mathsf{K}_3, \ldots$ be i.i.d. uniform r.v.s on $[\ell]$, where $[\ell]$ is the set of integers $\{1, 2, \ldots, \ell\}$. Let W be the q-ary channel we want to communicate over. Let W_0 be W. For each nonnegative integer n, let W_{n+1} be $(\mathsf{W}_n)^{(\mathsf{K}_{n+1})}$, which means $(\cdots(W^{(\mathsf{K}_1)})\cdots)^{(\mathsf{K}_{n+1})}$ in full. Recall that at the end of Section III-C we defined $W_\flat$, $W_\flat^{(i)}$, $(W_\flat^{(i)})^{(j)}$, $((W_\flat^{(i)})^{(j)})^{(k)}$, et seq. all with the same series of input distributions and bijections. Let V_0 be $W_\flat$; let V_{n+1} be $(\mathsf{V}_n)^{(\mathsf{K}_{n+1})}$, which means $(\cdots(W_\flat^{(\mathsf{K}_1)})\cdots)^{(\mathsf{K}_{n+1})}$ in full. These r.v.s provide a new family of randomness that does not appear in the encoding and decoding algorithms, but they help us understand the code rate $|\mathcal{I}|/\ell^n$ in this manner: Counting how many indices are in $\mathcal{I}$ is nothing more than measuring the probability $\mathsf{P}\{(\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_n) \in \mathcal{I}\}$. With the processes $\{\mathsf{W}_n\}$ and $\{\mathsf{V}_n\}$ defined earlier, it is further equivalent to measuring the probability $\mathsf{P}\{H(\mathsf{W}_n) < \theta_n \text{and } 1 - H(\mathsf{V}_n) < \theta_n\}$, where $\theta_n := \exp(-\ell^{\pi n} n)$. Moreover, it suffices to know how $\{H(\mathsf{W}_n)\}$ and $\{H(\mathsf{V}_n)\}$ behave as stochastic processes taking values in $[0, 1]$ without comprehending W_n and V_n themselves. The general fact is that $H(\mathsf{W}_n)$ is either very small (channel is reliable) or very close to 1 (channel is noisy). Arıkan called this phenomenon *channel polarization*. The following claim generalizes channel polarization and implies the main theorem.

Claim 11: Fix any $\pi, \rho > 0$ such that $\pi + 2\rho < 1$. We will choose an ℓ and a series of bijections of $\mathbb{F}_q^\ell$—namely, g^W, $g^{W^{(i)}}$, $g^{(W^{(i)})^{(j)}}$, $g^{((W^{(i)})^{(j)})^{(k)}}$, et seq.—such that

$$\mathsf{P}\{H(\mathsf{W}_n) < \exp(-\ell^{\pi n} n)\} > 1 - H(W) - \ell^{-\rho n + o(n)},$$
$$\mathsf{P}\{1 - H(\mathsf{W}_n) < \exp(-\ell^{\pi n} n)\} > H(W) - \ell^{-\rho n + o(n)},$$
$$\mathsf{P}\{H(\mathsf{V}_n) < \exp(-\ell^{\pi n} n)\} > 1 - H(W_\flat) - \ell^{-\rho n + o(n)},$$
$$\mathsf{P}\{1 - H(\mathsf{V}_n) < \exp(-\ell^{\pi n} n)\} > H(W_\flat) - \ell^{-\rho n + o(n)}.$$

Here, $o(n)$ is the little-o function in n; it is such that $o(n)/n \to 0$ as $n \to \infty$.

For polar codes over symmetric channels, the first inequality in Claim 11 alone implies that the code rate is $1 - H(W) - \ell^{-\rho n + o(n)} = I(W) - N^{-\rho + o(1)}$. The first two inequalities imply the polarization behavior that channels become either satisfactorily reliable (low $H(\mathsf{W}_n)$) or desperately noisy (high $H(\mathsf{W}_n)$). For asymmetric channels, however, we need to characterize $H(\mathsf{V}_n)$ alongside $H(\mathsf{W}_n)$. The last two inequalities in Claim 11 show that the same series of bijections polarize $W_\flat$ at the same time they polarize W. While $W_\flat$ contains no randomness form the channel W, what is polarized is that each input bit $U_{(k_1, k_2, \ldots, k_n)}$ either depends heavily on lexicographically earlier input bits (low $H(\mathsf{V}_n)$) or behaves like a free r.v. conditioned on earlier bits (high $H(\mathsf{V}_n)$). We then categorize the fate of indices in $[\ell]^n$ into the following three types. (A) Free and reliable: These are indices that will be in $\mathcal{I}$; they point to channels that transmit information bits. (B) Free but noisy: The sender *can* feed these channels with information only to find that the decoder will almost always make some mistakes. The sender should, instead, pad with some pseudo random numbers shared with the receiver. (C) Dependent and reliable. The inputs of these channels depends on previous inputs. Their main purpose is to *shape* the capacity-achieving input distribution. (D) Dependent but noisy is not possible because $H(\mathsf{V}_n) \geqslant H(\mathsf{W}_n)$. See Fig. 5 for an illustration of the asymptotic behavior. This is the key to [33, Theorem 1]. We reproduce their proof in the next subsection.

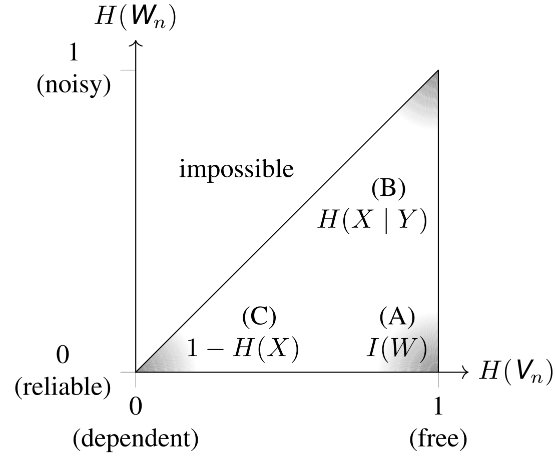

Fig. 5. The trichotomy of the fates of synthetic channels. Label (A) marks the corner of the free and reliable channels. Label $I(W)$ beneath (A) is the limit of the probability measure $\mathsf{P}(\mathsf{A}_n) = \mathsf{P}\{\mathsf{W}_n \text{is free and reliable}\}$ as $n \to \infty$. Labels (B) and (C) and the numbers beneath marks the corresponding fates and probability measures.

A. Claim 11 Implies the Main Theorem

As mentioned, $H(\mathsf{V}_n) \geqslant H(\mathsf{W}_n)$ so (D) dependent but noisy is not possible. Let A_n be the intersection event of free $\{1 - H(\mathsf{V}_n) < \exp(-\ell^{\pi n} n)\}$ and reliable $\{H(\mathsf{W}_n) < \exp(-\ell^{\pi n} n)\}$. Let B_n be the intersection event of free and noisy $\{1 - H(\mathsf{W}_n) < \exp(-\ell^{\pi n} n)\}$. Let C_n be the intersection event of dependent $\{H(\mathsf{V}_n) < \exp(-\ell^{\pi n} n)\}$ and reliable. Since noisy implies free, $\mathsf{P}(\mathsf{B}_n) > H(W) - \ell^{-\rho n + o(n)}$ follows the second inequality in Claim 11. Also since dependent implies reliable, $\mathsf{P}(\mathsf{C}_n) > 1 - H(W_\flat) - \ell^{-\rho n + o(n)}$ follows the third inequality in Claim 11. Note that A_n or B_n implies free but not "neither reliable nor noisy"; that is, $(\text{free} \wedge \text{reliable}) \vee (\text{free} \wedge \text{noisy}) \rightarrow \text{free} \wedge \neg(\neg\text{reliable} \vee \neg\text{noisy})$. We deduce that $P(A_n) + P(B_n) > H(W_\flat) - \ell^{-\rho n+o(n)} - 2\ell^{-\rho n+o(n)}$. Similarly, since A_n or C_n implies reliable but not "neither free nor dependent," we deduce that $P(A_n) + P(C_n) > 1 - H(W) - \ell^{-\rho n+o(n)} - 2\ell^{-\rho n+o(n)}$. In summary, we derive that

$$\begin{aligned} P(A_n) &\geqslant (P(A_n) + P(B_n)) + (P(A_n) + P(C_n)) - 1 \\ &> (H(W_\flat) - 3\ell^{-\rho n+o(n)}) + (1 - H(W) - 3\ell^{-\rho n+o(n)}) - 1 \\ &= H(X) - H(X \mid Y) - 6\ell^{-\rho n+o(n)} = I(W) - \ell^{-\rho n+o(n)} \end{aligned}$$

Finally, recall that $\mathcal{I}$ collects free and reliable indices, so the code rate is $|\mathcal{I}|/\ell^n = P(A_n) > I(W) - \ell^{-\rho n+o(n)}$. We almost finish the proof of the main theorem except that we claimed $I(W) - N^{-\rho} = I(W) - \ell^{-\rho n}$, without the little-$o$ term. It can be fixed by finding a slightly larger $\varrho > \rho$ such that $\pi + 2\varrho < 1$ still holds, and then rerunning the whole argument again with the new ϱ. The conclusion becomes that the code rate is at least $I(W) - \ell^{-\varrho n+o(n)}$. Since $-\varrho n + o(n) < -\rho n$ for sufficiently large n, this completes the proof of the main theorem. It remains to show that Claim 11 can be achieved.

VI. Global MDP Behavior Modulo Local Behaviors

In this section, we put constraints on an abstract process $\{H_n\}$ and show that they imply inequalities of the form

$$P\{H_n < \text{threshold}\} > \text{limit measure} - \text{decaying gap}$$

as those in Claim 11. Let $\mathcal{F}_n$ be the sigma-algebra generated by $K_1, K_2, \ldots, K_n$ for each n. Then $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ form a filtration of sigma-algebras. Let $\{H_n\}$, $\{Z_n\}$, and $\{S_n\}$ be three stochastic processes adapted to $\{\mathcal{F}_n\}$ (meaning that $K_1, K_2, \ldots, K_n$ determine H_n, Z_n, S_n). The following premises are elementary to verify when we reveal what those processes are: (pb) $0 \leqslant H_n, Z_n, S_n$ and $H_n \leqslant 1$; (pm) $\{H_n\}$ is a martingale, i.e., $E[H_{n+1} \mid \mathcal{F}_n] = H_n$; (pt) $H_n \leqslant q^3\sqrt{Z_n}$ and $Z_n \leqslant q^3\sqrt{H_n}$ along with $1 - H_n \leqslant q^3\sqrt{S_n}$ as well as $S_n \leqslant q^3\sqrt{1 - H_n}$ for all n. Furthermore, assume large kernels: (pl) $\ell \geqslant \max(3^q, e^4, q^5)$. Let $\alpha := \log(\log \ell)/\log \ell$ be a small number shrinking as ℓ increases. Define the potential function $h_\alpha\colon [0,1] \rightarrow [0,1]$ to be $h_\alpha(z) := \min(z, 1-z)^\alpha$. (Remark: h_2 is not a special case of h_α for $\alpha = 2$; we expect $\alpha \ll 1$ in practice.)

Here are the sufficient criteria for the main theorem. Two of them are difficult to verify; how to satisfy them is the main challenge this paper tackles.

Lemma 12 (Calculus Machinery for Global MDP): Assume premises (pb), (pm), (pt), and (pl). Assume the *local LDP behavior*:

$$\begin{aligned} Z_{n+1} &\leqslant \ell \exp(qZ_n\ell)(qZ_n)^{\lceil K_{n+1}^2/3\ell \rceil}, \\ S_{n+1} &\leqslant \ell \exp(qS_n\ell)(qS_n)^{\lceil (\ell+1-K_{n+1})^2/3\ell \rceil}. \end{aligned}$$

Assume the *local CLT behavior*:

$$E[h_\alpha(H_{n+1}) \mid \mathcal{F}_n] < 4\ell^{-1/2+\alpha}.$$

Then, for any constants $\pi, \rho > 0$ such that

$$\pi + 2\rho \leqslant 1 - 8\alpha, \tag{6}$$

the following holds:

$$P\{H_n < \exp(-\ell^{\pi n} n)\} > 1 - H_0 - \ell^{-\rho n+o(n)}. \tag{7}$$

We defer the proof until Appendix B. The term $K_{n+1}^2/3\ell$ in the lemma is to control the local LDP behavior of the process $\{H_n\}$—the behavior of H_{n+1} when H_n is close to 0 and the behavior that is closely related to the LDP behavior of polar codes. The term is chosen in a way such that $h_2((k^2/3\ell)/\ell) < k/\ell$ and such that $\sum_k (k^2/3\ell)^t$ is easy to handle. In [43, Theorem 7], a similar criterion is stated and is annotated as *faster polarization at the tails*. In [53, Definition 2.4], a similar criterion is stated and is annotated as *strong suction at the low end*. The *eigenfunction* h_α in the lemma is to control the local CLT behavior of the process $\{H_n\}$—the behavior of H_n when it is away from 0 and the behavior that is closely related to the CLT behavior of polar codes. In [43, Theorem 7], a similar criterion is annotated as *near optimal polarization in the middle* with $h_{\text{FHMV}}(z) := (z(1-z))^\alpha$ for positive but small α at most $\log(\log \ell)/\log \ell$. In [53, Definition 2.3], a similar criterion is annotated as *variance in the middle* with $h_{\text{BGS}}(z) := \sqrt{\min(z, 1-z)}$. Note how our choice of $h_\alpha(z) := \min(z, 1-z)^\alpha$ resembles theirs. In both cases, the criteria are *local* because they refer to a small slice of the process, focusing on how H_{n+1} (or Z_{n+1}) behaves in terms of H_n (or Z_n). This perspective frees [43], [53] from considering the (*global*) process $\{H_n\}$ as a whole and simplifies the analysis. We specifically benefit from the fact that we can choose the bijection $g^{(\cdots(W^{(k_1)})\cdots)^{(k_n)}}$ solely according to the channels $(\cdots(W^{(k_1)})\cdots)^{(k_n)}$ and $(\cdots(W_\flat^{(k_1)})\cdots)^{(k_n)}$ instead of the complete channel family-tree. This is also the approach taken in [46].

A. Lemma 12 Helps Achieve Claim 11

The formulation and the choice of the variables make it clear how Lemma 12 will be applied to support Claim 11. For instance, if we let $\{H_n\}$, $\{Z_n\}$ and $\{S_n\}$ be $\{H(W_n)\}$, $\{Z_{\text{mad}}(W_n)\}$, and $\{S_{\max}(W_n)\}$, respectively, then Lemma 12 supports the first of the four inequalities in Claim 11. Moreover, if we let $\{H_n\}$, $\{Z_n\}$, and $\{S_n\}$ be $\{1 - H(W_n)\}$, $\{S_{\max}(W_n)\}$, and $\{Z_{\text{mad}}(W_n)\}$, then Lemma 12 supports the second inequality of Claim 11. If $\{H_n\}$, $\{Z_n\}$, and $\{S_n\}$ are let to be $\{H(V_n)\}$, $\{Z_{\text{mad}}(V_n)\}$, and $\{S_{\max}(V_n)\}$, that supports the third inequality. If $\{H_n\}$, $\{Z_n\}$, and $\{S_n\}$ are let to be $\{1 - H(V_n)\}$, $\{S_{\max}(V_n)\}$, and $\{Z_{\text{mad}}(V_n)\}$, that supports the forth inequality. The premises (pb), (pm), (pt), and (pl) listed above Lemma 12 are easy to verify; for instance, Lemma 8 implies (pt) for all four cases. It remains to show that for each of the four triples of processes, the local LDP behavior and the local CLT behavior hold.

To do so, one advantage is that the two desired behaviors are local. They only involve how H_{n+1}, Z_{n+1} and S_{n+1} behave conditioned on the history $\mathcal{F}_n$. A potentially tedious aspect is that for each candidate of the bijection $g^{(\cdots(W^{(k_1)})\cdots)^{(k_n)}}$, we have to verify the two behaviors four times, once for each of the four triples of channel parameters. Luckily, within random coding theory, we are in the situation that to choose an object that satisfies multiple criteria, it suffices to choose the object from an ensemble and compute the probabilities that each criterion fails; as long as the sum of failing probabilities is small, most objects satisfy. Even more luckily, when we choose a bijection $g^{(\cdots(W^{(k_1)})\cdots)^{(k_n)}}$ from some ensemble, we only have to compute the probability that the local CLT or LDP behavior fails for $\{\mathsf{H}_n\}$, $\{\mathsf{Z}_n\}$, and $\{\mathsf{S}_n\}$ being $\{H(\mathsf{W}_n)\}$, $\{Z_{\rm mad}(\mathsf{W}_n)\}$, and $\{S_{\max}(\mathsf{W}_n)\}$ but not the other three triples. This is because other three triples are the special case and/or the *dual* of this triple. Elaboration: Since V_n are q-ary channels just like W_n are, inequalities hold true for arbitrary W_n should hold true for any V_n. Also, since $Z_{\rm mad}$ and $S_{\max}$ are in duality, inequalities hold true for $H, Z_{\rm mad}, S_{\max}$ hold true for $1 - H, S_{\max}, Z_{\rm mad}$. The duality is due to the duality between FTPCZ and FTPCS, within the explicit Hölder tolls, and within the ensemble of bijections we are to choose $g^{(\cdots(W^{(k_1)})\cdots)^{(k_n)}}$ from.

B. Random Linear Isomorphisms as Bijections

Fix q and ℓ. Let $\mathrm{GL}(\ell, q)$ be the group of ℓ-by-ℓ invertible matrices over $\mathbb{F}_q$ together with the ordinary matrix multiplication. Select an element $\mathbb{G} \in \mathrm{GL}(\ell, q)$ uniformly at random. Let $g^W\colon \mathbb{F}_q^\ell \to \mathbb{F}_q^\ell$ be the multiplication of $\mathbb{G}$ from the right, namely $g^W(u_1^\ell) := u_1^\ell \mathbb{G}$. This map is bijective since $\mathbb{G}$ is invertible. Let W be a q-ary channel. Recall that we defined q-ary channels $W^{(1)}, W^{(2)}, \ldots, W^{(\ell)}$ in Section III. To emphasis that these imaginary channels depend heavily on the randomness source $\mathbb{G}$, we call them $W_{\mathbb{G}}^{(1)}, W_{\mathbb{G}}^{(2)}, \ldots, W_{\mathbb{G}}^{(\ell)}$ instead. The following two lemmas help verify the two criteria in Lemma 12. Proofs are given in upcoming sections, VII and VIII.

Lemma 13 (Local LDP Behavior): Fix an $\ell \geqslant 30$. Let $\mathbb{G}$ vary; with probability less than $3q^{-\sqrt{\ell}/13}$, each of the following fails for each $i \in [\ell]$:

$$Z_{\rm mad}(W_{\mathbb{G}}^{(i)}) \leqslant \ell \exp(qZ_{\rm mad}(W)\ell)(qZ_{\rm mad}(W))^{\lceil i^2/3\ell \rceil}, \quad (8)$$

$$S_{\max}(W_{\mathbb{G}}^{(i)}) \leqslant \ell \exp(qS_{\max}(W)\ell)(qS_{\max}(W))^{\lceil (\ell+1-i)^2/3\ell \rceil}. \quad (9)$$

Lemma 14 (Local CLT Behavior): Fix an $\ell \geqslant 20$. Recall $\alpha := \log(\log \ell)/\log \ell$ and $h_\alpha(z) := \min(z, 1 - z)^\alpha$. Let $\mathbb{G}$ vary; with probability less than $2\ell^{-\log(\ell)/20}$, this fails:

$$\frac{1}{\ell}\sum_{i=1}^{\ell} h_\alpha(H(W_{\mathbb{G}}^{(i)})) < 4\ell^{-1/2+\alpha}. \quad (10)$$

C. Local Behaviors Imply Claim 11 (and Hence the Main theorem)

We now are able to see how Lemmas 12 to 14 imply that Claim 11 is achievable for the right choice of ℓ and bijections g^W, $g^{W^{(i)}}$, et seq.: For any given q-ary channel W, let ℓ be $\max(3^q, e^4, q^5)$. For any given $\pi, \rho > 0$ such that $\pi + 2\rho < 1$, enlarge ℓ such that Inequality (6) holds, given $\alpha := \log(\log \ell)/\log \ell$. Consider a random kernel $\mathbb{G}$ as a candidate of the bijection g^W. Increase ℓ further so that the failing probabilities—Lemma 13's $3q^{-\sqrt{\ell}/13}$ and Lemma 14's $2\ell^{-\log(\ell)/20}$—amount to $1/3$ or less. Recall the flattening channel $W_\flat$. The probability that any of the inequalities in Lemmas 13 and 14 fails for $W_\flat$ is less than $1/3$, too. Invoke the union bound; $1/3 + 1/3 < 1$. Hence there exists a solid choice of g^W as the multiplication of some proper instance of $\mathbb{G}$ from the right.

With this g^W determined, we define $W^{(i)}$ and $W_\flat^{(i)}$ for all $i \in [\ell]$. Consider first $i = 1$, anything that has been done to W now applies to $W^{(i)}$. That is, let $g^{W^{(i)}}$ be the multiplication of a random kernel $\mathbb{G}$ from the right. With W, i, and $W_{\mathbb{G}}^{(i)}$ replaced by $W^{(i)}$, j, and $(W^{(i)})_{\mathbb{G}}^{(j)}$, the probabilities that inequalities in Lemmas 13 and 14 fail add up to $1/3$ or less. So is the flattening ($\flat$) counterpart. Hence there is a solid choice of $g^{W^{(i)}}$. Repeat this for every other $i = 2, 3, \ldots, \ell$. Once finished, proceed to choosing $g^{(W^{(i)})^{(j)}}$ for all $i, j \in [\ell]$. And so on and so forth for cases beyond depth-2. Notice that we always make a solid choice of a bijection before we proceed to the next level of channels, hence the failing probabilities of Lemmas 13 and 14 do not accumulate as the depth increases.

By how we select bijections in the previous paragraph, the criteria in Lemma 12 hold for $(\{\mathsf{H}_n\}, \{\mathsf{Z}_n\}, \{\mathsf{S}_n\})$ being the four triples listed below: $(\{H(\mathsf{W}_n)\}, \{Z_{\rm mad}(\mathsf{W}_n)\}, \{S(\mathsf{W}_n)\})$ and $(\{1 - H(\mathsf{W}_n)\}, \{S_{\max}(\mathsf{W}_n)\}, \{Z_{\rm mad}(\mathsf{W}_n)\})$ as well as $(\{H(\mathsf{V}_n)\}, \{Z_{\rm mad}(\mathsf{V}_n)\}, \{S_{\max}(\mathsf{V}_n)\})$ in addition to $(\{1 - H(\mathsf{V}_n)\}, \{S_{\max}(\mathsf{V}_n)\}, \{Z_{\rm mad}(\mathsf{V}_n)\})$. Hence the process $\{\mathsf{H}_n\}$ satisfies Inequality (7) for these four processes: $\{H(\mathsf{W}_n)\}$ and $\{1 - H(\mathsf{W}_n)\}$ along with $\{H(\mathsf{V}_n)\}$ as well as $\{1 - H(\mathsf{V}_n)\}$. This results in the four inequalities in Claim 11. And we are done. It remains to prove Lemmas 12 to 14 in order to prove the main theorem.

VII. Local LDP Behavior (Proof of Lemma 13)

In this section, we will first prove the two fundamental theorems of polar coding in Section VII-A (for the Z-end) and in Section VII-B (for the S-end). And then we will target that the following inequalities hold with high probability:

$$Z_{\rm mad}(W_{\mathbb{G}}^{(i)}) \leqslant \ell \exp(qZ_{\rm mad}(W)\ell)(qZ_{\rm mad}(W))^{\lceil i^2/3\ell \rceil}, \quad \text{((8)'s copy)}$$

$$S_{\max}(W_{\mathbb{G}}^{(\ell+1-i)}) \leqslant \ell \exp(qS_{\max}(W)\ell)(qS_{\max}(W))^{\lceil i^2/3\ell \rceil}. \quad \text{((9)'s copy)}$$

By the duality between the two fundamental theorems and between the two targeted inequalities, it is not hard to see that it suffices to prove the $Z_{\rm mad}$-case and the $S_{\max}$-case follows immediately. We will prove that the first targeted inequality, for each $i \in [\ell]$, holds with probability $1 - 3q^{-\sqrt{\ell}/13}$ in Section VII-D, closing this section.

A. Proof of FTPCZ (Theorem 9)

As is promised in Section IV-C, we prove the two fundamental theorems of polar coding. We first go for the Z-end.

Recall that $f_{GZ}^{(i)}(z) := \sum_{u_{i+1}^\ell} z^{\mathrm{wt}(0_1^{i-1}1_i u_{i+1}^\ell G)}$ is the weight enumerator of the i-th coset code. Theorem 9 claims that

$Z_{\mathrm{mad}}(W^{(i)}) \leqslant f_{GZ}^{(i)}(Z_{\mathrm{mad}}(W))$. By the definition of $W^{(i)}$ and the definition of the Bhattacharyya parameter, $Z_{\mathrm{mad}}(W^{(i)})$ is

$$\max_{0\neq d_i\in\mathbb{F}_q} \sum_{u_i\in\mathbb{F}_q} \sum_{u_1^{i-1}y_1^\ell\in\mathbb{F}_q^i\times\mathcal{Y}^\ell} \sqrt{W^{(i)}(u_i,u_1^{i-1}y_1^\ell)W^{(i)}(u_i+d_i,u_1^{i-1}y_1^\ell)}.$$

By the nature of $\max_{0\neq d_i\in\mathbb{F}_q}$, it suffices to show that the double sum is at most $f_{GZ}^{(i)}(Z_{\mathrm{mad}}(W))$ for arbitrary nonzero d_i.

In the upcoming argument, tuple concatenation takes precedence over vector-matrix multiplication and vector addition. Fix a d_i, we argue that

$$\begin{aligned}
&\sum_{u_i\in\mathbb{F}_q}\sum_{u_1^{i-1}y_1^\ell\in\mathbb{F}_q^i\times\mathcal{Y}^\ell}\sqrt{W^{(i)}(u_i,u_1^{i-1}y_1^\ell)W^{(i)}(u_i+d_i,u_1^{i-1}y_1^\ell)}\\
&=\sum_{u_1^iy_1^\ell}\sqrt{W^{(i)}(u_i,u_1^{i-1}y_1^\ell)W^{(i)}(u_i+d_i,u_1^{i-1}y_1^\ell)}\\
&=\sum_{u_1^iy_1^\ell}\sqrt{\begin{matrix}\sum_{u_{i+1}^\ell\in\mathbb{F}_q^{\ell-i}}W^\ell(u_1^iu_{i+1}^\ell G,y_1^\ell)\\ \times\sum_{v_{i+1}^\ell\in\mathbb{F}_q^{\ell-i}}W^\ell(u_1^{i-1}(u_i+d_i)v_{i+1}^\ell G,y_1^\ell)\end{matrix}}\\
&\leqslant\sum_{u_1^iy_1^\ell}\sum_{u_{i+1}^\ell}\sum_{v_{i+1}^\ell}\sqrt{\begin{matrix}W^\ell(u_1^iu_{i+1}^\ell G,y_1^\ell)\\ \times W^\ell(u_1^{i-1}(u_i+d_i)v_{i+1}^\ell G,y_1^\ell)\end{matrix}}\\
&=\sum_{y_1^\ell}\sum_{u_1^\ell}\sum_{d_{i+1}^\ell\in\mathbb{F}_q^{\ell-i}}\sqrt{W^\ell(u_1^\ell G,y_1^\ell)W^\ell(u_1^{i-1}(u_i^\ell+d_i^\ell)G,y_1^\ell)}\\
&=\sum_{y_1^\ell}\sum_{x_1^\ell\in\mathbb{F}_q^\ell}\sum_{d_{i+1}^\ell}\sqrt{W^\ell(x_1^\ell,y_1^\ell)W^\ell(x_1^\ell+0_1^{i-1}d_i^\ell G,y_1^\ell)}\\
&=\sum_{d_{i+1}^\ell}\sum_{y_1^\ell}\sum_{x_1^\ell}\sqrt{W^\ell(x_1^\ell,y_1^\ell)W^\ell(x_1^\ell+e_1^\ell,y_1^\ell)}\\
&=\sum_{d_{i+1}^\ell}\sum_{y_1^\ell}\sum_{x_1^\ell}\prod_{j\in[\ell]}\sqrt{W(x_j,y_j)W(x_j+e_j,y_j)}\\
&=\sum_{d_{i+1}^\ell}\sum_{y_1^\ell}\sum_{x_1^\ell}\prod_{j\in J}\sqrt{\begin{matrix}W(x_j,y_j)\\ \times W(x_j+e_j,y_j)\end{matrix}}\prod_{k\notin J}W(x_k,y_k)\\
&=\sum_{d_{i+1}^\ell}\prod_{j\in J}\Bigl(\sum_{x_jy_j}\sqrt{\begin{matrix}W(x_j,y_j)\\ \times W(x_j+e_j,y_j)\end{matrix}}\Bigr)\prod_{k\notin J}\Bigl(\sum_{x_ky_k}W(x_k,y_k)\Bigr)\\
&=\sum_{d_{i+1}^\ell}\prod_{j\in J}\Bigl(\sum_{x_jy_j}\sqrt{W(x_j,y_j)W(x_j+e_j,y_j)}\Bigr)\\
&\leqslant\sum_{d_{i+1}^\ell}\prod_{j\in J}\max_{0\neq e_j\in\mathbb{F}_q}\Bigl(\sum_{x_jy_j}\sqrt{W(x_j,y_j)W(x_j+e_j,y_j)}\Bigr)\\
&=\sum_{d_{i+1}^\ell}\prod_{j\in J}Z_{\mathrm{mad}}(W)=\sum_{d_{i+1}^\ell}Z_{\mathrm{mad}}(W)^{|J|}\\
&=\sum_{d_{i+1}^\ell}Z_{\mathrm{mad}}(W)^{\mathrm{wt}(0_1^{i-1}d_id_{i+1}^\ell G)}\\
&=\sum_{d_{i+1}^\ell}Z_{\mathrm{mad}}(W)^{\mathrm{wt}(0_1^{i-1}1_id_{i+1}^\ell G)}=f_{GZ}^{(i)}(Z_{\mathrm{mad}}(W)).
\end{aligned}$$

The first equality abbreviates the summation. The next equality expands $W^{(i)}$ by the very definition, where u_{i+1}^ℓ and v_{i+1}^ℓ are free variables in $\mathbb{F}_q$. The next inequality is by the sub-additivity of the square root. In the next equality we define $d_{i+1}^\ell := v_{i+1}^\ell - u_{i+1}^\ell$; so summing over v_{i+1}^ℓ is equivalent to summing over d_{i+1}^ℓ. In the next equality we define $x_1^\ell := u_1^\ell G$; so summing over u_1^ℓ is equivalent to summing over x_1^ℓ as G is invertible. In the next equality we substitute $e_1^\ell := 0_1^{i-1}d_i^\ell G$ and reorder the summation. The next equality expands the product of the memoryless channels. The next equality classifies the indices into two classes—$j \in J$ are those such that $e_j \neq 0$ and $k \notin J$ are such that $e_k = 0$. The next equality is the distributive law $ax + ay + bx + by = (a + b)(x + y)$. The next equality uses the fact that $W(x, y)$ sum to 1. In the next inequality we replace e_j by a nonzero element that maximizes the sum in the parentheses. In the next equality we realize that the maximum is the Bhattacharyya parameter (surprisingly). The second last equality uses the fact that multiplying a vector by a scalar preserves its Hamming weight. And quod erat demonstrandum.

Experienced readers may find that all but the last inequality follows the proof strategy of [27, Lemma 10]. An instant improvement made here is that we stop at the weight enumerator $f_{GZ}^{(i)}(Z_{\mathrm{mad}})$ instead of loosening to $q^{\ell-i}Z_{\mathrm{mad}}^{\text{minimum distance}}$. This is a decent gain as the majority of codewords has Hamming weights much higher than the minimum distance. Another remark is that [29, Lemma 33] deals with $Z_{\max}$ (not our Z_{mad}). While the lemma itself is fine, the parameter $Z_{\max}$ is not bi-Hölder to H. This suppresses the application of Lemma 12, which requires (pt). To rephrase it, Z_{mad} is the *correct generalization* of Z for $q > 2$.

B. Proof of FTPCS (Theorem 10)

We now go for the S-end of the fundamental theorem of polar coding. Recall the character $\chi(x) := \exp(2\pi i\,\mathrm{tr}(x)/p)$. We need these properties: (xa) $\chi(0) = 1$; (xb) $|\chi(x)| = 1$ for all $x \in \mathbb{F}_q$; (xc) $\chi(x)\chi(z) = \chi(x + z)$ for all $x, z \in \mathbb{F}_q$; (xd) $\sum_{x\in\mathbb{F}_q}\chi(x) = 0$. See also [29, Definition 24] or a dedicated book [82]. To prove the theorem, we first verify that Fourier coefficients recover the origin: Let $M(w, y) := W_{\mathrm{out}}(y)M(w \mid y) = \sum_{z\in\mathbb{F}_q} W(z, y)\chi(wz)$, then

$$\begin{aligned}
\sum_{w\in\mathbb{F}_q}M(w,y)\chi(-xw)&=\sum_{w\in\mathbb{F}_q}\sum_{z\in\mathbb{F}_q}W(z,y)\chi(wz)\chi(-xw)\\
&=\sum_{z\in\mathbb{F}_q}W(z,y)\sum_{w\in\mathbb{F}_q}\chi(w(z-x))\\
&=\sum_{z\in\mathbb{F}_q}W(z,y)q\mathbb{I}\{z-x=0\}=qW(x,y).
\end{aligned}$$

The first equality expands $M(w, y)$ by the definition. The next equality uses that χ is an additive character (xc), and reorders the summation. The next equality uses $\sum_{w\in\mathbb{F}_q}\chi(w) = 0$ (xd) and $\sum_{w\in\mathbb{F}_q}\chi(0) = q$ (xa); and $\mathbb{I}$ is the indicator function.

Knowing $W(x_j, y_j) = q^{-1}\sum_{w_j\in\mathbb{F}_q} M(w_j, y_j)\chi(-x_jw_j)$, we proceed to

$$\begin{aligned}
W^{(i)}(u_i,u_1^{i-1}y_1^\ell)&=\sum_{u_{i+1}^\ell}W^\ell(u_1^\ell G,y_1^\ell)\\
&=\sum_{u_{i+1}^\ell\in\mathbb{F}_q^{\ell-i}}W^\ell(x_1^\ell,y_1^\ell)=\sum_{u_{i+1}^\ell}\prod_{j\in[\ell]}W(x_j,y_j)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{u_{i+1}^\ell} \prod_{j\in[\ell]} \Big(\frac{1}{q} \sum_{w_j\in\mathbb{F}_q} M(w_j,y_j)\chi(-x_jw_j)\Big) \\
&= \frac{1}{q^\ell} \sum_{u_{i+1}^\ell} \sum_{w_1^\ell} \prod_{j\in[\ell]} M(w_j,y_j)\chi(-x_jw_j) \\
&= \frac{1}{q^\ell} \sum_{u_{i+1}^\ell} \sum_{w_1^\ell} \chi(-x_1^\ell(w_1^\ell)^\top) \prod_{j\in[\ell]} M(w_j,y_j) \\
&= \frac{1}{q^\ell} \sum_{u_{i+1}^\ell} \sum_{w_1^\ell} \chi(-x_1^\ell(w_1^\ell)^\top) M^\ell(w_1^\ell,y_1^\ell) \\
&= \frac{1}{q^\ell} \sum_{u_{i+1}^\ell} \sum_{w_1^\ell} \chi(-u_1^\ell G(w_1^\ell)^\top) M^\ell(w_1^\ell,y_1^\ell) \\
&= \frac{1}{q^\ell} \sum_{u_{i+1}^\ell} \sum_{w_1^\ell} \chi(-u_1^\ell (w_1^\ell G^\top)^\top) M^\ell(w_1^\ell,y_1^\ell) \\
&= \frac{1}{q^\ell} \sum_{u_{i+1}^\ell} \sum_{v_1^\ell} \chi(-u_1^\ell (v_1^\ell)^\top) M^\ell(v_1^\ell G^{-\top},y_1^\ell) \\
&= \frac{1}{q^\ell} \sum_{v_1^\ell} \chi(-u_1^i (v_1^i)^\top) M^\ell(v_1^\ell G^{-\top},y_1^\ell) \sum_{u_{i+1}^\ell} \chi(-u_{i+1}^\ell (v_{i+1}^\ell)^\top) \\
&= \frac{1}{q^\ell} \sum_{v_1^\ell} \chi(-u_1^i (v_1^i)^\top) M^\ell(v_1^\ell G^{-\top},y_1^\ell) q^{\ell-i}\mathbb{I}\{v_{i+1}^\ell = 0\} \\
&= \frac{1}{q^i} \sum_{v_1^i} \chi(-u_1^i (v_1^i)^\top) M^\ell(v_1^i 0_{i+1}^\ell G^{-\top},y_1^\ell).
\end{aligned}
$$

The first equality expands the definition of $W^{(i)}$. In the next equality, we substitute $x_1^\ell \coloneqq u_1^\ell G$. The next equality expand the definition of W^ℓ down to W. The next two equalities Fourier expand W and reorder the operators. The next equality merges all $\chi(-x_jw_j)$ into one term by additivity (xc). In the next equality we define $M^\ell(w_1^\ell, y_1^\ell)$ to be the product of all $M(w_j,y_j)$. The next two equalities use $x_1^\ell(w_1^\ell)^\top = u_1^\ell G(w_1^\ell)^\top = u_1^\ell(w_1^\ell G^\top)^\top$. In the next equality we define $v_1^\ell \coloneqq w_1^\ell G^\top$; so summing over w_1^ℓ is equivalent to summing over v_1^ℓ. (Recall that $G^{-\top}$ is the notation of the inverse transpose of G.) The last three equalities sum over u_{i+1}^ℓ to force $v_{i+1}^\ell = 0$.

Having that $W^{(i)}(u_i, u_1^{i-1}y_1^\ell) = q^{-i}\sum_{v_1^i}\chi(-u_1^i(v_1^i)^\top)\times M^\ell(v_1^i 0_{i+1}^\ell G^{-\top}, y_1^\ell)$ in mind, we move on to

$$
\begin{aligned}
&M^{(i)}(\omega_i, u_1^{i-1}y_1^\ell) \coloneqq \sum_{z_i\in\mathbb{F}_q} W^{(i)}(z_i, u_1^{i-1}y_1^\ell)\chi(\omega_i z_i) \\
&= \sum_{z_i\in\mathbb{F}_q} \frac{1}{q^i} \sum_{v_1^i} \chi(-u_1^{i-1}z_i(v_1^i)^\top) M^\ell(v_1^i 0_{i+1}^\ell G^{-\top},y_1^\ell)\chi(\omega_i z_i) \\
&= \frac{1}{q^i} \sum_{v_1^i} \chi(-u_1^{i-1}(v_1^{i-1})^\top) M^\ell(v_1^i 0_{i+1}^\ell G^{-\top},y_1^\ell) \sum_{z_i\in\mathbb{F}_q} \chi(z_i(\omega_i - v_i)) \\
&= \frac{1}{q^i} \sum_{v_1^i} \chi(-u_1^{i-1}(v_1^{i-1})^\top) M^\ell(v_1^i 0_{i+1}^\ell G^{-\top},y_1^\ell) q\mathbb{I}\{\omega_i = v_i\} \\
&= \frac{q}{q^i} \sum_{v_1^{i-1}} \chi(-u_1^{i-1}(v_1^{i-1})^\top) M^\ell(v_1^{i-1}\omega_i 0_{i+1}^\ell G^{-\top},y_1^\ell).
\end{aligned}
$$

In the first line we let $M^{(i)}$ be the Fourier coefficient of $W^{(i)}$. The next equality plugs in what we have about $W^{(i)}$ in mind. The next three equalities sum over z_i to force $v_i = \omega_i$.

With $M^{(i)}(\omega_i, u_1^{i-1}y_1^\ell) = q^{1-i}\sum_{v_1^{i-1}}\chi(-u_1^{i-1}(v_1^{i-1})^\top)\times M^\ell(v_1^{i-1}\omega_i 0_{i+1}^\ell G^{-\top}, y_1^\ell)$ in place, we obtain that with arbitrary $0 \neq \omega_i \in \mathbb{F}_q$,

$$
\begin{aligned}
&\sum_{u_1^{i-1}y_1^\ell \in \mathbb{F}^{i-1}\times\mathcal{Y}^\ell} |M^{(i)}(\omega_i, u_1^{i-1}y_1^\ell)| \qquad (11)\\
&= \sum_{u_1^{i-1}y_1^\ell} \Big| \frac{q}{q^i} \sum_{v_1^{i-1}} \chi(-u_1^{i-1}(v_1^{i-1})^\top) M^\ell(v_1^{i-1}\omega_i 0_{i+1}^\ell G^{-\top},y_1^\ell)\Big| \\
&\leqslant \sum_{u_1^{i-1}y_1^\ell} \frac{q}{q^i} \sum_{v_1^{i-1}} |M^\ell(v_1^{i-1}\omega_i 0_{i+1}^\ell G^{-\top},y_1^\ell)| \\
&= \sum_{y_1^\ell} \sum_{v_1^{i-1}} |M^\ell(v_1^{i-1}\omega_i 0_{i+1}^\ell G^{-\top},y_1^\ell)| \\
&= \sum_{y_1^\ell} \sum_{v_1^{i-1}} \prod_{j\in[\ell]} |M(w_j,y_j)| \\
&= \sum_{y_1^\ell} \sum_{v_1^{i-1}} \prod_{j\in J} |M(w_j,y_j)| \prod_{k\notin J} |M(w_k,y_k)| \\
&= \sum_{v_1^{i-1}} \prod_{j\in J} \Big(\sum_{y_j} |M(w_j,y_j)|\Big) \prod_{k\notin J} \Big(\sum_{y_k} |M(w_k,y_k)|\Big) \\
&= \sum_{v_1^{i-1}} \prod_{j\in J} \Big(\sum_{y_j} |M(w_j,y_j)|\Big) \leqslant \sum_{v_1^{i-1}} \prod_{j\in J} S_{\max}(W) \\
&= \sum_{v_1^{i-1}} S_{\max}(W)^{|J|} = \sum_{v_1^{i-1}} S_{\max}(W)^{\mathrm{wt}(v_1^{i-1}\omega_i 0_{i+1}^\ell G^{-\top})} \\
&= \sum_{v_1^{i-1}} S_{\max}(W)^{\mathrm{wt}(v_1^{i-1}1_i 0_{i+1}^\ell G^{-\top})} = f_{GS}^{(i)}(S_{\max}(W)).
\end{aligned}
$$

The first equality expands the Fourier coefficients. The next inequality is triangle plus (xb). The next equality cancels the summation over u_1^{i-1} with q^{1-i}. In the next equality we substitute $w_1^\ell \coloneqq v_1^{i-1}\omega_i 0_{i+1}^\ell G^{-\top}$; slightly different from the free w_1^ℓ before, they are now restricted to a proper subspace. The next equality classifies the indices into two classes—$j \in J$ are those such that $w_j \neq 0$ and $k \notin J$ are such that $w_k = 0$. The next two equalities reorder the operators and simplify $\sum_{y_k}|M(0,y_k)| = \sum_{y_k} W_{\mathrm{out}}(y_k) = 1$. The next inequality replaces w_j by one that maximizes $\sum_{y_k}|M(w_j,y_k)|$. The rest is trivial.

Theorem 10 claims that $S_{\max}(W^{(i)}) \leqslant f_{GS}^{(i)}(S_{\max}(W))$, where $f_{GS}^{(i)}$ is the weight enumerator of the i-th dual coset code. Since $S_{\max}(W^{(i)})$ is merely the maximum of Formula (11) over $0 \neq \omega_i \in \mathbb{F}_q$, we arrive at $S_{\max}(W^{(i)}) \leqslant f_{GS}^{(i)}(S_{\max}(W))$. And quod erat demonstrandum.

Experienced readers may find that all but the last inequality is a duplicate of [29, Lemma 34]. An instant improvement made here is to stop at $f_{GS}^{(i)}(S_{\max})$ instead of ending at $q^{i-1}S_{\max}^{\text{minimum distance}}$.

C. An Upper Bound on Entropy Functions

For all $z \in [0,1]$,

$$h_2(z) \leqslant \sqrt{ez}.$$

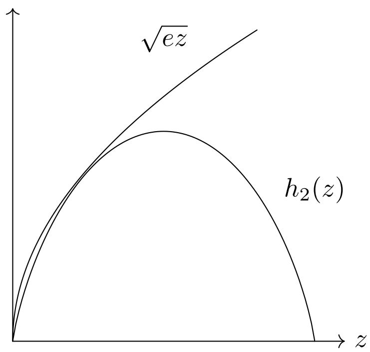

Fig. 6. Binary entropy function $h_2(z)$ and an upper bound of $\sqrt{ez}$.

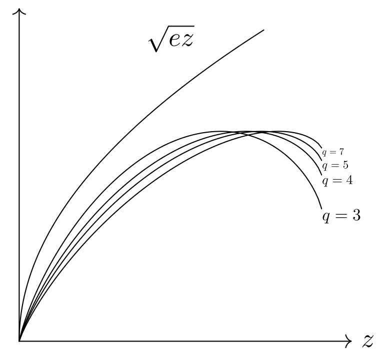

Fig. 7. $1 - \mathbb{D}(z \parallel 1 - 1/q)/\log q$ for $q = 3, 4, 5, 7$ and an upper bound of $\sqrt{ez}$.

See Fig. 6 for evidence. Continue this paragraph for a proof. There are three cases. Whenever $0 \leqslant z \leqslant 1/6$, replace $(1 - z)\log_2(1 - z)$ by the tangent line $z/\log 2$; it suffices to show $-z\log_2 z + z/\log 2 \leqslant \sqrt{ez}$. Substitute $z \mapsto \zeta^2$; want to show $-2\zeta^2\log_2\zeta + \zeta^2/\log 2 \leqslant \zeta\sqrt{e}$. Clean up; it remains to show $-2\zeta\log\zeta + \zeta \leqslant \sqrt{e}\log 2$. The left-hand side is monotonically increasing and does not reach $\sqrt{e}\log 2$ so long as $z \leqslant 1/6$ (i.e., $\zeta \leqslant 1/\sqrt{6}$). This closes the first case. Whenever $1/6 \leqslant z \leqslant 2/7$, replace $h_2(z)$ by the tangent line at $z = 2/9$. This line is of the form $f(z) := h_2'(2/9)(z-2/9)+h_2(2/9)$. It remains to show $f(z)-\sqrt{ez} \leqslant 0$. The left-hand side is a quadratic function in $\sqrt{z}$ whose roots can be computed algebraically. The roots lie outside $1/\sqrt{6} \leqslant \sqrt{z} \leqslant \sqrt{2/7}$, which close the second case. Whenever $2/7 \leqslant z \leqslant 1$, similarly, replace $h_2(z)$ by the tangent line at $z = 1/3$; that is, $f(z) := h_2'(1/3)(z-1/3)+h_2(1/3)$. The roots of $f(z)-\sqrt{ez}$, again, lie outside the concerned interval. This closes the third and last case.

More generally, for all prime powers q,

$$\begin{aligned} 1 - \frac{1}{\log q}\mathbb{D}\Big(z \,\Big\|\, 1 - \frac{1}{q}\Big) &= 1 - z\log_q\frac{z}{1-1/q} - (1-z)\log_q\frac{1-z}{1/q} \\ &= -z\log_q\frac{z}{q-1} - (1-z)\log_q(1-z) \leqslant \sqrt{ez} \end{aligned}$$

for all $z \in [0, 1]$, where $\mathbb{D}$ is the Kullback–Leibler divergence with the natural logarithm. This falls back to the h_2 case when $q = 2$. Figure 7 plots for $q = 3, 4, 5, 7$. It can be observed that as $q \to \infty$ the function tends to a line connecting $(0, 0)$ and $(1, 1)$, hence the upper bound should hold. Taking derivative in q shows that the left-hand side decreases as q increases and $z < 1/2$.

It will be seen later that $\sqrt{ez}$ acts as an easy-to-manipulate alternative of the Gilbert–Varshamov bound. We have not seen any relaxation like this in other works.

D. On the Weight Distribution of Random Linear Codes

This subsection contains the nontrivial part of the proof of Lemma 13. Fix any $i \in [\ell]$. We want to prove that when $\mathbb{G} \in \mathrm{GL}(\ell, q)$ is selected uniformly at random, the inequality

$$Z_{\mathrm{mad}}(W_{\mathbb{G}}^{(i)}) \leqslant \ell\exp(qZ_{\mathrm{mad}}(W)\ell)(qZ_{\mathrm{mad}}(W))^{\lceil i^2/3\ell\rceil} \qquad \text{((8)'s copy)}$$

holds with probability $1-3q^{-\sqrt{\ell}/13}$. In bounding the left-hand side, the fundamental theorem of polar coding—Z-end reads $Z_{\mathrm{mad}}(W_{\mathbb{G}}^{(i)}) \leqslant f_{\mathbb{G}Z}^{(i)}(Z_{\mathrm{mad}}(W))$, where $f_{\mathbb{G}Z}^{(i)}$ is the weight enumerator of codewords of the form $0_1^{i-1}1u_{i+1}^\ell\mathbb{G}$. Thus it remains to show the inequality with the left-hand side replaced

$$f_{\mathbb{G}Z}^{(i)}(z) \leqslant \ell\exp(qz\ell)(qz)^{\lceil i^2/3\ell\rceil}$$

where $z := Z_{\mathrm{mad}}(W)$ for short. This inequality is in fact a consequence of

$$f_{\mathbb{G}Z}^{(i)}(z) \leqslant \ell(1+(q-1)z)^{\ell-\lceil i^2/3\ell\rceil}((q-1)z)^{\lceil i^2/3\ell\rceil} \qquad (12)$$

because $(1+a)^b \leqslant \exp(ab)$. We will show the last inequality. Now divide i into two cases: $1 \leqslant i \leqslant \sqrt{3\ell}$ and $\sqrt{3\ell} < i \leqslant \ell$.

For $i = 1, 2, \ldots, \sqrt{3\ell}$, the exponent $\lceil i^2/3\ell\rceil$ is nothing but 1, thus the inequality to be proven reads $f_{\mathbb{G}Z}^{(i)}(z) \leqslant \ell(1+(q-1)z)^{\ell-1}(q-1)z$. The right-hand side overcounts all nonzero codewords by choosing a nonzero position (ℓ), assigning a nonzero symbol ($(q-1)z$), and filling in the rest of $\ell - 1$ blanks arbitrarily ($(1+(q-1)z)^{\ell-1}$). On the left-hand side, $f_{\mathbb{G}Z}^{(i)}$ enumerates only codewords of the form $0_1^{i-1}1_iu_{i+1}^\ell\mathbb{G}$, which are all nonzero as $\mathbb{G}$ is invertible. Hence Inequality (12) holds for $i \leqslant \sqrt{3\ell}$ and nonnegative z regardless of what kernel $\mathbb{G}$ is in effect.

For $i = \sqrt{3\ell}+1, \sqrt{3\ell}+2, \ldots, \ell$, let $k := \ell - i$ and let $d := i^2/3\ell$. These variables resemble the dimension and the minimal distance of a linear block code as in the notation *an* $[\ell, k, d]$*-code* in classical (algebraic) coding theory. To make Inequality (12) hold, we execute a two-phase procedure to avoid all codewords of weight less than d and to eliminate kernels with poor overall scores. In further detail, we will reject a kernel $\mathbb{G}$ if there exists u_{i+1}^ℓ such that $\mathrm{wt}(0_1^{i-1}1u_{i+1}^\ell\mathbb{G}) < d$ and call it phase I. Afterwards, among surviving kernels with only *heavy* (high weight) codewords, we will reject a kernel if its overall score $f_{\mathbb{G}Z}^{(i)}(z)$ is too low and call it phase II. The failing probability $3q^{-\sqrt{\ell}/13}$ is the price we pay for rejecting. Up to this point, two things remain to be analyzed: how much probability we pay for rejecting *light* (low weight) codewords in phase I (answer: $q^{-\sqrt{\ell}/13}$), and what is the Markov cutoff that honors Inequality (12) in phase II (answer: $2q^{-\sqrt{\ell}/13}$).

Phase I analysis is as follows: Fix u_{i+1}^ℓ and vary $\mathbb{G} \in \mathrm{GL}(\ell, q)$; the codeword $\mathbb{X}_1^\ell := 0_1^{i-1}1_iu_{i+1}^\ell\mathbb{G}$ is a nonzero vector distributed uniformly on $\mathbb{F}_q^\ell \setminus \{0_1^\ell\}$. This distribution is almost identical to the uniform distribution on $\mathbb{F}_q^\ell$. Assume $\mathbb{X}_1^\ell$ follows the latter; this makes $\mathbb{X}_1^\ell$ lighter, which is compatible with the direction of the inequalities we want. Then the probability that $\mathbb{X}_1^\ell$ has weight less than d is the probability that ℓ Bernoulli trials—each $\mathbb{X}_j$ is "zero" with probability $1/q$ and "nonzero" with probability $(q-1)/q$—result in less than d "nonzero"s. By the large deviations theory [83, Exercise 2.2.23(b)], $\mathrm{wt}(\mathbb{X}_1^\ell) < d$ holds with probability less than

$$\exp\Bigl(-\ell\mathbb{D}\Bigl(\frac{d}{\ell}\Bigm\|\frac{1}{2}\Bigr)\Bigr) = 2^{-\ell(1-h_2(d/\ell))}$$

for the $q = 2$ case, where $\mathbb{D}$ is the Kullback–Leibler divergence. For general q, similarly, $\mathrm{wt}(\mathbb{X}_1^\ell) < d$ holds with probability less than

$$\exp\Bigl(-\ell\mathbb{D}\Bigl(\frac{d}{\ell}\Bigm\|1-\frac{1}{q}\Bigr)\Bigr) \leqslant q^{-\ell(1-h_2(d/\ell))}.$$

It is less than $q^{-\ell(1-h_2(d/\ell))}$ by comparing Figs. 6 and 7 (meaning that $q = 2$ is the most difficult case). By Fig. 6, $h_2(d/\ell) < \sqrt{ed/\ell} = \sqrt{ei^2/3\ell^2} = (\sqrt{e/3})i/\ell < 0.952i/\ell$. So the single rejecting probability is less than $q^{-\ell(1-h_2(d/\ell))} < q^{-\ell+0.952i}$. Take into account that there are $q^{\ell-i}$ possibilities of u_{i+1}^ℓ. The union bound yields $q^{\ell-i}q^{-\ell+0.952i} = q^{-0.048i} < q^{-0.048\sqrt{3\ell}} < q^{-\sqrt{\ell}/13}$. Therefore, the total rejecting probability is $q^{-\sqrt{\ell}/13}$. Phase I ends here.

Phase II analysis is as follows: After we reject some $\mathbb{G}$ in phase I, some codewords will disappear; particularly, this includes all light codewords. Therefore, the expectation of $f_{\mathbb{G}Z}^{(i)}(z)$ is bounded by the weight enumerator of all heavy codewords rescaled by the number of codewords. In detail, start from

$$\begin{aligned}&\mathbb{E}[f_{\mathbb{G}Z}^{(i)}(z) \mid \mathbb{G}\text{survives phase I}]\\ &\quad= \mathbb{E}[f_{\mathbb{G}Z}^{(i)}(z)\mathbb{I}\{\mathbb{G}\text{survives}\}]/\mathbb{P}\{\mathbb{G}\text{survives}\}\\ &\quad\leqslant \mathbb{E}[f_{\mathbb{G}Z}^{(i)}(z)\mathbb{I}\{\mathbb{G}\text{survives}\}]/(1-q^{-\sqrt{\ell}/13}). \qquad (13)\end{aligned}$$

$\mathbb{I}$ is the indicator function. In the denominator, $1-q^{-\sqrt{\ell}/13} > 1/4$ as $\ell \geqslant 30$. Put that aside and redefine $d := \lceil i^2/3\ell\rceil$. The expected value part is bounded from above by

$$\begin{aligned}&\mathbb{E}[f_{\mathbb{G}Z}^{(i)}(z)\mathbb{I}\{\mathbb{G}\text{survives}\}] = \mathbb{E}\Bigl[\sum_{u_{i+1}^\ell} z^{\mathrm{wt}(u_{i+1}^\ell\mathbb{G})}\mathbb{I}\{\mathbb{G}\text{survives}\}\Bigr]\\ &\leqslant \mathbb{E}\Bigl[\sum_{u_{i+1}^\ell} z^{\mathrm{wt}(u_{i+1}^\ell\mathbb{G})}\mathbb{I}\{\mathrm{wt}(u_{i+1}^\ell\mathbb{G}) \geqslant d\}\Bigr]\\ &= \sum_{u_{i+1}^\ell} \mathbb{E}[z^{\mathrm{wt}(u_{i+1}^\ell\mathbb{G})}\mathbb{I}\{\mathrm{wt}(u_{i+1}^\ell\mathbb{G}) \geqslant d\}]\\ &\leqslant q^{\ell-i}\mathbb{E}[z^{\mathrm{wt}(\mathbb{X}_1^\ell)}\mathbb{I}\{\mathrm{wt}(\mathbb{X}_1^\ell) \geqslant d\}]\\ &= q^{\ell-i}q^{-\ell}\sum_{x_1^\ell} z^{\mathrm{wt}(x_1^\ell)}\mathbb{I}\{\mathrm{wt}(x_1^\ell) \geqslant d\}\\ &= q^{-i}\sum_{w\geqslant d}\binom{\ell}{w}z^w(q-1)^w \leqslant q^{-i}\sum_{w\geqslant d}\binom{\ell}{d}\binom{\ell-d}{w-d}z^w(q-1)^w\\ &= q^{-i}\binom{\ell}{d}\sum_{w\geqslant d}\binom{\ell-d}{w-d}z^{w-d}(q-1)^{w-d}((q-1)z)^d.\\ &= q^{-i}\binom{\ell}{d}(1+(q-1)z)^{\ell-d}((q-1)z)^d\\ &(\text{overestimate the scalar } q^{-i}\tbinom{\ell}{d})\\ &\leqslant (q^{-\sqrt{\ell}/13}\ell/2)(1+(q-1)z)^{\ell-d}((q-1)z)^d.\end{aligned}$$

The first equality expands the definition. The next inequality replaces $\mathbb{G}$ surviving phase I by a weaker condition. The next equality swaps $\mathbb{E}$ and $\sum$. The next inequality replaces the ensemble of $u_{i+1}^\ell\mathbb{G}$ by a uniform $\mathbb{X}_1^\ell \in \mathbb{F}_q^\ell$. The next equality expands the definition of the expectation over $\mathbb{X}_1^\ell$. The next equality counts codewords. The next inequality selects w positions by first selecting d and then selecting $w-d$. The next two equalities factor and apply the binomial theorem. The rest is by a series of inequalities that overestimate the scalar: $q^{-i}\binom{\ell}{d} = q^{-i}\binom{\ell}{\lceil i^2/4\ell\rceil} < q^{-i}\binom{\ell}{i^2/4\ell}\ell/2 < q^{-i}2^{\ell h_2(i^2/4\ell^2)}\ell/2 \leqslant q^{-i+\ell h_2(i^2/4\ell^2)}\ell/2$. Similar to the end of phase I, the exponent part is $-i+\ell h_2(i^2/3\ell^2) < -i+\ell\sqrt{ei^2/3\ell^2} = -i+i\sqrt{e/3} < -0.048i < -0.048\sqrt{3\ell} < -\sqrt{\ell}/13$. Hence the scalar part is less than $q^{-\sqrt{\ell}/13}\ell/2$. Put $1-q^{-\sqrt{\ell}/13} > 1/4$ back to the denominator as in Inequality (13); $\mathbb{E}[f_{\mathbb{G}Z}^{(i)}(z) \mid \mathbb{G}\text{survives phase I}]$ has an upper bound of

$$2q^{-\sqrt{\ell}/13}\ell(1+(q-1)z)^{\ell-d}((q-1)z)^d.$$

By Markov's inequality, Inequality (12) holds with probability $1-2q^{-\sqrt{\ell}/13}$, i.e., the rejecting probability is $2q^{-\sqrt{\ell}/13}$. Phase II ends here. The sum of the two rejecting probabilities is $3q^{-\sqrt{\ell}/13}$ as claimed in Lemma 13, hence the lemma settled.

E. Bibliographic Remarks

Concerning the fundamental theorems: Nonlinear g^W is not taken into consideration for that it is harder to deal with the MacWilliams duality of nonlinear codes. Also the S-parameter does not generalize to non-field input alphabet. Concerning random linear codes: [9, Section II.C] portrays a clear picture of the weight distributions of binary random linear codes. Section VII-D accommodates and extends their argument to general prime power q. Concerning the LDP behavior: [27, Theorem 22] showed that $\pi < 1$ can be arbitrary close to 1 over binary alphabet utilizing the Bose–Chaudhuri–Hocquenghem codes. Our Lemma 13, on the other hand, implies that almost all kernels make π close to 1.

We are halfway through the clarification of (cr). It remains to prove Lemmas 12 and 14.

VIII. Local CLT Behavior (Proof of Lemma 14)

We are to prove that the following inequality holds with high probability (w.r.t. the random kernel $\mathbb{G}$):

$$\sum_{i=1}^{\ell} h_\alpha(H(W_{\mathbb{G}}^{(i)})) < 4\ell^{1/2+\alpha}. \qquad ((10)\text{'s copy})$$

The target inequality is the sum of the following three inequalities:

$$\sum_{i=\lceil H(W)\ell+\ell^{1/2+\alpha}\rceil+1}^{\ell} h_\alpha(H(W_{\mathbb{G}}^{(i)})) < \ell^{1/2+\alpha}, \qquad (14)$$

$$\sum_{i=\lfloor H(W)\ell-\ell^{1/2+\alpha}\rfloor+1}^{\lceil H(W)\ell+\ell^{1/2+\alpha}\rceil} h_\alpha(H(W_{\mathbb{G}}^{(i)})) < 2\ell^{1/2+\alpha},$$
$$\sum_{i=1}^{\lfloor H(W)\ell-\ell^{1/2+\alpha}\rfloor} h_\alpha(H(W_{\mathbb{G}}^{(i)})) < \ell^{1/2+\alpha}. \quad (15)$$

The second one is trivial as $h_\alpha(z) \leqslant (1/2)^\alpha$. The first one will be proven in Section VIII-C with failing probability $\ell^{-\log(\ell)/20}$. The third one will be proven in Section VIII-D with failing probability $\ell^{-\log(\ell)/20}$. Before the main proofs, we devote Section VIII-A to introduce the symmetrization trick, which will reduce our proof to the case of symmetric q-ary channels. A channel W being symmetric means that for any affine shifting $\xi \in \mathbb{F}_q$, there exists a permutation σ on $\mathcal{Y}$ such that $W(y \mid \xi + x) = W(\sigma(y) \mid x)$ holds for all $x \in \mathbb{F}_q$ and $y \in \mathcal{Y}$. It also means that the uniform input achieves the Shannon capacity. This justifies the usage of linear codes. In Section VIII-B, we invoke some *universal bound* on entropies and exponents from Chang, Draper, and Sahai's works. Finally, we will be abusing the theory of random (linear) codes in Section VIII-C for robustness over noisy channels and in Section VIII-D for secrecy over wiretap channels.

A. Symmetrize Channel and Uniformize Input

Let $W\colon \mathbb{F}_q \to \mathcal{Y}$ be any q-ary channel; let X and Y be some input and the corresponding output. Symmetrize the channel as follows: Let $\Xi \in \mathbb{F}_q$ be a uniform r.v. independent of X, Y. Let $\bar{W}\colon \mathbb{F}_q\times(\mathbb{F}_q\times\mathcal{Y}) \to [0,1]$ be the probability mass function of this combination of r.v.s $(\Xi + X, (X,Y)) \in \mathbb{F}_q \times (\mathbb{F}_q \times \mathcal{Y})$. This $\bar{W}$ behaves like a channel such that, quote, unquote, $\bar{W}((x,y) \mid z) = W(x,y)/q$ for all inputs $z \in \mathbb{F}_q$ and outputs $(x,y) \in \mathbb{F}_q \times \mathcal{Y}$. Despite that this channel might be properly simulated by a symmetric channel with feedback Ξ to the sender, all that matters is that the biased input X is neutralized by the uniform r.v. Ξ, and becomes uniform. Let g^W be the multiplication of an invertible matrix G from the right. Let $\bar{W}^{(i)}(u_i, u_1^{i-1}x_1^\ell y_1^\ell)$ be the probability mass function of the tuple $(U_i, U_1^{i-1}X_1^\ell Y_1^\ell)$, where $U_1^\ell G = \Xi_1^\ell + X_1^\ell$. This definition is compatible with the channel transformation of $\bar{W}$ as if $\bar{W}$ was an actual channel in the first place. Let $H(\bar{W}^{(i)})$ be $H(U_i \mid U_1^{i-1}X_1^\ell Y_1^\ell)$; this is also compatible. The following lemma justifies why $\bar{W}$ is useful in theory.

Lemma 15 (Channel Symmetrization): $\bar{W}$ is a symmetric q-ary channel, $H(\bar{W}) = H(W)$, and $H(\bar{W}^{(i)}) = H(W^{(i)})$ for all $i \in [\ell]$.

This lemma is by [29, Definition 6 and Lemmas 7 and 8], plus the arguments in between. See also [33, Theorem 2] where they cared about whether $Z(\bar{W}^{(i)}) = Z(W^{(i)})$. One could also expand all definitions to verify the identities.

The consequence of this lemma is that $\bar{W}$ behaves like a shadow copy of W, but is symmetric. All inequalities involving entropies of W and $W^{(i)}$ are reduced to inequalities involving entropies of $\bar{W}$ and $\bar{W}^{(i)}$. Subsequently, passing statements to $\bar{W}$ is effectively assuming that the channel W is symmetric with the uniform input to begin with. In the upcoming subsections, we will prove that the targeted inequalities, (14) and (15), hold for any symmetric q-ary channel W with the uniform input with high probability. We conjecture that the symmetrization technique is optional as it seems like a wrapper of complicated Bayesian formulas.

B. Chang–Sahai's Universal Quadratic Bound

This and the next two subsections contain the most convoluted part of the proof of Lemma 14. This subsection prepares a universal upper bound on Gallager's E-null function, which ultimately evolves into a universal lower bound on Gallager's error exponent.

Let $W\colon \mathcal{X} \to \mathcal{Y}$ be a q-ary channel. Symmetry is not required in this subsection but it is in the next two. Assume the uniform input distribution $W_{\text{in}}(x) = 1/q$ for all $x \in \mathcal{X}$. Define Gallager's E-null function and its complement [73, Formula (1)]:

$$E_0(t) := -\log \sum_{y\in\mathcal{Y}} \Big(\sum_{x\in\mathcal{X}} W_{\text{in}}(x) W(y \mid x)^{\frac{1}{1+t}}\Big)^{1+t},$$
$$\bar{E}_0(t) := \log \sum_{y\in\mathcal{Y}} \Big(\sum_{x\in\mathcal{X}} W(x,y)^{\frac{1}{1+t}}\Big)^{1+t}.$$

By complement we mean that under the uniform input, $\bar{E}_0(t)$ degenerates to

$$\begin{aligned}\bar{E}_0(t) &= \log \sum_{y\in\mathcal{Y}} \Big(\sum_{x\in\mathcal{X}} (q^{-1}W(y \mid x))^{\frac{1}{1+t}}\Big)^{1+t} \\ &= t\log q + \log \sum_{y\in\mathcal{Y}} \Big(\sum_{x\in\mathcal{X}} q^{-1}W(y \mid x)^{\frac{1}{1+t}}\Big)^{1+t} \\ &= t\log q - E_0(t).\end{aligned}$$

Equivalently, $E_0(t) + \bar{E}_0(t) = t\log q$. For non-uniform inputs, $W_{\text{in}}(x)$ does not penetrate the summations.

The E-null function and its complement deeply associate to the following family of measures. For any $t \in [-2/5, 1]$, define the t-tilted probability mass function $W^t\colon \mathcal{X} \times \mathcal{Y} \to [0,1]$ as in [73, Definition 1]:

$$W^t(x,y) := \frac{\Big(\sum_{\xi\in\mathcal{X}} W(\xi,y)^{\frac{1}{1+t}}\Big)^{1+t}}{\sum_{\eta\in\mathcal{Y}}\Big(\sum_{\xi\in\mathcal{X}} W(\xi,\eta)^{\frac{1}{1+t}}\Big)^{1+t}} \times \frac{W(x,y)^{\frac{1}{1+t}}}{\sum_{\xi\in\mathcal{X}} W(\xi,y)^{\frac{1}{1+t}}}$$

Do not confuse W^ℓ with W^t, the latter W is tilted. When $t = 0$, the tilted falls back to its italic origin $W^0(x,y) = W(x,y)$. These measures can be interpreted as follows: W^t behaves like a channel with a dedicated input distribution. The first fraction in the definition specifies the output distribution $W^t_{\text{out}}(y)$. The second fraction specifies the a posteriori distribution $W^t(x \mid y)$ when y is known. As W^t is not an actual channel, it is not meaningful to alter the input distribution and ask for the corresponding output. Like the symmetrization trick, all that matters is that we can compute entropies, and what not, as if they were real channels. Quantities we are interested in are listed below: Let H_e be the base-e entropy. Let $H_e(W^t)$ be $H_e(X^t \mid Y^t)$ where (X^t, Y^t) is a tuple r.v. that follows W^t. Let $H_e(X^t{\upharpoonright}y)$ be the entropy of the a posteriori distribution of X^t given $Y^t = y$; to be specific, $H_e(X^t{\upharpoonright}y) = \sum_{x\in X} W^t(x \mid y) \log W^t(x \mid y)$.

[73, Formula (13) and (19)] have that the following hold for $t \in [0, 1]$:

$$\frac{d}{dt}\bar{E}_0(t) = \bar{E}_0'(t) = H_e(W^t), \quad \text{and}$$
$$\frac{d^2}{dt^2}\bar{E}_0(t) = \bar{E}_0''(t) = \frac{d}{dt}H_e(W^t)$$
$$= \frac{1}{1+t}\sum_{y\in\mathcal{Y}} W^t_{\text{out}}(y)\sum_{x\in\mathcal{X}} W^t(x \mid y)\log(W^t(x \mid y))^2$$
$$+ \frac{t}{1+t}\sum_{y\in\mathcal{Y}} W^t_{\text{out}}(y)H_e(X^t{\upharpoonright}y)^2 - H_e(W^t)^2. \quad (16)$$

Careful readers may verify them by hand or follow [73, Formulas (13) to (19)] and [74, Lemmas 9 and 10]. Similar computations are also carried out by [19], [21].

Notice that $\bar{E}_0(t)$, $H_e(W^t)$, and every other term in Equation (16) are all holomorphic functions in t on the half-plane $\operatorname{Re} t > -1$ (there is a singularity at $1/(1+t) = \infty$). By the identity theorem in complex analysis [84, Corollary 8.16], [85, page 127], Equation (16) holds for all $t \in [-2/5, 1]$. Dropping the nonpositive square, we deduce an upper bound for each $t \in [-2/5, 1]$:

$$\bar{E}_0''(t) \leqslant \frac{1}{1+t}\sum_{y\in\mathcal{Y}} W^t_{\text{out}}(y)\sum_{x\in\mathcal{X}} W^t(x \mid y)\log(W^t(x \mid y))^2$$
$$+ \frac{\max(0,t)}{1+t}\sum_{y\in\mathcal{Y}} W^t_{\text{out}}(y)H_e(X^t{\upharpoonright}y)^2. \quad (17)$$

This upper bound on $\bar{E}_0''(t)$ is a positive combination of

$$\sum_{x\in\mathcal{X}} W^t(x \mid y)\log(W^t(x \mid y))^2 \quad \text{and} \quad H_e(X^t{\upharpoonright}y)^2$$

parametrized by $y \in \mathcal{Y}$, so it remains to bound them separately. For the second kind of constituents, the entropy cannot exceed $\log q$ so $H_e(X^t{\upharpoonright}y)^2 \leqslant \log(q)^2$. For the first kind of constituents, the following lemma adapted from [73, Lemma 1] helps.

Lemma 16 (Second Moment): If $w_1, w_2, , \ldots, w_q$ are positive numbers of sum 1, then

$$\sum_i w_i\log(w_i)^2 \leqslant \begin{Bmatrix} \log(q)^2 & \text{for } q \geqslant 3 \\ 0.563 & \text{for } q = 2 \end{Bmatrix} \leqslant 1.2\log(q)^2.$$

With the lemma, $\sum_{x\in\mathcal{X}} W^t(x \mid y)(\log W^t(x \mid y))^2 \leqslant 1.2\log(q)^2$ can be stated. Now Inequality (17) becomes

$$\bar{E}_0''(t) \leqslant \frac{1}{1+t}\sum_{y\in\mathcal{Y}} W^t_{\text{out}}(y)\cdot 1.2\log(q)^2$$
$$+ \frac{\max(0,t)}{1+t}\sum_{y\in\mathcal{Y}} W^t_{\text{out}}(y)\log(q)^2$$
$$\leqslant \frac{1}{1+t}\cdot 1.2\log(q)^2 + \frac{\max(0,t)}{1+t}\log(q)^2$$
$$\leqslant 2\log(q)^2$$

for all $t \in [-2/5, 1]$. Since $E_0(t)$ is a linear function $t\log q$ minus $\bar{E}_0(t)$, their first derivatives sum to $\log q$ while their second derivatives are opposite. Hence the following lemma.

Lemma 17 (Universal Quadratic Bound): [73, Theorem 2]. Cf. [6, Theorem 5.6.3]. Let W be a q-ary channel. Assume the uniform input distribution. Then Gallager's E-null function satisfies

$$E_0(0) = 0,$$
$$E_0'(0) = I(W)\log q,$$
$$E_0''(t) \geqslant -2\log(q)^2$$

for all $t \in [-2/5, 1]$. In particular, it satisfies

$$E_0(t) \geqslant I(W)t\log q - t^2\log(q)^2.$$

C. Gallager's Argument at Bob's End

This subsection take advantage of the universal bound developed three lines ago and starts actually proving Lemma 14. This subsection deals with

$$\sum_{i=\lceil H(W)\ell+\ell^{1/2+\alpha}\rceil+1}^{\ell} h_\alpha(H(W^{(i)}_{\mathbb{G}})) < \ell^{-1/2+\alpha} \quad \text{((14)'s copy)}$$

by passing it to an inequality that captures the performance of noisy-channel coding. Owing to h_α's concavity, the left-hand side of Inequality (14) is

$$\sum_{i=j+1}^{\ell} h_\alpha(H(W^{(i)}_{\mathbb{G}})) \leqslant (\ell-j)h_\alpha\Big(\frac{1}{\ell-j}\sum_{i=j+1}^{\ell} H(W^{(i)}_{\mathbb{G}})\Big)$$

where $j := \lceil H(W)\ell + \ell^{1/2+\alpha}\rceil$ for short. It suffices to prove that the right-hand side is less than $\ell^{-1/2+\alpha}$. In the spirit of the motivational Chain Rule (1), the sum of the chain of $H(W^{(i)}_{\mathbb{G}})$ on the right-hand side is $H(U^\ell_{j+1} \mid U^j_1 Y^\ell_1)$. In order to prove Inequality (14), we will show

$$(\ell-j)h_\alpha\Big(\frac{1}{\ell-j}H(U^\ell_{j+1} \mid U^j_1 Y^\ell_1)\Big) < \ell^{-1/2+\alpha}. \quad (18)$$

But what is $H(U^\ell_{j+1} \mid U^j_1 Y^\ell_1)$? It measures the equivocation at Bob's end when U^j_1 is known to Bob. In other words, we may as well pretend that there is a random rectangular full-rank matrix $\mathbb{G}'$ with ℓ columns and only $k := \ell - j = \lfloor \ell - H(W)\ell - \ell^{1/2+\alpha}\rfloor$ rows, that Alice computes and sends $X^\ell_1 := U^k_1\mathbb{G}'$ to Bob, and that Bob attempts to decode $\hat{U}^k_1$ upon receiving Y^ℓ_1 using the MAP decoder. The equivocation is thus, by Fano's inequality, bounded in terms of the probability that Bob fails to decode U^k_1:

$$H(U^\ell_{j+1} \mid U^j_1 Y^\ell_1)$$
$$\leqslant -P_{\mathrm{e}}\log_q P_{\mathrm{e}} - (1-P_{\mathrm{e}})\log_q(1-P_{\mathrm{e}}) + P_{\mathrm{e}}\log_q(q^k-1)$$
$$\leqslant -P_{\mathrm{e}}\log_q P_{\mathrm{e}} + \frac{P_{\mathrm{e}}}{\log q} + P_{\mathrm{e}} = P_{\mathrm{e}}\Big(\frac{1-\log P_{\mathrm{e}}}{\log q} + k\Big). \quad (19)$$

Here P_{e} is the probability that Bob fails to decode, $\hat{U}^k_1 \neq U^k_1$.

The following is how to compute Bob's decoder block error probability. The generator matrix $\mathbb{G}'$ Alice uses is selected uniformly from the ensemble of full-rank k-by-ℓ matrices. The difference of every pair of codewords distributes uniformly on $\mathbb{F}^\ell_q \setminus \{0^\ell_1\}$. Over symmetric channels, the difference alone determines the likelihood ratios because $W^\ell(y^\ell_1 \mid \xi^\ell_1 + x^\ell_1) = W^\ell(\sigma^\ell_1(y^\ell_1) \mid x^\ell_1)$ for some component-wise permutation σ^ℓ_1

on $\mathcal{Y}^\ell$ depending on ξ_1^ℓ. So Gallager's bound applies. To elaborate, let $t \in [0,1]$. Then Bob's average error probability satisfies [6, Inequalities (5.6.2) to (5.6.14)]

$$\begin{aligned}
&\mathbb{E}P\{\text{Bob fails to decode } U_1^k \text{given } \mathbb{G}'\}\\
&= \mathbb{E}\sum_{u_1^k}\frac{1}{q^k}\sum_{y_1^\ell} W^\ell(y_1^\ell \mid u_1^k\mathbb{G}')\mathbb{I}\begin{Bmatrix}\text{Bob has } \hat{U}_1^k \neq u_1^k\\ \text{given } \mathbb{G}', u_1^k, y_1^\ell\end{Bmatrix}\\
&= \mathbb{E}\sum_{y_1^\ell} W^\ell(y_1^\ell \mid 0_1^\ell)\mathbb{I}\{\text{Bob has } \hat{U}_1^k \neq 0_1^k \text{given } \mathbb{G}', 0_1^k, y_1^\ell\}\\
&\leqslant \mathbb{E}\sum_{y_1^\ell} W^\ell(y_1^\ell \mid 0_1^\ell)\Bigl(\sum_{v_1^k\neq 0_1^k}\mathbb{I}\begin{Bmatrix}\text{Bob prefers } v_1^k\\ \text{over } 0_1^k\text{given } \mathbb{G}'\end{Bmatrix}\Bigr)^t\\
&\leqslant \mathbb{E}\sum_{y_1^\ell} W^\ell(y_1^\ell \mid 0_1^\ell)\Bigl(\sum_{v_1^k\neq 0_1^k}\frac{W^\ell(y_1^\ell \mid v_1^k\mathbb{G}')^{\frac{1}{1+t}}}{W^\ell(y_1^\ell \mid 0_1^\ell)^{\frac{1}{1+t}}}\Bigr)^t\\
&= \mathbb{E}\sum_{y_1^\ell} W^\ell(y_1^\ell \mid 0_1^\ell)^{\frac{1}{1+t}}\Bigl(\sum_{v_1^k\neq 0_1^k} W^\ell(y_1^\ell \mid v_1^k\mathbb{G}')^{\frac{1}{1+t}}\Bigr)^t\\
&\leqslant \sum_{y_1^\ell} W^\ell(y_1^\ell \mid 0_1^\ell)^{\frac{1}{1+t}}\Bigl(\mathbb{E}\sum_{v_1^k\neq 0_1^k} W^\ell(y_1^\ell \mid v_1^k\mathbb{G}')^{\frac{1}{1+t}}\Bigr)^t\\
&= \sum_{y_1^\ell} W^\ell(y_1^\ell \mid 0_1^\ell)^{\frac{1}{1+t}}\Bigl(\sum_{x_1^\ell\neq 0_1^\ell}\frac{q^k-1}{q^\ell-1} W^\ell(y_1^\ell \mid x_1^\ell)^{\frac{1}{1+t}}\Bigr)^t\\
&\leqslant q^{kt}\sum_{y_1^\ell} W^\ell(y_1^\ell \mid 0_1^\ell)^{\frac{1}{1+t}}\Bigl(\sum_{x_1^\ell\neq 0_1^\ell}\frac{1}{q^\ell} W^\ell(y_1^\ell \mid x_1^\ell)^{\frac{1}{1+t}}\Bigr)^t\\
&\leqslant q^{kt}\sum_{y_1^\ell} W^\ell(y_1^\ell \mid 0_1^\ell)^{\frac{1}{1+t}}\Bigl(\sum_{x_1^\ell}\frac{1}{q^\ell} W^\ell(y_1^\ell \mid x_1^\ell)^{\frac{1}{1+t}}\Bigr)^t\\
&= q^{kt}\sum_{y_1^\ell}\Bigl(\sum_{x_1^\ell}\frac{1}{q^\ell} W^\ell(y_1^\ell \mid x_1^\ell)^{\frac{1}{1+t}}\Bigr)\Bigl(\sum_{x_1^\ell}\frac{1}{q^\ell} W^\ell(y_1^\ell \mid x_1^\ell)^{\frac{1}{1+t}}\Bigr)^t\\
&= q^{kt}\sum_{y_1^\ell}\Bigl(\sum_{x_1^\ell}\frac{1}{q^\ell} W^\ell(y_1^\ell \mid x_1^\ell)^{\frac{1}{1+t}}\Bigr)^{1+t}\\
&= q^{kt}\sum_{y_1^\ell}\Bigl(\sum_{x_1^\ell} W_{\text{in}}^\ell(x_1^\ell) W^\ell(y_1^\ell \mid x_1^\ell)^{\frac{1}{1+t}}\Bigr)^{1+t}\\
&= \exp(kt\log q - (\textit{the E-null function of } W^\ell)(t))\\
&= \exp(kt\log q - \ell E_0(t)).
\end{aligned}$$

In summary, $\mathbb{E}P\{\text{Bob fails to decode } U_1^k\text{given } \mathbb{G}'\}$ is less than $\exp(kt\log q - \ell E_0(t))$ whenever $0 \leqslant t \leqslant 1$. Recall the universal quadratic bound $E_0(t) \geqslant I(W)t\log q - t^2\log(q)^2$ derived in Lemma 17. We obtain that the exponent is

$$\begin{aligned}
&kt\log q - \ell E_0(t) = (\ell - j)\log q - \ell E_0(t)\\
&\quad= (\ell - H(W)\ell - \ell^{1/2+\alpha})t\log q - \ell E_0(t)\\
&\quad= (I(W)\ell - \ell^{1/2+\alpha})t\log q - \ell E_0(t)\\
&\quad\leqslant (I(W)\ell - \ell^{1/2+\alpha})t\log q - \ell(I(W)t\log q - t^2\log(q)^2)\\
&\quad= (\ell t\log q - \ell^{1/2+\alpha})t\log q
\end{aligned}$$

(redeem the infimum at $t = \ell^{-1/2+\alpha}/2\log q$)

$$\begin{aligned}
&\quad\mapsto (\ell\ell^{-1/2+\alpha}/2 - \ell^{1/2+\alpha})\ell^{-1/2+\alpha}/2\\
&\quad= -\ell^{2\alpha}/4 = -\ell^{2\log(\log\ell)/\log\ell}/4 = -\log(\ell)^2/4.
\end{aligned}$$

So far we obtain that the average error probability is less than $\exp(-\log(\ell)^2/4) = \ell^{-\log(\ell)/4}$.

Run Markov's inequality with cutoff $\ell^{-\log(\ell)/20}$. To put it another way, we sample a random kernel and reject it if $P\{\text{Bob fails to decode } U_1^k\text{given } \mathbb{G}'\} \geqslant \ell^{-\log(\ell)/5}$. Then the rejecting probability is $\ell^{-\log(\ell)/20}$ because $1/20 + 1/5 = 1/4$. An upper bound on Bob's error probability being $P_\text{e} < \ell^{-\log(\ell)/5}$, an upper bound on Bob's equivocation is

$$\begin{aligned}
H(U_{j+1}^\ell \mid U_1^j Y_1^\ell) &\leqslant \ell^{-\log(\ell)/5}\Bigl(\frac{1-\log\ell^{-\log(\ell)/5}}{\log q} + k\Bigr)\\
&= \ell^{-\log(\ell)/5}\Bigl(\frac{1+\log(\ell)^2/5}{\log q} + k\Bigr)
\end{aligned}$$

by Inequality (19). Plugging the latter into $kh_\alpha(\text{this place}/k)$, we derive that the left-hand side of Inequality (18) is less than

$$\begin{aligned}
&kh_\alpha\Bigl(\frac{\ell^{-\log(\ell)/5}}{k}\Bigl(\frac{1+\log(\ell)^2/5}{\log q} + k\Bigr)\Bigr)\\
&\quad= k\Bigl(\ell^{-\log(\ell)/5}\Bigl(\frac{1+\log(\ell)^2/5}{k\log q} + 1\Bigr)\Bigr)^\alpha\\
&\quad= \ell^{-\alpha\log(\ell)/5}k\Bigl(\frac{1+\log(\ell)^2/5}{k\log q} + 1\Bigr)^\alpha\\
&\quad< \ell^{-\alpha\log(\ell)/5}\ell\Bigl(\frac{1+\log(\ell)^2/5}{\ell\log q} + 1\Bigr)^\alpha\\
&\quad< \ell^{-\alpha\log(\ell)/5}\cdot\ell\cdot 2^\alpha = 2^\alpha\ell\log(\ell)^{-\log(\ell)/5}.
\end{aligned}$$

The first inequality uses that the left-hand side increases monotonically in k and $k := \ell - j = \lfloor \ell - H(W)\ell - \ell^{1/2+\alpha}\rfloor < \ell$. The second inequality uses the assumption $\ell \geqslant 2$. In any regard, the quantity at the end of the inequalities decays to 0 as $\ell \to \infty$, so eventually it becomes less than $\ell^{1/2+\alpha}$, the right-hand side of Inequality (18). This proves that Inequality (14) holds with failing probability $\ell^{-\log(\ell)/20}$ as soon as ℓ is large enough. The lower bound on ℓ in the statement of Lemma 14 is large enough, hence the first half of Lemma 14 settled.

D. Hayashi's Argument at Eve's End

This subsection contains the very last ingredient of the proof of Lemma 14. We dealt with Inequality (14) in the last subsection. We now deal with

$$\sum_{i=1}^{\lfloor H(W)\ell - \ell^{1/2+\alpha}\rfloor} h_\alpha(H(W_\mathbb{G}^{(i)})) < \ell^{1/2+\alpha}. \qquad \text{((15)'s copy)}$$

Similar to how we motivated Inequality (18), we hereby apply Jensen's inequality and the chain rule of conditional entropy to simplify Inequality (15). The left-hand side becomes $jh_\alpha(H(U_1^j \mid Y_1^\ell)/j)$ where $j := \lfloor H(W)\ell - \ell^{1/2+\alpha}\rfloor$ for short. (This is not the same j as in the last subsection.) The input uniform, the argument of h_α is $H(U_1^j \mid Y_1^\ell)/j = 1 - I(U_1^j\,;Y_1^\ell)/j$, which can be replaced by $I(U_1^j \mid Y_1^\ell)/j$ thanks to the evenness $h_\alpha(1-z) = h_\alpha(z)$. We will show

$$jh_\alpha\Bigl(\frac{1}{j}I(U_1^j\,;Y_1^\ell)\Bigr) < \ell^{1/2+\alpha}. \tag{20}$$

But what is $I(U_1^j\,;Y_1^\ell)$? It is the amount of information Eve learns from wiretapping Y_1^ℓ if they know that U_{j+1}^ℓ are junk. In other words, we may pretend that Alice transmits $X_1^\ell := U_1^j V_{j+1}^\ell\mathbb{G}$ with confidential bits U_1^j and obfuscating bits V_{j+1}^ℓ,

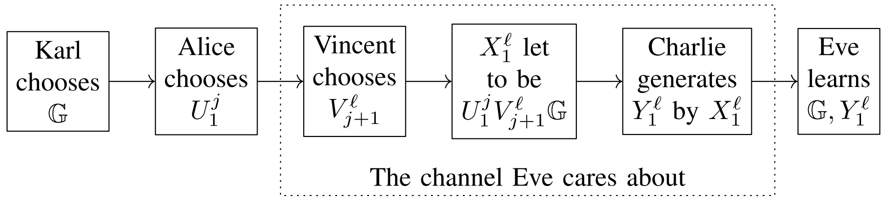

Fig. 8. A finer setup for Hayashi's secrecy exponent. Charlie generates Y_1^ℓ such that $X_1^\ell \coloneqq U_1^j V_{j+1}^\ell \mathbb{G}$ and Y_1^ℓ follow W^ℓ. Despite of the seemingly sequential structure, Karl, Alice, and Vincent work independently.

Bob receives X_1^ℓ in full, and Eve learns Y_1^ℓ. This context falls back to (a special case of) the traditional setup of wiretap channels [86] where various bounds are studied, some in terms of Gallager's E-null function.

Here are some preliminaries to control the information leaked to Eve. We follow the blueprint of how Hayashi derived the secrecy exponent in [87, Inequality (21)]. Consider the communication protocol depicted in Fig. 8: Karl fixes a kernel $\mathbb{G} \in \mathrm{GL}(\ell, q)$ and everyone knows $\mathbb{G}$. Alice chooses the confidential message U_1^ℓ. Vincent chooses the obfuscating bits V_{j+1}^ℓ. Charlie generates Y_1^ℓ by plugging $X_1^\ell \coloneqq U_1^j V_{j+1}^\ell \mathbb{G}$ into a simulator of W^ℓ. Eve learns Y_1^ℓ and is interested in knowing U_1^j alone. So the channel on topic is the composition of Vincent and Charlie. Notation: Running out of symbols, we all use $\mathbb{P}$ with proper subscriptions to indicate the corresponding probability measures. That said, indices in the subscription will be omitted. As Eve is interested in the relation between U_1^j and Y_1^ℓ, let $Y_1^\ell \upharpoonright G u_1^j$ be the r.v. that follows the a posteriori distribution of Y_1^ℓ given $\mathbb{G} = G$ and $U_1^j = u_1^j$. More formally, $\mathbb{P}_{Y \upharpoonright Gu}(y_1^\ell) = \mathbb{P}_{Y|\mathbb{G}U}(y_1^\ell \mid G, u_1^\ell) = \mathbb{P}_{\mathbb{G}UY}(G, u_1^j, y_1^\ell)/\mathbb{P}_{\mathbb{G}U}(G, u_1^j)$. We could have defined $Y_1^\ell \upharpoonright G$ to be the a posteriori distribution of Y_1^ℓ given $\mathbb{G} = G$; but it is simply the same distribution as Y_1^ℓ since $U_1^j V_{j+1}^\ell G$ traverses all inputs uniformly regardless of the choice of G. That is, $\mathbb{P}_{Y|\mathbb{G}}(y_1^\ell \mid G) = \mathbb{P}_Y(y_1^\ell)$ for all $y_1^\ell \in \mathcal{Y}^\ell$.

Fix G as an instance of $\mathbb{G}$. Let I_e be the base-e mutual information. The channel Eve cares about leaks information of this amount:

$$
\begin{aligned}
&I_e(U_1^j \,;\, Y_1^\ell \mid G) \\
&= \sum_{u_1^j y_1^\ell} \mathbb{P}_{UY|\mathbb{G}}(u_1^j, y_1^\ell \mid G) \log \frac{\mathbb{P}_{Y|\mathbb{G}U}(y_1^\ell \mid G, u_1^j)}{\mathbb{P}_{Y|\mathbb{G}}(y_1^\ell \mid G)} \\
&= \sum_{u_1^j} \mathbb{P}_U(u_1^j) \sum_{y_1^\ell} \mathbb{P}_{Y|\mathbb{G}U}(y_1^\ell \mid G, u_1^j) \log \frac{\mathbb{P}_{Y|\mathbb{G}U}(y_1^\ell \mid G, u_1^j)}{\mathbb{P}_{Y|\mathbb{G}}(y_1^\ell \mid G)} \\
&= \sum_{u_1^j} \mathbb{P}_U(u_1^j) \sum_{y_1^\ell} \mathbb{P}_{Y \upharpoonright Gu}(y_1^\ell) \log \frac{\mathbb{P}_{Y \upharpoonright Gu}(y_1^\ell)}{\mathbb{P}_Y(y_1^\ell)} \\
&= \sum_{u_1^j} \mathbb{P}_U(u_1^j) \mathbb{D}(Y_1^\ell \upharpoonright G u_1^j \parallel Y_1). \qquad (21)
\end{aligned}
$$

$\mathbb{D}(Y_1^\ell \upharpoonright G u_1^j \parallel Y_1^\ell)$ is the Kullback–Leibler divergence from the a posteriori distribution of Y_1^ℓ given G, u_1^j to the coarsest distribution Y_1^ℓ. We are to take expectation over $\mathbb{G}$ to find the average information leak since we are interested in Markov's inequality. Equality (21) yields

$$
\begin{aligned}
\mathbb{E} I_e(U_1^j \,;\, Y_1^\ell \mid \mathbb{G}) &= \sum_G \mathbb{P}_{\mathbb{G}}(G) I_e(U_1^j \,;\, Y_1^\ell \mid G) \\
&= \sum_G \mathbb{P}_{\mathbb{G}}(G) \sum_{u_1^j} \mathbb{P}_U(u_1^j) \mathbb{D}(Y_1^\ell \upharpoonright G u_1^j \parallel Y_1^\ell). \qquad (22)
\end{aligned}
$$

We now discover that there are redundancies in traversing all G and u_1^ℓ: After all, X_1^j is $u_1^j V_{j+1}^\ell G = u_1^j 0_{j+1}^\ell G + 0_1^j V_{j+1}^\ell G$, which is a fixed linear combination of the first j rows plus a random vector from the span of the bottom $\ell - j$ rows. When V_1^ℓ varies, the track of X_1^ℓ forms an affine subspace of $\mathbb{F}_q^\ell$, a *coset code* as in the context of the fundamental theorems. So what matters is the distribution of this coset code.

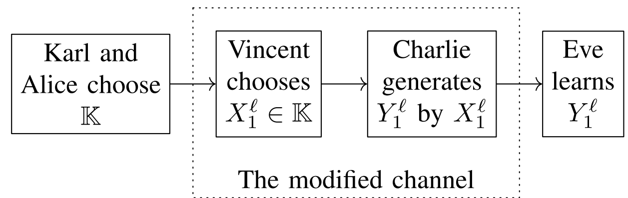

Fig. 9. A simplified setup for Hayashi's secrecy exponent. Charlie generates Y_1^ℓ such that X_1^ℓ and Y_1^ℓ follow W^ℓ.

In this regard, we replace the uniform ensemble of $(\mathbb{G}, U_1^j)$ by the uniform ensemble of $\mathbb{K}$ a rank-$(\ell - j)$ affine subspace of $\mathbb{F}_q^\ell$, where $j \coloneqq \lfloor H(W)\ell - \ell^{1/2+\alpha} \rfloor$. Karl and Alice together choose $\mathbb{K}$ uniformly. Vincent chooses $X_1^\ell \in \mathbb{K}$ uniformly. Charlie generates Y_1^ℓ by throwing X_1^ℓ into a simulator of W^ℓ. See Fig. 9 for the depiction of the new scheme. Hence Equality (22) becomes

$$
\mathbb{E} I_e(U_1^j \,;\, Y_1^\ell \mid \mathbb{G}) = \sum_K \mathbb{P}_{\mathbb{K}}(K) \mathbb{D}(Y_1^\ell \upharpoonright K \parallel Y_1^\ell)
$$

where $Y_1^\ell \upharpoonright K$ is the a posteriori distribution of Y_1^ℓ given $\mathbb{K} = K$. Suddenly, the quantity $\mathbb{E} I_e(U_1^j \,;\, Y_1^\ell \mid \mathbb{G})$ we are interested in turns into the mutual information $I_e(\mathbb{K} \,;\, Y_1^\ell)$ between $\mathbb{K}$ and Y_1^ℓ as $\mathbb{K}$ replaces the role of U_1^j in Formula (21). Recall that in Lemma 17 the mutual information is the derivative of Gallager's E-null function. We exploit this. Define the double-stroke E-null function for $(\mathbb{K}, Y_1^\ell)$ as follows

$$
\mathbb{E}_0(t) \coloneqq -\log \sum_{y_1^\ell} \Bigl(\sum_K \mathbb{P}_{\mathbb{K}}(K) \mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)^{\frac{1}{1+t}} \Bigr)^{1+t}.
$$

Then $\mathbb{E}_0'(0) = I_e(\mathbb{K} \,;\, Y_1^\ell) = \mathbb{E} I_e(U_1^j \,;\, Y_1^\ell \mid \mathbb{G})$. Owing to the concavity of the E-null function, $\mathbb{E}_0'(0) \leqslant \mathbb{E}_0(t)/t$

whenever $-2/5 \leqslant t < 0$. Recap: To bound the average leaked information $\mathbb{E}I_e(U_1^j ; Y_1^\ell \mid \mathbb{G})$ it suffices to bound $I_e(\mathbb{K} ; Y_1^\ell)$, which is then morphing to bounding $\mathbb{E}_0'(0)$ from above and to bounding $\mathbb{E}_0(t)$ from below.

The double-stroke E-null function is bounded as below. Assume $-2/5 \leqslant t < 0$. Let s be $-t/(1+t)$; so $0 < s \leqslant 2/3$ and $(1+s)(1+t) = 1$. For any fixed K and fixed $x_1^\ell \in K$, the base of the $(1+t)$-th root in the definition of the double-stroke E-null function is

$$\begin{aligned}\mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K) &= \sum_{\xi_1^\ell \in K} \mathbb{P}_{X|\mathbb{K}}(\xi_1^\ell \mid K)\mathbb{P}_{Y|X}(y_1^\ell \mid \xi_1^\ell) \\ &= \sum_{\xi_1^\ell \in K} q^j \mathbb{P}_X(\xi_1^\ell)\mathbb{P}_{Y|X}(y_1^\ell \mid \xi_1^\ell) = \sum_{\xi_1^\ell \in K} q^j \mathbb{P}_{XY}(\xi_1^\ell, y_1^\ell) \\ &= q^j\Big(\mathbb{P}_{XY}(x_1^\ell, y_1^\ell) + \sum_{x_1^\ell \neq \xi_1^\ell \in K} \mathbb{P}_{XY}(\xi_1^\ell, y_1^\ell)\Big) \\ &= q^j\Big(\mathbb{P}_{XY}(x_1^\ell, y_1^\ell) + \mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)\Big).\end{aligned}$$

Here $\mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)$ is a temporary shorthand for the summation of $\mathbb{P}_{XY}(\xi_1^\ell, y_1^\ell)$ over $\xi_1^\ell \in K$ that excludes x_1^ℓ. Raise $\mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)$ to the power of s; it becomes $q^{js}(\mathbb{P}_{XY}(x_1^\ell, y_1^\ell) + \mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell))^s \leqslant q^{js}(\mathbb{P}_{XY}(x_1^\ell, y_1^\ell)^s + \mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)^s)$ by sub-additivity. Put that aside; raise $\mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)$ to the power of $1+s = 1/(1+t)$:

$$\begin{aligned}&\mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)^{1+s} = \mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)\mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)^s \\ &= \sum_{x_1^\ell \in K} q^j \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)\mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)^s \\ &\leqslant \sum_{x_1^\ell \in K} q^j \mathbb{P}_{XY}(x_1^\ell, y_1^\ell) q^{js}\Big(\mathbb{P}_{XY}(x_1^\ell, y_1^\ell)^s + \mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)^s\Big) \\ &= q^{j+js}\Big(\sum_{x_1^\ell \in K} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)^{1+s} \\ &\qquad + \sum_{x_1^\ell \in K} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)\mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)^s\Big).\end{aligned}$$

The inequality rewrites the s-th power of $\mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)$. Then the inner sum of the E-null function morphs as follows

$$\begin{aligned}&\sum_K \mathbb{P}_\mathbb{K}(K)\mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)^{1+s} \\ &\leqslant \sum_K \mathbb{P}_\mathbb{K}(K) q^{j+js}\Big(\sum_{x_1^\ell \in K} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)^{1+s} \\ &\qquad + \sum_{x_1^\ell \in K} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)\mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)^s\Big) \\ &= q^{j+js}\sum_K \mathbb{P}_\mathbb{K}(K)\sum_{x_1^\ell \in K} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)^{1+s} \qquad \text{(diagonal arc)} \\ &+ q^{j+js}\sum_K \mathbb{P}_\mathbb{K}(K)\sum_{x_1^\ell \in K} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)\mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)^s. \text{ (off)}\end{aligned}$$

The inequality rewrites the $(s+1)$-th power of $\mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)$. Divide and conquer—the inner sum of the double-stroke E-null function is split into two arcs as labeled.

The diagonal arc is exactly

$$\begin{aligned}&q^{j+js}\sum_K \mathbb{P}_\mathbb{K}(K)\sum_{x_1^\ell \in K} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)^{1+s} \\ &= q^{j+js}\frac{1}{q^j}\sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)^{1+s} \\ &= q^{js}\sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_X(x_1^\ell)^{1+s}\mathbb{P}_{Y|X}(y_1^\ell \mid x_1^\ell)^{1+s} \\ &= q^{js-\ell s}\sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_X(x_1^\ell)\mathbb{P}_{Y|X}(y_1^\ell \mid x_1^\ell)^{1+s}.\end{aligned}$$

The off-diagonal arc is

$$\begin{aligned}&q^{j+js}\sum_K \mathbb{P}_\mathbb{K}(K)\sum_{x_1^\ell \in K} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)\mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)^s \\ &= q^{j+js}\sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)\sum_{K \ni x_1^\ell} \mathbb{P}_\mathbb{K}(K)\mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)^s.\end{aligned}$$

The inner sum is loosen to

$$\begin{aligned}&\sum_{K \ni x_1^\ell} \mathbb{P}_\mathbb{K}(K)\mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)^s \\ &= \frac{1}{q^j}\sum_{K \ni x_1^\ell} \mathbb{P}_{\mathbb{K}|X}(K \mid x_1^\ell)\mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)^s \\ &\leqslant \frac{1}{q^j}\Big(\sum_{K \ni x_1^\ell} \mathbb{P}_{\mathbb{K}|X}(K \mid x_1^\ell)\mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)\Big)^s \\ &= \frac{1}{q^j}\Big(\sum_{K \ni x_1^\ell} \mathbb{P}_{\mathbb{K}|X}(K \mid x_1^\ell)\sum_{x_1^\ell \neq \xi_1^\ell \in K} \mathbb{P}_{XY}(\xi_1^\ell, y_1^\ell)\Big)^s \\ &= \frac{1}{q^j}\Big(\frac{q^{\ell-j}-1}{q^\ell - 1}\sum_{x_1^\ell \neq \xi_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_{XY}(\xi_1^\ell, y_1^\ell)\Big)^s. \\ &\leqslant \frac{1}{q^{j+js}}\Big(\sum_{x_1^\ell \neq \xi_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_{XY}(\xi_1^\ell, y_1^\ell)\Big)^s\end{aligned}$$

The last equality counts the multiplicity of ξ_1^ℓ. So the off-diagonal arc is loosen to

$$\begin{aligned}&q^{j+js}\sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)\sum_{K \ni x_1^\ell} \mathbb{P}_\mathbb{K}(K)\mathbb{P}_{XY}(K\backslash x_1^\ell, y_1^\ell)^s \\ &\leqslant \sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)\Big(\sum_{x_1^\ell \neq \xi_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_{XY}(\xi_1^\ell, y_1^\ell)\Big)^s \\ &\leqslant \sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)\Big(\sum_{\xi_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_{XY}(\xi_1^\ell, y_1^\ell)\Big)^s \\ &= \sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_{XY}(x_1^\ell, y_1^\ell)\mathbb{P}_Y(y_1^\ell)^s \\ &= \mathbb{P}_Y(y_1^\ell)\mathbb{P}_Y(y_1^\ell)^s = \mathbb{P}_Y(y_1^\ell)^{1+s}.\end{aligned}$$

Both the diagonal and off-diagonal arcs conquered, merge them and raise to the $(1+t)$-th power. The summand for any fixed y_1^ℓ in the definition of the double-stroke E-null function is

$$\begin{aligned}&\Big(\sum_K \mathbb{P}_\mathbb{K}(K)\mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)^{\frac{1}{1+t}}\Big)^{1+t} \\ &\leqslant (\text{off-diagonal} + \text{diagonal})^{1+t} \leqslant \text{off}^{1+t} + \text{diagona}^{1+t} \\ &\leqslant \Big(\mathbb{P}_Y(y_1^\ell)^{1+s}\Big)^{1+t} + \text{diagon}^{1+t} \leqslant \mathbb{P}_Y(y_1^\ell) + \text{diago}^{1+t} \\ &= \mathbb{P}_Y(y_1^\ell) + \Big(q^{js-\ell s}\sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_X(x_1^\ell)\mathbb{P}_{Y|X}(y_1^\ell \mid x_1^\ell)^{1+s}\Big)^{1+t}\end{aligned}$$

$$= \mathbb{P}_Y(y_1^\ell) + q^{\ell t - jt}\Big(\sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_X(x_1^\ell)\mathbb{P}_{Y|X}(y_1^\ell \mid x_1^\ell)^{1+s}\Big)^{1+t}$$

We can finally bound the double-stroke E-null function per se:

$$\exp(-\mathbb{E}_0(t)) = \sum_{y_1^\ell}\Big(\sum_K \mathbb{P}_{\mathbb{K}}(K)\mathbb{P}_{Y|\mathbb{K}}(y_1^\ell \mid K)^{\frac{1}{1+t}}\Big)^{1+t}$$
$$\leqslant \sum_{y_1^\ell} \mathbb{P}_Y(y_1^\ell) + q^{\ell t - jt}\Big(\sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_X(x_1^\ell)\mathbb{P}_{Y|X}(y_1^\ell \mid x_1^\ell)^{1+s}\Big)^{1+t}$$
$$= 1 + q^{\ell t - jt}\sum_{y_1^\ell}\Big(\sum_{x_1^\ell \in \mathbb{F}_q^\ell} \mathbb{P}_X(x_1^\ell)\mathbb{P}_{Y|X}(y_1^\ell \mid x_1^\ell)^{1+s}\Big)^{1+t}$$
$$= 1 + q^{\ell t - jt}\exp(-(\textit{the E-null function of } W^\ell)(t))$$
$$= 1 + q^{\ell t - jt}\exp(-\ell E_0(t)).$$

All efforts we spent on bounding $I_e(U_1^j\ ;\ Y_1^\ell)$ are for three creeds: First, we see Gallager's bound possessing innate elegance. Second, it fits the paradigm that solving the primary (noisy channel) and the dual (wiretap channel) problems as a whole is easier than solving the primary problem alone. Third, the universal quadratic bound is waiting ahead for the E-null function. We infer that

$$\mathbb{E}I_e(U_1^j\ ; Y_1^\ell \mid \mathbb{G}) = I_e(\mathbb{K}\ ; Y_1^\ell) = \mathbb{E}_0'(0)$$
$$\leqslant \frac{1}{t}\mathbb{E}_0(t) = \frac{1}{-t}\log\Big(\exp(-\mathbb{E}_0(t))\Big)$$
$$\leqslant \frac{1}{-t}\log\Big(1 + q^{\ell t - jt}\exp(-\ell E_0(t))\Big)$$
$$< \frac{1}{-t}q^{\ell t - jt}\exp(-\ell E_0(t))$$
$$= \exp(-\log(-t) + (\ell - j)t\log q - \ell E_0(t)).$$

Recall the universal quadratic bound $E_0(t) \geqslant I(W)t\log q - t^2\log(q)^2$ as stated in Lemma 17 and used in the previous subsection. But this time $-2/5 \leqslant t < 0$. We obtain that the exponent is

$$-\log(-t) + (\ell - j)t\log q - \ell E_0(t)$$
$$= -\log(-t) + (\ell - H(W)\ell + \ell^{1/2+\alpha})t\log q - \ell E_0(t)$$
$$= -\log(-t) + (I(W)\ell + \ell^{1/2+\alpha})t\log q - \ell E_0(t)$$
$$\leqslant -\log(-t) + (I(W)\ell + \ell^{\frac{1}{2}+\alpha})t\log q - \ell(I(W)t\log q - t^2\log(q)^2)$$
$$= -\log(-t) + (\ell t\log q + \ell^{1/2+\alpha})t\log q$$

(redeem the infimum at $t = -\ell^{-1/2+\alpha}/2\log q$)

$$\mapsto -\log\Big(\frac{\ell^{-1/2+\alpha}}{2\log q}\Big) - \Big(-\frac{\ell\ell^{-1/2+\alpha}}{2} + \ell^{1/2+\alpha}\Big)\frac{\ell^{-1/2+\alpha}}{2}$$
$$= \frac{\log\ell}{2} - \alpha\log\ell + \log 2 + \log\log q - \frac{\ell^{2\alpha}}{4}$$
$$= \frac{\log\ell}{2} - \log\log\ell + \log 2 + \log\log q - \frac{\ell^{2\log(\log\ell)/\log\ell}}{4}$$
$$< \frac{\log\ell}{2} + \log\log q - \frac{\log(\ell)^2}{4}.$$

The first inequality uses $\ell - j = \ell - H(W)\ell + \ell^{1/2+\alpha}$. The last inequality uses the assumption $\ell \geqslant e^2$. With the last line we conclude that $\mathbb{E}I_e(U_1^j\ ; Y_1^\ell \mid \mathbb{G}) < \exp(\log(\ell)/2 + \log\log q - \log(\ell)^2/4) = \ell^{1/2-\log(\ell)/4}\log q$. Switch back to the base-q mutual information $\mathbb{E}I(U_1^j\ ; Y_1^\ell \mid \mathbb{G}) < \ell^{1/2-\log(\ell)/4}$.

We now reject kernels such that $I(U_1^j\ ;\ Y_1^\ell \mid \mathbb{G}) \geqslant \ell^{1/2-\log(\ell)/5}$. By Markov's inequality, the opposite direction ($<$) holds with probability $1 - \ell^{-\log(\ell)/20}$ because $1/5 + 1/20 = 1/4$. Plug this upper bound into h_α. The left-hand side of Inequality (20) is less than

$$jh_\alpha\Big(\frac{1}{j}\ell^{1/2-\log(\ell)/5}\Big) = jj^{-\alpha}\ell^{\alpha/2-\alpha\log(\ell)/5}$$
$$< \ell^{1-\alpha}\ell^{\alpha/2-\alpha\log(\ell)/5} = \ell^{1-\alpha/2-\alpha\log(\ell)/5}$$
$$= \ell\log(\ell)^{-1/2-\log(\ell)/5}.$$

The inequality uses that the left-hand side increases monotonically in j and $j \coloneqq H(W)\ell - \ell^{1/2+\alpha} < \ell$. In any regard, the quantity at the end of the inequalities decays to 0 as $\ell \to \infty$, so eventually it becomes less than $\ell^{1/2+\alpha}$, the right-hand side of Inequality (20). This proves that Inequality (15) holds with failing probability $\ell^{-\log(\ell)/20}$ as soon as ℓ is large enough. The lower bound on ℓ in the statement of Lemma 14 is large enough, hence the second half of Lemma 14 settled. That means the proof of the whole lemma has finally come to an end.

E. Bibliographic Remarks

Concerning the second moment bound: [78, Lemma 1] has a looser bound comparing to Lemma 16. A similar bound for the third moment is [18, Lemma 46], wherein Inequality (468) looks dubious. In general, Gallager's E-null function is the cumulant generating function (the logarithm of the moment generating function), and bounding E-null is equivalent to bounding higher moments. Concerning the group symmetry: On both Bob and Eve's ends, we use heavily the 2-transitive nature of $\mathrm{GL}(\ell, q)$'s action on $\mathbb{F}_q^\ell$. Interestingly enough, 2-transitivity is the main ingredient to prove that Reed–Muller codes achieve capacity over BECs [88] as well. Concerning the secrecy bound: According to Hayashi [87, Remark 4], this technique of bounding secrecy exponent via the resolvability exponent and then the E-null function dated back to Oohama's conference paper [89], although no formal proof was found there. See [90], [91] for alternative descriptions and approaches on the same topic.

For readers who took Lemma 12 as granted or went through Appendices B and C in advance, this is the last sentence of the proof of the main theorem—polar codes' simplicity, random codes' durability.

IX. CONCLUSION

Shannon introduced what we now understand as discrete memoryless channels seventy-two years ago. In the beginning, Shannon had no tool but developed their own theory of typical set, proved the noisy-channel coding theory, and justified the notion of capacity. Gallager brought in error exponents. Capacities and error exponents quantify the first and second order terms in the asymptotic performance of codes. Only in 2010 we are revealed the complete second order term. It was

TABLE II

POLAR CODING WORKS ARRANGED BY THEIR CONTRIBUTION IN TERMS OF TARGETED CHANNELS AND TARGETED BEHAVIORS. SEE SECTION IX FOR DETAILS

	BEC	BDMC	p-ary	q-ary	finite	asym.
LLN	[25]	[25]	[92]	[92]	[92]	[32]
wLDP	[26]	[26]	[92]	[77]	[31]	[33]
wCLT	[34]	[36]	[51]	Thm. 1	Thm. 1	Thm. 1
wMDP	[48]	[48]	[53]	Thm. 1	Thm. 1	Thm. 1
LDP	[27]	[27]	[77]	[77]	Thm. 1	Thm. 1
CLT	[43]	[46]	Thm. 1	Thm. 1	Thm. 1	Thm. 1
MDP	[50]	Thm. 1	Thm. 1	Thm. 1	Thm. 1	Thm. 1

around the time that polar coding as a graceful instrument to explore the limits at low cost was discovered when Arıkan experimented with the channel transformation and with error exponents. Another ten years it took to grow variants and proof techniques of polar coding. Ultimately, it is feasible, and done by us coincidentally, to piece the puzzle together to show the mere possibility to achieve the second order limits at low cost.

An overall comparison is integrated in Table II. Columns are classes of channels; from left to right: (BEC) binary erasure channels; (BDMC) binary-input discrete-output memoryless channels; (p-ary) channels of prime input size; (q-ary) channels of prime power input size; (finite) channels of discrete input. Columns to the right are wider than columns to the left. The last column is exceptional; (asym.) is about whether we can achieve the true Shannon capacity, instead of the symmetric capacity. Rows are goals; from top to bottom: (LLN) to achieve (symmetric) capacity; (wLDP) there exists $\pi > 0$ such that $P_e < \exp(-N^\pi)$; (wCLT) there exists $\rho > 0$ such that $R > I - N^{-\rho}$; (wMDP) there exist $\pi, \rho > 0$ such that $P_e < \exp(-N^\pi)$ and $R > I - N^{-\rho}$ at once; (LDP) the π in (wLDP) can be arbitrarily close to 1; (CLT) the ρ in (wCLT) can be arbitrarily close to $1/2$; (MDP) the (π, ρ)-pair can be arbitrarily close to $\pi + 2\rho = 1$. Row (MDP) implies every other row; row (CLT) implies (wCLT); row (LDP) implies (wLDP); and every other row implies (LLN). Rows (LDP) and (CLT) together almost imply (MDP) (need the partial distance profile). Cells represent how various goals are achieved over various channels. The greenish background means it is possible using Arıkan's kernel $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$. The purplish background means it is possible using other kernels. The orangish background means it is only possible using dynamic kernels.

The following works made critical progresses but our classification fails to include them: Rate-dependent result in LDP paradigm [28]. Optimal relations among channel parameters [29]. First family of (π, ρ) pairs in MDP paradigm [38]. AWGNCs intersecting MDP [39].

We did our best to excavate the archive but throughout the course of manuscript preparation we found ourselves underestimating early works multiple times so the record kept updating. We sincerely hope to hear about possible references to add to the table.

Potential improvements include but are not limited to what follow: (Tolls) Tighten the explicit Hölder tolls. The current toll between any pair of parameters H, P_e, Z, Z_{mad}, T, S, and $S_{\max}$ is roughly the total of the tolls collected when traveling through the spanning tree illustrated in Lemma 7. Any improvement on the $S_{\max}$–S–P–H path will tighten the bounds in Lemma 12. (FTPC) Tighten the two fundamental theorems such that they degenerate to equalities over erasure channels. Once done, $\{Z_n\}$ is a supermartingale and Appendix C-A is obsolete. (Symmetry) Generalize the arguments presented in Sections VIII-B to VIII-D to asymmetric channels. Once done, Section VIII-A is obsolete. Note that the proof of the fundamental theorems applies to asymmetric channels. (Bijection) Early works on polar coding over arbitrary alphabets introduced arbitrary bijections g^W. Generalize the two fundamental theorems to include arbitrary bijections. (Dynamic) Achieve the main theorem with a large, but fixed, kernel. This does not immediately make the code practical. But the answer should shed light on our understanding of coding. (Alphabet) Achieve the main theorem without the reduction to prime power alphabets. This is currently not an option because *linear codes* are barely defined over non-fields. Plus the S-parameter—and thus FTPCS—would simply break. (Dispersion) Recall Proposition 2. Weird things happens when the channel dispersion vanishes $V = 0$. Can we describe those channels better? One example of such channels is this:

$$\begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{bmatrix}.$$

(Quantization) Control the cardinality of the output alphabets of the synthetic channels [46], [64], [93]. This step is a necessity for application because real numbers are not real—rounding errors emerges in the a posteriori probabilities in Section III. We anticipate a generalization of our theory that meets [46]'s standard, namely polynomially ($N^{O(1)}$) sized output alphabet.

We look forward to generalizations of the main theorem to non-identical channels (i.e., non-stationary) [94], dependent channels (i.e., with memory) [95], [96], deletion channels [97], [98], channels with restrictions on input distributions (e.g., due to energy constrains) [99], wiretap channels [100], [101], rate-distortion problem [56], Wyner-Ziv problem [56], Slepian-Wolf problem [102], broadcast channels [103], [104], and multiple access channels [105], [106]. We focus on noisy-channel coding in this work for its historical significance.

APPENDIX A
EXPLICIT HÖLDER TOLLS (PROOF OF LEMMA 8)

As was promised in Section IV-A, we prove the explicit Hölder toll. Let W be a q-ary channel. In the upcoming arguments, H, P_e, Z, Z_{mad}, S, and $S_{\max}$ mean $H(W)$, $P_e(W)$, $Z(W)$, $Z_{\text{mad}}(W)$, $S(W)$, and $S_{\max}(W)$, respectively. Also q' means $q-1$, and q'' means $q-2$. Furthermore, lg means the base-2 logarithm; this is handy when we jump back and forth between nats, bits, and q-bits.

First we show

$$Z_{\mathrm{mad}} \leqslant q\sqrt{H \log_4 q}. \qquad \text{((2)'s copy)}$$

Start from Z_{mad}: By the definition $Z_{\mathrm{mad}} \leqslant q'Z$. Move on to Z: By Lemma 3, $q'q^{-2}(\sqrt{1+q'Z} - \sqrt{1-Z})^2 \leqslant P_{\mathrm{e}}$ so $\sqrt{1+q'Z} - \sqrt{1-Z} \leqslant q\sqrt{P_{\mathrm{e}}/q'}$. Multiplying by the conjugate yields $(1+q'Z)-(1-Z) \leqslant q\sqrt{P_{\mathrm{e}}/q'}(\sqrt{1+q'Z}+\sqrt{1-Z})$. The left-hand side is qZ; in the right-hand side $\sqrt{1+q'z}+\sqrt{1-z}$ has maximum $q/\sqrt{q'}$ at $z = q''/q'$ by calculus. So $Z \leqslant \sqrt{P_{\mathrm{e}}/q'}(q/\sqrt{q'}) = q\sqrt{P_{\mathrm{e}}}/q'$. Move on to P_{e}: By Lemma 6 (the first lower bound), $2P_{\mathrm{e}} \leqslant H \lg q$ or equivalently $P_{\mathrm{e}} \leqslant H \log_4 q$. Now we chain the inequalities $Z_{\mathrm{mad}} \leqslant q'Z \leqslant q\sqrt{P_{\mathrm{e}}} \leqslant q\sqrt{H \log_4 q}$. This completes Inequality (2). That being proven, we use a weaker form

$$Z_{\mathrm{mad}} \leqslant q^3\sqrt{H}$$

in the calculus machinery for global MDP.

Second we show

$$H \leqslant \sqrt{eq'Z_{\mathrm{mad}}/2}. \qquad \text{((3)'s copy)}$$

Start from H: By Lemma 6 (the upper bound, Fano's inequality), $H \lg q \leqslant h_2(P_{\mathrm{e}}) + P_{\mathrm{e}} \lg q'$. By Fig. 6, $h_2(P_{\mathrm{e}}) + P_{\mathrm{e}} \lg q' \leqslant \sqrt{eP_{\mathrm{e}}} + P_{\mathrm{e}} \lg q' = \sqrt{P_{\mathrm{e}}}(\sqrt{e} + \sqrt{P_{\mathrm{e}}} \lg q')$. What is inside parentheses is less than $\sqrt{e} + \sqrt{q'/q} \lg q'$. Hence $H \leqslant \sqrt{P_{\mathrm{e}}}(\sqrt{e} + \sqrt{q'/q} \lg q')/\lg q$. Focus on the scalar—$(\sqrt{e} + \sqrt{q'/q} \lg q')/\lg q$ has maximum $\sqrt{e}$ at $q = 2$ (remember that $q \geqslant 2$). So $H \leqslant \sqrt{eP_{\mathrm{e}}}$. Move on to P_{e}: By Lemma 3, $P_{\mathrm{e}} \leqslant q'Z/2$. Move on to Z: By definition $Z \leqslant Z_{\mathrm{mad}}$. Now we chain the inequalities $H \leqslant \sqrt{eP_{\mathrm{e}}} \leqslant \sqrt{eq'Z/2} \leqslant \sqrt{eq'Z_{\mathrm{mad}}/2}$. This completes Inequality (3). That being proven, we use a weaker form

$$H \leqslant q^3\sqrt{Z_{\mathrm{mad}}}$$

in the calculus machinery for global MDP.

Third we show (notice the logarithm is natural)

$$S_{\max} \leqslant q'q\sqrt{(1-H)\log(q)/2}. \qquad \text{((4)'s copy)}$$

Start from $S_{\max}$: By definition $S_{\max} \leqslant q'S$. Move on to S: By Lemma 5, $S \leqslant q'q(q'/q - P_{\mathrm{e}})\sqrt{1 - \frac{q}{q'}\frac{q''}{q'}}$. The square root simplifies to $\sqrt{1/(q')^2} = 1/q'$ as $qq'' = (q')^2 - 1$. So $S \leqslant q' - qP_{\mathrm{e}}$. Move on to $q' - qP_{\mathrm{e}}$: By Lemma 6 (the upper bound, Fano's inequality), $H \lg q \leqslant h_2(P_{\mathrm{e}}) + P_{\mathrm{e}} \lg q'$. We claim that $h_2(P_{\mathrm{e}}) + P_{\mathrm{e}} \lg q' \leqslant \lg q - 2(q'/q - P_{\mathrm{e}})^2/\log 2$. To prove the claim, Taylor expand both sides at $P_{\mathrm{e}} = q'/q$. Verify that both evaluate to $\lg q$ at $P_{\mathrm{e}} = q'/q$; verify that both have derivative 0 at $P_{\mathrm{e}} = q'/q$; and verify that the acceleration of the left-hand side, $-1/(P_{\mathrm{e}}(1-P_{\mathrm{e}})\log 2)$, is more negative than the acceleration of the right-hand side, $-4/\log 2$. By Taylor's theorem, mean value theorem, or Euler method, the function with greater acceleration is greater; hence the claim. See also [80, Fig. 1]; the Φ-curve seems parabolic at the upper right corner. Now we have $H \lg q \leqslant \lg q - 2(q'/q - P_{\mathrm{e}})^2/\log 2$, which is equivalent to $2(q'/q - P_{\mathrm{e}})^2/\log q \leqslant 1 - H$ and to $q' - qP_{\mathrm{e}} \leqslant q\sqrt{(1-H)\log(q)/2}$. Now we chain the inequalities $S_{\max} \leqslant q'S \leqslant q'(q' - qP_{\mathrm{e}}) \leqslant q'q\sqrt{(1-H)\log(q)/2}$. This completes Inequality (4). That being proven, we use a weaker form

$$S_{\max} \leqslant q^3\sqrt{1-H}$$

in the calculus machinery for global MDP.

Fourth we show

$$1 - H \leqslant q'S_{\max}/\log q. \qquad \text{((5)'s copy)}$$

Start from $1-H$: By Lemma 6 (the second lower bound), $H \lg q \geqslant q'q \lg(q/q')(P_{\mathrm{e}} - q''/q') + \lg q'$. The right-hand side is $\lg q - q' \lg(q/q')(q' - qP_{\mathrm{e}})$ by matching the (rational) coefficients of $P_{\mathrm{e}} \lg q$, $P_{\mathrm{e}} \lg q'$, $\lg q$, and $\lg q'$, respectively. As $H \lg q \geqslant \lg q - q' \lg(q/q')(q' - qP_{\mathrm{e}})$ we bound $\lg(q/q') = \lg(1 + 1/q') \leqslant 1/q'$ by the tangent line at $1/q' = 0$. So $H \lg q \geqslant \lg q - (q' - qP_{\mathrm{e}})$ and hence $1 - H \leqslant (q' - qP_{\mathrm{e}})/\lg q$. Move on to $q' - qP_{\mathrm{e}}$: By Lemma 5, $1 - qP_{\mathrm{e}}/q' \leqslant S$ so $q' - qP_{\mathrm{e}} \leqslant q'S$. Move on to S: By definition $S \leqslant S_{\max}$. Now we chain the inequalities $1 - H \leqslant (q' - qP_{\mathrm{e}})/\lg q \leqslant q'S/\lg q \leqslant q'S_{\max}/\lg q$. This completes Inequality (5). That being proven, we use a weaker form

$$1 - H \leqslant q^3\sqrt{S_{\max}}$$

in the calculus machinery for global MDP.

This is end of the proof of Lemma 8. The proof of Lemma 7 follows the same logic, only shorter.

Appendix B
Calculus Machinery for Global MDP (Proof of Lemma 12)

We are to prove that

$$\boldsymbol{P}\{\boldsymbol{H}_n < \exp(-\ell^{\pi n} n)\} > 1 - \boldsymbol{H}_0 - \ell^{-\rho n + o(n)} \quad \text{((7)'s copy)}$$

given premises (pb), (pm), (pt), and (pl), the local LDP behavior, the local CLT behavior, and that $\pi + 2\rho \leqslant 1 - 8\alpha$. The proof is split into several stepping stones. We will prove each of the following inequalities (including two equalities) in each of the upcoming subsections. This will be proven in Appendix B-A: The *eigen behavior* reads

$$\boldsymbol{E}[h_\alpha(\boldsymbol{H}_{n+1}) \mid \mathcal{F}_n] \leqslant 4\ell^{-1/2+3\alpha} h_\alpha(\boldsymbol{H}_n). \qquad (23)$$

This will be proven in Appendix B-B: As a lemma, $\{\boldsymbol{H}_n\}$ and $\{\boldsymbol{Z}_n\}$ converges to 0 with probability $1 - \boldsymbol{H}_0$, i.e.,

$$\boldsymbol{P}\{\boldsymbol{Z}_n \to 0\} = \boldsymbol{P}\{\boldsymbol{H}_n \to 0\} = 1 - \boldsymbol{H}_0. \qquad (24)$$

This will be proven in Appendix B-C: The *en23 behavior* reads

$$\boldsymbol{P}\{\boldsymbol{Z}_n < \exp(-n^{2/3})\} > 1 - \boldsymbol{H}_0 - \ell^{(-1/2+4\alpha)n + o(n)}. \qquad (25)$$

This will be proven in Appendix C-A: As a lemma, process $\{\min(\ell^{-2}, \sqrt[4]{\boldsymbol{Z}_n})\}$ is a supermartingale, i.e.,

$$\boldsymbol{E}[\min(\ell^{-2}, \sqrt[4]{\boldsymbol{Z}_{n+1}}) \mid \mathcal{F}_n] \leqslant \min(\ell^{-2}, \sqrt[4]{\boldsymbol{Z}_n}). \qquad (26)$$

This will be proven in Appendix C-B: As a lemma, the following hold when $\boldsymbol{Z}_n < \ell^{-8}$:

$$\boldsymbol{Z}_{n+1} \leqslant \boldsymbol{Z}_n^{\lceil \boldsymbol{K}_{n+1}^2/3\ell \rceil \cdot 3/4}, \qquad (27)$$

$$\boldsymbol{E}[(\lceil \boldsymbol{K}_{n+1}^2/3\ell \rceil \cdot 3/4)^{-1/2} \mid \mathcal{F}_n] < \ell^{-1/2+2\alpha}. \qquad (28)$$

This will be proven in Appendix C-C: The *een13 behavior* reads

$$P\{Z_n < \exp(-e^{n^{1/3}})\} > 1 - H_0 - \ell^{(-1/2+4\alpha)n+o(n)}. \quad (29)$$

This will be proven in Appendix C-D: The *elpin behavior* reads for any constants $\pi, \rho > 0$ such that $\pi + \rho \leqslant 1 - 8\alpha$,

$$P\{Z_n < \exp(-\ell^{\pi n} n^2)\} > 1 - H_0 - \ell^{-\rho n+o(n)}. \quad (30)$$

The last inequality is a bi-Hölder toll away from

$$P\{H_n < \exp(-\ell^{\pi n} n)\} > 1 - H_0 - \ell^{-\rho n+o(n)}, \ \text{((7)'s copy)}$$

our destination. This finishes the proof of Lemma 12.

The eigen, en23, een13, and elpin behaviors are intermediate checkpoints pinned in a way that moving from one to the next is easy while skipping any of them makes the next unreachable. Their entire purpose is to form a chain that connects the local LDP and CLT behaviors to the global MDP behavior and we do not specify if any of them falls inside the LDP, CLT, or MDP paradigm.

A. The Eigen Behavior

We want to prove Inequality (23), $E[h_\alpha(H_{n+1}) \mid \mathcal{F}_n] \leqslant 4\ell^{-1/2+3\alpha} h_\alpha(H_n)$, given the local LDP behavior and the local CLT behavior. The idea is that, for H_n that is close to $1/2$, the local CLT behavior provides a measurement of the dichotomy/bifurcation behavior of H_{n+1}. For H_n that are close to 0, the Z-part of the local LDP behavior provides a measurement of the attraction toward 0. For H_n that is close to 1, the S-part handles it dually. The formal proof is below.

Inequality (23) is a local statement so we may assume $n = 0$. To prove that $E[h_\alpha(H_1)] \leqslant 4\ell^{-1/2+3\alpha} h_\alpha(H_0)$, we divide it into three cases per how H_0 compares to ℓ^{-2} and $1 - \ell^{-2}$. The mediocre case: if $\ell^{-2} \leqslant H_0 \leqslant 1 - \ell^{-2}$ then $h_\alpha(H_0) \geqslant \ell^{-2\alpha}$. The local CLT behavior implies $E[h_\alpha(H_1)] < 4\ell^{-1/2+\alpha} = 4\ell^{-1/2+3\alpha}\ell^{-2\alpha} \leqslant 4\ell^{-1/2+3\alpha} h_\alpha(H_0)$ and we are done with this case. The noisy case: if $H_0 > 1 - \ell^{-2}$, we replace (H, Z, S) by $(1 - H, S, Z)$ to dual it to the reliable case dealt below and we are done with this case. (This is the only place in the proof where we ever mentioned S explicitly. Nevertheless, every statement concerning Z concerns S by duality.)

The last case—the reliable case: when $H_0 < \ell^{-2}$, we further split it into two subcases per how K_1 compares to $k := \ell^{1/2+5\alpha/2}$. For the small K_1 subcase, the martingale property fits; for the large K_1 subcase, the local LDP behavior fits:

$$\begin{aligned}
&E[h_\alpha(H_1)] \\
&= E[h_\alpha(H_1) \mid K_1 \leqslant k]\frac{k}{\ell} + E[h_\alpha(H_1) \mid K_1 > k]\frac{\ell - k}{\ell} \\
&\leqslant h_\alpha(E[H_1 \mid K_1 \leqslant k])\frac{k}{\ell} + h_\alpha(E[H_1 \mid K_1 > k])\frac{\ell - k}{\ell}. \quad (31)
\end{aligned}$$

For the $K_1 \leqslant k := \ell^{1/2+5\alpha/2}$ subcase, the martingale property $E[H_1] = H_0$ implies $E[H_1 \mid K_1 \leqslant k] \leqslant H_0\ell/k$. Therefore, $h_\alpha(E[H_1 \mid K_1 \leqslant k])k/\ell \leqslant h_\alpha(H_0\ell/k)k/\ell = h_\alpha(H_0)\ell^\alpha k^{-\alpha} k \ell^{-1} = h_\alpha(H_0)\ell^\alpha \ell^{-\alpha/2-5\alpha^2/2}\ell^{1/2+5\alpha/2}\ell^{-1} \leqslant \ell^{-1/2+3\alpha} h_\alpha(H_0)$. And the $K_1 \leqslant k$ subcase is closed.

For the $K_1 > k := \ell^{1/2+5\alpha/2}$ subcase, pay the explicit Hölder toll: $Z_0 \leqslant q^3\sqrt{H_0} < q^3/\ell < 1$. Invoke the local LDP behavior:

$$\begin{aligned}
E[Z_1 \mid K_1 > k] &\leqslant E[\ell \exp(qZ_0\ell)(qZ_0)^{\lceil K_1^2/3\ell \rceil} \mid K_1 > k] \\
&\leqslant \ell \exp(qZ_0\ell)(qZ_0)^{k^2/3\ell} \leqslant \ell \exp(q^4)(q^4\sqrt{H_0})^{k^2/3\ell} \\
&= \ell \exp(q^4)(q^8 H_0)^{\log(\ell)^5/6}.
\end{aligned}$$

Pay the explicit Hölder toll for the return-trip: $H_1 \leqslant q^3\sqrt{Z_1} \leqslant q^3\ell^{1/2}\exp(q^4/2)(q^8H_0)^{\log(\ell)^5/12}$. Now we claim and prove that the following quantity is less than 1: (there is nothing to show if $h_\alpha(H_0) = 0$)

$$\begin{aligned}
&(h_\alpha(H_1)/\ell^{-1/2+3\alpha}h_\alpha(H_0))^{12/\alpha} \\
&\quad = H_1^{12}\ell^{6/\alpha-36}H_0^{-12} < H_1^{12}\ell^{6\log\ell-30}H_0^{-12} \\
&\quad \leqslant q^{36}\ell^6\exp(6q^4)(q^8H_0)^{\log(\ell)^5}\ell^{6\log\ell-30}H_0^{-12} \\
&\quad = q^{36+8\log(\ell)^5}e^{6q^4}\ell^{6\log\ell-24}H_0^{\log(\ell)^5-12} \\
&\quad < q^{36+8\log(\ell)^5}e^{6q^4}\ell^{6\log\ell-24}\ell^{-2\log(\ell)^5+24} \\
&\quad = q^{36+8\log(\ell)^5}e^{6q^4}\ell^{6\log\ell-1.6\log(\ell)^5-0.4\log(\ell)^5} \\
&\quad \leqslant q^{36+8\log(\ell)^5}e^{6q^4}\ell^{6\log\ell-8\log(q)\log(\ell)^4-0.4\log(\ell)^5} \\
&\quad = q^{36}e^{6q^4}\ell^{6\log\ell-0.4\log(\ell)^5} \\
&\quad = q^{36}e^{6q^4}\ell^{6\log\ell-0.1\log(\ell)^5}e^{-0.3\log(\ell)^6} \\
&\quad < q^{36}e^{6q^4}\ell^{6\log\ell-0.1\log(\ell)^5}e^{-0.3\log(41)^2(q\log 3)^4} \\
&\quad < q^{36}e^{6q^4}\ell^{6\log\ell-0.1\log(\ell)^5}e^{-6.02q^4} \\
&\quad < q^{36}\ell^{6\log\ell-0.1\log(\ell)^5} \\
&\quad = e^{36\log q+6\log(\ell)^2-\log(\ell)^6/30-\log(\ell)^6/15} \\
&\quad \leqslant e^{36\log q+6\log(\ell)^2-5\log(q)\log(22)^5/30-\log(22)^4\log(\ell)^2/15} \\
&\quad < e^0 \leqslant 1.
\end{aligned}$$

The inequality involving 1.6 uses $\ell \geqslant q^5$. The inequality involving 0.3 uses $\ell \geqslant \max(41, 3^q)$. The inequality involving 15 uses $\ell \geqslant \max(22, q^5)$. We have just showed that the quotient $h_\alpha(H_1)/\ell^{-1/2+3\alpha}h_\alpha(H_0)$ is less than 1, with and hence without the power of $12/\alpha$. Therefore, $E[h_\alpha(H_1) \mid K_1 > k] \leqslant \ell^{-1/2+3\alpha}h_\alpha(H_0)$. And the $K_1 > k := \ell^{1/2+5\alpha/2}$ subcase is closed.

To sum up the reliable case: We bound separately the two terms in Formula (31) by considering two subcases. Both of them are at most $\ell^{-1/2+3\alpha}h_\alpha(H_1)$, hence their sum is at most $2\ell^{-1/2+3\alpha}h_\alpha(H_1)$. Since Inequality (23) allows 4, let alone 2, the reliable case is closed. And the proof of the eigen behavior, Inequality (23), is sound when combining the three cases.

Bibliographic remarks: [43, Theorem 7] also cut the cases at ℓ^{-2} and $1 - \ell^{-2}$. In contrast, [46, Theorem 5.1] cut at ℓ^{-4} and $1 - \ell^{-4} + \varepsilon$. A potential improvement is, when $\ell^{-2} \leqslant H_0 < \ell^{-1}$, Inequality (15) will simply evaporate. Similarly, Inequality (14) evaporates when $1 - \ell^{-1} < H_0 \leqslant 1 - \ell^{-2}$. They tighten the right-hand side of Inequality (10). The lesson here is that the hard transition between local LDP and CLT behaviors weakens the bounds.

B. Polarization in Mean

We want to prove Equality (24), namely $P\{Z_0 \to 0\} = P\{H_0 \to 0\} = 1 - H_0$, given the martingale property and the eigen behavior. The idea is that the eigen behavior expels H_n from being close to $1/2$, so the only reasonable limits are 0 and 1. The formal proof is below.

As a bounded martingale $\{H_n\}$ converges to an r.v.—which we call H_∞—a.s. (almost surely). This is Doob's martingale convergence theorem [107, Theorem 4.2.11]. Owing to h_α's continuity, $h_\alpha(H_n) \to h_\alpha(H_\infty)$ a.s. Point-wise convergence and (uniform) boundedness imply convergence in L^1, i.e., $E[h_\alpha(H_n)] \to E[h_\alpha(H_\infty)]$ as $n \to \infty$. This is Lebesgue's dominated convergence theorem [107, Theorem 1.6.7]. By the eigen behavior, $E[h_\alpha(H_n)]$ decays toward 0 by a constant factor every time n increases, thus $E[h_\alpha(H_\infty)]$ is 0. This forces $h_a(H_\infty) = 0$ a.s. and hence $H_\infty \in \{0, 1\}$ a.s. Since H_∞ is Bernoulli $P\{H_\infty = 0\} = E[\mathbb{I}\{H_\infty = 0\}] = E[1 - H_\infty] \leftarrow E[1 - H_n] = 1 - H_0$. So $P\{H_n \to 0\} = 1 - H_0$. By the implicit bi-Hölder toll, $H_n \to 0$ if and only if $Z_n \to 0$, thus the latter has the same probability measure. And the proof of Equality (24) is sound.

Bibliographic remarks: The statement $H_n \to H_\infty \in \{0, 1\}$ is usually referred to as *channel polarization* in spite of that it does not guarantee the corresponding codes to be capacity-achieving. See also [25, Proposition 10], [29, Definition 3], [31, Lemma 3.8]. This lemma should have been bestowed upon the (zeroth) fundamental theorem. But it is not mandatory if some sort of CLT behavior is present; see [38, Lemma 1], [43, Lemma 4], [46, Lemma 9.5]. Recently, Reed–Muller codes' channels are shown to polarize [108]; Reed–Muller codes achieving capacity is not a consequence, but a different story.

C. The en23 Behavior

We want to prove $P\{Z_n < \exp(-n^{2/3})\} < 1 - H_0 - \ell^{(-1/2+4\alpha)n+o(n)}$, namely Inequality (25), given the eigen behavior and the polarization in mean. The idea is to read off the behavior of $\{H_n\}$ from the behavior of $\{h_\alpha(H_n)\}$ in the eigen behavior. The formal proof is below.

$E[h_\alpha(H_{n+1}) \mid \mathcal{F}_n] \leqslant \ell^{-1/2+4\alpha} h_\alpha(H_n)$ by $\ell \geqslant e^4$ and the eigen behavior. This simplifies the eigenvalue. Without loss of generality, we rescale h_α such that $h_\alpha(H_0) = 1$. Let ε_n be $\exp(-n^{3/4})$; note that $\varepsilon_n \leqslant H_0 \leqslant 1 - \varepsilon_n$ for n large enough. Owing to h_α's concavity, that $h_\alpha(0) = h(1) = 0$, and that $h_\alpha(H_0) = 1$, we deduce that $h_\alpha(z) \geqslant \varepsilon_n$ whenever $\varepsilon_n \leqslant z \leqslant 1 - \varepsilon_n$. Consider these three events as a partition: let A_n be $\{H_n < \varepsilon_n\}$; let B_n be $\{\varepsilon_n \leqslant H_n \leqslant 1 - \varepsilon_n\}$; let C_n be $\{1 - \varepsilon_n < H_n\}$. Note that B_n implies $h_\alpha(H_n) \geqslant \varepsilon_n$.

Next we want to show that $P(B_n) < \ell^{(-1/2+4\alpha)n+o(n)}$: Telescoping leads to $\mathbb{E}[h_\alpha(H_n)] \leqslant h_a(H_0)\ell^{(-1/2+4\alpha)n} = \ell^{(-1/2+4\alpha)n}$. Markov's inequality leads to $P\{h(H_n) \geqslant \varepsilon_n\} \leqslant \mathbb{E}[h(H_n)]/\varepsilon_n \leqslant \ell^{(-1/2+4\alpha)n}/\varepsilon_n = \ell^{(-1/2+4\alpha)n+O(n^{3/4})} < \ell^{(-1/2+4\alpha)n+o(n)}$. Therefore $P(B_n) \leqslant P\{h(H_n) \geqslant \varepsilon_n\} < \ell^{(-1/2+4\alpha)n+o(n)}$, as desired. Moreover, summing the geometric series leads to $\sum_{m \geqslant n} P(B_m) < \ell^{(-1/2+4\alpha)n+o(n)}$.

Next we show $1 - H_0 - P(A_n) < \ell^{(-1/2+4\alpha)n+o(n)}$: The left-hand side is at most the probability measure of $\{H_\infty = 0\} \setminus A_n$. That is the probability that H_n was not small (not in A_n) but $H_{n+1}, H_{n+1}, \ldots$ will end up converging to 0. Note that being a martingale causes $1 - H_{n+1} \leqslant \ell(1 - H_n)$, which forbids H_n jumping from C_n directly into A_{n+1}—it must pass by B_m for some $m \geqslant n$ before ever landing in A_{m+1}. From the summation of $P(B_m)$ over $m \geqslant n$ we know that very few descendants of H_n can do that; the probability measure of $\{H_\infty = 0\} \setminus A_n$ is less than $\ell^{(-1/2+4\alpha)n+o(n)}$. Therefore $1 - H_0 - P(A_n) < \ell^{(-1/2+4\alpha)n+o(n)}$ and $P\{H_n < \exp(-n^{3/4})\} = P(A_n) > 1 - H_0 - \ell^{(-1/2+4\alpha)n+o(n)}$. Pay the implicit bi-Hölder toll $P\{Z_n < \exp(-n^{2/3})\} > 1 - H_0 - \ell^{(-1/2+4\alpha)n+o(n)}$. And the proof of the en23 behavior, Inequality (25), is sound.

Bibliographic remarks: The functionality of this step is translating the eigen behavior into an estimate of this general form $P\{H_n < \text{threshold}\} > \text{limit} - \text{decay}$. In this vein are [38, Theorem 1], [43, Lemma 4], [53, Theorem 2.5], and [46, Lemma 9.5]. The proof presented here is not the shortest one, but it applies to processes that are not martingale, e.g., $\{Z_n\}$. This is the case if we are given the eigen behavior of $\{Z_n\}$.

Appendix C
The een13 and Elpin Behaviors

In this section, we continue proving Lemma 12. The previous section covers (23) to (25). We are left with (26) to (30).

A. A Supermartingale

We want to show that a certain monotonic function in Z_n is a supermartingale so we can control how frequently does Z_n stay in the turf where the local LDP behavior dominates. Making it a supermartingale, we are able to cite Ville's inequality [109] (or Doob's optional stopping theorem [107, Theorem 4.8.4]) later. The formal proof is below.

Inequality (26) is a local statement so we may assume $n = 0$. To prove that $E[\min(\ell^{-2}, \sqrt[4]{Z_1})] \leqslant \sqrt[4]{Z_0}$, we may assume $Z_0 < \ell^{-8}$ or the inequality becomes trivial. Invoke the local LDP behavior $Z_1 \leqslant \ell \exp(qZ_0\ell)(qZ_0)^{\lceil K_1^2/3\ell \rceil} \leqslant \ell e(qZ_0)^{\lceil K_1^2/3\ell \rceil}$. The last inequality uses $q \leqslant \ell$. When $K_1 \leqslant \sqrt{3\ell}$, we do nothing but apply the last-resort exponent 1:

$$E[\min(\ell^{-2}, \sqrt[4]{Z_1}) \mid K_1 \leqslant \sqrt{3\ell}] \leqslant \sqrt[4]{\ell eqZ_0}.$$

When $K_1 > \sqrt{3\ell}$, a stronger exponent applies:

$$\begin{aligned} &E[\min(\ell^{-2}, \sqrt[4]{Z_1}) \mid K_1 > \sqrt{3\ell}] \\ &\quad \leqslant E\Big[\sqrt[4]{\ell e(qZ_0)^{\lceil K_1^2/3\ell \rceil}} \mid K_1 > \sqrt{3\ell}\Big] \\ &\quad \leqslant \sqrt[4]{\ell e(qZ_0)^2} = \sqrt[4]{\ell eq^2 Z_0 Z_0} \\ &\quad \leqslant \sqrt[4]{\ell eq^2 Z_0/\ell^8} \leqslant \sqrt[4]{eq^2 Z_0/\ell^7}. \end{aligned}$$

Combining the two cases that are cut per how K_1 compares to $\sqrt{3\ell}$, we infer that

$$\begin{aligned} &E[\min(\ell^{-2}, \sqrt[4]{Z_1})] \\ &= E[\sqrt[4]{Z_1} \mid K_1 \leqslant \sqrt{3\ell}]\frac{\sqrt{3\ell}}{\ell} + E[\sqrt[4]{Z_1} \mid K_1 > \sqrt{3\ell}]\frac{\ell - \sqrt{3\ell}}{\ell} \\ &\leqslant \sqrt[4]{\ell eqZ_0} \cdot \frac{\sqrt{3\ell}}{\ell} + \sqrt[4]{eq^2 Z_0/\ell^3} \cdot \frac{\ell}{\ell} \\ &= \Big(\sqrt[4]{9eq/\ell} + \sqrt[4]{eq^2/\ell^7}\Big)\sqrt[4]{Z_0} \leqslant \sqrt[4]{Z_0}. \end{aligned}$$

The last inequality uses $\ell \geqslant \max(50, q^5)$. And the proof of Inequality (26) is sound.

Bibliographic remarks: This lemma is inspired by [25, Proposition 9]. In [36, Lemma 22] Arıkan's lemma is overlooked and another is reinvented that serves the same purpose. The latter lemma also served in [38, Theorem 3]. We generalized the idea to non-binary cases in [50, Lemma 1]. The quartic root here is an aesthetic choice; $\min(\ell^{-2}, \sqrt[2+\varepsilon]{Z_n})$ is also a supermartingale but only for astronomic ℓ (depending on q). For any non-random kernel, a small enough power works provided that the kernel polarizes channels in the first place.

B. A Cramér–Chernoff Gadget

Let D_{n+1} be $\lceil K_{n+1}^2/3\ell\rceil \cdot 3/4$. We want to prove Inequalities (27) and (28), that $Z_n < \ell^{-8}$ implies $Z_{n+1} \leqslant Z_n^{D_{n+1}}$ and $E[D_{n+1}^{-1/2} \mid \mathcal{F}_n] < \ell^{-1/2+2\alpha}$, given the local LDP behavior. The motivation is to reformat the local inequalities so that it is easy to telescope for future reference. The formal proof is below.

Both of them are local statements, hence we may assume $n = 0$. When $Z_0 < \ell^{-8}$, invoke the local LDP behavior

$$\begin{aligned} Z_1 &\leqslant \ell \exp(qZ_0\ell)(qZ_0)^{\lceil K_1^2/3\ell\rceil} \leqslant \ell e (qZ_0)^{\lceil K_1^2/3\ell\rceil} \\ &\leqslant \ell e (q^4 Z_0)^{\lceil K_1^2/3\ell\rceil/4} Z_0^{\lceil K_1^2/3\ell\rceil\cdot 3/4} \\ &\leqslant \ell e (\ell^{-7})^{\lceil K_1^2/3\ell\rceil/4} Z_0^{\lceil K_1^2/3\ell\rceil\cdot 3/4} \leqslant Z_0^{\lceil K_1^2/3\ell\rceil\cdot 3/4}. \end{aligned}$$

The fourth inequality uses $\ell \geqslant q^4$. That validates Inequality (27). For the second inequality,

$$\begin{aligned} E[D_1^{-1/2}] &= \frac{1}{\ell}\sum_{k=1}^{\ell}\Big(\Big\lceil\frac{k^2}{3\ell}\Big\rceil\cdot\frac{3}{4}\Big)^{-1/2} \\ &< \frac{1}{\ell}\sum_{k=1}^{\sqrt{3\ell}}\Big(\frac{3}{4}\Big)^{-1/2} + \frac{1}{\ell}\sum_{k=\sqrt{3\ell}+1}^{\ell}\Big(\frac{k^2}{4\ell}\Big)^{-1/2} \\ &< \frac{1}{\ell}\sqrt{3\ell}\frac{2}{\sqrt{3}} + \frac{1}{\ell}\sqrt{4\ell}\int_{\sqrt{3\ell}}^{\ell}\frac{dk}{k} \\ &= 2\ell^{-1/2} + 2\ell^{-1/2}\log k\Big|_{\sqrt{3\ell}}^{\ell} < 2\ell^{-1/2} + 2\ell^{-1/2}\log\ell \\ &= 2\ell^{-1/2} + 2\ell^{-1/2+\alpha} < 4\ell^{-1/2+\alpha} < \ell^{-1/2+2\alpha}. \end{aligned}$$

The last inequality uses $\ell \geqslant e^4$. This validates Inequality (28). And the proof of Inequalities (27) and (28) is sound.

C. The een13 Behavior

We want to prove $P\{Z_n < \exp(-e^{n^{1/3}})\} > 1 - H_0 - \ell^{(-1/2+4\alpha)n+o(n)}$, namely Inequality (29), given the en23 behavior, the supermartingale property, and the Cramér–Chernoff gadget. The idea is to apply the gadget consecutively to show that Z_n becomes smaller and smaller as n increases. To reach the goal $\exp(-e^{n^{1/3}})$, we apply $\sqrt{n}$ times to avoid losing too much code rate.

(Define events.) Let E_0^0 be the empty event. For every $m = \sqrt{n}, 2\sqrt{n}, \ldots, n - \sqrt{n}$, we hereby define five series of events A_m, B_m, C_m, E_m, and E_0^m inductively as below: Let A_m be $\{Z_m < \exp(-m^{2/3})\} \setminus E_0^{m-\sqrt{n}}$. Let B_m be a subevent of A_m where $Z_k \geqslant \ell^{-8}$ for some $k \geqslant m$. Let C_m a subevent of A_m where

$$D_{m+1}D_{m+2}\cdots D_{m+\sqrt{n}} \leqslant \ell^{2\alpha\sqrt{n}}. \tag{32}$$

Let E_m be $A_m \setminus (B_m \cup C_m)$. Let E_0^m be $E_0^{m-\sqrt{n}} \cup E_m$. Let a_m, b_m, c_m, e_m, and e_0^m be the probability measures of the corresponding capital letter events. Moreover, let g_m be $1 - H_0 - e_0^m$.

(Bound b_m/a_m from above.) Conditioning on A_m, we want to estimate the probability that $Z_k \geqslant \ell^{-8}$ for some $k \geqslant m$, which is equal to the probability that $\min(\ell^{-2}, \sqrt[4]{Z_k}) \geqslant \ell^{-2}$ for some $k \geqslant m$. Recall that $\min(\ell^{-2}, \sqrt[4]{Z_k})$ was made a supermartingale. Hence by Ville's inequality (see [109] or [107, Exercise 4.8.2]), $P\{\min(\ell^{-2}, \sqrt[4]{Z_k}) \geqslant \ell^{-2}\text{for some } k \geqslant m \mid A_m\} \leqslant \min(\ell^{-2}, \sqrt[4]{Z_m})\ell^2 < \exp(-m^{2/3}/4)\ell^2$. This is an upper bound on b_m/a_m and will be summoned in Formula (33).

(Bound c_m/a_m from above.) We want to estimate how often does Inequality (32) happen. That would be the probability of $(D_{m+1}D_{m+2}\cdots D_{m+\sqrt{n}})^{-1/2} \geqslant \ell^{-\alpha\sqrt{n}}$. This probability must not exceed $E[(D_{m+1}D_{m+2}\cdots D_{m+\sqrt{n}})^{-1/2}]\ell^{\alpha\sqrt{n}} = E[D_1^{-1/2}]^{\sqrt{n}}\ell^{\alpha\sqrt{n}} = (E[D_1^{-1/2}]\ell^{\alpha})^{\sqrt{n}} \leqslant \ell^{(-1/2+3\alpha)\sqrt{n}}$ by Markov's inequality. This is an upper bound on c_m/a_m and will be summoned in Formula (33).

(Bound $(g_{m-\sqrt{n}} - a_m)^+$ from above.) By the definition, $g_{m-\sqrt{n}} - a_m = 1 - H_0 - (e_0^{m-\sqrt{n}} + a_m)$. The definition of A_m forces it to be disjoint from $E_0^{m-\sqrt{n}}$, therefore $e_0^{m-\sqrt{n}} + a_m$ is the probability measure of $E_0^{m-\sqrt{n}} \cup A_m$. This union event must contain the event $\{Z_m < \exp(-m^{2/3})\}$ by how A_m was defined. From the en23 behavior $P\{Z_m < \exp(-m^{2/3})\} > 1 - H_0 - \ell^{(-1/2+4\alpha)m}$. Chaining all inequalities together, we deduce that $g_{m-\sqrt{n}} - a_m < \ell^{(-1/2+4\alpha)m+o(m)}$. Let $(g_{m-\sqrt{n}} - a_m)^+$ be $\max(0, g_{m-\sqrt{n}} - a_m)$ so we can write $(g_{m-\sqrt{n}} - a_m)^+ < \ell^{(-1/2+4\alpha)m+o(m)}$. This upper bound will be summoned in Formula (34).

(Bound e_0^n from below.) We start rewriting g_m with g_m^+ being $\max(0, g_m)$:

$$\begin{aligned} g_m &= 1 - H_0 - e_0^m = 1 - H_0 - (e_0^{m-\sqrt{n}} + e_m) \\ &= 1 - H_0 - e_0^{m-\sqrt{n}} - e_m = g_{m-\sqrt{n}} - e_m \\ &= g_{m-\sqrt{n}}\Big(1 - \frac{e_m}{a_m}\Big) + \frac{e_m}{a_m}(g_{m-\sqrt{n}} - a_m) \\ &\leqslant g_{m-\sqrt{n}}^+\Big(1 - \frac{e_m}{a_m}\Big) + \frac{e_m}{a_m}(g_{m-\sqrt{n}} - a_m)^+ \\ &\leqslant g_{m-\sqrt{n}}^+\Big(1 - \frac{e_m}{a_m}\Big) + (g_{m-\sqrt{n}} - a_m)^+ \\ &\leqslant g_{m-\sqrt{n}}^+\Big(\frac{b_m}{a_m} + \frac{c_m}{a_m}\Big) + (g_{m-\sqrt{n}} - a_m)^+ \\ &< g_{m-\sqrt{n}}^+\Big(\exp(-m^{2/3}/4)\ell^2 + \ell^{(-1/2+3\alpha)\sqrt{n}}\Big) \qquad (33) \\ &\qquad + \ell^{(-1/2+4\alpha)m+o(m)} \qquad (34) \end{aligned}$$

The first four equalities are by the definitions of g_m and E_0^m. The next equality is simple algebra. The next two inequalities are by $0 \leqslant e_m/a_m \leqslant 1$. The next inequality is by the definition of E_m. The last inequality summons upper bounds derived in the last few paragraphs. The last line contains two terms in the big parentheses. Between them $\ell^{(-1/2+3\alpha)\sqrt{n}}$ dominates $\exp(-m^{2/3}/4)\ell^2$ once m is greater than $O(n^{3/4})$.

Subsequently, we obtain this recurrence relation

$$\begin{cases} g_{O(n^{3/4})} \leqslant 1; \\ g_m \leqslant 2g^+_{m-\sqrt{n}}\ell^{(-1/2+4\alpha)\sqrt{n}} + \ell^{(-1/2+4\alpha)m+o(m)}. \end{cases}$$

Solve it (cf. the master theorem); we get that $g_{n-\sqrt{n}} < \ell^{(-1/2+4\alpha)n+o(n)}$. By the relation between $e_{n-\sqrt{n}}$ and $g_{n-\sqrt{n}}$, we immediately get $e_0^{n-\sqrt{n}} > 1 - H_0 - \ell^{(-1/2+4\alpha)n+o(n)}$.

(Analyze $E_0^{n-\sqrt{n}}$.) We want to estimate H_n when $E_0^{n-\sqrt{n}}$ happens. More precisely, we attempt to bound $Z_{m+\sqrt{n}}$ when E_m happens for each $m = \sqrt{n}, 2\sqrt{n}, \ldots, n-\sqrt{n}$. When E_m happens, its superevent A_m happens, so we know that $Z_m < \exp(-m^{2/3})$. But B_m does not happen, so $Z_k < \ell^{-8}$ for all $k \geqslant m$. This implies that $Z_{k+1} \leqslant Z_k^{D_{k+1}}$ for those k. Telescope; $Z_{m+\sqrt{n}}$ is less than Z_m raised to the power of $D_{m+1}D_{m+2}\cdots D_{m+\sqrt{n}}$. But C_m does not happen, so the product is greater than $\ell^{2\alpha\sqrt{n}}$. Jointly we have $Z_{m+\sqrt{n}} \leqslant Z_m^{\ell^{2\alpha\sqrt{n}}} < \exp(-m^{2/3}\ell^{2\alpha\sqrt{n}})$. Recall that $Z_{k+1} \leqslant \ell eq Z_k$ for all $k \geqslant m+\sqrt{n}$ so long as Z_k stays below ℓ^{-8}, which it does because B_m is excluded. Then telescope again; $Z_n \leqslant (\ell eq)^{n-m-\sqrt{n}} Z_{m+\sqrt{n}} < (\ell eq)^n \exp(-m^{2/3}\ell^{2\alpha\sqrt{n}}) < \exp(-e^{n^{1/3}})$ provided that n is sufficiently large. In other words, $E_0^{n-\sqrt{n}}$ implies $Z_n < \exp(-e^{n^{1/3}})$.

(Summary.) Now we conclude $P\{Z_n < \exp(-e^{n^{1/3}})\} \geqslant P(E_0^{n-\sqrt{n}}) = e_0^n > 1 - H_0 - \ell^{(-1/2+4\alpha)n+o(n)}$. And hence the proof of the een13 behavior, Inequality (29), is sound.

This subsection is parallel to [50, Section V] and also to [46, Section 10.2]. Do not confuse this subsection with the next. The subtlety is explained in [50, Section III].

D. The Elpin Behavior

Recall $\pi, \rho > 0$ is such that $\pi + 2\rho \leqslant 1 - 8\alpha$. We want to prove $P\{Z_n < \exp(-\ell^{\pi n}n^2)\} > 1 - H_0 - \ell^{-\rho n+o(n)}$, namely Inequality (30), given the een13 behavior, the supermartingale property, and the Cramér–Chernoff gadget. The idea is to apply the gadget consecutively to show that Z_n becomes smaller and smaller as n increases. To reach the goal $\exp(-\ell^{\pi n})$, we apply as many times as possible before we run out of depth n.

(Define events.) Let A_0^0 and E_0^0 be the empty event. For every $m = \sqrt{n}, 2\sqrt{n}, \ldots, n-\sqrt{n}$, we hereby define six series of events A_m, A_0^m, B_m, C_m, E_m, and E_0^m inductively as follows: Let A_m be $\{Z_m < \exp(-e^{m^{1/3}})\} \setminus A_0^{m-\sqrt{n}}$. Let A_0^m be $A_0^{m-\sqrt{n}} \cup A_m$. Let B_m be a subevent of A_m where $Z_k \geqslant \ell^{-8}$ for some $k \geqslant m$. Let C_m a subevent of A_m where

$$D_{m+1}D_{m+2}\cdots D_n \leqslant \ell^{\pi n}. \tag{35}$$

Let E_m be $A_m \setminus (B_m \cup C_m)$. Let E_0^m be $E_0^{m-\sqrt{n}} \cup E_m$. Let a_m, a_0^m, b_m, c_m, e_m, and e_0^m be the probability measures of the corresponding capital letter events. Moreover, let f_m be $1 - H_0 - a_0^m$ and let g_m be $1 - H_0 - e_0^m$.

(Bound b_m/a_m from above.) Conditioning on A_m, we want to estimate the probability that $Z_k \geqslant \ell^{-8}$ for some $k \geqslant m$, which is equal to the probability that $\min(\ell^{-2}, \sqrt[4]{Z_k}) \geqslant \ell^{-2}$ for some $k \geqslant m$. Recall that $\min(\ell^{-2}, \sqrt[4]{Z_k})$ was made a supermartingale. Hence by Ville's inequality (see [109] or [107, Exercise 4.8.2]), $P\{\min(\ell^{-2}, \sqrt[4]{Z_k}) \geqslant \ell^{-2}$ for some $k \geqslant m \mid A_m\} \leqslant \min(\ell^{-2}, \sqrt[4]{Z_m})\ell^2 < \exp(-e^{m^{1/3}}/4)\ell^2$. This is an upper bound on b_m/a_m and will be summoned in Formula (36).

(Bound c_m/a_m from above.) We want to estimate how often does Inequality (35) happen. That would be the probability of $(D_{m+1}D_{m+2}\cdots D_n)^{-1/2} \geqslant \ell^{-\pi n/2}$. By Markov's inequality, this probability is at most $E[D_1^{-1/2}]^{n-m}\ell^{\pi n/2} < \ell^{(-1/2+2\alpha)(n-m)}\ell^{\pi n/2} = \ell^{(1/2-2\alpha)m-(1/2-2\alpha-\pi/2)n}$, and finally, $\leqslant \ell^{(1/2-2\alpha)m-(\rho+2\alpha)n}$. The final inequality uses Inequality (6), $\pi + 2\rho \leqslant 1 - 8\alpha$. This is an upper bound on c_m/a_m and will be summoned in Formula (36).

(Bound f_m^+ from above.) The definition of f_m reads $1 - H_0 - a_0^m$. Here a_0^m is the probability measure of A_0^m, and A_0^m is a superevent of A_m by how the former is defined. Event A_0^m must contain $\{Z_m < \exp(-e^{m^{1/3}})\}$ by how A_m was defined. By the een13 behavior, $P\{Z_m < \exp(-e^{m^{1/3}})\} > 1 - H_0 - \ell^{(-1/2+4\alpha)m+o(m)}$. Chaining all inequalities together, we infer that $f_m < \ell^{(-1/2+4\alpha)m+o(m)}$. Let f_m^+ be $\max(0, f_{m+\sqrt{n}})$ so we can write $f_m^+ < \ell^{(-1/2+4\alpha)m+o(m)}$. This upper bound will be summoned in Formula (37).

(Bound e_0^n from below.) We start rewriting $g_m - f_m^+$ with $(f_{m-\sqrt{n}} - a_m)^+$ being $\max(0, f_{m-\sqrt{n}} - a_m)$:

$$\begin{aligned} g_m - f_m^+ &= 1 - H_0 - e_0^m - (1 - H_0 - a_0^m)^+ \\ &= 1 - H_0 - e_0^{m-\sqrt{n}} - e^m - (1 - H_0 - a_0^{m-\sqrt{n}} - a_m)^+ \\ &= g_{m-\sqrt{n}} - e_m - (f_{m-\sqrt{n}} - a_m)^+ \\ &\leqslant g_{m-\sqrt{n}} - e_m - \frac{e_m}{a_m}(f_{m-\sqrt{n}} - a_m)^+ \\ &\leqslant g_{m-\sqrt{n}} - e_m - \frac{e_m}{a_m}(f^+_{m-\sqrt{n}} - a_m) \\ &= g_{m-\sqrt{n}} - f^+_{m-\sqrt{n}} + f^+_{m-\sqrt{n}}\Big(1 - \frac{e_m}{a_m}\Big) \\ &\leqslant g_{m-\sqrt{n}} - f^+_{m-\sqrt{n}} + f^+_{m-\sqrt{n}}\Big(\frac{b_m}{a_m} + \frac{c_m}{a_m}\Big) \\ &< g_{m-\sqrt{n}} - f^+_{m-\sqrt{n}} + \ell^{(-1/2+4\alpha)(m-\sqrt{n})+o(m-\sqrt{n})} \qquad (36) \\ &\qquad \times \Big(\exp(-e^{m^{1/3}}/4)\ell^2 + \ell^{(1/2-2\alpha)m-(\rho+2\alpha)n}\Big) \qquad (37) \end{aligned}$$

The first three equalities are by the definitions of g_m and f_m. The next inequality is by $0 \leqslant e_m/a_m \leqslant 1$. The next inequality is by $\max(0, f-a) = \max(a, f) - a \geqslant \max(0, f) - a$. The next equality is simple algebra. The next inequality is by the definition of E_m. The last inequality summons upper bounds derived in the last few paragraphs. Now the last line contains two terms in the big parentheses. Between them, $\ell^{(1/2-2\alpha)m-(\rho+2\alpha)n}$ dominates $\exp(-e^{m^{1/3}}/4)\ell^2$ once $n \to \infty$. Subsequently, we obtain this recurrence relation

$$\begin{cases} g_0 - f_0^+ = 0; \\ g_m - f_m^+ \leqslant g_{m-\sqrt{n}} - f^+_{m-\sqrt{n}} + 2\ell^{-\rho n+o(n)}. \end{cases}$$

Solve it (cf. the Cesàro summation); we get that $g_{n-\sqrt{n}} - f^+_{n-\sqrt{n}} < \ell^{-\rho n+o(n)}$. Once again we summon $f^+_{n-\sqrt{n}} < \ell^{(-1/2+4\alpha)(n-\sqrt{n})+o(n)} < \ell^{-\rho n+o(n)}$; therefore $g_{n-\sqrt{n}} < \ell^{-\rho n+o(n)}$. Based on the relation between $e_{n-\sqrt{n}}$ and $g_{n-\sqrt{n}}$ we immediately get $e_0^{n-\sqrt{n}} > 1 - H_0 - \ell^{-\rho n+o(n)}$.

(Analyze $E_0^{n-\sqrt{n}}$.) We want to estimate Z_n when $E_0^{n-\sqrt{n}}$ happens. More precisely, we attempt to bound Z_n when E_m

happens for each $m = \sqrt{n}, 2\sqrt{n}, \dots, n - \sqrt{n}$. When E_m happens, its superevent A_m happens, so we know that $\mathsf{Z}_m < \exp(-e^{m^{1/3}})$. But B_m does not happen, so $\mathsf{Z}_k < \ell^{-8}$ for all $k \geqslant m$. This implies $\mathsf{Z}_{k+1} \leqslant \mathsf{Z}_k^{\mathsf{D}_{k+1}}$ for those k. Telescope; Z_n is less than Z_m raised to the power of $\mathsf{D}_{m+1}\mathsf{D}_{m+2}\cdots\mathsf{D}_n$. But C_m does not happen, so the product is greater than $\ell^{\pi n}$. Jointly we have $\mathsf{Z}_n \leqslant \mathsf{Z}_m^{\ell^{\pi n}} < \exp(-e^{m^{1/3}}\ell^{\pi n}) < \exp(-\ell^{\pi n} n^2)$. In other words, $\mathsf{E}_0^{n-\sqrt{n}}$ implies $\mathsf{Z}_n < \exp(-\ell^{\pi n} n^2)$.

(Summary.) Now we conclude $P\{\mathsf{Z}_n < \exp(-\ell^{\pi n} n^2)\} \geqslant P(\mathsf{E}_0^{n-\sqrt{n}}) = \mathsf{e}_0^n > 1 - \mathsf{H}_0 - \ell^{-\rho n + o(n)}$. And hence the proof of the elpin behavior, Inequality (30), is sound.

This subsection is parallel to [50, Section VI] and also to [46, Section 10.3]. Do not confuse this subsection with the previous. The subtlety is explained in [50, Section III].

As we finish proving Inequalities (23) to (30), we finish the proof of Lemma 12. Lemmas 12 to 14 are all finished. This is the last sentence of the proof of the main theorem.

Appendix D
When Binary and Symmetric

In this appendix, we list some simplifications of the notations and the proofs when W is BDMC and symmetric (or asymmetric but with uniform input).

We have $H(W) + I(W) = 1$. Since there is no need to add dummy symbols, W remains what it was after Section II.

In Section III, the DUs in the decoder can either pass a posteriori probabilities or log-likelihood ratios, whichever is easier to imagine. Since the input is uniform, the flattening channel $W_\flat$ is completely noisy. Ergo, its synthetic descendants $W_\flat^{(i)}$, $(W_\flat^{(i)})^{(j)}$, $((W_\flat^{(i)})^{(j)})^{(k)}$, et seq. are all completely noisy. The frozen bits are either fair, independent coin tosses or some fixed values (for asymmetric channels the choice matters). The encoder falls back into the usual matrix multiplication.

The selection of information bits, $\mathcal{I}$, depends solely on whether $H\Big((\cdots(W^{(k_1)})\cdots)^{(k_n)}\Big) < \theta_n$ or not, and not on $1 - H\Big((\cdots(W_\flat^{(k_1)})\cdots)^{(k_n)}\Big) < \theta_n$ anymore because the test always passes. For the same reason, the channel process $\{\mathsf{V}_n\}$ is constantly, completely noisy. Thus the last two inequalities of Claim 11 are pointless.

Z_{mad} becomes Z; so FTPCZ is just [27, Lemma 10]. To the S-end, the character χ is the alternating character $\chi(x) = (-1)^x$. The Fourier coefficients are:

$$\begin{bmatrix} M(0 \mid y) \\ M(1 \mid y) \end{bmatrix} := \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} W(0 \mid y) \\ W(1 \mid y) \end{bmatrix}$$

The inverse of the Hadamard matrix is half of itself. Note that $M(0 \mid y)$ is always (for all channels) 1 because that is the sum of probabilities. Note also that $|M(1 \mid y)| = \sum_{x\in\mathbb{F}_2}|W(x \mid y) - 1/2| = 1 - 2\max_{x\in\mathbb{F}_2} W(x \mid y)$; so $S = S_{\max}$ coincide with $T = 1 - 2P_{\mathrm{e}}$. Beyonds these, the proof of FTPCS does not change too much.

In the local LDP behavior, the q's in Eq. (8) can be dropped: $Z_{\mathrm{mad}}(W_{\mathbb{G}}^{(i)}) \leqslant \ell\exp(Z_{\mathrm{mad}}(W)\ell)(Z_{\mathrm{mad}}(W))^{\lceil i^2/3\ell\rceil}$. Same applies to Eq. (9). This is because they are, in fact, $(q-1)$ in the proof; we prefer q over $q-1$ to save spaces. As for the proof, since $q = 2$ is the tightest case to begin with, the proof does not simplify too much. (Besides that we do not have to mention Kullback–Leibler divergence.)

Other than that the symmetrization trick is unnecessary, both the statement and the proof of the local CLT behavior stays unchanged. None of Chang–Sahai, Gallager, or Hayashi's argument is designed for any particular q, we believe. Likewise, the global MDP machinery is designed to deal with abstract processes $\{\mathsf{H}_n\}$, $\{\mathsf{Z}_n\}$, and $\{\mathsf{S}_n\}$; it has little to do with q and symmetry.

Appendix E
Constants Dependence Summary

Take a discrete memoryless channel W. The sender chooses the message alphabet size $\varsigma \geqslant 2$. Depending on the factorization of ς, we choose q to be a certain prime power or alternate between $q_2, q_3, q_5, \dots$ (a finite list depending on ς). Fix a q. Given constants $\pi, \rho > 0$ such that $\pi + 2\rho < 1$; fix them. Choose ℓ; this also determines $\alpha := \log(\log\ell)/\log\ell$. The choice of ℓ is such that $\pi + 2\rho \leqslant 1 - 8\alpha$ and such that the failing probabilities in Lemmas 13 and 14 do not sum to 1. It depends on q, π, ρ. Once ℓ is fixed, the complexity is a function in n (or in $N = \ell^n$). The asymptotic complexity $O(N\log N)$ hides the scalar term that is determined by q and ℓ. The decaying gap $\ell^{-\rho n + o(n)}$ in Claim 11 and Lemma 12 hides two things: A scalar term in front of ℓ determined by q and ℓ alongside with a $O(n^{1-\varepsilon})$ term in the exponent determined by the choice of en23 and een13 checkpoints. This ε is fixed throughout the paper and is irrespective of $\varsigma, \pi, \rho, q, \ell$.

References

[1] R. W. Hamming, "Error detecting and error correcting codes," *Bell Syst. Tech. J.*, vol. 29, no. 2, pp. 147–160, Apr. 1950.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[3] R. Fano, *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA, USA: MIT Press, 1961. [Online]. Available: https://books.google.com.tw/books?id=VSYIAQAAIAAJ

[4] R. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 1, pp. 3–18, Jan. 1965.

[5] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. I," *Inf. Control*, vol. 10, no. 1, pp. 65–103, Jan. 1967.

[6] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[7] R. Gallager, "The random coding bound is tight for the average code (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 2, pp. 244–246, Mar. 1973.

[8] R. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 4, pp. 405–417, Jul. 1974.

[9] 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.

[10] A. G. I. Fabregas, I. Land, and A. Martinez, "Extremes of random coding error exponents," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2011, pp. 2896–2898.

[11] Y. Domb, R. Zamir, and M. Feder, "The random coding bound is tight for the average linear code or lattice," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 121–130, Jan. 2016.

[12] J. Wolfowitz, "The coding of messages subject to chance errors," *Illinois J. Math.*, vol. 1, no. 4, pp. 591–606, Dec. 1957.

[13] L. Weiss, "On the strong converse of the coding theorem for symmetric channels without memory," *Quart. Appl. Math.*, vol. 18, no. 3, pp. 209–214, 1960. [Online]. Available: http://www.jstor.org/stable/43636330

[14] R. L. Dobrushin, "Mathematical problems in the Shannon theory of optimal coding of information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab. Contrib. Statist.*, vol. 1. Berkeley, CA, USA: Univ. of California Press, 1961, pp. 211–252. [Online]. Available: https://projecteuclid.org/euclid.bsmsp/1200512168

[15] V. Strassen, "Asymptotische abschätzungen in Shannons informationstheorie," in *Proc. Trans. 3rd Prague Conf. Inf. Theory*. New York, NY, USA: Academic of Sciences, 1962, pp. 689–723. [Online]. Available: https://www.math.cornell.edu/ pmlut/strassen.pdf

[16] D. Baron, M. A. Khojastepour, and R. G. Baraniuk, "How quickly can we approach channel capacity?" in *Proc. Conf. Rec. 38th Asilomar Conf. Signals, Syst. Comput.*, vol. 1, Nov. 2004, pp. 1096–1100.

[17] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.

[18] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[19] 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.

[20] Y. Polyanskiy and S. Verdu, "Channel dispersion and moderate deviations limits for memoryless channels," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2010, pp. 1334–1339.

[21] Y. Altug and A. B. Wagner, "Moderate deviations in channel coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4417–4426, Aug. 2014.

[22] E. Arikan, "A packing lemma for polar codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 2441–2445.

[23] 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.

[24] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 1173–1177.

[25] 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.

[26] E. Arikan and E. Telatar, "On the rate of channel polarization," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2009, pp. 1493–1495.

[27] S. B. Korada, E. Sasoglu, and R. Urbanke, "Polar codes: Characterization of exponent, bounds, and constructions," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6253–6264, Dec. 2010.

[28] S. H. Hassani, R. Mori, T. Tanaka, and R. L. Urbanke, "Rate-dependent analysis of the asymptotic behavior of channel polarization," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2267–2276, Apr. 2013.

[29] 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.

[30] E. Sasoglu, E. Telatar, and E. Arikan, "Polarization for arbitrary discrete memoryless channels," in *Proc. IEEE Inf. Theory Workshop*, Oct. 2009, pp. 144–148.

[31] E. Sasoglu, "Polar coding theorems for discrete systems," Ph.D. dissertation, School Comput. Commun. Sci., École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, Sep. 2011.

[32] D. Sutter, J. M. Renes, F. Dupuis, and R. Renner, "Achieving the capacity of any DMC using only polar codes," in *Proc. IEEE Inf. Theory Workshop*, Sep. 2012, pp. 114–118.

[33] 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.

[34] 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.

[35] A. Goli, S. H. Hassani, and R. Urbanke, "Universal bounds on the scaling behavior of polar codes," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2012, pp. 1957–1961.

[36] 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.

[37] D. Goldin and D. Burshtein, "Improved bounds on the finite length scaling of polar codes," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 6966–6978, Nov. 2014.

[38] 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.

[39] 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

[40] A. Fazeli and A. Vardy, "On the scaling exponent of binary polarization kernels," in *Proc. 52nd Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2014, pp. 797–804.

[41] H. Yao, A. Fazeli, and A. Vardy, "Explicit polar codes with small scaling exponent," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1757–1761.

[42] H. D. Pfister and R. Urbanke, "Near-optimal finite-length scaling for polar codes over large alphabets," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 215–219.

[43] A. Fazeli, S. H. Hassani, M. Mondelli, and A. Vardy, "Binary linear codes with optimal scaling and quasi-linear complexity," *CoRR*, vol. abs/1711.01339, pp. 1–27, Nov. 2017. [Online]. Available: http://arxiv.org/abs/1711.01339

[44] A. Fazeli, H. Hassani, M. Mondelli, and A. Vardy, "Binary linear codes with optimal scaling: Polar codes with large kernels," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[45] A. Fazeli, H. Hassani, M. Mondelli, and A. Vardy, "Binary linear codes with optimal scaling: Polar codes with large kernels," *IEEE Trans. Inf. Theory*, to be published. [Online]. Available: https://ieeexplore.ieee.org/document/9261460

[46] V. Guruswami, A. Riazanov, and M. Ye, "Arikan meets Shannon: Polar codes with near-optimal convergence to channel capacity," *CoRR*, vol. abs/1911.03858, pp. 1–78, Jun. 2019. [Online]. Available: http://arxiv.org/abs/1911.03858

[47] V. Guruswami, A. Riazanov, and M. Ye, "Arikan meets Shannon: Polar codes with near-optimal convergence to channel capacity," in *Proc. 52nd Annu. ACM SIGACT Symp. Theory Comput.* New York, NY, USA: Association for Computing Machinery, Jun. 2020, pp. 552–564, doi: 10.1145/3357713.3384323.

[48] 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.

[49] V. Guruswami and P. Xia, "Polar codes: Speed of polarization and polynomial gap to capacity," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 3–16, Jan. 2015.

[50] H. Wang and I. M. Duursma, "Polar-like codes and asymptotic tradeoff among block length, code rate, and error probability," *CoRR*, vol. abs/1812.08112, pp. 1–26, Dec. 2018. [Online]. Available: http://arxiv.org/abs/1812.08112

[51] J. Blasiok, V. Guruswami, P. Nakkiran, A. Rudra, and M. Sudan, "General strong polarization," *CoRR*, vol. abs/1802.02718, pp. 1–45, Jun. 2018. [Online]. Available: http://arxiv.org/abs/1802.02718

[52] J. Błasiok, V. Guruswami, P. Nakkiran, A. Rudra, and M. Sudan, "General strong polarization," in *Proc. 50th Annu. ACM SIGACT Symp. Theory Comput. (STOC)*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 485–492, doi: 10.1145/3188745.3188816.

[53] J. Blasiok, V. Guruswami, and M. Sudan, "Polar codes with exponentially small error at finite block length," *CoRR*, vol. abs/1810.04298, pp. 1–18, Oct. 2018. [Online]. Available: http://arxiv.org/abs/1810.04298

[54] J. Blasiok, 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)* (Leibniz International Proceedings in Informatics), vol. 116, E. Blais, K. Jansen, J. D. P. Rolim, and D. Steurer, Eds. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018, pp. 34:1–34:17. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2018/9438

[55] S. B. Korada, "Polar codes for channel and source coding," Ph.D. dissertation, School Comput. Commun. Sci., École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2009. [Online]. Available: http://infoscience.epfl.ch/record/138655

[56] N. Hussami, S. B. Korada, and R. Urbanke, "Performance of polar codes for channel and source coding," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2009, pp. 1488–1492.

[57] S. B. Korada and R. L. Urbanke, "Polar codes are optimal for lossy source coding," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1751–1768, Apr. 2010.

[58] E. Arikan, "Source polarization," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 899–903.

[59] H. S. Cronie and S. B. Korada, "Lossless source coding with polar codes," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 904–908.

[60] R. Mori, "Properties and construction of polar codes," *CoRR*, vol. abs/1002.3521, pp. 1–30, Feb. 2010. [Online]. Available: http://arxiv.org/abs/1002.3521

[61] A. G. Sahebi and S. S. Pradhan, "Multilevel polarization of polar codes over arbitrary discrete memoryless channels," in *Proc. 49th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2011, pp. 1718–1725.

[62] E. Sasoglu, "Polarization and polar codes," *Found. Trends Commun. Inf. Theory*, vol. 8, no. 4, pp. 259–381, 2011, doi: 10.1561/0100000041.

[63] E. Sasoglu, "Polar codes for discrete alphabets," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2012, pp. 2137–2141.

[64] I. Tal and A. Vardy, "How to construct polar codes," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6562–6582, Oct. 2013.

[65] S. H. Hassani, "Polarization and spatial coupling two techniques to boost performance," Ph.D. dissertation, School Comput. Commun. Sci., École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2013. [Online]. Available: http://infoscience.epfl.ch/record/188275

[66] W. C. Park and A. Barg, "Polar codes for Q-ary channels, $q = 2^r$," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 955–969, Feb. 2013.

[67] I. Tal and A. Vardy, "List decoding of polar codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2213–2226, May 2015.

[68] M. Mondelli, "From polar to Reed-Muller codes: Unified scaling, non-standard channels, and a proven conjecture," Ph.D. dissertation, School Comput. Commun. Sci., École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2016. [Online]. Available: http://infoscience.epfl.ch/record/222865

[69] R. Nasser, "An ergodic theory of binary operations—Part I: Key properties," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 6931–6952, Dec. 2016.

[70] R. Nasser, "An ergodic theory of binary operations—Part II: Applications to polarization," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 1063–1083, Feb. 2017.

[71] Y. Polyanskiy, "Channel coding: Non-asymptotic fundamental limits," Ph.D. dissertation, Dept. Elect. Eng., Princeton Univ., Princeton, NJ, USA, 2010.

[72] M. Ye and A. Barg, "Polar codes using dynamic kernels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 231–235.

[73] C. Chang and A. Sahai, "Universal quadratic lower bounds on source coding error exponents," in *Proc. 41st Annu. Conf. Inf. Sci. Syst.*, Mar. 2007, pp. 714–719.

[74] S. C. Draper, C. Chang, and A. Sahai, "Lossless coding for distributed streaming sources," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1447–1474, Mar. 2014.

[75] N. Presman, O. Shapira, and S. Litsyn, "Mixed-kernels constructions of polar codes," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 2, pp. 239–253, Feb. 2016.

[76] M. El-Khamy, H. Mahdavifar, G. Feygin, J. Lee, and I. Kang, "Relaxed polar codes," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 1986–2000, Apr. 2017.

[77] R. Mori and T. Tanaka, "Channel polarization on Q-ary discrete memoryless channels by arbitrary kernels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 894–898.

[78] E. Arikan, "Varentropy decreases under the polar transform," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3390–3400, Jun. 2016.

[79] M. Karzand and E. Telatar, "Polar codes for Q-ary source coding," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 909–912.

[80] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 259–266, Jan. 1994.

[81] T. S. Jayram and E. Arikan, "A note on some inequalities used in channel polarization and polar coding," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5767–5768, Aug. 2018.

[82] A. Terras, *Fourier Analysis on Finite Groups and Applications* (London Mathematical Society Student Texts), vol. 43. Cambridge, U.K.: Cambridge Univ. Press, 1999. [Online]. Available: https://doi.org/10.1017/CBO9780511626265

[83] A. Dembo and O. Zeitouni, *Large deviations Techniques and Applications* (Stochastic Modelling and Applied Probability), vol. 38. Berlin, Germany: Springer-Verlag, 2010. [Online]. Available: https://doi.org/10.1007/978-3-642-03311-7

[84] M. Beck, G. Marchesi, D. Pixton, and L. Sabalka. *A First Course in Complex Analysis*. [Online]. Available: https://aimath.org/textbooks/approved-textbooks/beck-marchesi-pixton-sabalka/

[85] L. V. Ahlfors, *Complex Analysis: An Introduction to the Theory of Analytic Functions of One Complex Variable* (International Series in Pure and Applied Mathematics), 3rd ed. New York, NY, USA: McGraw-Hill, 1978.

[86] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Oct. 1975. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1975.tb02040.x

[87] M. Hayashi, "General nonasymptotic and asymptotic formulas in channel resolvability and identification capacity and their application to the wiretap channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1562–1575, Apr. 2006.

[88] S. Kudekar, S. Kumar, M. Mondelli, H. D. Pfister, E. Şaşoğlu, 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.

[89] Y. Oohama, "Error probability of identification via channels at rates above capacity," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2002, p. 26.

[90] M. Hayashi and R. Matsumoto, "Universally attainable error and information exponents, and equivocation rate for the broadcast channels with confidential messages," in *Proc. 49th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2011, pp. 439–444.

[91] M. Bastani Parizi, E. Telatar, and N. Merhav, "Exact random coding secrecy exponents for the wiretap channel," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 509–531, Jan. 2017.

[92] E. Sasoglu, E. Telatar, and E. Arikan, "Polarization for arbitrary discrete memoryless channels," *CoRR*, vol. abs/0908.0302, pp. 1–12, Oct. 2009. [Online]. Available: http://arxiv.org/abs/0908.0302

[93] Y. Inan and E. Telatar, "Finite-level quantization procedures for construction and decoding of polar codes," 2020, *arXiv:2001.04009*. [Online]. Available: https://arxiv.org/abs/2001.04009

[94] H. Mahdavifar, "Fast polarization for non-stationary channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 849–853.

[95] R. Wang, J. Honda, H. Yamamoto, R. Liu, and Y. Hou, "Construction of polar codes for channels with memory," in *Proc. IEEE Inf. Theory Workshop Fall (ITW)*, Oct. 2015, pp. 187–191.

[96] E. Sasoglu and I. Tal, "Polar coding for processes with memory," *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 1994–2003, Apr. 2019.

[97] I. Tal, H. D. Pfister, A. Fazeli, and A. Vardy, "Polar codes for the deletion channel: Weak and strong polarization," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1362–1366.

[98] Y. Li and V. Y. F. Tan, "On the capacity of channels with deletions and states," 2019, *arXiv:1911.04473*. [Online]. Available: https://arxiv.org/abs/1911.04473

[99] S. L. Fong and V. Y. F. Tan, "On the scaling exponent of polar codes for binary-input energy-harvesting channels," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3540–3551, Dec. 2016.

[100] H. Mahdavifar and A. Vardy, "Achieving the secrecy capacity of wiretap channels using polar codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6428–6443, Oct. 2011.

[101] E. Sasoglu and A. Vardy, "A new polar coding scheme for strong security on wiretap channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 1117–1121.

[102] E. Abbe, "Randomness and dependencies extraction via polarization, with applications to Slepian–Wolf coding and secrecy," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2388–2398, May 2015.

[103] N. Goela, E. Abbe, and M. Gastpar, "Polar codes for broadcast channels," *IEEE Trans. Inf. Theory*, vol. 61, no. 2, pp. 758–782, Feb. 2015.

[104] M. Mondelli, S. Hamed Hassani, I. Sason, and R. L. Urbanke, "Achieving Marton's region for broadcast channels using polar codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 2, pp. 783–800, Feb. 2015.

[105] E. Abbe and E. Telatar, "Polar codes for the m-user multiple access channel," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5437–5448, Aug. 2012.

[106] R. Nasser and E. Telatar, "Polar codes for arbitrary DMCs and arbitrary MACs," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 2917–2936, Jun. 2016.

[107] R. Durrett, *Probability: Theory and Examples* (Cambridge Series in Statistical and Probabilistic Mathematics), 5th ed. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[108] E. Abbe and M. Ye, "Reed-muller codes polarize," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7311–7332, Dec. 2020.

[109] J. Ville. *Étude Critique de la Notion de Collectif*. NUMDAM. Accessed: 1939. [Online]. Available: http://www.numdam.org/item?id=THESE_1939__218__1_0

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.