Electronic Colloquium on Computational Complexity, Revision 3 of Report No. 154 (2019)

Arıkan meets Shannon: Polar codes with near-optimal convergence to channel capacity

Venkatesan Guruswami* Andrii Riazanov† Min Ye‡

Abstract

Let W be a binary-input memoryless symmetric (BMS) channel with Shannon capacity $I(W)$ and fix any $\alpha > 0$. We construct, for any sufficiently small $\delta > 0$, binary linear codes of block length $O(1/\delta^{2+\alpha})$ and rate $I(W) - \delta$ that enable reliable communication on W with quasi-linear time encoding and decoding. Shannon's noisy coding theorem established the *existence* of such codes (without efficient constructions or decoding) with block length $O(1/\delta^2)$. This quadratic dependence on the gap δ to capacity is known to be best possible. Our result thus yields a constructive version of Shannon's theorem with near-optimal convergence to capacity as a function of the block length. This resolves a central theoretical challenge associated with the attainment of Shannon capacity. Previously such a result was only known for the erasure channel.

Our codes are a variant of Arıkan's polar codes based on multiple carefully constructed local kernels, one for each intermediate channel that arises in the decoding. A crucial ingredient in the analysis is a strong converse of the noisy coding theorem when communicating using random linear codes on arbitrary BMS channels. Our converse theorem shows extreme unpredictability of even a single message bit for random coding at rates slightly above capacity.

Keywords:

Polar codes, capacity-achieving codes, scaling exponent, finite blocklength

An extended abstract of this paper was presented at the 2020 ACM Symposium on Theory of Computing (STOC) [GRY20].

*Computer Science Departmemt, Carnegie Mellon University, Pittsburgh, PA 15213. Email: venkatg@cs.cmu.edu. Research supported in part by NSF grants CCF-1422045, CCF-1563742, and CCF-1814603, and a Google Research Award.

†Computer Science Departmemt, Carnegie Mellon University, Pittsburgh, PA 15213. Email: riazanov@cs.cmu.edu. Research supported in part by NSF grants CCF-1422045, CCF-1563742, and CCF-1814603.

‡Data Science and Information Technology Research Center, Tsinghua-Berkeley Shenzhen Institute, Shenzhen, China. Email: yeemmi@gmail.com. Some of this research was carried out when the author was visiting Carnegie Mellon University.

ISSN 1433-8092

Contents

1 Introduction

We construct binary linear codes that achieve the Shannon capacity of the binary symmetric channel, and indeed any binary-input memoryless symmetric (BMS) channel, with a near-optimal scaling between the code length and the gap to capacity. Further, our codes have efficient (quasi-linear time) encoding and decoding algorithms. Let us now describe the context of our result and its precise statement in more detail.

The binary symmetric channel (BSC) is one of the most fundamental and well-studied noise models in coding theory. The BSC with crossover probability $p \in (0, 1/2)$ (BSC_p) flips each transmitted bit independently with probability p. By Shannon's seminal noisy coding theorem [Sha48], we know that the capacity of BSC_p is $1 - h(p)$, where $h(\cdot)$ is the binary entropy function. This means that reliable communication over BSC_p is possible at information rates approaching $1 - h(p)$, and at rates above $1 - h(p)$ this is not possible. More precisely, for any $\delta > 0$, there *exist* codes of rate $1 - h(p) - \delta$ using which one can achieve miscommunication probability at most $2^{-\Omega(\delta^2 n)}$ where n is the block length of the code. In fact, random linear codes under maximum likelihood decoding offer this guarantee with high probability. Thus Shannon's theorem implies the existence of codes of block length $O(1/\delta^2)$ that can achieve small error probability on BSC_p at rates within δ of capacity. Conversely, by several classical results [Wol57, Str62, Str09, PPV10], we know that the block length has to be at least $\Omega(1/\delta^2)$ in order to approach capacity within δ.

Shannon's theorem is based on the probabilistic method and does not describe the codes that approach capacity or give efficient algorithms to decode them from errors caused by BSC_p. Thus the codes with rates $1 - h(p) - \delta$ take at least time exponential in $1/\delta^2$ to construct as well as decode. This is also true for concatenated coding schemes [For67] as the inner codes have to be decoded by brute-force, and either have to also be found by a brute-force search or allowed to vary over an exponentially large ensemble (leading to exponentially large block length).

The theoretical challenge of constructing codes of rate $1 - h(p) - \delta$ with construction/decoding complexity scaling polynomially in $1/\delta$ in fact remained wide open for a long time. Finally, around 2013, two independent works [GX15, HAU14] gave an effective finite-length analysis of Arıkan's remarkable polar codes construction [Arı09]. (Arıkan's original analysis, as well as follow-ups like [AT09], proved convergence to capacity as the block length grew to infinity but did not quantify the speed of convergence.) Based on this, a construction of polar codes with block length, construction, and decoding complexity all bounded by a polynomial in $1/\delta$ to capacity was obtained in [GX15, HAU14]. The result also applies to any BMS channel, not just the BSC.

If the block length of the code scales as $O(1/\delta^\mu)$ as a function of the gap δ to capacity, we say that μ is the *scaling exponent*. The above results established that the scaling exponent of polar codes is finite. It is worth pointing out that polar codes are the *only* known efficiently decodable capacity-achieving family proven to have a finite scaling exponent. The work [GX15] did not give an explicit upper bound on the scaling exponent of polar codes, whereas [HAU14] showed the bound $\mu \leq 6$. Following some improvements in [GB14, MHU16], the current best known upper bound on μ for the BSC (and any BMS channel) is 4.714.

Note that random linear codes have optimal scaling exponent 2. The above results thus raise the intriguing challenge of constructing codes with scaling exponent close to 2, a goal we could not even dream of till the recent successes of polar codes.

Arıkan's original polar coding construction is based on a large tensor power of a simple 2×2 matrix, which is called the *kernel* of the construction. For this construction, it was shown in [HAU14] that the scaling exponent μ for Arıkan's original polar code construction is *lower bounded*

by 3.579, even for the simple binary erasure channel. Given this limitation, one approach to improve μ is to consider polar codes based on $\ell \times \ell$ kernels for larger ℓ. However, better upper bounds on the scaling exponent of polar codes based on larger kernels have not been established except for the simple case of the binary erasure channel (BEC).[1] For the BEC, using large kernels, polar codes with scaling exponent $2+\alpha$ for any desired $\alpha > 0$ were given in the very nice paper [FHMV17] which spurred our work. (We will discuss this and other related works in more detail in Sections 2.2–2.3.)

Our main result in this work is a polynomial time construction of polar codes based on large kernels that approach the optimal scaling exponent of 2 for every BMS channel. Specifically, for any desired small $\alpha > 0$, by picking sufficiently large kernels (as a function of α), the block length N can be made as small as $O_\alpha(1/\delta^{2+\alpha})$ for codes of rate $I(W) - \delta$ (the notation $O_\alpha(\cdot)$ hides a constant that depends only on α). The encoding and decoding complexity will be *quasi-linear* in N, and thus can also have a near-quadratic growth with $1/\delta$.

Theorem 1.1 (Main)**.** *Let W be an arbitrary BMS channel with Shannon capacity $I(W)$. For any desired $\alpha \in \left(0, \frac{1}{36}\right)$, if we choose a large enough constant $\ell \geq \ell_0(\alpha)$ to be a power of 2, then there is a code $\mathcal{C}$ generated by the polar coding construction using kernels of size $\ell \times \ell$ such that the following four properties hold when N is the code length:*

1. *the code construction has $N^{O_\alpha(1)}$ complexity;*
2. *both encoding and decoding have $O_\alpha(N \log N)$ complexity;*
3. *the rate of $\mathcal{C}$ is at least $I(W) - N^{-1/2+18\alpha}$; and*
4. *the block decoding error probability is bounded by $\exp(-N^\alpha)$ when $\mathcal{C}$ is used for channel coding over W.*

The above "constructivizes" the quantitative finite-length version of Shannon's theorem with a small α slack in the speed of convergence to capacity. The lower bound on ℓ can be chosen as $\ell_0(\alpha) = \exp(\Omega(\alpha^{-1.01}))$. More precisely, it should satisfy $\log \ell_0(\alpha) \geq \frac{11}{\alpha}$ and $\frac{\log \ell_0(\alpha)}{\log\log \ell_0(\alpha)+2} \geq \frac{3}{\alpha}$. Note that a similar lower bound on ℓ also appears in the aforementioned result for the BEC from [FHMV17]. Due to the requirement of extremely large ℓ, our result is thus primarily theoretical in nature, and meant to illustrate that the polar coding framework is powerful enough to achieve asymptotically optimal rate of convergence to Shannon capacity with efficient algorithms.

We would like to point out that in the conference version of this work [GRY20] we only proved inverse polynomial decoding error probability, as opposed to the inverse sub-exponential $\exp(-N^\alpha)$ bound which we show here. This improvement uses the subsequent analysis of polarization due to Wang and Duursma in [WD19], where they extended the results of Theorem 1.1 to arbitrary discrete memoryless channels, possibly non-binary and asymmetric, and proved the $\exp\left(-N^{O(\alpha)}\right)$ bound on the decoding error probability. However, this was done at a cost of losing the polynomial-time construction complexity of the code. We are able to non-trivially combine the analysis from [WD19] with our approach of constructing the code to achieve *both* polynomial time construction and sub-exponentially small decoding error probability simultaneously. Getting a polynomial-time constructible version of the results of [WD19] for general channels with arbitrary input alphabet remains a challenging and interesting open question.

[1] Polar codes based on $\ell \times \ell$ kernels have much larger block length ℓ^t compared to 2^t for the 2×2 case. So to get an improvement in μ, one has to compensate for the increasing block length via better bounds on the local behavior of the kernel.

2 Overview of our construction and analysis

In order to better explain our work and situate it in the rich backdrop of related works on polar codes, we begin with some context and background on the phenomenon of channel polarization that lies at the heart of Arıkan's polar coding approach.

2.1 Channel transforms, entropy polarization, and polar codes

Consider an arbitrary binary-input memoryless symmetric (BMS)[2] channel $W : \{0,1\} \to \mathcal{Y}$, and an $\ell \times \ell$ invertible binary matrix K (referred to as the *kernel*). Suppose that we are transmitting a binary vector $\boldsymbol{U} = (U_1, U_2, \ldots, U_\ell)$ uniformly chosen from $\{0,1\}^\ell$ in the following way: first, it is transformed into $\boldsymbol{X} = \boldsymbol{U}K$, which is then transmitted through ℓ copies of the channel W to get the output $\boldsymbol{Y} = W^\ell(\boldsymbol{X}) \in \mathcal{Y}^\ell$.

Now imagine decoding the input bits U_i successively in the order of increasing i. This naturally leads to a binary-input channel $W_i : \{0,1\} \to \mathcal{Y}^\ell \times \{0,1\}^{i-1}$, for each $i \in [\ell]$, which is the channel "seen" by the bit U_i when all the previous bits $\boldsymbol{U}_{<i}$ and all the channel outputs $\boldsymbol{Y} \in \mathcal{Y}^\ell$ are known. Formally, the transition probabilities of this channel are

$$W_i(\boldsymbol{Y}, \boldsymbol{U}_{<i} \,|\, U_i) = \frac{1}{2^{\ell-1}} \sum_{\boldsymbol{V} \in \{0,1\}^{\ell-i}} W^\ell\Big(\boldsymbol{Y} \,|\, (\boldsymbol{U}_{<i}, U_i, \boldsymbol{V})K\Big), \tag{1}$$

where $\boldsymbol{U}_{<i} \in \{0,1\}^{i-1}$ are the first $(i-1)$ bits of $\boldsymbol{U}$, and the sum is over all possible values $\boldsymbol{V} \in \{0,1\}^{\ell-i}$ that the last $(\ell - i)$ bits of $\boldsymbol{U}$ can take. In this paper we will address the channel W_i as "Arıkan's bit-channel of W with respect to K."

A *polarization transform* associated with the kernel K is then defined as a transformation that maps ℓ copies of the channel W to the channels $W_1, W_2, \ldots, W_\ell$. For a BMS channel W, we define $H(W)$ as the conditional entropy of the channel input random variable given the channel output random variable when the channel input has uniform distribution. Since K is invertible, a direct implication of the chain rule for entropy gives *entropy conservation property*, which is

$$\ell \cdot H(W) = H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{U}|\boldsymbol{Y}) = \sum_{i=1}^{\ell} H(U_i|\boldsymbol{U}_{<i}, \boldsymbol{Y}) = \sum_{i=1}^{\ell} H(W_i). \tag{2}$$

If K is invertible and is not upper-triangular under any column permutation (which we refer to as a *mixing matrix*), then the bit-channels $W_1, W_2, \ldots, W_\ell$ start *polarizing* – some of them become better than W (have smaller entropy), and some become worse [KSU10, Lemma 1 and Theorem 2]. The standard approach is then to recursively apply the polarization transform of K to these bit-channels. This naturally leads to an ℓ-ary tree of channels. The t'th level of the tree corresponds to the linear transformation $K^{\otimes t}$, the t-fold Kronecker product of K. [3]

In his landmark paper [Arı09], Arıkan proved that when $K = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$, at the t'th level, all but a $o(1)$ fraction of the channels (as $t \to \infty$) are either almost noiseless (have tiny entropy) or completely useless (have entropy very close to 1). To get capacity-achieving codes from polarization, the idea is to use the almost-noiseless channels, which will constitute $\approx I(W)$ fraction by conservation

[2] We say that a channel $W : \{0,1\} \to \mathcal{Y}$ is a BMS channel if there is a permutation π on the output alphabet $\mathcal{Y}$ satisfying i) $\pi^{-1} = \pi$ and ii) $W(y|1) = W(\pi(y)|0)$ for all $y \in \mathcal{Y}$.

[3] Actually, the analysis is more convenient if one applies a bit-reversal permutation of the U_i's, and indeed we do so also in this paper, but this is not important for our current discussion.

of entropy, to carry the message bits, and "freeze" the bits in the remaining positions to predetermined values (eg. all 0s). Thus the generator matrix of the code will consist of those rows of $K^{\otimes t}$ that correspond to the almost-noiseless positions. Arıkan presented a successive cancellation (SC) decoder and proved that it can be implemented using $O(N \log N)$ operations where $N = \ell^t$ is the code length, thanks to the nice recursive structure of $K^{\otimes t}$.

For the parameters of the code, if one shows that at most δ_t fraction of the channels at the t'th level have entropies in the range $(\zeta_t, 1 - \zeta_t)$, then one (roughly) gets codes of length 2^t, rate $I(W) - \delta_t - \zeta_t$, for which the SC decoder achieves decoding error probability $\zeta_t \ell^t$ for noise caused by W (see, for example [BGN$^+$18, Theorem A.3]). Thus, one needs ζ_t sub-exponentially small in t (i.e., at most $\exp(-\omega(t))$) to achieve good decoding error. For Arıkan's original 2×2 kernel, this was shown in [AT09]. Korada, Sasoglu and Urbanke extended the analysis to arbitrary $\ell \times \ell$ mixing matrices over the binary field [KSU10], and Mori and Tanaka established a similar claim over all finite fields [MT14].

The fraction δ_t of *unpolarized* channels (whose entropies fail to be sub-exponentially close to 0 or 1) governs the gap to capacity of polar codes. The above works established that $\lim_{t\to\infty} \delta_t = 0$, and thus polar codes achieve capacity asymptotically as the block length grows to infinity. However, they did not provide any bounds on the speed at which $\delta_t \to 0$ as a function t, much less quantify a scaling exponent. Note that one would need to show $\delta_t \leq O(\ell^{-t/\mu})$ to establish a scaling exponent of μ, since the code length is ℓ^t.

2.2 Scaling exponents: prior work

For Arıkan's original kernel $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$, two independent works [HAU14, GX15] proved that δ_t drops to 0 exponentially fast in t. This proved that Arıkan's polar codes have finite scaling exponent (i.e., converge to capacity polynomially fast in the block length), the first codes with this important feature. Blasiok *et al* generalized this result significantly [BGN$^+$18], proving that the entire class of polar codes, based on arbitrary mixing matrices over any prime field as kernels, has finite scaling exponent.

For concrete upper bounds on the scaling exponent, the work of Hassani, Alishahi, and Urbanke [HAU14] had proved $\mu \leq 6$ for Arıkan's original kernel. This was improved to $\mu \leq 5.702$ in [GB14]. Mondelli, Hassani, and Urbanke [MHU16] showed that $\mu \leq 4.714$ for any BMS channel W, and showed a better upper bound $\mu \leq 3.639$ for the case when W is a binary erasure channel (BEC). A *lower bound* $\mu \geq 3.579$ appears in [HAU14] for the case when successive cancellation decoder is used and analyzed using standard methods, which suggests that polar codes based on Arıkan's original 2×2 kernel fall short of the optimal scaling exponent of 2.

For larger kernels, effective upper bounds on the scaling exponent are harder to establish as the local evolution of the channels is more complex. In fact, to the best of our knowledge, there is no such explicit bound in the literature, for any[4] kernel of size bigger than 2. The analysis of polar codes is a lot more tractable for the case of erasure channels, where symbols get erased (replaced by a "?" but never corrupted). Next we describe some results for erasure channels as well as the difficulty in extending these results to channels such as the BSC.

[4]Here we exclude special cases such as a block diagonal matrix with blocks of size at most 2 which can be reduced to the 2×2 case but will only have a worse scaling exponent.

2.3 Polar codes for erasure channels

For the erasure channel, we have analyses of larger kernels and even codes with scaling exponent approaching 2. Binary $\ell \times \ell$ kernels for powers of two $\ell \leq 64$ optimized for the binary erasure channel appear in [MT12, FV14, YFV19]; a 64×64 kernel achieving $\mu < 3$ is reported in [YFV19].

Pfister and Urbanke proved in [PU16] that the optimal scaling exponent $\mu = 2$ can be approached if one considers transmission over the q-ary erasure channel for large alphabet size q. They used polar codes based on $q \times q$ kernels. Fazeli, Hassani, Mondelli, and Vardy [FHMV17] then established a similar result for the more challenging and also more interesting case of $q = 2$, i.e., for the binary erasure channel, using $\ell \times \ell$ kernels for large ℓ. They pose proving an analogous result for arbitrary BMS channels as an important challenge. Their conjecture that this can be accomplished provided some of the impetus for our work. Our analysis structure follows a similar blueprint to [FHMV17] though the technical ingredients become significantly more complex for channels other than the BEC, as explained next.

The polarization process for erasure channels has a particularly nice structure. If the initial channel W is the binary erasure channel with erasure probability z (denoted BEC(z)), then the Arıkan channels W_i, $i \in [\ell]$, arising from the linear transformation by the kernel are also binary erasure channels (specifically, BEC($p_i(z)$) where $p_i(\cdot)$ are some polynomials of degree at most ℓ). Crucially, *all* the channels in the recursive tree remain BEC. Therefore it suffices to prove the existence of a good polarizing kernel for the class of binary erasure channels, which is parameterized by a single number, the erasure probability, which also equals the entropy of the channel. As shown in [FHMV17], a random kernel works with good probability for all BEC universally. However, fundamentally the calculations for BEC revolve around the rank of various random subspaces, as decoding under the BEC is a linear-algebraic task. Moving beyond the BEC takes us outside the realm of linear algebra into information-theoretic settings where tight quantitative results are much harder to establish.

2.4 The road to BSC: Using multiple kernels

For the case when the initial channel W is a BSC, a fundamental difficulty (among others) is that the channels in the recursion tree will no longer remain BSC (even after the first step). Further, to the best of our knowledge, the various channels that arise do not share a nice common exploitable structure. Therefore, we have to think of the intermediate channels as arbitrary BMS channels, a very large and diverse class of channels. It is not clear (to us) if there exists a single kernel to universally polarize *all* BMS channels at a rapid rate. Even if such a kernel exists, proving so seems out of reach of current techniques. Finally, even for a specific BMS, proving that a random kernel polarizes it fast enough requires some very strong quantitative bounds on the performance and limitations of random linear codes for channel coding. We next describe these issues dealing with which constitutes the core of our contributions.

Since we are not able to establish that a single kernel could work for the whole construction universally, our idea is to pick different kernels, which depend on the channels that we face during the procedure. That way, by picking a suitable kernel for each channel in the tree, we can ensure that polarization is fast enough throughout the whole process.

Though we use different kernels in the code construction, all of them have the same size $\ell \times \ell$. We say that a kernel is *good* if all but a $\widetilde{O}(\ell^{-1/2})$ fraction of the bit-channels obtained after polar transform by this kernel have entropy $\ell^{-\Omega(\log \ell)}$-close to either 0 or 1. Given a BMS channel W, the code construction consists of t steps, from Step 0 to Step $t-1$. At Step 0, we find a good kernel

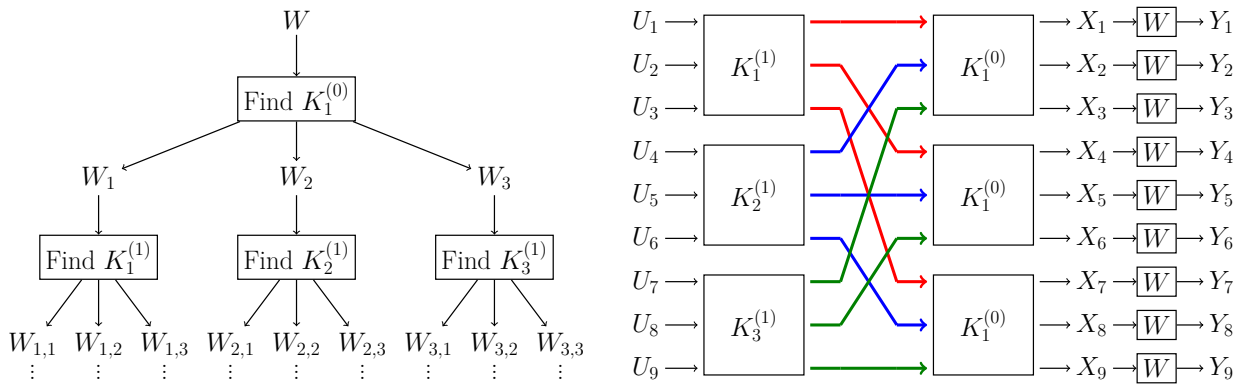

Figure 1: The left figure illustrates the code construction, and the right figure shows the encoding procedure for the special case of $\ell = 3$ and $t = 2$. All the kernels in this figure have size 3×3. One can show that the bit-channel $W_{i,j}$ in the left figure is exactly the channel mapping from $U_{3(i-1)+j}$ to $(\boldsymbol{U}_{[1:3(i-1)+j-1]}, \boldsymbol{Y}_{[1:9]})$ in the right figure.

$K_1^{(0)}$ for the original channel W. After the polar transform of W using kernel $K_1^{(0)}$, we obtain ℓ bit-channels $W_1, \ldots, W_\ell$. Then in Step 1, we find good kernels for each of these ℓ bit-channels. More precisely, the good kernel for W_i is denoted as $K_i^{(1)}$, and the bit-channels obtained by polar transforms of W_i using kernel $K_i^{(1)}$ are denoted as $\{W_{i,j} : j \in [\ell]\}$; see Figure 1 for an illustration. At Step j, we will have ℓ^j bit-channels $\{W_{i_1,\ldots,i_j} : i_1, \ldots, i_j \in [\ell]\}$. For each of them, we find a good kernel $K_{i_1,\ldots,i_j}^{(j)}$. After polar transform of $\{W_{i_1,\ldots,i_j} : i_1, \ldots, i_j \in [\ell]\}$ using these kernels, we will obtain ℓ^{j+1} bit-channels. Finally, after the last step (Step $t-1$), we will obtain $N = \ell^t$ bit-channels $\{W_{i_1,\ldots,i_t} : i_1, \ldots, i_t \in [\ell]\}$. Using the good kernels we obtained in this code construction procedure, we can build an $N \times N$ matrix (or we can view it as a large kernel) $M^{(t)}$ such that the N bit-channels resulting from the polar transform of the original channel W using this large kernel $M^{(t)}$ are exactly $\{W_{i_1,\ldots,i_t} : i_1, \ldots, i_t \in [\ell]\}$. We will say a few more words about this in Section 2.7 and provide all the details in Section 9.

Define now a random process by $\mathsf{W}_0 = W$ and $\mathsf{W}_{j+1} = (\mathsf{W}_j)_i$ for i uniformly chosen from $[\ell]$, where $(\mathsf{W}_j)_i$ is the i^{th} Arıkan bit-channel of W_j with respect to the appropriate kernel chosen in the construction phase. In other words, this is a random process of going down the tree of channels, where a uniformly random child of a current node is taken at each step. Finally, define another random process $\mathsf{H}_j \coloneqq H(\mathsf{W}_j)$. Since every kernel in the construction is chosen to be invertible, H_j is a martingale due to the conservation of entropy property (2). It is clear that W_j marginally is a uniformly random channel of the j^{th} level of channel tree, and then H_j is the entropy of such a randomly chosen channel.

2.5 Analysis of polarization via recursive potential function

The principle behind polarization is that for large enough t, almost all of the channels on the t-th level of the tree from Figure 1 will be close to either the useless or noiseless channel, i.e., their entropy is very close to 1 or 0, correspondingly. To estimate how fast such polarization actually happens, one needs to bound the fraction of channels on the t-th level that are not yet sufficiently polarized, i.e., $\mathbb{P}\left[\mathsf{H}_t \in (\zeta, 1-\zeta)\right]$ for some tiny threshold ζ, and show that this fraction vanishes rapidly with increasing t.

Specifically, we have the following result (stated explicitly in [BGN+18, Theorem A.3]) already alluded to in Section 2.1: if for all t

$$\mathbb{P}[\mathsf{H}_t \in (p\ell^{-t}, 1-p\ell^{-t})] \le D \cdot \beta^t, \tag{3}$$

then this corresponds to a polar code with block length $N = \ell^t$, rate $(D \cdot \beta^t + p\ell^{-t})$-close to the capacity of the channel, and decoding error probability at most p under the successive cancellation decoder [5].

To track the fractions of polarized and non-polarized channels at each level of the construction, we use a potential function

$$g_\alpha(h) = (h(1-h))^\alpha, \tag{4}$$

where $\alpha > 0$ is some small fixed parameter. This α corresponds to the gap to the scaling exponent in Theorem 1.1, and in this paper we always consider $\alpha < \frac{1}{12}$ (and smaller in some cases). Such a potential function was also used for example in [MHU16] and [FHMV17]. We are going to track the expected value $\mathbb{E}[g_\alpha(\mathsf{H}_t)]$ as t increases, since Markov's inequality implies

$$\mathbb{P}[\mathsf{H}_t \in (p\ell^{-t}, 1-p\ell^{-t})] = \mathbb{P}[g_\alpha(\mathsf{H}_t) \ge g_\alpha(p\ell^{-t})] \le \frac{\mathbb{E}[g_\alpha(\mathsf{H}_t)]}{g_\alpha(p\ell^{-t})} \le 2\left(\ell^t/p\right)^\alpha \cdot \mathbb{E}[g_\alpha(\mathsf{H}_t)]. \tag{5}$$

To upper bound $\mathbb{E}[g_\alpha(\mathsf{H}_t)]$ we choose kernels in our construction so that the average of the potential function of all the children of any channel in the tree decreases significantly with respect to the potential function of this channel. Precisely, we want for any channel W' in the tree

$$\mathop{\mathbb{E}}_{i\sim[\ell]}\left[g_\alpha\left(H(W_i')\right)\right] \le \lambda_\alpha \cdot g_\alpha\left(H(W')\right), \tag{6}$$

where W_i' are the children of W' in the construction tree for $i \in [\ell]$, and the constant λ_α only depends on α and ℓ, but is universal for all the channels in the tree (and thus for all the kernels chosen during the construction). If one can guarantee that (6) holds throughout the construction process, then for the martingale process H_t obtain

$$\begin{aligned}
\mathbb{E}\left[g_\alpha\left(\mathsf{H}_t\right)\right] &= \mathbb{E}\left[\mathop{\mathbb{E}}_{j\sim[\ell]}\left[g_\alpha\left(H\left((\mathsf{W}_{t-1})_j\right)\right)\right] \,\middle|\, \mathsf{W}_{t-1}\right] \\
&= \mathbb{E}\left[\frac{1}{\ell}\frac{\sum_{j=1}^{\ell} g_\alpha\left(H\left((\mathsf{W}_{t-1})_j\right)\right)}{g_\alpha\left(H(\mathsf{W}_{t-1})\right)} \cdot g_\alpha\left(H(\mathsf{W}_{t-1})\right) \,\middle|\, \mathsf{W}_{t-1}\right] \\
&\le \lambda_\alpha \cdot \mathbb{E}\left[g_\alpha\left(\mathsf{H}_{t-1}\right)\right],
\end{aligned} \tag{7}$$

and then inductively

$$\mathbb{E}\left[g_\alpha\left(\mathsf{H}_t\right)\right] \le \lambda_\alpha \cdot \mathbb{E}\left[g_\alpha\left(\mathsf{H}_{t-1}\right)\right] \le \lambda_\alpha^2 \cdot \mathbb{E}\left[g_\alpha\left(\mathsf{H}_{t-2}\right)\right] \le \cdots \le \lambda_\alpha^t \mathsf{H}_0 \le \lambda_\alpha^t. \tag{8}$$

Then (5) and (3) imply existence of code with rate $O\left((N/p)^\alpha \cdot \lambda_\alpha^t\right)$-close to capacity of the channel. Since our main task is to achieve a gap which is close to $N^{-1/2} = \ell^{-t/2}$, we need to argue that it is possible to choose kernels at each step in the construction so that (6) always holds for some $\alpha \to 0$ and $\lambda_\alpha \approx \ell^{-1/2}$.

[5] For this part the reader should think of p as being inverse polynomial (of fixed degree) in N. We will discuss improving the decoding error probability in Section 2.8.

2.6 Sharp transition in polarization

The main technical contribution of this paper consists in showing that if ℓ is large enough, it is possible to choose kernels in the construction process for which λ_α is close to $\ell^{-1/2}$. Specifically, we prove that if ℓ is a power of 2 such that $\log \ell = \Omega\left(\frac{1}{\alpha^{1.01}}\right)$, then it is possible to achieve

$$\lambda_\alpha \leq \ell^{-1/2+5\alpha}. \tag{9}$$

To obtain such a behavior, while choosing the kernel for the current channel W' during the recursive process we differentiate between two cases:

Case 1: W' is already very noisy or almost noiseless. Such regime is called *suction at the ends* (following [BGN$^+$18]), and it is known that polarization happens (much) faster for this case. So in this case we take a standard Arıkan's polarization kernel $K = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)^{\otimes \log \ell}$ and prove (6) with a geometric decrease factor $\lambda_\alpha \leq \ell^{-1/2}$.

Case 2: W' is neither very noisy nor almost noiseless. We refer to this case as *variance in the middle* regime (following [BGN$^+$18] again). For such a channel we adopt the framework from [FHMV17] and show a *sharp transition in polarization* for a random kernel K and W'. Specifically, we prove that with high probability over $K \sim \{0,1\}^{\ell \times \ell}$ (for ℓ large enough) it holds

$$\begin{aligned} H(W_i'(K)) &\leq \ell^{-\Omega(\log \ell)} \quad \text{for} \quad i \geq \ell \cdot H(W') + \Omega(\ell^{1/2} \log^3 \ell), \\ H(W_i'(K)) &\geq 1 - \ell^{-\Omega(\log \ell)} \quad \text{for} \quad i \leq \ell \cdot H(W') - \Omega(\ell^{1/2} \log^3 \ell). \end{aligned} \tag{10}$$

It then follows that only about $\widetilde{O}(\ell^{-1/2})$ fraction of bit-channels are not polarized for some kernel K, which then easily translates into the bound (9) on λ_α that we desire. Note that we can always ensure that we take an invertible kernel K, since a random binary matrix is invertible with at least some constant probability.

Proving such a sharp transition constitutes the bulk of the technical work in this paper. In Section 5.2 we show that inequalities in (10) correspond to decoding a single bit of a message which is transmitted through W' using a random linear code. The first set of inequalities in (10) then correspond to saying that one can decode this single bit with low error probability with high probability over the randomness of the code, if the rate of the code is at least approximately $\ell^{-1/2}$ smaller than the capacity of the channel (where ℓ is the blocklength of the code). This follows from the well-studied fact that random linear codes achieve Shannon's capacity over any BMS ([Gal65], [BF02]).

The second set of inequalities, on the other hand, corresponds to saying that for random linear codes with rate exceeding capacity by at least $\approx \ell^{-1/2}$, even predicting a single bit of the message with tiny advantage over a uniform guess is not possible. While it follows from the converse Shannon's coding theorem that decoding the *entire* message is not possible (even with small success probability) for *any* code above capacity, it does not follow that one cannot recover *a particular message bit* with accuracy noticeably better than random guessing. In fact, if we only want to decode a specific message bit and we do not put any constraints on the code, then we can easily construct codes with rate substantially above capacity that still allow us to decode this specific message bit with high probability. All we need to do here is to repeat the message bit sufficiently many times in the codeword, decode each copy based on the corresponding channel output, and then take a majority vote. The overal code rate does not even figure in this argument. Therefore, one can only hope that the converse theorem for bit-decoding holds for certain code ensembles, and for the purpose of this paper, we restrict ourselves to random linear code ensemble. While the

converse for bit-decoding in this case is surely intuitive, establishing it in the strong quantitative form (10) that we need, and also for all BMS channels, turns out to be a challenging task. We describe some of the ideas behind our strong converse theorem for bit-decoding in Section 3.

2.7 Encoding and decoding

Once we have obtained the kernels in the code construction (see Section 2.4), the encoding procedure is pretty standard; see [PSL15, YB15, GBLB17, BBGL17, WD18] for discussions on polar codes using multiple kernels. As mentioned in Section 2.4, we can build an $N \times N$ matrix $M^{(t)} := D^{(t-1)}Q^{(t-1)}D^{(t-2)}Q^{(t-2)}\dots D^{(1)}Q^{(1)}D^{(0)}$, where the matrices $Q^{(1)}, Q^{(2)}, \dots, Q^{(t-1)}$ are some permutation matrices, and $D^{(0)}, D^{(1)}, \dots, D^{(t-1)}$ are block diagonal matrices. In particular, all the blocks on the diagonal of $D^{(j)}$ are the kernels that we obtained in Step j of the code construction. We illustrate the special case of $\ell = 3$ and $t = 2$ in Figure 1. We take a random vector $\boldsymbol{U}_{[1:N]}$ consisting of $N = \ell^t$ i.i.d. Bernoulli-1/2 random variables and we transmit the random vector $\boldsymbol{X}_{[1:N]}$ through N independent copies of W. Denote the output vector as $\boldsymbol{Y}_{[1:N]}$. Then for every $i \in [N]$, the bit-channel mapping from U_i to $(\boldsymbol{U}_{[1:i-1]}, \boldsymbol{Y}_{[1:N]})$ is exactly $W_{i_1,\dots,i_t}$, where $(i_1, \dots, i_t)$ is ℓ-ary expansion of i.

We have shown that almost all of the N bit-channels $\{W_{i_1,\dots,i_t} : i_1, \dots, i_t \in [\ell]\}$ become either noiseless or completely noisy. In the code construction, we can track $H(W_{i_1,\dots,i_t})$ for every $(i_1, \dots, i_t) \in [\ell]^t$, and this allows us to identify which U_i's are noiseless under successive decoding. Then in the encoding procedure, we only put information in these noiseless U_i's and set all the other U_i's to be some "frozen" value, e.g., 0. This is equivalent to saying that the generator matrix of our code is the submatrix of $M^{(t)}$ consisting of rows corresponding to indices i of the noiseless U_i's. In Section 9, we will show that similarly to the classical polar codes, both the encoding and decoding of our code also have $O(N \log N)$ complexity.

As a final remark, we mention that we need to quantize every bit-channel we obtain in every step of the code construction. More precisely, we merge the output symbols whose log-likelihood ratios are close to each other, so that after the quantization, the output alphabet size of every bit-channel is always polynomial in N. This allows us to construct the code in polynomial time. Without quantization, the output alphabet size would eventually be exponential in N. We will provide more details about this aspect, and how it affects the code construction and the analysis of decoding, in Section 5.1 and Section 9.

2.8 Inverse sub-exponential decoding error probability

Up to this moment, the described construction only achieved inverse polynomial decoding error probability. One reason for this restriction comes from the quantization of the bit-channels that we do, which leads to only having approximations of the actual bit-channels. In particular this means that we only track the parameters (entropy and Bhattacharyya parameter) of the bit-channels approximately, with an additive error inverse polynomial in the blocklength. This directly translates to only claiming inverse polynomial decoding error probability.

It a recent work Wang and Duursma [WD19] show that it is possible to achieve a good scaling exponent $(2 + O(\alpha))$ and inverse sub-exponential decoding error probability $(\exp(-N^{\alpha}))$ for polar codes simultaneously, using the idea of multiple kernels in the construction. However, the construction phase in [WD19] tracked the exact bit-channels that are obtained in the ℓ-ary tree of channels (without quantization), which means that the construction of such polar codes is no longer doable in polynomial time. This is because (most of) the exact bit-channels cannot even be described in

a tractable way, since they have exponential size of output alphabet.

We combine our approach of using Arıkan's kernels for polarized bit-channels (Case 1 in Section 2.6) with a strong analysis of polarization from [WD19] to achieve good scaling exponent, inverse sub-exponential decoding error probability, and polynomial time complexity of construction, all at the same time. The main idea behind our argument is that even though we cannot track the exact bit-channels in the construction, we know how basic Arıkan's kernel evolves the parameters of the bit-channels. Then, if we start with a slightly polarized bit-channel, and take a sufficient amount of "good" branches of Arıkan's 2×2 kernels, we end up with a strongly polarized channel. The crucial observation here is that it suffices to only track the approximation of the bit-channel to verify that it is slightly polarized, and no additional computation is needed to check how many "good" branches were taken in the tree of bit-channels. In such a way, we show that it is possible to prove very strong polarization for bit-channels, which leads to good decoding error probability, while still only tracking the approximations of the bit-channels, which keeps the construction complexity polynomial. All of these arguments, which lead to the main result of this paper, are made precise and proven in Section 10.

3 Outline of strong converse for random linear codes

In this section we describe the plan of the proof for the strong converse theorem for bit-decoding random linear codes under the binary symmetric channel. In particular, we need to show the sharp transition as in (10), when the channel is BSC. The proof for the general BMS channel case follows the same blueprint by using the fact that a BMS channel can be represented as a convex combination of BSC subchannels, but executing it involves overcoming several additional technical hurdles. Let us fist formulate the precise theorem for the binary symmetric channel.

Theorem 3.1. *Let W be the BSC_p channel, and let ℓ, k be integers that satisfy $\ell \geq k \geq \ell(1 - H(W)) + \Omega(\ell^{1/2} \log \ell)$ and $\ell \geq 8$. Let G be a random binary matrix uniform over $\{0,1\}^{k \times \ell}$. Suppose a message $\boldsymbol{V} \cdot G$ is transmitted through ℓ copies of the channel W, where $\boldsymbol{V}$ is uniformly random over $\{0,1\}^k$, and let $\boldsymbol{Y}$ be the output vector, i.e. $\boldsymbol{Y} = W^\ell(\boldsymbol{V} \cdot G)$. Then, with probability at least $1 - \ell^{-\Omega(\log \ell)}$ over the choice of G it holds $H\left(V_1 \mid \boldsymbol{Y}\right) \geq 1 - \ell^{-\Omega(\log \ell)}$.*

We want to point out two quantitative features of the above theorem. First, it applies at rates $\approx \Omega(\ell^{-1/2})$ above capacity. Second, it rules out predicting the bit V_1 with advantage $\ell^{-\omega(1)}$ over random guessing. Both these features are important to guarantee the desired bound $\lambda_\alpha \lesssim \ell^{-1/2}$.

Proof plan. We prove the lower bound on $H\left(V_1 \mid \boldsymbol{Y}\right)$ by lower bounding $\mathop{\mathbb{E}}_{g \sim G}\left[H\left(V_1 \mid \boldsymbol{Y}\right)\right]$ and using Markov's inequality. Thus we write

$$\mathop{\mathbb{E}}_{g \sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] = \sum_g \mathbb{P}(G = g) H^{(g)}(V_1|\boldsymbol{Y}) = \sum_g \mathbb{P}(G = g) \left(\sum_{\boldsymbol{y} \in \mathcal{Y}^\ell} \mathbb{P}^{(g)}(\boldsymbol{Y} = \boldsymbol{y}) H^{(g)}(V_1|\boldsymbol{Y} = \boldsymbol{y}) \right),$$

where the summation of g is over $\{0,1\}^{k \times \ell}$, and by $\mathbb{P}^{(g)}(\cdot)$ and $H^{(g)}(\cdot)$ we denote probability and entropy over the randomness of the message $\boldsymbol{V}$ and channel noise *for a fixed matrix g*.

1: Restrict to zero-input. The first step is to use the linearity of the (random linear) code and the additive structure of BSC to prove that we can change $\mathbb{P}^{(g)}(\boldsymbol{Y} = \boldsymbol{y})$ to $\mathbb{P}^{(g)}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{V} = \boldsymbol{0})$ in the above summation, where $\boldsymbol{0}$ is the all-zero vector. This observation is crucial for our arguments, since it allows to only consider the outputs which are "typical" for the all-zero codeword, and there

is no dependence on g in this case. Formally, in Appendix A we prove (Lemma 6.1):

$$\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] = \sum_{\boldsymbol{y}\in\mathcal{Y}^\ell} \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0}) \cdot \mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right].$$

2: Define a typical set of outputs. We define a typical output set for the zero-input as $\mathcal{F} := \left\{\boldsymbol{y}\in\mathcal{Y}^\ell \,:\, |wt(\boldsymbol{y})-\ell p| \le 2\sqrt{\ell}\log\ell\right\}$. It is clear that if zero-vector is transmitted through the channel, the output will be a vector from $\mathcal{F}$ with high probability. It means that we do not lose too much in terms of accuracy if we restrict our attention only to this typical set, so the following inequality suffices as a good lower bound on the expectation.

$$\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] \ge \sum_{\boldsymbol{y}\in\mathcal{F}} \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0}) \cdot \mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right]. \tag{11}$$

3: Fix a typical output $\boldsymbol{y}\in\mathcal{F}$. For a fixed choice of $\boldsymbol{y}\in\mathcal{F}$, we express $H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y}) = h(\mathbb{P}^{(g)}(V_1=0|\boldsymbol{Y}=\boldsymbol{y})) = h\left(\frac{\mathbb{P}^{(g)}(V_1=0,\boldsymbol{Y}=\boldsymbol{y})}{\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y})}\right)$. It suffices to show that the ratio of these probabilities is very close to $1/2$ w.h.p. To this end, we will show that both denominator and numerator are highly concentrated around their respective means for $g\sim G$, and that the means have a ratio nearly $1/2$. Focusing on the denominator (the argument for the numerator is very similar), we have:

$$2^k\cdot\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y}) = \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}\mid\boldsymbol{V}=\boldsymbol{0}) + \sum_{d=0}^{\ell} B_g(d,\boldsymbol{y})p^d(1-p)^{\ell-d}, \tag{12}$$

where $B_g(d,\boldsymbol{y})$ is equal to the number of nonzero codewords in the code spanned by the rows of g at Hamming distance d from $\boldsymbol{y}$. We proceed with proving concentration on the summation above by splitting it into two parts.

3a: Negligible part. It is very unlikely that an input codeword $\boldsymbol{x}$ such that $|\mathrm{dist}(\boldsymbol{x},\boldsymbol{y})-\ell p| \ge 6\sqrt{\ell}\log\ell$ was transmitted, if $\boldsymbol{y}$ was received as the output. It is then possible to show that the expectation (over $g\sim G$) of $\sum_{d\,:\,|d-\ell p|\ge 6\sqrt{\ell}\log\ell} B_g(d,\boldsymbol{y})p^d(1-p)^{\ell-d}$ is negligible with respect to the expectation of the whole summation. Markov's inequality implies then that this sum is negligible with high probability over $g\sim G$.

3b: Substantial part. On the other hand, for any d such that $|d-\ell p| \le 6\sqrt{\ell}\log\ell$, the expectation of $B_g(d,\boldsymbol{y})$ is going to be extremely large for the above-capacity regime. We can apply Chebyshev's inequality to prove concentration on every single weight coefficient $B_g(d,\boldsymbol{y})$ with d in such a range. A union bound then implies that they are all concentrated around their means simultaneously.

This proves that the summation over d is concentrated around its mean in (12). Finally, since $|wt(\boldsymbol{y})-\ell p| \le 2\sqrt{\ell}\log\ell$ for $\boldsymbol{y}\in\mathcal{F}$ and we leave enough room above the capacity of the channel, w.h.p. over choice of g we have $B_g(wt(\boldsymbol{y}),\boldsymbol{y}) \gg 1$, and consequently $\mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}\mid\boldsymbol{V}=\boldsymbol{0}) = p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})}$ is negligible compared to the second sum term in (12).

4: Concentration of entropy. Proving in the same way concentration on $\mathbb{P}^{(g)}(V_1=0,\boldsymbol{Y}=\boldsymbol{y})$, we derive that $\frac{\mathbb{P}^{(g)}(V_1=0,\boldsymbol{Y}=\boldsymbol{y})}{\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y})}$ is close to $\frac{1}{2}$ with high probability for any typical $\boldsymbol{y}\in\mathcal{F}$, and thus $\mathbb{E}_{g\sim G}[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})]$ is close to 1 with high probability for such $\boldsymbol{y}$. Recalling that the probability to receive $\boldsymbol{y}\in\mathcal{F}$ is overwhelming for zero-vector input, out of (11) obtain the desired lower bound on $\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right]$.

The full proof for the BSC case is presented in Section 6. In order to generalize the proof to general BMS channels we need to track and prove concentration bounds for many more parameters (in the BSC case, we had a single parameter d that was crucial). More specifically, in the BSC case we have to deal with a single binomial distribution when trying to estimate the expectation of $B_g(d, \boldsymbol{y})$. For general BMS channels, however, we have to cope with a multinomial distribution and an ensemble of binomially distributed variables that depend on the particular realization of that multinomial distribution. Moreover, we emphasize that Theorem 3.1 and its analogue for BMS must hold in the *non-asymptotic regime*, namely for all code lengths above some absolute constant which does not depend on the channel. (In contrast, in typical coding theorems in information theory one fixes the channel and lets the block length grow to infinity.) We show how to overcome all these technical challenges for the general BMS case in Section 7.

Organization of rest of the paper

The rest of the paper, which contains all the formal theorem statements and full proofs, is organized as follows. In Section 5, we describe how to find a good polarizing kernel for any BMS, and reduce its analysis to a strong coding theorem and its converse for bit-decoding of random linear codes. The case when the BMS has entropy already reasonably close to either 0 or 1 is handled in Section 8. Also, the analysis of the complexity of the kernel finding algorithm is deferred to Section 9.

Turning to the converse coding theorem for random codes, as a warmup this is first proven for the case of the binary symmetric channel in Section 6. We then present the proof for general BMS channels in Section 7. Finally, Section 9 has the complete details of our code construction based on the multiple kernels found at various levels, and a sketch of the encoding and decoding algorithms, which when all combined yield Theorem 9.6, which is almost our main result, but with decoding error probability proven to be only inverse polynomial in the blocklength.

Lastly, in Section 10 we show how to combine the tight analysis of the polarization from [WD19] and our construction of codes from Section 9 to obtain our final result, also stated in the introductory section as Theorem 1.1, with inverse sub-exponential $\exp(-N^{\alpha})$ decoding error probability.

4 Useful entropic facts

4.1 Binary entropy function

All the logarithms in this paper are to the base 2. The binary entropy function is defined as $h(x) = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x}$ for $x \in [0,1]$, where $0 \log 0$ is taken to be 0. We will use a simple fact that $h(x) \leq 2x \log \frac{1}{x}$ for $x \in [0, 1/2)$ several times in the proofs. The following proposition follows from the facts that $h(x)$ is concave, increasing for $x \in [0, 1/2)$, and symmetric around $1/2$, i.e. $h(x) = h(1-x)$ for $x \in [0,1]$.

Proposition 4.1. *For any $x, y \in [0,1]$, $|h(x) - h(y)| \leq h(|x-y|)$.*

Proof. The inequality is trivial when x or y is equal to 0 or $h(x) = h(y)$. Without loss of generality, assume $x > y$. Further, consider first the case $h(x) > h(y)$. We have two cases:

(a) $0 < y \leq (x-y) < x$. By the mean value theorem, we can write $(h(x) - h(x-y)) = h'(\xi_1)y$ for some $\xi_1 \in (x-y, x)$, and $h(y) = h(y) - h(0) = h'(\xi_2)y$ for some $\xi_2 \in (0, y)$. Then $\xi_2 \leq \xi_1$, and since h is concave, it follows that $h'(\xi_2) \geq h'(\xi_1)$, thus $h(x) - h(x-y) \leq h(y)$. Rearranging, obtain the desired inequality.

(b) $0 < (x-y) \le y < x$. By the same argument, one has $(h(x)-h(y)) = h'(\xi_1)(x-y)$ for $\xi_1 \in (y,x)$ and $h(x-y) = h(x-y)-h(0) = h'(\xi_2)(x-y)$ for some $\xi_2 \in (0, x-y)$, and so $\xi_2 \le \xi_1$, therefore $h'(\xi_2) \ge h'(\xi_1)$ by concavity. Thus $h(x)-h(y) \le h(x-y)$.

Next, if $h(x) < h(y)$, define $x' = 1-y$ and $y' = 1-x$. It follows that $x' > y'$ and $h(x') = h(y) > h(x) = h(y')$, so the inequality in the proposition holds for x' and y' by the cases (a)-(b) above. But clearly $|h(x)-h(y)| = |h(x')-h(y')| \le h(|x'-y'|) = h(|x-y|)$ by symmetry of h around $\frac{1}{2}$. □

Proposition 4.2. *$h(x) \le 2x \log \frac{1}{x}$ for $x \in [0, 1/2]$.*

Proof. Consider the function $f(x) = 2x \log \frac{1}{x} - h(x) = x \log \frac{1}{x} - (1-x) \log \frac{1}{1-x}$ on $[0, 1/2]$. We have $f''(x) = \dfrac{2x-1}{x(1-x)\ln 2} < 0$ on $(0, 1/2)$, so f is strictly concave on this interval, and further $f(0) = f(1/2) = 0$. Therefore, $f(x)$ is positive on $(0, 1/2)$. □

4.2 Channel degradation

Definition 4.3. *Let $W : \{0,1\} \to \mathcal{Y}$ and $\widetilde{W} : \{0,1\} \to \widetilde{\mathcal{Y}}$ be two BMS channels. We say that $\widetilde{W}$ is* degraded *with respect to W, or, correspondingly, W is* upgraded *with respect to $\widetilde{W}$, denoted as $\widetilde{W} \preceq W$, if there exists a discrete memoryless channel $W_1 : \mathcal{Y} \to \widetilde{\mathcal{Y}}$ such that*

$$\widetilde{W}(\widetilde{y} \,|\, x) = \sum_{y \in \mathcal{Y}} W(y \,|\, x) W_1(\widetilde{y} \,|\, y) \qquad \forall\ x \in \{0,1\},\ \widetilde{y} \in \widetilde{\mathcal{Y}}.$$

Note that this is equivalent to saying that $\widetilde{W}(x)$ and $W_1(W(x))$ are identically distributed for any $x \in \{0,1\}$. In other words, one can simulate the usage of $\widetilde{W}$ by first using the channel W and then applying some other channel W_1 to the output of W to get a final output.

We will use some useful facts from [TV13, Lemma 3] and [YB15, Lemma IV.1]. Note that Proposition 4.5 below was first proved in [KU10, Lemma 21] for the special case of Arıkan kernel and then generalized in [YB15, Lemma IV.1] to general kernels.

Proposition 4.4. *Let W and $\widetilde{W}$ be two BMS channels, such that $\widetilde{W} \preceq W$. Then $H(\widetilde{W}) \ge H(W)$.*

Proposition 4.5. *Let W and $\widetilde{W}$ be BMS channels, such that $\widetilde{W} \preceq W$, and $K \in \{0,1\}^{\ell \times \ell}$ be any invertible matrix. Denote by W_i, $\widetilde{W}_i$ the Arıkan's bit-channels of W and $\widetilde{W}$ with respect to the kernel K for any $i \in [\ell]$. Then for any $i \in [\ell]$, we have $\widetilde{W}_i \preceq W_i$, and consequently $H(\widetilde{W}_i) \ge H(W_i)$.*

5 Give me a channel, I'll give you a kernel

In this section we show that for any given binary-input memoryless symmetric (BMS) channel W we can find a kernel K of size $\ell \times \ell$, such that the Arıkan bit-channels of W with respect to this kernel will be highly polarized. By this we mean that the multiplicative decrease λ_α defined in (6) will be sufficiently close to $\ell^{-1/2}$. The algorithm (Algorithm A) to find such a kernel is as follows: if the channel is already almost noiseless or too noisy (entropy is very close to 0 or 1), we take this kernel to be a tensor power of original Arıkan's kernel for polar codes, $A_2 = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. Otherwise, the algorithm will just try out all the possible invertible kernels in $\{0,1\}^{\ell \times \ell}$, until a "good" kernel is found, which means that conditions (13) should be satisfied. Before proving that Algorithm A achieves our goals of bringing λ_α close to $\ell^{-1/2}$, we discuss several details about it.

5.1 Local kernel construction

Algorithm A: Kernel search

Input: BMS channel $\widetilde{W}$ with output size $\leq \mathsf{Q}$, error parameter Δ, and number ℓ
Output: invertible kernel $K \in \{0,1\}^{\ell \times \ell}$

if $H(\widetilde{W}) < \ell^{-4}$ **or** $H(\widetilde{W}) > 1 - \ell^{-4} + \Delta$ **then**
  return $K = A_2^{\otimes \log \ell}$
else
  for $K \in \{0,1\}^{\ell \times \ell}$, **if** K *is invertible* **do**
    Compute Arıkan's bit-channels $\widetilde{W}_i(K)$ of $\widetilde{W}$ with respect to the kernel K, as in (1)
    if

$$\begin{aligned} H(\widetilde{W}_i(K)) &\leq \ell^{-(\log \ell)/4} \quad \text{for} \quad i \geq \ell \cdot H(\widetilde{W}) + \ell^{1/2} \log^3 \ell \\ H(\widetilde{W}_i(K)) &\geq 1 - \ell^{-(\log \ell)/20} \quad \text{for} \quad i \leq \ell \cdot H(\widetilde{W}) - 14\ell^{1/2} \log^3 \ell \end{aligned} \tag{13}$$

    then
      return K
    end
  end
end

As briefly discussed at the end of Section 2.7, we are unable to efficiently track all the bit-channels in the ℓ-ary recursive tree *exactly*. This is because the size of the output alphabet of the channels increase *exponentially* after each step deeper into the tree (this simply follows from the definition of bit-channels (1)). Thus computing all the channels (and their entropies) cannot be done in poly(N) time. To overcome this issue we follow the approach of [TV13], with subsequent simplification in [GX15], of approximating the channels in the tree by degrading (see Definition 4.3) them. Degradation is achieved via the procedure of merging the output symbols, which (a) decreases the output alphabet size, and (b) does not change the entropy of the channel too much. This implies (with all the details worked out in Section 9) that we can substitute all the channels in the tree of depth t by their *degraded approximations*, such that all the channels have output alphabet size at most Q (a parameter depending on $N = \ell^t$ to be chosen), and that if $\widetilde{W}$ is a degraded approximation of the channel W in the tree, than $H(W) \leq H(\widetilde{W}) \leq H(W) + \Delta$ for some Δ depending on Q. Moreover, in Theorem 5.1 which we formulate and prove shortly, we show that when we apply Algorithm A to a degraded approximation $\widetilde{W}$ of W with small enough Δ, then, even though conditions (13) only dictate a sharp transition for $\widetilde{W}$, the same kernel will induce a sharp transition in polarization for W.

The second issue which such degraded approximation resolves is the running time of Algorithm A. Notice that we are only going to apply it for channels with output size bounded by Q, and recall that we think of ℓ as of a constant (though very large). First of all, trying out all the possible kernels will then also take a constant number of iterations. Finally, within each iteration, just calculating all the Arıkan's bit-channels and their entropies in a straightforward way will take poly(Q^ℓ) time, which is just poly(Q) when we treat ℓ as a constant. Therefore by choosing Q to be polynomial in N, the algorithm indeed works in poly(N) time.

We now leave the full details concerning the complexity of the algorithm to be handled in Section 9, and proceed with showing that Algorithm A always returns a kernel which makes λ_α from (6) close to $\ell^{-1/2}$.

Theorem 5.1. *Let $\alpha \in \left(0, \frac{1}{12}\right)$ be a small fixed constant. Let ℓ be an even power of 2 such that $\log \ell \geq \frac{11}{\alpha}$ and $\frac{\log \ell}{\log\log \ell + 2} \geq \frac{3}{\alpha}$. Let $W : \{0,1\} \to \mathcal{Y}$ and $\widetilde{W} : \{0,1\} \to \widetilde{\mathcal{Y}}$ be two BMS channels, such that $\widetilde{W} \preceq W$, $H(\widetilde{W}) - \Delta \leq H(W) \leq H(\widetilde{W})$ for some $0 \leq \Delta \leq \ell^{-\log \ell}$, and $|\widetilde{\mathcal{Y}}| \leq \mathsf{Q}$. Then Algorithm* A *on inputs $\widetilde{W}$, Δ, and ℓ returns a kernel $K \in \{0,1\}^{\ell \times \ell}$ that satisfies*

$$\frac{1}{\ell \cdot g_\alpha(H(W))} \sum_{i=1}^{\ell} g_\alpha\left(H(W_i)\right) \leq \ell^{-\frac{1}{2}+5\alpha}, \tag{14}$$

where $W_1, W_2, \ldots, W_\ell$ are the Arıkan's bit-channels of W with respect to the kernel K, and $g_\alpha(\cdot)$ is the potential function $g_\alpha(h) = (h(1-h))^\alpha$ for any $h \in [0,1]$, as defined in (4).

Proof. As we discussed above, we consider two cases:

Suction at the ends. If $H(\widetilde{W}) \notin (\ell^{-4}, 1 - \ell^{-4} + \Delta)$, Algorithm A returns a standard Arıkan's kernel $K = A_2^{\otimes \log \ell}$ on input $\widetilde{W}$ and Δ. For this case $H(W) \notin (\ell^{-4}, 1 - \ell^{-4})$, and fairly standard arguments imply that the polarization under such a kernel is much faster when the entropy is close to 0 or 1. For completeness, we present the full proofs for this case in a deferred Section 8. Specifically, Lemma 8.1 immediately implies the result of the theorem for this regime, as we pick $\log \ell \geq \frac{1}{\alpha}$.

Variance in the middle. Otherwise, if $H(\widetilde{W}) \in (\ell^{-4}, 1 - \ell^{-4} + \Delta)$, it holds $H(W) \in (\ell^{-4} - \Delta, 1 - \ell^{-4} + \Delta)$, thus $H(W) \in \left(\ell^{-4}/2, 1 - \ell^{-4}/2\right)$ since $0 \leq \Delta \leq \ell^{-\log \ell}$ and $\log \ell$ is large by the conditions of the theorem.

We first need to argue that the algorithm will at least return some kernel. This argument is one of the main technical contributions of this work, and we formulate it as Theorem 5.3 in Section 5.2. The theorem essentially claims that for any $\widetilde{W}$ an overwhelming fraction of possible kernels $K \in \{0,1\}^{\ell \times \ell}$ satisfies the conditions in (13) for $\widetilde{W}$ and K (note that we do not use any conditions on the size of $\widetilde{\mathcal{Y}}$ or the entropy $H(\widetilde{W})$ at all at this point). Clearly then, there is a decent fraction of *invertible* kernels from $\{0,1\}^{\ell \times \ell}$ which also satisfy these conditions. Therefore, the algorithm will indeed terminate and return such a good kernel. Moreover, since the theorem claims that a random kernel from $\{0,1\}^{\ell \times \ell}$ will satisfy (13) with high probability, and it is also known that it will be invertible with at least some constant probability. It means that instead of iterating through all possible kernels in step 4 of Algorithm A, we could take a random kernel and check it, and then the number of iterations needed to find a good kernel would be very small with high probability. However, to keep everything deterministic, we stick to the current approach.

Suppose now the algorithm returned an invertible kernel $K \in \{0,1\}^{\ell \times \ell}$, which means that relations (13) hold for $\widetilde{W}$ and Arıkan's bit-channels $\widetilde{W}_1, \widetilde{W}_2, \ldots, \widetilde{W}_\ell$ (we omit dependence on K from now on). Denote also $W_i = W_i(K)$ as an Arıkan's bit-channels of W with respect to K. First, since degradation is preserved after considering Arıkan's bit-channels according to Proposition 4.5, $\widetilde{W}_i \preceq W_i$, thus $H(W_i) \leq H(\widetilde{W}_i)$ for all $i \in [\ell]$. Now, similarly to the proof of Proposition 9.3, since K is invertible, conservation of entropy implies $\sum_{i=1}^{\ell} \left(H(\widetilde{W}_i) - H(W_i)\right) = \ell \left(H(\widetilde{W}) - H(W)\right) \leq \ell \cdot \Delta$, therefore derive $H(W_i) \leq H(\widetilde{W}_i) \leq H(W_i) + \ell \cdot \Delta$ for any $i \in [\ell]$. Then deduce from (13)

$$\begin{aligned} &H(W_i) \leq H(\widetilde{W}_i) \leq \ell^{-(\log \ell)/4} && \text{for} \quad i \geq \ell \cdot H(\widetilde{W}) + \ell^{1/2} \log^3 \ell \\ &H(W_i) \geq H(\widetilde{W}_i) - \ell \cdot \Delta \geq 1 - \ell^{-(\log \ell)/21} && \text{for} \quad i \leq \ell \cdot H(\widetilde{W}) - 14 \cdot \ell^{1/2} \log^3 \ell, \end{aligned} \tag{15}$$

where we used that $\Delta \leq \ell^{-\log \ell}$ and ℓ is large in the condition of the theorem.

Recall that $H(W) \in (\ell^{-4}/2, 1-\ell^{-4}/2)$ for variance in the middle regime, and note that this implies

$$g_\alpha(H(W)) \geq g_\alpha(\ell^{-4}/2) = \left(\frac{1}{2}\cdot(1-\ell^{-4}/2)\right)^\alpha \cdot \ell^{-4\alpha} \geq \left(\frac{1}{4}\right)^\alpha \ell^{-4\alpha} \geq \frac{1}{2}\ell^{-4\alpha}, \tag{16}$$

since g_α is increasing on $(0,1/2)$ and $\alpha < 1/2$. Using (15) and the trivial bound $g_\alpha(x) \leq 1$ for all the indices i close to $\ell \cdot H(\widetilde{W})$ obtain that the LHS of the desired inequality (14) is at most

$$\begin{aligned}
&\frac{1}{\ell \cdot g_\alpha(H(W))}\Bigg(\sum_{i=1}^{\ell\cdot H(\widetilde{W})-14\cdot\ell^{1/2}\log^3\ell} g_\alpha\left(1-\ell^{-(\log\ell)/21}\right) + 15\ell^{1/2}\log^3\ell \\
&\qquad + \sum_{i=\ell\cdot H(\widetilde{W})+\ell^{1/2}\log^3\ell}^{\ell} g_\alpha\left(\ell^{-(\log\ell)/4}\right)\Bigg) \\
&\overset{(a)}{<} 2\ell^{4\alpha-1}\left(15\ell^{1/2}\log^3\ell + \ell\cdot H(\widetilde{W})\cdot\ell^{-(\alpha\log\ell)/21} + (\ell-\ell\cdot H(\widetilde{W}))\cdot\ell^{-(\alpha\log\ell)/4}\right) \\
&< 30\ell^{-\frac{1}{2}+4\alpha}\log^3\ell + 2\ell^{-(\alpha\log\ell)/21+4\alpha} \\
&\overset{(b)}{\leq} \ell^{-\frac{1}{2}+4\alpha}\left(30\log^3\ell + 2\ell^{-1/42}\right) < \ell^{-\frac{1}{2}+4\alpha}\cdot 32\log^3\ell \\
&\overset{(c)}{\leq} \ell^{-\frac{1}{2}+5\alpha}
\end{aligned}$$

where (a) follows from (16) and the fact that $g_\alpha(x) = g_\alpha(1-x) \leq x^\alpha$ for $x \in (0,1)$; (b) uses the condition $\log\ell \geq \dfrac{11}{\alpha}$, and (c) uses $\dfrac{\log\ell}{\log\log\ell+2} \geq \dfrac{3}{\alpha}$ from the requirements that we have on ℓ in the conditions of this theorem. □

Remark 5.2. *In this paper, we are interested in the cases where α is very close to 0. For such α, we can absorb the two conditions on ℓ in Theorem 5.1 into one condition $\log\ell \geq \Omega(\alpha^{-1.01})$ for convenience of notation.*

5.2 Strong channel coding and converse theorems

In this section we will show that Algorithm A, which is used to prove the multiplicative decrease of almost $\ell^{-1/2}$ as in (14) in the settings of Theorem 5.1, indeed always returns some kernel for the regime when the entropy of the channel is not close to 0 or 1. While the analysis of suction at the ends regime, deferred to Section 8, follows standard methods in the literature and only relies on the fact that polarization becomes much faster when the channel is noiseless or useless, in this section we will follow the ideas from [FHMV17] and prove a *sharp transition in the polarization behaviour*, when we use a random and sufficiently large kernel.

The sharp transition stems from the fact that when the kernel K is large enough, with high probability (over randomness of K) all the Arıkan's bit-channel with respect to K, except for approximately $\ell^{1/2}$ of them in the middle, are guaranteed to be either very noisy or almost noiseless. We formulate the main result of this section in the following theorem, which was used in the proof of Theorem 5.1:

Theorem 5.3. *Let W be any BMS channel. Let $W_1, W_2, \ldots, W_\ell$ be the Arıkan's bit-channels defined in (1) with respect to the kernel K chosen uniformly at random from $\{0,1\}^{\ell\times\ell}$, where ℓ is a large integer such that $\log\ell > 40$. Then for the following inequalities all hold with probability $(1-o_\ell(1))$ over the choice of K:*

(a) $H(W_i) \leq \ell^{-(\log \ell)/4}$ *for* $i \geq \ell \cdot H(W) + \ell^{1/2} \log^3 \ell$;

(b) $H(W_i) \geq 1 - \ell^{-(\log \ell)/20}$ *for* $i \leq \ell \cdot H(W) - 14 \cdot \ell^{1/2} \log^3 \ell$.

Remark 5.4. *One can notice that the above theorem is stated for any BMS channel W, independent of the value of $H(W)$.*

The proof of this theorem relies on results concerning bit-decoding for random linear codes that are interesting beyond the connection to polar codes. The following proposition shows how to connect Arıkan's bit-channels to this context.

Proposition 5.5. *Let W be a BMS channel, $K \in \{0,1\}^{\ell \times \ell}$ be an invertible matrix, and $i \in [\ell]$. Set $k = \ell - i + 1$, and let G be a matrix which is formed by the last k rows of K. Let $\mathbf{U}$ be a random vector uniformly distributed over $\{0,1\}^\ell$, and $\mathbf{V}$ be a random vector uniformly distributed over $\{0,1\}^k$. Then*

$$H\left(U_i \,\middle|\, W^\ell(\mathbf{U} \cdot K), \mathbf{U}_{<i}\right) = H\left(V_1 \,\middle|\, W^\ell(\mathbf{V} \cdot G)\right). \tag{17}$$

We are implicitly using the concept of coset codes [Gal68, Section 6.2] in this proposition, and the proof technique here is quite standard in the polar coding literature. For example, the same proof technique is used to show that the values of the frozen bits do not matter for polar codes [Arı09, KU10]. The proof of this proposition only uses basic properties of BMS channels and linear codes, and is deferred to Appendix A. Notice now that the LHS of (17) is exactly the entropy $H(W_i)$ of the i-th Arıkan's bit-channel of W with respect to the kernel K, by definition of this bit-channel. On the other hand, one can think of the RHS of (17) in the following way: look at G as a generator matrix for a linear code of blocklength ℓ and dimension k, which is transmitted through the channel W. Then $H\left(V_1 \,\middle|\, W^\ell(\mathbf{V} \cdot G)\right)$ in some sense corresponds to how well one can decode the first bit of the message, given the output of the channel. Since in Theorem 5.3 we are interested in random kernels, the generator matrix G is also random, and thus we are indeed interested in understanding bit-decoding of random linear codes.

5.2.1 The BEC case

When W is the binary erasure channel, a statement very similar to Theorem 5.3 was established in [FHMV17]. The situation for the BEC is simpler and we now describe the intuition behind this.

Suppose we map uniformly random bits $\boldsymbol{U} \in \{0,1\}^\ell$ to $\boldsymbol{X} = \boldsymbol{U}K$ for a *random* $\ell \times \ell$ binary matrix K. We will observe $\approx (1-z)\ell$ bits of $\boldsymbol{X}$ after it passes through BEC(z); call these bits $\boldsymbol{Z}$. For a random K, with high probability the first $\approx z\ell$ bits of $\boldsymbol{U}$ will be almost independent of these observed bits $\boldsymbol{Z}$. When this happens we will have $H(W_i) = 1$ for $i \lesssim z\ell$. On the other hand, w.h.p. over the choice of K, the remaining bits U_i for $i \gtrsim z\ell$ can be uniquely determined as linear combinations of $\boldsymbol{Z}$ and $U_i, i \lesssim z\ell$, making the corresponding conditional entropies $H(W_i) = 0$. Thus except for a few exceptional indices around $i \approx z\ell$, the entropy $H(W_i)$ will be very close to 0 or 1. The formal details and quantitative aspects are non-trivial as the argument has to handle the case when z is itself close to 0 or 1, and one has to show the number of exceptional indices to be $\lesssim \sqrt{\ell}$ (which is the optimal bound). But ultimately the proof amounts to understanding the ranks of various random subspaces. When W is a BMS channel, the analysis is no longer linear-algebraic, and becomes more intricate. This is the subject of the rest of this section as well as Sections 6 and 7.

5.2.2 Part (a): channel capacity theorem

Part (a) of Theorem 5.3 corresponds to transmitting through W random linear codes with rates *below* the capacity of the channel. For this regime, it turns out that we can use the classical result that random linear codes achieve the capacity of the channel with *low error decoding probability*. Trivially, the bit-decoding error probability is even smaller, making the corresponding conditional entropy also very small. Therefore, the following theorem follows from classical Shannon's theory:

Theorem 5.6. *Let W be any BMS channel and $k \le \ell(1 - H(W)) - \ell^{1/2}\log^3 \ell$, where $\ell \ge 4$. Let G be a random binary matrix uniform over $\{0,1\}^{k\times\ell}$. Suppose a codeword $\mathbf{V}\cdot G$ is transmitted through ℓ copies of the channel W, where $\mathbf{V}$ is uniformly random over $\{0,1\}^k$, and let $\mathbf{Y}$ be the output vector, i.e. $\mathbf{Y} = W^\ell(\mathbf{V}\cdot G)$. Then with high probability over the choice of G it holds $H\left(V_1 \mid \mathbf{Y}\right) \le \ell^{-(\log \ell)/4}$.*

Proof. The described communication is just a transmission of a random linear code $C = \{\mathbf{v}G,\ \mathbf{v} \in \{0,1\}^k\}$ through W^ℓ, where the rate of the code is $R = \frac{k}{\ell} \le I(W) - \ell^{-1/2}\log^3 \ell$, so it is separated from the capacity of the channel. It is a well-studied fact that random (linear) codes achieve capacity for BMS, and moreover a tight error exponent was described by Gallager in [Gal65] and analyzed further in [BF02], [For05], [DZF16]. Specifically, one can show $\overline{P_e} \le \exp(-\ell E_r(R,W))$, where $\overline{P_e}$ is the probability of decoding error, averaged over the ensemble of all linear codes of rate R, and $E_r(R,W)$ is the so-called *random coding exponent*. It is proven in [iFLM11, Theorem 2.3] that for any BMS channel W, one has $E_r(R,W) \ge E_r^{\mathrm{BSC}}(R, I(W))$ where the latter is the error exponent for the BSC channel with the same capacity $I(W)$ as W. But the optimal scaling exponent for BSC channels for the regime when the rate is close to the capacity of the channel is given by the so-called sphere-packing exponent $E_r^{\mathrm{BSC}}(R,I) = E_{\mathrm{sp}}(R,I)$ (see, for instance, [For05, Section 1.2], which is easily shown to be almost quadratic in $(I-R)$. Specifically, we use the following

Lemma 5.7. *$E_{sp}(R,I) \ge \frac{2\log^4 \ell}{\ell}$ for $R \le I - \ell^{-1/2}\log^3 \ell$.*

Proof. For the sphere-packing exponent we use the expression from [For05, eq (1.4)]

$$E_{\mathrm{sp}}(R,I) = D_{\mathrm{KL}}\Big(\delta_{\mathrm{GV}}(R)\,||\,p\Big),$$

where $I = I(W) = 1 - H(W) = 1 - h(p)$ is the capacity of the BSC_p channel (with $p < \frac{1}{2}$), D_{KL} stands for the Kullback–Leibler divergence, and $\delta_{GV}(R)$ is the relative Gilbert-Varshamov distance, which is defined as the solution to $1 - h(\delta) = R$ for $\delta \in \left(0, \frac{1}{2}\right)$. For convenience, we will just write δ instead of $\delta_{\mathrm{GV}}(R)$ below.

For $R \le I - \ell^{-1/2}\log^3 \ell = 1 - h(p) - \ell^{-1/2}\log^3 \ell$, we then have $1 - h(\delta) \le 1 - h(p) - \ell^{-1/2}\log^3 \ell$, and so $h(\delta) - h(p) \ge \ell^{-1/2}\log^3 \ell$. Using Proposition 4.1, obtain $h(\delta - p) \ge h(\delta) - h(p) \ge \ell^{-1/2}\log^3 \ell$. Next, since $h(x)$ in increasing on $\left(0, \frac{1}{2}\right)$ and by Proposition 4.2

$$h(\ell^{-1/2}\log^2 \ell) \le 2\ell^{-1/2}\log^2 \ell \cdot \log \frac{\ell^{1/2}}{\log^2 \ell} \le 2\ell^{-1/2}\log^2 \ell \cdot \frac{1}{2}\log \ell = \ell^{-1/2}\log^3 \ell,$$

we conclude that $\delta - p \ge \ell^{-1/2}\log^2 \ell$.

Finally, we use Pinsker's inequality $D_{\mathrm{KL}}\left(P\,||\,Q\right) \ge 2\Delta^2(P,Q)$ between the KL divergence and the total variation distance $\Delta(P,Q) = \frac{1}{2}||P - Q||_1$ of two distributions P and Q over the same

probability space. Abusing the notation and denoting $\Delta(\delta, p)$ as the distance between $\mathrm{Bern}(\delta)$ and $\mathrm{Bern}(p)$, we have $\Delta(\delta, p) = |\delta - p|$, and so obtain

$$E_{\mathrm{sp}}(R, I) = D_{\mathrm{KL}}(\delta \,||\, p) \geq 2\Delta^2(\delta, p) = 2(\delta - p)^2 \geq \frac{2\log^4 \ell}{\ell}. \qquad \square$$

Therefore using this lemma we have $\overline{P_e} \leq \exp(-\ell E_r(R, W)) \leq \exp(-\ell E_{\mathrm{sp}}(R, I(W))) \leq \exp(-2\log^4 \ell)$. Then Markov's inequality implies that if we take a random linear code (i.e. choose a random binary matrix G), then with probability at least $1 - \ell^{-2}$ the decoding error is going to be at most $\ell^2 \exp(-2\log^4 \ell) \leq \exp(-\log^4 \ell) \leq \ell^{-\log \ell}$. Consider such a good linear code (matrix G), and then $\boldsymbol{V}$ can be decoded from $\boldsymbol{Y}$ with high probability, thus, clearly, V_1 can be recovered from $\boldsymbol{Y}$ with at least the same probability. Then Fano's inequality and Proposition 4.2 gives us:

$$\begin{aligned} H(V_1 \mid \boldsymbol{Y}) \leq h_2(\ell^{-\log \ell}) &\leq 2\ell^{-\log \ell} \cdot \log\left(\frac{1}{\ell^{-\log \ell}}\right) \\ &= 2\ell^{-\log \ell} \cdot \log^2 \ell \quad \leq \quad \ell^{-(\log \ell)/4}, \end{aligned}$$

where the last inequality follows from $2\log^2 \ell \leq 2^{\frac{3\log^2 \ell}{4}}$, which holds for $\ell \geq 4$. Thus we indeed obtain that the above holds with high probability (at least $1 - \ell^{-2}$, though this is very loose) over the random choice of G. $\square$

5.2.3 Part (b): strong converse for bit-decoding under noisy channel coding

On the other hand, part (b) of Theorem 5.3 concerns bit-decoding of linear codes with rates *above* the capacity of the channel. We prove that with high probability, for a random linear code with rate slightly above capacity of a BMS channel, any single bit of the input message is highly unpredictable based on the outputs of the channel on the transmitted codeword. Formally, we have the following theorem.

Theorem 5.8. *Let W be any BMS channel, and ℓ and k be integers that satisfy $\ell \geq k \geq \ell(1 - H(W)) + 14\ell^{1/2}\log^3 \ell$, and let ℓ be large enough so that $\log \ell \geq 20$. Let G be a random binary matrix uniform over $\{0,1\}^{k \times \ell}$. Suppose a message $\boldsymbol{V} \cdot G$ is transmitted through ℓ copies of the channel W, where $\boldsymbol{V}$ is uniformly random over $\{0,1\}^k$, and let $\boldsymbol{Y}$ be the output vector, i.e. $\boldsymbol{Y} = W^\ell(\boldsymbol{V} \cdot G)$. Then, with probability at least $1 - \ell^{-(\log \ell)/20}$ over the choice of G it holds $H\left(V_1 \mid \boldsymbol{Y}\right) \geq 1 - \ell^{-(\log \ell)/20}$.*

Since the theorem is of independent interest and of a fundamental nature, we devote a separate Section 7 to present a proof for it.

The above statements make the proof of Theorem 5.3 immediate:

Proof of Theorem 5.3. Denote $k = \ell - i + 1$, then by Proposition 5.5 $H(W_i) = H\left(V_1 \,\middle|\, W^\ell(\boldsymbol{V} \cdot G_k)\right)$, where $\boldsymbol{V} \sim \{0,1\}^k$ and G_k is formed by the last k rows of K. Note that since K is uniform over $\{0,1\}^{\ell \times \ell}$, this makes G_k uniform over $\{0,1\}^{k \times \ell}$ for any k. Then:

(a) For any $i \geq \ell \cdot H(W) + \ell^{1/2}\log^3 \ell$, we have $k \leq \ell(1 - H(W)) - \ell^{1/2}\log^3 \ell$, and therefore Theorem 5.6 applies, giving $H(W_i) \leq \ell^{-(\log \ell)/4}$ with probability at least $1 - \ell^{-2}$ over K.

(b) Analogically, if $i \leq \ell \cdot H(W) - 14 \cdot \ell^{1/2}\log^3 \ell$, then $k \geq \ell(1 - H(W)) + 14\ell^{1/2}\log^3 \ell$, and Theorem 5.8 gives $H(W_i) \geq 1 - \ell^{-(\log \ell)/20}$ with probability at least $1 - \ell^{-(\log \ell)/20}$ over K.

It only remains to take the union bound over all indices i as in (a) and (b) and recall that we took ℓ large enough so that $\log \ell > 40$. This implies that all of the bounds on the entropies will hold simultaneously with probability at least $1 - \ell \cdot \ell^{-2} \geq 1 - \ell^{-1}$ over the random kernel K. □

6 Strong converse for BSC_p

We present a proof of Theorem 5.8 in the next two sections. It is divided into three parts: first, we prove it for a special case of W being a BSC channel in this section. The analysis for this case is simpler (but already novel), and it provides the roadmap for the argument for the case of general BMS channel. Next, in Section 7.1 we prove Theorem 5.8 for the case when the output alphabet size of W is bounded by $2\sqrt{\ell}$, which is the main technical challenge in the paper. The proof will mimic the approach for the BSC case to some extent. Finally, in Section 7.2, we show how the case of general BMS channel can be reduced to the case of the channel with bounded alphabet via "upgraded binning" to merge output symbols.

Throughout this section consider the channel W to be BSC with the crossover probability $p \leq \frac{1}{2}$. Denote $H = H(W) = h(p)$, where $h(\cdot)$ is the binary entropy function. For the BSC case we will actually only require $k \geq \ell(1-H) + 8\sqrt{\ell}\log\ell$ in the condition of the Theorem 5.8. Thus we are in fact proving Theorem 3.1 here.

Proof of Theorem 3.1. We will follow the plan described in Section 3. As we discussed there, we prove that $H(V_1 \mid \boldsymbol{Y})$ is very close to 1 with high probability over G by showing that its expectation over G is already very close to 1 and then using Markov inequality. So we want to prove a lower bound on

$$\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] = \sum_g \mathbb{P}(G=g)H^{(g)}(V_1|\boldsymbol{Y}),$$

where $H^{(g)}(V_1|\boldsymbol{Y})$ is the conditional entropy for the fixed matrix g. Similarly, in the remainder of this section, $\mathbb{P}^{(g)}(\cdot)$ denotes probabilities of certain events *for a fixed matrix* g. By $\sum_g$ we denote the summation over all binary matrices from $\{0,1\}^{k\times\ell}$.

Restrict to zero-input. We rewrite

$$\begin{aligned}\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] &= \sum_g \mathbb{P}(G=g)\left(\sum_{\boldsymbol{y}\in\mathcal{Y}^\ell}\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y})H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right)\\ &= \sum_{\boldsymbol{y}\in\mathcal{Y}^\ell}\sum_g \mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y})\cdot\mathbb{P}(G=g)H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y}).\end{aligned}$$

Our first step is to prove that in the above summation we can change $\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y})$ to $\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0})$, where $\boldsymbol{0}$ is the all-zero vector. This observation is crucial for our arguments, since it allows us to only consider the outputs $\boldsymbol{y}$ which are "typical" for the all-zero codeword when approximating $\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right]$. Precisely, we prove

Lemma 6.1. *Let W be a BMS channel, ℓ and k be integers such that $k \leq \ell$. Let G be a random binary matrix uniform over $\{0,1\}^{k\times\ell}$. Suppose a message $\boldsymbol{V}\cdot G$ is transmitted through ℓ copies of W, where $\boldsymbol{V}$ is uniformly random over $\{0,1\}^k$, and let $\boldsymbol{Y}$ be the output vector $\boldsymbol{Y} = W^\ell(\boldsymbol{V}\cdot G)$. Then*

$$\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] = \sum_{\boldsymbol{y}\in\mathcal{Y}^\ell}\sum_g \mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0})\cdot\mathbb{P}(G=g)H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y}). \tag{18}$$

Note that the above lemma is formulated for any BMS channel, and we will also use it for the proof of the general case in Section 7. The proof of this lemma uses the symmetry of linear codes with respect to shifting by a codeword and additive structure of BSC, together with the fact that a BMS channel can be represented as a convex combination of several BSC subchannels. We defer the proof to Appendix A.

Note that $\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0})$ does not in fact depend on the matrix g, since $\boldsymbol{0}\cdot g=\boldsymbol{0}$, and so randomness here only comes from the usage of the channel W. Specifically, $\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0}) = p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})}$, where we denote by $wt(\boldsymbol{y})$ the Hamming weight of $\boldsymbol{y}$. Then in (18) we obtain

$$\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] = \sum_{\boldsymbol{y}\in\mathcal{Y}^\ell} p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})} \mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right].$$

Define a typical set. The above expression allows us to only consider "typical" outputs $\boldsymbol{y}$ for the all-zero input while approximating $\mathbb{E}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right]$. For the BSC case, we consider $\boldsymbol{y}$ to be typical when $|wt(\boldsymbol{y})-\ell p| \le 2\sqrt{\ell}\log\ell$. Then we can write:

$$\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] \ge \sum_{|wt(\boldsymbol{y})-\ell p|\le 2\sqrt{\ell}\log\ell} p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})} \mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right]. \tag{19}$$

Fix a typical output. Let us fix any typical $\boldsymbol{y}\in\mathcal{Y}^\ell$ such that $|wt(\boldsymbol{y})-\ell p| \le 2\sqrt{\ell}\log\ell$, and show that $\mathbb{E}_{g\sim G}[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})]$ is very close to 1. To do this, we first notice that

$$H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y}) = h\left(\frac{\mathbb{P}^{(g)}(V_1=0,\boldsymbol{Y}=\boldsymbol{y})}{\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y})}\right). \tag{20}$$

Denote $\widetilde{\boldsymbol{V}} = \boldsymbol{V}^{[2:k]}$ to be bits 2 to k of vector $\boldsymbol{V}$, and by $\widetilde{g} = g[2:k]$ the matrix g without its first row. Next we define the shifted weight distributions of the codebooks generated by g and $\widetilde{g}$:

$$\begin{aligned} B_g(d,\boldsymbol{y}) &:= |\{\boldsymbol{v}\in\{0,1\}^k\setminus\boldsymbol{0} \quad : wt(\boldsymbol{v}g+\boldsymbol{y})=d\}|,\\ \widetilde{B}_g(d,\boldsymbol{y}) &:= |\{\widetilde{\boldsymbol{v}}\in\{0,1\}^{k-1}\setminus\boldsymbol{0} : wt(\widetilde{\boldsymbol{v}}\widetilde{g}+\boldsymbol{y})=d\}|. \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\mathbb{P}^{(g)}(V_1=0,\boldsymbol{Y}=\boldsymbol{y})}{\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y})} &= \frac{\sum_{\widetilde{\boldsymbol{u}}}\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y}|V_1=0,\widetilde{\boldsymbol{V}}=\widetilde{\boldsymbol{u}})}{\sum_{\boldsymbol{u}}\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{u})}\\ &= \frac{p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})}+\sum_{d=0}^{\ell}\widetilde{B}_g(d,\boldsymbol{y})p^d(1-p)^{\ell-d}}{p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})}+\sum_{d=0}^{\ell}B_g(d,\boldsymbol{y})p^d(1-p)^{\ell-d}}. \end{aligned} \tag{21}$$

We will prove a concentration of the above expression around $1/2$, which will then imply that $H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})$ is close to 1 with high probability by (20). To do this, we will prove concentrations around means for both numerator and denominator of the above ratio. Since the following arguments work in exactly the same way, let us only consider the denominator for now.

By definition,

$$B_g(d,\boldsymbol{y}) = \sum_{\boldsymbol{v}\neq\boldsymbol{0}} \mathbb{1}[wt(\boldsymbol{v}g+\boldsymbol{y})=d]. \tag{22}$$

The expectation and variance of each summand is

$$\mathop{\mathrm{Var}}_{g\sim G}\mathbb{1}\big[wt(\boldsymbol{v}g+\boldsymbol{y})=d\big] \le \mathop{\mathbb{E}}_{g\sim G}\mathbb{1}\big[wt(\boldsymbol{v}g+\boldsymbol{y})=d\big] = \binom{\ell}{d}2^{-\ell} \qquad \forall\boldsymbol{v}\in\{0,1\}^k\setminus\boldsymbol{0}.$$

Clearly, the summands in (22) are pairwise independent. Therefore,

$$\operatorname*{Var}_{g\sim G}[B_g(d,\boldsymbol{y})] \le \mathop{\mathbb{E}}_{g\sim G}[B_g(d,\boldsymbol{y})] = (2^k-1)\binom{\ell}{d}2^{-\ell}, \tag{23}$$

and then

$$\mathop{\mathbb{E}}_{g\sim G}\left[\sum_{d=0}^{\ell} B_g(d,\boldsymbol{y})p^d(1-p)^{\ell-d}\right] = (2^k-1)2^{-\ell}\left(\sum_{d=0}^{\ell}\binom{\ell}{d}p^d(1-p)^{\ell-d}\right) = (2^k-1)2^{-\ell}.$$

Let us now show that $\sum_{d=0}^{\ell} B_g(d,\boldsymbol{y})p^d(1-p)^{\ell-d}$ is tightly concentrated around its mean for $g\sim G$. To do this, we split the range of d into two parts: when $|d-\ell p| > 6\sqrt{\ell}\log\ell$, and when $|d-\ell p| \le 6\sqrt{\ell}\log\ell$:

$$\sum_{d=0}^{\ell} B_g(d,\boldsymbol{y})p^d(1-p)^{\ell-d} = \sum_{|d-\ell p|>6\sqrt{\ell}\log\ell} B_g(d,\boldsymbol{y})p^d(1-p)^{\ell-d} + \sum_{|d-\ell p|\le 6\sqrt{\ell}\log\ell} B_g(d,\boldsymbol{y})p^d(1-p)^{\ell-d}.$$

In the proof below we will use the following multiplicative form of Chernoff bound applied to a binomial random variable:

$$\mathop{\mathbb{P}}_{X\sim\mathrm{Binom}(\ell,p)}[|X-\ell p| \ge \delta\ell p] \le 2e^{-\ell p\delta^2/3} \qquad \text{for all } 0\le\delta\le 1. \tag{24}$$

Applying this for $\delta = \frac{6\log\ell}{p\ell^{1/2}}$, we have

$$\mathop{\mathbb{P}}_{X\sim\mathrm{Binom}(\ell,p)}\left[|X-\ell p| \ge 6\sqrt{\ell}\log\ell\right] = \sum_{|d-\ell p|\ge 6\sqrt{\ell}\log\ell}\binom{\ell}{d}p^d(1-p)^{\ell-d} \le 2e^{\frac{-12\log^2\ell}{p}} < 2\ell^{-12\log\ell}. \tag{25}$$

Negligible part. Denote $Z_g(\boldsymbol{y}) = \sum_{|d-\ell p|>6\sqrt{\ell}\log\ell} B_g(d,\boldsymbol{y})p^d(1-p)^{\ell-d}$, and notice that

$$\mathop{\mathbb{E}}_{g\sim G}[Z_g(\boldsymbol{y})] = (2^k-1)2^{-\ell}\sum_{|d-\ell p|>6\sqrt{\ell}\log\ell}\binom{\ell}{d}p^d(1-p)^{\ell-d} \le (2^k-1)2^{-\ell}\cdot 2\ell^{-12\log\ell},$$

where we used (23) and (25). Then Markov's inequality gives $\mathbb{P}_{g\sim G}\left[Z_g(\boldsymbol{y}) \ge \mathop{\mathbb{E}}_{g\sim G}[Z_g(\boldsymbol{y})]\ell^{2\log\ell}\right] \le \ell^{-2\log\ell}$, and so

$$\mathbb{P}\left[Z_g(\boldsymbol{y}) < 2(2^k-1)2^{-\ell}\ell^{-10\ell\log\ell}\right] \ge 1-\ell^{-2\log\ell}.$$

Define the set

$$\mathcal{G}_1 := \{g\in\{0,1\}^{k\times\ell} \,:\, Z_g(\boldsymbol{y}) < 2(2^k-1)2^{-\ell}\ell^{-10\ell\log\ell}\}, \tag{26}$$

and then $\mathop{\mathbb{P}}_{g\sim G}[g\in\mathcal{G}_1] \ge 1-\ell^{-2\log\ell}$.

Substantial part. Now we deal with the part when $|d-\ell p| \le 6\sqrt{\ell}\log\ell$. For now, let us fix any d in this interval, and use Chebyshev's inequality together with (23):

$$\begin{aligned}\mathop{\mathbb{P}}_{g\sim G}\left[\left|B_g(d,\boldsymbol{y}) - \mathbb{E}[B_g(d,\boldsymbol{y})]\right| \ge \ell^{-2\log\ell}\,\mathbb{E}[B_g(d,\boldsymbol{y})]\right] &\le \frac{\operatorname{Var}[B_g(d,\boldsymbol{y})]}{\ell^{-4\log\ell}\,\mathbb{E}^2[B_g(d,\boldsymbol{y})]} \\ &\le \frac{\ell^{4\log\ell}}{\mathop{\mathbb{E}}_{g\sim G}[B_g(d,\boldsymbol{y})]} \le \ell^{4\log\ell}\,\frac{2^{\ell-k+1}}{\binom{\ell}{d}}.\end{aligned} \tag{27}$$

We use the following bound on the binomial coefficients

Fact 6.2 ([MS77], Chapter 10, Lemma 7)**.** *For any integer* $0 \le d \le \ell$*,*

$$\frac{1}{\sqrt{2\ell}} 2^{\ell h(d/\ell)} \le \binom{\ell}{d} \le 2^{\ell h(d/\ell)} \tag{28}$$

Since we fixed $|d - \ell p| \le 6\sqrt{\ell}\log\ell$, Propositions 4.1 and 4.2 imply

$$\left| h(p) - h\left(\frac{d}{\ell}\right) \right| \le h(6\ell^{-1/2}\log\ell) \le 12\ell^{-1/2}\log\ell \cdot \log\frac{\ell^{1/2}}{6\log\ell} \le 6\ell^{-1/2}\log^2\ell. \tag{29}$$

Recalling that we consider the above-capacity regime with $k \ge \ell(1-h(p)) + 8\sqrt{\ell}\log^2\ell$, we derive from (28) and (29)

$$\frac{2^{\ell-k+1}}{\binom{\ell}{d}} \le \sqrt{2\ell} \cdot 2^{\ell\left[h(p) - h\left(\frac{d}{\ell}\right) - 8\ell^{-1/2}\log^2\ell\right]} \le \sqrt{2\ell} \cdot 2^{-2\ell^{1/2}\log^2\ell}.$$

Therefore, we get in (27):

$$\mathop{\mathbb{P}}_{g\sim G}\left[\left| B_g(d,\boldsymbol{y}) - \mathbb{E}[B_g(d,\boldsymbol{y})] \right| \ge \ell^{-2\log\ell}\,\mathbb{E}[B_g(d,\boldsymbol{y})] \right] \le \sqrt{2\ell} \cdot \ell^{4\log\ell}\, 2^{-2\ell^{1/2}\log^2\ell} \le \ell^{-\sqrt{\ell}-1}, \tag{30}$$

where the last inequality holds since $\ell \ge 8$. Finally, denote

$$\mathcal{G}_2 := \left\{ g \in \{0,1\}^{k\times\ell} \,:\, \left| B_g(d,\boldsymbol{y}) - \mathbb{E}[B_g(d,\boldsymbol{y})] \right| \le \ell^{-2\log\ell}\,\mathbb{E}[B_g(d,\boldsymbol{y})] \quad \text{for all } |d-\ell p| \le 6\sqrt{\ell}\log\ell \right\}. \tag{31}$$

Then by a simple union bound applied to (30) for all d such that $|d - \ell p| \le 6\sqrt{\ell}\log\ell$ we obtain

$$\mathop{\mathbb{P}}_{g\sim G}[g \in \mathcal{G}_2] \ge 1 - \ell^{-\sqrt{\ell}}.$$

We are now ready to combine these bounds to get the needed concentration.

Lemma 6.3. *Fix* $\boldsymbol{y}$*. With probability at least* $1 - 2\ell^{-2\log\ell}$ *over the choice of* $g \sim G$*, it holds that*

$$(2^k-1)2^{-\ell}(1 - 2\ell^{-2\log\ell}) \le \sum_{d=0}^{\ell} B_g(d,\boldsymbol{y}) p^d (1-p)^{\ell-d} \le (2^k-1)2^{-\ell}(1 + 2\ell^{-2\log\ell}). \tag{32}$$

Proof. Indeed, by union bound $\mathbb{P}_{g\sim G}[g \in \mathcal{G}_1 \cap \mathcal{G}_2] \ge 1 - \ell^{-2\log\ell} - \ell^{-\sqrt{\ell}} \ge 1 - 2\ell^{-2\log\ell}$. But for any $g \in \mathcal{G}_1 \cap \mathcal{G}_2$ we derive

$$\begin{aligned}
\sum_{d=0}^{\ell} B_g(d,\boldsymbol{y}) p^d (1-p)^{\ell-d} &\ge \sum_{|d-\ell p| \le 6\sqrt{\ell}\log\ell} B_g(d,\boldsymbol{y}) p^d (1-p)^{\ell-d} \\
&\overset{(a)}{\ge} (2^k-1)2^{-\ell}(1-\ell^{-2\log\ell}) \sum_{|d-\ell p| \le 6\sqrt{\ell}\log\ell} \binom{\ell}{d} p^d (1-p)^{\ell-d} \\
&\overset{(b)}{\ge} (2^k-1)2^{-\ell}(1-\ell^{-2\log\ell})(1-2\ell^{-12\log\ell}) \\
&\ge (2^k-1)2^{-\ell}(1-2\ell^{-2\log\ell}),
\end{aligned}$$

where (a) follows from (31) (since $g \in \mathcal{G}_2$) and the expression in (23) for $\mathbb{E}[B_g(d, \boldsymbol{y})]$, and (b) uses the concentration inequality for binomial r.v. from (25). On the other hand, we can upper bound this expression as

$$\begin{aligned}
&\sum_{d=0}^{\ell} B_g(d,\boldsymbol{y}) p^d (1-p)^{\ell-d} \\
&\qquad = \sum_{|d-\ell p| \le 6\sqrt{\ell}\log\ell} B_g(d,\boldsymbol{y}) p^d (1-p)^{\ell-d} + \sum_{|d-\ell p| > 6\sqrt{\ell}\log\ell} B_g(d,\boldsymbol{y}) p^d (1-p)^{\ell-d} \\
&\qquad \overset{(a)}{\le} (2^k-1)2^{-\ell}(1+\ell^{-2\log\ell}) \sum_{|d-\ell p| \le 6\sqrt{\ell}\log\ell} \binom{\ell}{d} p^d (1-p)^{\ell-d} + Z_g(\boldsymbol{y}) \\
&\qquad \overset{(b)}{\le} (2^k-1)2^{-\ell}(1+\ell^{-2\log\ell}) + 2(2^k-1)2^{-\ell}\ell^{-10\log\ell} \\
&\qquad \le (2^k-1)2^{-\ell}(1+2\ell^{-2\log\ell}),
\end{aligned}$$

where (a) is again from (31) and (23) and the notation $Z_g(\boldsymbol{y})$ for the negligible part, and (b) is from (26) (as g is in $\mathcal{G}_1$). □

We similarly obtain the concentration for the sum in the numerator of (21): with probability at least $1 - 2\ell^{-2\log\ell}$ over the choice of g, it holds

$$(2^{k-1}-1)2^{-\ell}(1-2\ell^{-2\log\ell}) \le \sum_{d=0}^{\ell} \widetilde{B}_g(d,\boldsymbol{y}) p^d (1-p)^{\ell-d} \le (2^{k-1}-1)2^{-\ell}(1+2\ell^{-2\log\ell}). \tag{33}$$

Next, let us use the fact that we took a typical output $\boldsymbol{y}$ with $|wt(\boldsymbol{y}) - \ell p| \le 2\sqrt{\ell}\log\ell$ to show that the terms $p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})}$ are negligible in both numerator and denominator of (21). We have

$$p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})} = \left(\frac{1-p}{p}\right)^{\ell p - wt(\boldsymbol{y})} \cdot p^{\ell p}(1-p)^{\ell-\ell p} = 2^{(\ell p - wt(\boldsymbol{y}))\cdot\log\left(\frac{1-p}{p}\right)} \cdot 2^{-\ell h(p)}. \tag{34}$$

Simple case analysis gives us:

(a) If $p < \frac{1}{\sqrt{\ell}}$, then $(\ell p - wt(\boldsymbol{y}))\cdot\log\left(\frac{1-p}{p}\right) \le \ell p \log\frac{1}{p} < \ell\frac{1}{\sqrt{\ell}}\log\sqrt{\ell} < \sqrt{\ell}\log^2\ell$;

(b) In case $p \ge \frac{1}{\sqrt{\ell}}$, obtain $(\ell p - wt(\boldsymbol{y}))\cdot\log\left(\frac{1-p}{p}\right) \le 2\sqrt{\ell}\log\ell\cdot\log\frac{1}{p} \le \sqrt{\ell}\log^2\ell$.

Using the above in (34) we derive for $k \ge \ell(1-h(p)) + 8\sqrt{\ell}\log^2\ell$

$$p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})} \le 2^{\sqrt{\ell}\log^2\ell - \ell h(p)} \le 2^{2\sqrt{\ell}\log^2\ell - \ell h(p) - 2\log^2\ell - 2} \le \ell^{-2\log\ell}(2^{k-1}-1)2^{-\ell}.$$

Combining this with (32) and (33) and using a union bound we derive that with probability at least $1 - 4\ell^{-2\log\ell}$ it holds

$$\left|\left(p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})} + \sum_{d=0}^{\ell} B_g(d,\boldsymbol{y}) p^d (1-p)^{\ell-d}\right) - (2^k-1)2^{-\ell}\right| \le 3\ell^{-2\log\ell}\cdot(2^k-1)2^{-\ell},$$

$$\left|\left(p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})} + \sum_{d=0}^{\ell} \widetilde{B}_g(d,\boldsymbol{y}) p^d (1-p)^{\ell-d}\right) - (2^{k-1}-1)2^{-\ell}\right| \le 3\ell^{-2\log\ell}\cdot(2^{k-1}-1)2^{-\ell}.$$

Therefore, with probability at least $1-4\ell^{-2\log\ell}$ the expression in (21) is bounded as

$$\frac{(1-3\ell^{-2\log\ell})(2^{k-1}-1)2^{-\ell}}{(1+3\ell^{-2\log\ell})(2^k-1)2^{-\ell}} \le \frac{\mathbb{P}^{(g)}(V_1=0,\boldsymbol{Y}=\boldsymbol{y})}{\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y})} \le \frac{(1+3\ell^{-2\log\ell})(2^{k-1}-1)2^{-\ell}}{(1-3\ell^{-2\log\ell})(2^k-1)2^{-\ell}}. \tag{35}$$

We can finally derive:

$$\begin{aligned}\frac{(1-3\ell^{-2\log\ell})(2^{k-1}-1)}{(1+3\ell^{-2\log\ell})(2^k-1)} &\ge (1-6\ell^{-2\log\ell})\left(\frac{1}{2}-2^{-k}\right) \ge (1-6\ell^{-2\log\ell})\left(\frac{1}{2}-\ell^{-8\sqrt{\ell}\log\ell}\right)\\ &\ge \frac{1}{2}-\ell^{-\log\ell}, \qquad (36)\\ \frac{(1+3\ell^{-2\log\ell})(2^{k-1}-1)}{(1-3\ell^{-2\log\ell})(2^k-1)} &\le (1+9\ell^{-2\log\ell})\frac{1}{2} \le \frac{1}{2}+\ell^{-\log\ell}.\end{aligned}$$

Therefore, with probability at least $1-4\ell^{-2\log\ell}$ over $g\sim G$ it holds

$$\left|\frac{\mathbb{P}^{(g)}(V_1=0,\boldsymbol{Y}=\boldsymbol{y})}{\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y})}-\frac{1}{2}\right| \le \ell^{-\log\ell}. \tag{37}$$

Since $h(1/2+x)\ge 1-4x^2$ for any $x\in[-1/2,1/2]$ ([Top01, Theorem 1.2]), we then derive:

$$\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right] = \mathop{\mathbb{E}}_{g\sim G}\left[h\left(\frac{\mathbb{P}^{(g)}(V_1=0,\boldsymbol{Y}=\boldsymbol{y})}{\mathbb{P}^{(g)}(\boldsymbol{Y}=\boldsymbol{y})}\right)\right] \ge (1-4\ell^{-2\log\ell})(1-4\ell^{-2\log\ell}) \ge 1-8\ell^{-2\log\ell}.$$

Concentration of entropy. We are now ready to plug this into (19):

$$\begin{aligned}\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] &\ge (1-8\ell^{-2\log\ell}) \sum_{|wt(\boldsymbol{y})-\ell p|\le 2\sqrt{\ell}\log\ell} p^{wt(\boldsymbol{y})}(1-p)^{\ell-wt(\boldsymbol{y})}\\ &= (1-8\ell^{-2\log\ell}) \sum_{|d-\ell p|\le 2\sqrt{\ell}\log\ell} \binom{\ell}{d} p^d(1-p)^{\ell-d}\\ &= (1-8\ell^{-2\log\ell}) \mathop{\mathbb{P}}_{X\sim\mathrm{Binom}(\ell,p)}\left[|X-\ell p|\le 2\sqrt{\ell}\log\ell\right]\\ &\ge (1-8\ell^{-2\log\ell})(1-2e^{-(4\log^2\ell)/3p})\\ &\ge (1-8\ell^{-2\log\ell})(1-2\ell^{-2\log\ell})\\ &\ge 1-10\ell^{-2\log\ell}, \qquad (38)\end{aligned}$$

where the second inequality is obtained from the Chernoff bound (24) with $\delta=\frac{2\log\ell}{p\ell^{1/2}}$, and the third inequality follows from $p\le 1/2$ and $e^{-8/3}<2^{-2}$. Finally, using the fact that $H^{(g)}(V_1|\boldsymbol{Y})\le 1$, Markov's inequality, and (38), we get

$$\mathop{\mathbb{P}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\le 1-\ell^{-\log\ell}\right] = \mathop{\mathbb{P}}_{g\sim G}\left[1-H^{(g)}(V_1|\boldsymbol{Y})\ge \ell^{-\log\ell}\right] \le \frac{\mathop{\mathbb{E}}_{g\sim G}\left[1-H^{(g)}(V_1|\boldsymbol{Y})\right]}{\ell^{-\log\ell}} \le 10\ell^{-\log\ell}.$$

Thus we conclude that with probability at least $1-10\ell^{-\log\ell}$ over the choice of the kernel G it holds that $H(V_1\,|\,\boldsymbol{Y})\ge 1-\ell^{-\log\ell}$ when $k\ge \ell(1-h(p))+8\sqrt{\ell}\log^2\ell$ and the underlying channel is BSC. This completes the proof of Theorem 3.1, which is a version of Theorem 5.8 for the BSC case. □

7 Strong converse for BMS channel

To make this section completely self-contained, we restate the theorem here:

Theorem 5.8. *Let W be any BMS channel, and ℓ and k be integers that satisfy $\ell \geq k \geq \ell(1 - H(W)) + 14\ell^{1/2}\log^3 \ell$, and let ℓ be large enough so that $\log \ell \geq 20$. Let G be a random binary matrix uniform over $\{0,1\}^{k\times\ell}$. Suppose a message $\boldsymbol{V}\cdot G$ is transmitted through ℓ copies of the channel W, where $\boldsymbol{V}$ is uniformly random over $\{0,1\}^k$, and let $\boldsymbol{Y}$ be the output vector, i.e. $\boldsymbol{Y} = W^\ell(\boldsymbol{V}\cdot G)$. Then, with probability at least $1 - \ell^{-(\log\ell)/20}$ over the choice of G it holds $H\left(V_1 \mid \boldsymbol{Y}\right) \geq 1 - \ell^{-(\log\ell)/20}$.*

7.1 Bounded alphabet size

This section is devoted to proving Theorem 5.8 for the case when $W : \{0,1\} \to \mathcal{Y}$ is a BMS channel which has a bounded output alphabet size, specifically we consider $|\mathcal{Y}| \leq 2\sqrt{\ell}$. We will use the fact that any BMS can be viewed as a convex combination of BSCs (see for example [LH06, Kor09]), and generalize the ideas of the previous section. Namely, think of the channel W as follows: it has m possible underlying BSC subchannels $W^{(1)}, W^{(2)}, \ldots, W^{(m)}$. On any input, W randomly chooses one of the subchannels it is going to use with probabilities $q_1, q_2, \ldots, q_m$ respectively. The subchannel $W^{(j)}$ has crossover probability p_j, and without loss of generality $0 \leq p_1 \leq p_2 \leq \cdots \leq p_m \leq \frac{1}{2}$. The subchannel $W^{(j)}$ has two possible output symbols $z_j^{(0)}$ or $z_j^{(1)}$, corresponding to 0 and 1, respectively (i.e. 0 goes to $z_j^{(0)}$ with probability $1-p_j$, or to $z_j^{(1)}$ with probability p_j under $W^{(j)}$). Then the whole output alphabet is $\mathcal{Y} = \{z_1^{(0)}, z_1^{(1)}, z_2^{(0)}, z_2^{(1)}, \ldots, z_m^{(0)}, z_m^{(1)}\}$, $|\mathcal{Y}| = 2m \leq 2\sqrt{\ell}$. For the conditional entropy of the BMS channel W we have $H(W) = \sum\limits_{i=1}^{m} q_i h(p_i)$, i.e. it is a convex combination of entropies of the subchannels $W^{(1)}, W^{(2)}, \ldots, W^{(m)}$ with the corresponding coefficients $q_1, q_2, \ldots, q_m$.

Remark 7.1. *Above we ignored the case when some of the subchannels have only one output (i.e. BEC subchannels), see [TV13, Lemma 4] for a proof that we can do this without loss of generality.*

Notations and settings. In this section the expectation is only going to be taken over the kernel $g \sim G$, so we omit this in some places. As in the BSC case, by $\mathbb{P}^{(g)}[\cdot]$ and $H^{(g)}(\cdot)$ we denote the probability and entropy only over the randomness of the channel and the message, *for a fixed kernel g.*

For any possible output $\boldsymbol{y} \in \mathcal{Y}^\ell$ we denote by d_i the number of symbols from $\{z_i^{(0)}, z_i^{(1)}\}$ it has (i.e. the number of uses of the $W^{(i)}$ subchannel), so $\sum_{i=1}^m d_i = \ell$. Let also t_i be the number of symbols $z_i^{(1)}$ in $\boldsymbol{y}$. Then

$$\mathbb{P}[\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{V} = \boldsymbol{0}] = \prod_{i=1}^{m} q_i^{d_i} p_i^{t_i} (1-p_i)^{d_i - t_i}. \tag{39}$$

For this case of bounded output alphabet size, we will consider the above-capacity regime when $k \geq \ell(1 - H(W)) + 13\ell^{1/2}\log^3\ell$ (note that this is made intentionally weaker than the condition in Theorem 5.8).

We will follow the same blueprint of the proof for BSC from Section 3, however all the technicalities along the way are going to be more challenging. In particular, while we were dealing with one

binomial distribution in Section 6, here we will face a multinomial distribution of $(d_1, d_2, \ldots, d_m)$ as a choice of which subchannels to use, as well as binomial distributions $t_i \sim \text{Binom}(d_i, p_i)$ which correspond to "flips" within one subchannel.

Proof of Theorem 5.8. As in the BSC case, we are going to lower bound the expectation of $H^{(g)}(V_1|\boldsymbol{Y})$ and use Markov's inequality afterwards.

Restrict to zero-input. We use Lemma 6.1 to write

$$\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] = \sum_{\boldsymbol{y}\in\mathcal{Y}^\ell} \mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0}] \mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right]. \tag{40}$$

Notice that there is no dependence of $\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0}]$ on the kernel g, since the output for the zero-input depends only on the randomness of the channel.

Typical output set

As for the binary case, we would like to consider the set of "typical" outputs (for input $\boldsymbol{0}$) from $\mathcal{Y}^\ell$. We define $\boldsymbol{y}\in\mathcal{Y}^\ell$ to be typical if

$$\sum_{i=1}^{m}(\ell\cdot q_i - d_i)h(p_i) \le 2\sqrt{\ell}\log\ell, \tag{41}$$

$$\sum_{i=1}^{m}(p_i d_i - t_i)\log\left(\frac{1-p_i}{p_i}\right) \le 3\sqrt{\ell}\log^2\ell. \tag{42}$$

By typicality of this set we mean the following

Lemma 7.2. $\sum\limits_{\boldsymbol{y}\ typical} \mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0}] \ge 1-\ell^{-\log\ell}$. *In other words, on input* $\boldsymbol{0}$*, the probability to get the output string which is not typical is at most* $\ell^{-\log\ell}$.

We defer the proof of this lemma until Section 7.1.3, after we see why we are actually interested in these conditions on $\boldsymbol{y}$.

7.1.1 Fix a typical output

For this part, let us fix one $\boldsymbol{y}\in\mathcal{Y}^\ell$ which is typical and prove that $\mathbb{E}_g\left[H^{(g)}(V_1|\boldsymbol{Y})\right]$ is very close to 1. We have

$$H^{(g)}(V_1|\boldsymbol{Y}) = h\left(\frac{\mathbb{P}^{(g)}\left[V_1=0,\boldsymbol{Y}=\boldsymbol{y}\right]}{\mathbb{P}^{(g)}\left[\boldsymbol{Y}=\boldsymbol{y}\right]}\right). \tag{43}$$

Similarly to the BSC case, we will prove that both the denominator and numerator of the fraction inside the entropy function above are tightly concentrated around their means. The arguments for the denominator and the numerator are almost exactly the same, so we only consider denominator for now.

Concentration for $\mathbb{P}^{(g)}[\boldsymbol{Y} = \boldsymbol{y}]$

Define now the shifted weight distributions for the codebook g with respect to m different underlying BSC channels. First, for any $x \in \{0,1\}^\ell$ and $i = 1, 2, \ldots, m$, define

$$\mathrm{dist}_i(x, \boldsymbol{y}) = |\{\text{positions } j \text{ such that } (x_j = 0, \boldsymbol{y}_j = z_i^{(1)}) \text{ or } (x_j = 1, \boldsymbol{y}_j = z_i^{(0)})\}|.$$

That is, if you send x through W^ℓ and receive $\boldsymbol{y}$, then $\mathrm{dist}_i(x, \boldsymbol{y})$ is just the number of coordinates where the subchannel i was chosen, and the bit was flipped.

In our settings, we now need to think of "distance" between some binary vector $x \in \{0,1\}^\ell$ and $\boldsymbol{y}$ as of an integer vector $\boldsymbol{s} = (s_1, s_2, \ldots, s_m)$, where $0 \le s_i \le d_i$ for $i \in [m]$, where $s_i = \mathrm{dist}_i(x, \boldsymbol{y})$ is just the number of flips that occurred in the usage of i^{th} subchannel when going from x to $\boldsymbol{y}$. In other words, s_i is just the Hamming distance between the parts of x and $\boldsymbol{y}$ which correspond to coordinates j where $\boldsymbol{y}_j$ is $z_i^{(0)}$ or $z_i^{(1)}$ (coming from the subchannel $W^{(i)}$).

Now we can formally define shifted weight distributions for our fixed typical $\boldsymbol{y}$. For an integer vector $\boldsymbol{s} = (s_1, s_2, \ldots, s_m)$, where $0 \le s_i \le d_i$ define

$$B_g(\boldsymbol{s}, \boldsymbol{y}) = \Big|\boldsymbol{v} \in \{0,1\}^k \setminus \boldsymbol{0} \ : \ \mathrm{dist}_i(\boldsymbol{v} \cdot g, \boldsymbol{y}) = s_i \quad \text{for } i = 1, 2, \ldots, m\Big|.$$

We can express $\mathbb{P}^{(g)}[\mathcal{Y} = \boldsymbol{y}]$ in terms of $B_g(\boldsymbol{s}, \boldsymbol{y})$ as follows:

$$2^k \cdot \mathbb{P}^{(g)}[\mathcal{Y} = \boldsymbol{y}] = \mathbb{P}[\mathcal{Y} = \boldsymbol{y} | \boldsymbol{v} = \boldsymbol{0}] + \sum_{\substack{0 \le s_j \le d_j \\ j = 1,2,\ldots,m}} B_g(\boldsymbol{s}, \boldsymbol{y}) \prod_{i=1}^m q_i^{d_i} p_i^{s_i} (1-p_i)^{d_i - s_i}, \tag{44}$$

because $\prod_{i=1}^m q_i^{d_i} p_i^{s_i} (1-p_i)^{d_i - s_i}$ is exactly the probability to get output $\boldsymbol{y}$ if a $\boldsymbol{v}$ is sent that satisfies $\mathrm{dist}_i(\boldsymbol{v} \cdot g, \boldsymbol{y}) = s_i \quad \text{for } i = 1, 2, \ldots, m$.

We have:

$$B_g(\boldsymbol{s}, \boldsymbol{y}) = \sum_{\boldsymbol{v} \neq \boldsymbol{0}} \mathbb{1}\big[\mathrm{dist}_i(\boldsymbol{v} \cdot g, \boldsymbol{y}) = s_i, \quad \forall i = 1, 2, \ldots, m\big]. \tag{45}$$

For a fixed $\boldsymbol{v}$ but uniformly random binary matrix g, the vector $\boldsymbol{v} \cdot g$ is just a uniformly random vector from $\{0,1\}^\ell$. Now, the number of vectors x in $\{0,1\}^\ell$ such that $\mathrm{dist}_i(x, \boldsymbol{y}) = s_i \ \ \forall i = 1, 2, \ldots, m$ is $\prod_{i=1}^m \binom{d_i}{s_i}$, since for any $i = 1, 2, \ldots, m$, we need to choose which of the s_i coordinates amongst the d_i uses of the subchannel $W^{(i)}$, got flipped. So

$$\mathop{\mathbb{P}}_{g \sim G}\big[\mathrm{dist}_i(\boldsymbol{v} \cdot g, \boldsymbol{y}) = s_i, \ \ \forall i = 1, 2, \ldots, m\big] = 2^{-\ell} \prod_{i=1}^m \binom{d_i}{s_i}.$$

Then for the expectation of the shifted weight distributions we obtain

$$\mathop{\mathbb{E}}_{g \sim G}[B_g(\boldsymbol{s}, \boldsymbol{y})] = \sum_{\boldsymbol{v} \neq \boldsymbol{0}} \mathop{\mathbb{P}}_{g \sim G}\big[\mathrm{dist}_i(\boldsymbol{v} \cdot g, \boldsymbol{y}) = s_i, \ \ \forall i = 1, 2, \ldots, m\big] = \frac{2^k - 1}{2^\ell} \prod_{i=1}^m \binom{d_i}{s_i}. \tag{46}$$

Then for the expectation of the summation in the RHS of (44) we have:

$$\begin{aligned}
E \coloneqq & \mathop{\mathbb{E}}_{g\sim G}\left[\sum_{\substack{0\le s_j\le d_j\\ j=1,2,\dots,m}} B_g(\boldsymbol{s},\boldsymbol{y})\prod_{i=1}^{m} q_i^{d_i} p_i^{s_i}(1-p_i)^{d_i-s_i}\right] \\
=& \left(\prod_{i=1}^{m} q_i^{d_i}\right)\cdot \sum_{\substack{0\le s_j\le d_j\\ j=1,2,\dots,m}} \left(\mathop{\mathbb{E}}_{g\sim G}\left[B_g(\boldsymbol{s},\boldsymbol{y})\right]\cdot \prod_{i=1}^{m} p_i^{s_i}(1-p_i)^{d_i-s_i}\right) \\
=& \frac{2^k-1}{2^\ell}\left(\prod_{i=1}^{m} q_i^{d_i}\right)\cdot \sum_{\substack{0\le s_j\le d_j\\ j=1,2,\dots,m}} \prod_{i=1}^{m}\binom{d_i}{s_i} p_i^{s_i}(1-p_i)^{d_i-s_i} \\
=& \frac{2^k-1}{2^\ell}\prod_{i=1}^{m} q_i^{d_i}\cdot \prod_{i=1}^{m}\left(\underbrace{\sum_{s_i=0}^{d_i}\binom{d_i}{s_i} p_i^{s_i}(1-p_i)^{d_i-s_i}}_{=1}\right) = \frac{2^k-1}{2^\ell}\prod_{i=1}^{m} q_i^{d_i}.
\end{aligned} \tag{47}$$

Next, by (45) we can see that $B_g(\boldsymbol{s},\boldsymbol{y})$ is a sum of pairwise independent indicator random variables, since $\boldsymbol{v}_1\cdot g$ and $\boldsymbol{v}_2\cdot g$ are independent for distinct and non-zero $\boldsymbol{v}_1,\boldsymbol{v}_2$. Therefore

$$\mathop{\mathrm{Var}}_{g\sim G}[B_g(\boldsymbol{s},\boldsymbol{y})] \le \mathop{\mathbb{E}}_{g\sim G}[B_g(\boldsymbol{s},\boldsymbol{y})]. \tag{48}$$

Splitting the summation in (44)

We will split the summation in (44) into two parts: for the first part, we will show that the expectation of each term is very large, and then use Chebyshev's inequality to argue that each term is concentrated around its expectation. For the second part, its expectation is going to be very small, and Markov's inequality will imply that this part also does not deviate from its expectation too much with high probability (over the random kernel $g\sim G$). Putting these two arguments together, we will obtain that the sum in the RHS of (44) is concentrated around its mean.

To proceed, define a distribution $\Omega = \mathrm{Binom}(d_1,p_1)\times\mathrm{Binom}(d_2,p_2)\times\cdots\times\mathrm{Binom}(d_m,p_m)$, and consider a random vector $\chi\sim\Omega$. In other words, χ has m independent coordinates χ_i, $i=1,\dots,m$, where χ_i is a binomial random variable with parameters d_i and p_i. Note that by definition then for any vector $\boldsymbol{s}=(s_1,s_2,\dots,s_m)$, where $0\le s_i\le d_i$ and s_i is integer for any i, we have

$$\mathop{\mathbb{P}}_{\chi}[\chi=\boldsymbol{s}] = \prod_{i=1}^{m}\mathop{\mathbb{P}}_{\chi}[\chi_i=s_i] = \prod_{i=1}^{m}\binom{d_i}{s_i}p_i^{s_i}(1-p_i)^{d_i-s_i}.$$

Let now $\mathcal{T}$ be some subset of $\mathcal{S}=[0:d_1]\times[0:d_2]\times\cdots\times[0:d_m]$, where $[0:d]=\{0,1,2,\dots,(d-1),d\}$ for integer d. Let also $\mathcal{N}$ be $\mathcal{S}\setminus\mathcal{T}$. Then the summation in the RHS of (44) we can write as

$$\sum_{s\in\mathcal{S}} B_g(\boldsymbol{s},\boldsymbol{y})\prod_{i=1}^{m} q_i^{d_i}p_i^{s_i}(1-p_i)^{d_i-s_i} = \sum_{s\in\mathcal{T}} B_g(\boldsymbol{s},\boldsymbol{y})\prod_{i=1}^{m} q_i^{d_i}p_i^{s_i}(1-p_i)^{d_i-s_i} + \sum_{s\in\mathcal{N}} B_g(\boldsymbol{s},\boldsymbol{y})\prod_{i=1}^{m} q_i^{d_i}p_i^{s_i}(1-p_i)^{d_i-s_i}. \tag{49}$$

In the next section we describe how to choose $\mathcal{T}$.

7.1.1.(i) Substantial part

Exactly as in the binary case, using (48) and Chebyshev's inequality, we have for any $s \in \mathcal{S}$

$$\mathop{\mathbb{P}}_{g\sim G}\left[\left|B_g(\boldsymbol{s},\boldsymbol{y}) - \mathbb{E}[B_g(\boldsymbol{s},\boldsymbol{y})]\right| \geq \ell^{-2\log\ell}\,\mathbb{E}[B_g(\boldsymbol{s},\boldsymbol{y})]\right] \leq \frac{\mathrm{Var}[B_g(\boldsymbol{s},\boldsymbol{y})]}{\ell^{-4\log\ell}\,\mathbb{E}^2[B_g(\boldsymbol{s},\boldsymbol{y})]}$$
$$\leq \frac{\ell^{4\log\ell}}{\mathbb{E}_{g\sim G}[B_g(\boldsymbol{s},\boldsymbol{y})]} \leq \ell^{4\log\ell}\,\frac{2^{\ell-k+1}}{\prod_{i=1}^m \binom{d_i}{s_i}}. \quad (50)$$

We need the above to be upper bounded by $\ell^{-2\sqrt{\ell}}$ to be able to use union bound for all $\boldsymbol{s} \in \mathcal{T} \subset \mathcal{S}$, since $|\mathcal{S}| \leq \ell^{O(\sqrt{\ell})}$. Recall that we have $k \geq \ell(1-H(W)) + 13\ell^{1/2}\log^3\ell$, and then using a lower bound for binomial coefficients from Fact 6.2 we obtain for the RHS of (50)

$$\ell^{4\log\ell}\,\frac{2^{\ell-k+1}}{\prod_{i=1}^m \binom{d_i}{s_i}} \leq \ell^{4\log\ell}\cdot\left(2\prod_{i=1}^m\sqrt{2d_i}\right)\cdot 2^{\ell H(W) - \sum_{i=1}^m d_i h\left(\frac{s_i}{d_i}\right) - 13\ell^{1/2}\log^3\ell}. \quad (51)$$

We want to show that the term $2^{-\Omega(\ell^{1/2}\log^3\ell)}$ is the dominant one. First, it is easy to see that $\ell^{4\log\ell} = 2^{4\log^2\ell} \leq 2^{\ell^{1/2}\log^3\ell}$ for $\ell \geq 4$. To deal with the factor $2\prod_{i=1}^m\sqrt{2d_i}$, recall that $\sum_{i=1}^m d_i = \ell$ and $m \leq \sqrt{\ell}$ in this section (recall discussion at the beginning of Section 7.1), then AM-GM inequality gives us

$$2\prod_{i=1}^m\sqrt{2d_i} \leq 2\cdot 2^{m/2}\cdot\sqrt{\left(\frac{\sum_{i=1}^m d_i}{m}\right)^m} = 2\cdot\left(\frac{2\ell}{m}\right)^{m/2} \leq 2\cdot(2\sqrt{\ell})^{\sqrt{\ell}/2} \leq 2^{\ell^{1/2}\log^3\ell}, \quad (52)$$

where we used the fact that $(a/x)^x$ is increasing while $x \leq a/e$ and the condition $\ell \geq 4$. For the last factor of (51) we formulate a lemma.

Lemma 7.3. *There exists a set $\mathcal{T} \subseteq \mathcal{S} = [0:d_1]\times[0:d_2]\times\cdots\times[0:d_m]$, such that $\mathop{\mathbb{P}}_{\chi\sim\Omega}[\chi\in\mathcal{T}] \geq 1-\ell^{-(\log\ell)/4}$, and for any $\boldsymbol{s}\in\mathcal{T}$ it holds that*

$$\ell H(W) - \sum_{i=1}^m d_i h\left(\frac{s_i}{d_i}\right) \leq 8\,\ell^{1/2}\log^3\ell.$$

($\Omega = Binom(d_1,p_1)\times Binom(d_2,p_2)\times\cdots\times Binom(d_m,p_m)$ above)

Proof. Rearrange the above summation as follows:

$$\ell H(W) - \sum_{i=1}^m d_i h\left(\frac{s_i}{d_i}\right) = \sum_{i=1}^m\left(\ell q_i h(p_i) - d_i h\left(\frac{s_i}{d_i}\right)\right)$$
$$= \sum_{i=1}^m(\ell q_i - d_i)h(p_i) + \sum_{i=1}^m d_i\left(h(p_i) - h\left(\frac{s_i}{d_i}\right)\right).$$

Now recall that we took typical $\boldsymbol{y}$ for now, so by inequality (41) from the definition of the typicality of $\boldsymbol{y}$ we already have that the first part of the above sum is bounded by $\ell^{1/2}\log^3\ell$.

To deal with the second part, which is $\sum_{i=1}^m d_i\left(h(p_i) - h\left(\frac{s_i}{d_i}\right)\right)$, we use a separate Lemma 7.12, since the proof will be almost exactly similar for another concentration inequality we will need later. Lemma 7.12 claims that $\sum_{i=1}^m d_i\left(h(p_i) - h\left(\frac{\chi_i}{d_i}\right)\right) \leq 7\ell^{1/2}\log^3\ell$ with probability at least $1-\ell^{-(\log\ell)/4}$ over $\chi\sim\Omega$. Then the result of the current lemma follows by taking $\mathcal{T}$ to be the subset of $\mathcal{S}$ where this inequality holds. □

Fix now a set $\mathcal{T} \subseteq \mathcal{S}$ as in Lemma 7.3. Then using the arguments above we conclude that the RHS in (51), and therefore (50), is bounded above by $2^{-3\ell^{1/2}\log^3 \ell}$ for any $\boldsymbol{s} \in \mathcal{T}$. Thus we can apply union bound over $\boldsymbol{s} \in \mathcal{T}$ for (50), since $|\mathcal{T}| \leq |\mathcal{S}| = \prod_{i=1}^m (d_i+1) \leq \left(2\sqrt{\ell}\right)^{\sqrt{\ell}} \leq 2^{\ell^{1/2}\log^3 \ell}$ for $\ell \geq 4$, similarly to (52). Therefore, we derive

Corollary 7.4. *With probability at least $1 - 2^{-2\ell^{1/2}\log^3 \ell}$ (over the random kernel $g \sim G$) it holds simultaneously for all $\boldsymbol{s} \in \mathcal{T}$ that*

$$\left|B_g(\boldsymbol{s},\boldsymbol{y}) - \mathbb{E}[B_g(\boldsymbol{s},\boldsymbol{y})]\right| \leq \ell^{-2\log \ell}\,\mathbb{E}[B_g(\boldsymbol{s},\boldsymbol{y})].$$

Moreover, the set $\mathcal{N} = \mathcal{S} \setminus \mathcal{T}$ satisfies $\mathbb{P}_{\chi\sim\Omega}[\chi \in \mathcal{N}] \leq \ell^{-(\log \ell)/4}$, which we will use next section to bound the second part of (49).

7.1.1.(ii) Negligible part

Denote for convenience $Z_g(\boldsymbol{y}) = \sum_{s\in\mathcal{N}} B_g(\boldsymbol{s},\boldsymbol{y}) \prod_{i=1}^m q_i^{d_i} p_i^{s_i} (1-p_i)^{d_i - s_i}$, the second part of the RHS of (49). Recall the value of $\mathbb{E}_{g\sim G}[B_g(\boldsymbol{s}, \boldsymbol{Y})]$ from (46) and notation of E in (47). Then for the expectation of $Z_g(\boldsymbol{y})$ derive

$$\begin{aligned}
\mathop{\mathbb{E}}_{g\sim G}[Z_g(\boldsymbol{y})] &= \left(\prod_{i=1}^m q_i^{d_i}\right) \cdot \sum_{\boldsymbol{s}\in\mathcal{N}} \left(\mathop{\mathbb{E}}_{g\sim G}[B_g(\boldsymbol{s},\boldsymbol{y})] \prod_{i=1}^m p_i^{s_i}(1-p_i)^{d_i-s_i}\right) \\
&= \frac{2^k-1}{2^\ell}\left(\prod_{i=1}^m q_i^{d_i}\right) \cdot \sum_{\boldsymbol{s}\in\mathcal{N}} \prod_{i=1}^m \binom{d_i}{s_i} p_i^{s_i}(1-p_i)^{d_i-s_i} \\
&= E \cdot \mathop{\mathbb{P}}_{\chi\sim\Omega}[\chi \in \mathcal{N}] \\
&\leq E \cdot \ell^{-(\log \ell)/4}.
\end{aligned}$$

Thus Markov's inequality implies

Corollary 7.5. *With probability at least $1 - \ell^{-(\log \ell)/8}$ (over the random kernel $g \sim G$) it holds*

$$Z_g(\boldsymbol{y}) \leq \ell^{(\log \ell)/8}\,\mathbb{E}[Z_g(\boldsymbol{y})] \leq E \cdot \ell^{-(\log \ell)/8}.$$

7.1.1.(iii) Putting it together

Combining the Corollaries 7.4 and 7.5 together and using the union bound, we derive

Corollary 7.6. *With probability at least $1 - \ell^{-(\log \ell)/8} - 2^{-2\ell^{1/2}\log^3 \ell} \geq 1 - 2\ell^{-(\log \ell)/8}$ over the randomness of the kernel $g \sim G$ it simultaneously holds*

$$\begin{aligned}
&\left|B_g(\boldsymbol{s},\boldsymbol{y}) - \mathbb{E}[B_g(\boldsymbol{s},\boldsymbol{y})]\right| \leq \ell^{-2\log \ell}\,\mathbb{E}[B_g(\boldsymbol{s},\boldsymbol{y})], \qquad \text{for all } \boldsymbol{s} \in \mathcal{T}, \\
&\sum_{s\in\mathcal{N}} B_g(\boldsymbol{s},\boldsymbol{y}) \prod_{i=1}^m q_i^{d_i} p_i^{s_i}(1-p_i)^{d_i-s_i} \leq E \cdot \ell^{-(\log \ell)/8}.
\end{aligned} \tag{53}$$

We are finally ready to formulate the concentration result we need. The following lemma is an analogue of Lemma 6.3 from the BSC case:

Lemma 7.7. *With probability at least $1 - 2\ell^{-(\log \ell)/8}$ over the choice of $g \sim G$ it holds*

$$\left| \sum_{s \in \mathcal{S}} B_g(\boldsymbol{s}, \boldsymbol{y}) \prod_{i=1}^{m} q_i^{d_i} p_i^{s_i} (1-p_i)^{d_i - s_i} - E \right| \leq 2\ell^{-(\log \ell)/8} \cdot E.$$

Proof. Let us consider a kernel g such that the conditions (53) hold, which happens with probability at least $1 - 2\ell^{-(\log \ell)/8}$ according to Corollary 7.6. Then

$$\begin{aligned}
\sum_{s \in \mathcal{S}} B_g(\boldsymbol{s}, \boldsymbol{y}) \prod_{i=1}^{m} q_i^{d_i} p_i^{s_i} (1-p_i)^{d_i - s_i} &\geq \sum_{s \in \mathcal{T}} B_g(\boldsymbol{s}, \boldsymbol{y}) \prod_{i=1}^{m} q_i^{d_i} p_i^{s_i} (1-p_i)^{d_i - s_i} \\
&\overset{(53)}{\geq} \sum_{s \in \mathcal{T}} \left(1 - \ell^{-2\log \ell}\right) \mathbb{E}[B_g(\boldsymbol{s}, \boldsymbol{y})] \prod_{i=1}^{m} q_i^{d_i} p_i^{s_i} (1-p_i)^{d_i - s_i} \\
&\overset{(46)}{=} \left(1 - \ell^{-2\log \ell}\right) \frac{2^k - 1}{2^\ell} \prod_{i=1}^{m} q_i^{d_i} \cdot \sum_{s \in \mathcal{T}} \prod_{i=1}^{m} \binom{d_i}{s_i} p_i^{s_i} (1-p_i)^{d_i - s_i} \\
&= \left(1 - \ell^{-2\log \ell}\right) \cdot E \cdot \mathop{\mathbb{P}}_{\chi \sim \Omega} [\chi \in \mathcal{T}] \\
&\geq \left(1 - \ell^{-2\log \ell}\right) \left(1 - \ell^{-(\log \ell)/8}\right) E \\
&\geq \left(1 - 2\ell^{-(\log \ell)/8}\right) E.
\end{aligned}$$

For the other direction, we derive for such g

$$\begin{aligned}
\sum_{s \in \mathcal{S}} B_g(\boldsymbol{s}, \boldsymbol{y}) \prod_{i=1}^{m} q_i^{d_i} p_i^{s_i} (1-p_i)^{d_i - s_i} &= \left(\sum_{s \in \mathcal{T}} + \sum_{s \in \mathcal{N}} \right) B_g(\boldsymbol{s}, \boldsymbol{y}) \prod_{i=1}^{m} q_i^{d_i} p_i^{s_i} (1-p_i)^{d_i - s_i} \\
&\overset{(53)}{\leq} \sum_{s \in \mathcal{T}} \left(1 + \ell^{-2\log \ell}\right) \mathbb{E}[B_g(\boldsymbol{s}, \boldsymbol{y})] \prod_{i=1}^{m} q_i^{d_i} p_i^{s_i} (1-p_i)^{d_i - s_i} + E \cdot \ell^{-(\log \ell)/8} \\
&\leq \left(1 + \ell^{-2\log \ell}\right) \underbrace{\sum_{s \in \mathcal{S}} \mathbb{E}[B_g(\boldsymbol{s}, \boldsymbol{y})] \prod_{i=1}^{m} q_i^{d_i} p_i^{s_i} (1-p_i)^{d_i - s_i}}_{E} + E \cdot \ell^{-(\log \ell)/8} \\
&= \left(1 + \ell^{-2\log \ell} + \ell^{-(\log \ell)/8}\right) E \\
&\leq \left(1 + 2\ell^{-(\log \ell)/8}\right) E.
\end{aligned}$$

□

7.1.2 Concentration of entropy

We can now get a tight concentration for $\mathbb{P}^{(g)}[\boldsymbol{Y} = \boldsymbol{y}]$ using the relation (44). We already showed that the sum in RHS of (44) is tightly concentrated around its expectation, so it only remains to show that $\mathbb{P}[\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{v} = \boldsymbol{0}]$ is tiny compared to E. Here we will use that we picked $\boldsymbol{y}$ to be "typical" from the start so that (41) and (42) hold, and that we consider here the above-capacity regime. Recall (39), as well the the conditions (41) and (42) on $\boldsymbol{y}$ being typical. We derive

$$\begin{aligned}
\mathbb{P}[\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{V} = \boldsymbol{0}] = \prod_{i=1}^{m} q_i^{d_i} p_i^{t_i} (1-p_i)^{d_i - t_i} &= \prod_{i=1}^{m} \left[q_i^{d_i} \cdot p_i^{d_i p_i} (1-p_i)^{d_i (1-p_i)} \cdot \left(\frac{1-p_i}{p_i} \right)^{d_i p_i - t_i} \right] \\
&= \prod_{i=1}^{m} q_i^{d_i} \cdot \prod_{i=1}^{m} 2^{-d_i h(p_i)} \cdot \prod_{i=1}^{m} 2^{(d_i p_i - t_i) \log \left(\frac{1-p_i}{p_i} \right)}
\end{aligned}$$

$$= \prod_{i=1}^{m} q_i^{d_i} \cdot 2^{\sum_{i=1}^{m}(-\ell q_i h(p_i)+(\ell q_i - d_i)h(p_i))} \cdot 2^{\sum_{i=1}^{m}(d_i p_i - t_i)\log\left(\frac{1-p_i}{p_i}\right)}$$

$$\overset{(41),(42)}{\leq} \prod_{i=1}^{m} q_i^{d_i} \cdot 2^{-\ell H(W)+2\ell^{1/2}\log\ell+3\ell^{1/2}\log^2\ell}$$

$$\leq \prod_{i=1}^{m} q_i^{d_i} \cdot \frac{2^k-1}{2^\ell} \cdot \ell^{-\log\ell} \;=\; E \cdot \ell^{-\log\ell}, \tag{54}$$

where the last inequality follows from $k \geq \ell(1-H(W)) + 13\ell^{1/2}\log^3\ell$.

Now, combining this with Lemma 7.7, we obtain a concentration for (44):

Corollary 7.8. *With probability at least $1-2\ell^{-(\log\ell)/8}$ over the choice of kernel $g \sim G$ and for any typical $\boldsymbol{y}$*

$$\left|2^k \cdot \mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}] - E\right| \leq 3\ell^{-(\log\ell)/8} \cdot E,$$

where $E = \dfrac{2^k-1}{2^\ell}\prod_{i=1}^{m} q_i^{d_i}$.

Next, completely analogously we derive the concentration for $\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}|V_1=0]$, which is the numerator inside the entropy in (43). The only thing that changes is that we will have dimension $k-1$ instead of k for this case. We can state

Corollary 7.8′. *With probability at least $1-2\ell^{-(\log\ell)/8}$ over the choice of kernel $g \sim G$ and for any typical $\boldsymbol{y}$*

$$\left|2^k \cdot \mathbb{P}^{(g)}[V_1=0, \boldsymbol{Y}=\boldsymbol{y}] - \widetilde{E}\right| \leq 3\ell^{-(\log\ell)/8} \cdot \widetilde{E},$$

where $\widetilde{E} = \dfrac{2^{k-1}-1}{2^\ell}\prod_{i=1}^{m} q_i^{d_i}$.

Combining these two together and skipping the simple math, completely analogical to that of the BSC case in (35)–(37), we derive

Corollary 7.9. *With probability at least $1-4\ell^{-(\log\ell)/8}$ over the choice of kernel $g \sim G$ and for any typical $\boldsymbol{y}$,*

$$\left|\frac{\mathbb{P}^{(g)}[V_1=0, \boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}]} - \frac{1}{2}\right| \leq \ell^{-(\log\ell)/9}.$$

Since $h(1/2+x) \geq 1-4x^2$ for any $x \in [-1/2, 1/2]$ ([Top01, Theorem 1.2]), we then derive for a typical $\boldsymbol{y}$:

$$\mathop{\mathbb{E}}_{g}\left[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right] = \mathop{\mathbb{E}}_{g}\left[h\left(\frac{\mathbb{P}^{(g)}[V_1=0, \boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}]}\right)\right] \geq (1-4\ell^{-(\log\ell)/8}) \cdot (1-4\ell^{-(\log\ell)/9})$$
$$\geq 1 - 8\ell^{-(\log\ell)/9}.$$

Then in (40) we have

$$\begin{aligned}\mathop{\mathbb{E}}_{g}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] &= \sum_{\boldsymbol{y}\in\mathcal{Y}^\ell} \mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0}]\mathop{\mathbb{E}}_{g}\left[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right] \\ &\geq \sum_{\boldsymbol{y}\text{ typical}} \mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0}]\mathop{\mathbb{E}}_{g}\left[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right] \\ &\geq (1-\ell^{-\log\ell})\cdot(1-8\ell^{-(\log\ell)/9}). \\ &\geq 1-9\ell^{-(\log\ell)/9} \geq 1-\ell^{-(\log\ell)/10},\end{aligned} \tag{55}$$

where we used that the probability to get a typical output on a zero input is at least $1-\ell^{-\log\ell}$ by Lemma 7.2, as well as the condition $\log\ell \geq 20$.

Finally, using the fact that $H^{(g)}(V_1|\boldsymbol{Y}) \leq 1$, Markov's inequality, and (55), we get

$$\mathop{\mathbb{P}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y}) \leq 1-\ell^{-\frac{\log\ell}{20}}\right] = \mathbb{P}\left[1-H^{(g)}(V_1|\boldsymbol{Y}) \geq \ell^{-\frac{\log\ell}{20}}\right] \leq \frac{\mathbb{E}\left[1-H^{(g)}(V_1|\boldsymbol{Y})\right]}{\ell^{-(\log\ell)/20}} \leq \ell^{-(\log\ell)/20}.$$

This completes the proof of Theorem 5.8 for the case of BMS channel with bounded output alphabet size, asuming the typicality Lemma 7.2 and concentration Lemma 7.12 which we used in Lemma 7.3. We now proceed to proving these.

7.1.3 Proof that the typical set is indeed typical

Proof of Lemma 7.2. We start with proving that (41) is satisfied with high probability (over the randomness of the channel). Notice that $(d_1, d_2, \ldots, d_m)$ are multinomially distributed by construction, since for each of the ℓ bits transitioned, we choose independently the subchannel $W^{(i)}$ to use with probability q_i, for $i = 1, 2, \ldots, m$, and d_i represents the number of times the channel $W^{(i)}$ was chosen. So indeed $(d_1, d_2, \ldots, d_m) \sim \mathrm{Mult}(\ell, q_1, q_2, \ldots, q_m)$. The crucial property of multinomial random variables we are going to use is *negative association* ([JDP83], [DR96]). The (simplified version of the) fact we are going to use about negatively associated random variables can be formulated as follows:

Lemma 7.10 ([JDP83], Property P_2)**.** *Let $X_1, X_2, \ldots, X_m$ be negatively associated random variables. Then, for every set of m positive monotone non-decreasing functions $f_1, \ldots, f_m$ it holds*

$$\mathbb{E}\left[\prod_{i=1}^{m} f_i(X_i)\right] \leq \prod_{i=1}^{m} \mathbb{E}[f_i(X_i)].$$

We also use the fact that since $(d_1, d_2, \ldots, d_m)$ are negatively associated, then applying decreasing functions $g_i(x) = \ell q_i - x$ coordinate-wise to these random variables, we will also obtain negatively associated random variables ([DR96], Proposition 7). In other words, we argue that $(\ell q_1 - d_1, \ell q_2 - d_2, \ldots, \ell q_m - d_m)$ are also negatively associated, thus we can apply Lemma 7.10 to these random variables.

Let us now denote for convenience $\alpha_i = h(p_i)$ for $i = 1, 2, \ldots, m$, and so we have $0 \leq \alpha_i \leq 1$. Let also $X = \sum_{i=1}^{m}(\ell \cdot q_i - d_i)\alpha_i$, and we now can start with simple exponentiation and Markov's inequality: for any a and any $t > 0$

$$\mathbb{P}[X \geq a] = \mathbb{P}[e^{tX} \geq e^{ta}] \leq e^{-ta}\,\mathbb{E}\left[e^{tX}\right] = e^{-ta}\,\mathbb{E}\left[\prod_{i=1}^{m} e^{t\cdot\alpha_i(\ell q_i - d_i)}\right] \leq e^{-ta}\prod_{i=1}^{m}\mathbb{E}\left[e^{t\cdot\alpha_i(\ell q_i - d_i)}\right], \tag{56}$$

where in the last inequality we applied Lemma 7.10 for negatively associated random variables $(\ell q_1 - d_1, \ell q_2 - d_2, \ldots, \ell q_m - d_m)$, as discussed above, and positive non-decreasing functions $f_i(x) = e^{t \cdot \alpha_i \cdot x}$, since $\alpha_i, t \geq 0$.

Next, consider the following claim, which follows from standard Chernoff-type arguments:

Claim 7.11. *Let $Z \sim Binom(n, p)$, and let $b > 0$. Then $\mathbb{E}[e^{-b \cdot Z}] \leq e^{np \cdot (e^{-b} - 1)}$.*

Proof. We can write $Z = \sum_{j=1}^{n} Z_j$, where $Z_j \sim \mathrm{Bern}(p)$ are independent Bernoulli random variables. Then

$$\mathbb{E}\left[e^{-b \cdot Z}\right] = \mathbb{E}\left[\prod_{j=1}^{n} e^{-b \cdot Z_j}\right] = \prod_{j=1}^{n} \mathbb{E}\left[e^{-b \cdot Z_j}\right] = \left((1-p) + p \cdot e^{-b}\right)^n \leq e^{np(e^{-b}-1)}, \tag{57}$$

where the only inequality uses the fact that $1 + x \leq e^x$ for any x. □

Turning back to (56), we are going to bound the terms $\mathbb{E}\left[e^{t \cdot \alpha_i (\ell q_i - d_i)}\right]$ individually. It is clear that the marginal distribution of d_i is just $\mathrm{Binom}(\ell, q_i)$, so we are able to use Claim 7.11 for it. We derive:

$$\mathbb{E}\left[e^{t \cdot \alpha_i (\ell q_i - d_i)}\right] = e^{t \alpha_i \ell q_i} \cdot \mathbb{E}\left[e^{-t \alpha_i \cdot d_i}\right] \overset{(57)}{\leq} e^{t \cdot \alpha_i \ell q_i} \cdot e^{\ell q_i \left(e^{-t\alpha_i} - 1\right)} = e^{\ell q_i \left(t\alpha_i + e^{-t\alpha_i} - 1\right)} \leq e^{\ell q_i \left(t + e^{-t} - 1\right)}, \tag{58}$$

where the last inequality uses that $x + e^{-x}$ is increasing for $x \geq 0$ together with $0 \leq t\alpha_i \leq t$, as $t > 0$ and $0 \leq \alpha_i \leq 1$. Plugging the above into (56) and using $\sum_{i=1}^{m} q_i = 1$, we obtain

$$\mathbb{P}[X \geq a] \leq e^{-ta} \prod_{i=1}^{m} e^{\ell q_i \left(t + e^{-t} - 1\right)} = e^{-ta} \cdot e^{\ell \left(t + e^{-t} - 1\right)} \leq e^{-ta + \ell \frac{t^2}{2}}, \tag{59}$$

where we used $x + e^{-x} - 1 \leq \frac{x^2}{2}$ for any $x \geq 0$. Finally, by taking $a = 2\sqrt{\ell} \log \ell$, setting $t = a/\ell$, and recalling what we denoted by X and α_i above, we immediately deduce

$$\mathbb{P}\left[\sum_{i=1}^{m} (\ell \cdot q_i - d_i) h(p_i) \geq 2\sqrt{\ell} \log \ell\right] \leq e^{-\frac{a^2}{2\ell}} = e^{-2 \log^2 \ell} \leq \ell^{-2 \log \ell}.$$

This means that the first typicality requirement (41) holds with very high probability (over the randomness of the channel).

Let us now prove that the second typicality condition (42) holds with high probability. For that, we condition on the values of $d_1, d_2, \ldots, d_m$. We will see that (42) holds with high probability for all values of $d_1, d_2, \ldots, d_m$, and then it is clear that is will imply that it also holds with high probability overall.

So, fix the values of $d_1, d_2, \ldots, d_m$. Denote a random variable $Y = \sum_{i=1}^{m} (p_i d_i - t_i) \log\left(\frac{1-p_i}{p_i}\right)$, and then our goal it to show that Y is bounded above by $O(\sqrt{\ell} \log^2 \ell)$ with high probability (over the randomness of t_i's). Given the conditioning on $d_1, d_2, \ldots, d_m$, it is clear that $t_i \sim \mathrm{Binom}(d_i, p_i)$ for all $i = 1, 2, \ldots, m$, and they are all independent (recall that d_i corresponds to the number of times subchannel $W^{(i)}$ is chosen, while t_i corresponds to the number of "flips" within this subchannel).

We split the summation in Y into two parts: let $T_1 = \{i : p_i \leq \frac{1}{\ell}\}$ and $T_2 = [m] \setminus T_1$. Then for any realization of t_i's, we have $\sum_{i \in T_1} (p_i d_i - t_i) \log\left(\frac{1-p_i}{p_i}\right) \leq \sum_{i \in T_1} p_i d_i \log\left(\frac{1}{p_i}\right) \leq \sum_{i \in T_1} \frac{d_i \log \ell}{\ell} \leq \log \ell$.

Denote the second part of the summation as $Y_2 = \sum_{i \in T_2} (p_i d_i - t_i) \log\left(\frac{1-p_i}{p_i}\right)$. Notice that $\log\left(\frac{1-p_i}{p_i}\right) \le \log\left(\frac{1}{p_i}\right) \le \log \ell$ for $i \in T_2$. Denote then $\gamma_i = \log\left(\frac{1-p_i}{p_i}\right) / \log \ell$, and so $0 \le \gamma_i \le 1$ for $i \in T_2$. Finally, let $\widetilde{Y_2} = Y_2 / \log \ell = \sum_{i \in T_2} (p_i d_i - t_i) \cdot \gamma_i$.

We now prove the concentration on $\widetilde{Y_2}$ in almost exactly the same way as we did for X above. Similarly to (56) we obtain

$$\mathbb{P}\left[\widetilde{Y_2} > a\right] = \mathbb{P}\left[e^{t\widetilde{Y_2}} > e^{ta}\right] \le e^{-ta}\,\mathbb{E}\left[e^{t\widetilde{Y_2}}\right] = e^{-ta} \cdot \mathbb{E}\left[\prod_{i=1}^{m} e^{t \cdot \gamma_i \cdot (p_i d_i - t_i)}\right] = e^{-ta} \cdot \prod_{i=1}^{m} \mathbb{E}\left[e^{t \cdot \gamma_i \cdot (p_i d_i - t_i)}\right], \tag{60}$$

where the last equality holds because we conditioned on $d_1, d_2, \ldots, d_m$, and so $t_1, t_2, \ldots, t_m$ are independent, as discussed above. Next, Claim 7.11 applied for $t_i \sim \mathrm{Binom}(d_i, p_i)$ and $t \cdot \gamma_i > 0$ for any $t > 0$ gives $\mathbb{E}\left[e^{-t\gamma_i \cdot t_i}\right] \le e^{d_i p_i (e^{-t\gamma_i} - 1)}$, and so similarly to (57)–(59) derive from (60)

$$\mathbb{P}\left[\widetilde{Y_2} > a\right] \le e^{-ta} \cdot \prod_{i \in T_2} e^{p_i d_i \left(t\gamma_i + e^{-t\gamma_i} - 1\right)} \le e^{-ta} \cdot \prod_{i \in T_2} e^{p_i d_i \left(t + e^{-t} - 1\right)} \le e^{-ta + \sum_{i \in T_2} p_i d_i \cdot t^2/2} \le e^{-ta + \ell t^2/2}$$

for any $t > 0$, where we used $0 \le \gamma_i \le 1$ for $i \in T_2$, $p_i < 1$, and $\sum_{i \in T_2} d_i \le \ell$. Therefore, by taking again $a = 2\sqrt{\ell} \log \ell$ and $t = a/\ell$, obtain

$$\mathbb{P}\left[Y_2 \ge 2\sqrt{\ell} \log^2 \ell\right] = \mathbb{P}\left[\widetilde{Y_2} \ge 2\sqrt{\ell} \log \ell\right] \le \ell^{-2 \log \ell}.$$

Since $Y \le \log \ell + Y_2$, we conclude that $Y \le 3\sqrt{\ell} \log^2 \ell$ with probability at least $1 - \ell^{-2 \log \ell}$ over the randomness of the channel.

Since both (41) and (42) hold with probability at least $1 - \ell^{-2 \log \ell}$, the union bound implies that these two conditions hold simultaneously with probability at least $1 - 2\ell^{-2 \log \ell} \ge 1 - \ell^{-\log \ell}$. □

7.1.4 Concentration Lemma

Lemma 7.12. *Let $\chi \sim \Omega = Binom(d_1, p_1) \times Binom(d_2, p_2) \times \cdots \times Binom(d_m, p_m)$, where d_i's are nonnegative integers for $i \in [m]$, $p_i \le 1/2$, $\sum_{i=1}^{m} d_i = \ell$, and $m \le \sqrt{\ell}$. Let also ℓ be large so that $\log \ell \ge 8$. Then the following holds with probability at least $1 - \ell^{-(\log \ell)/4}$:*

$$\sum_{i=1}^{m} d_i \left(h(p_i) - h\left(\frac{\chi_i}{d_i}\right)\right) \le 7\ell^{1/2} \log^3 \ell. \tag{61}$$

Proof. First, notice that we can disregard all the indices i for which $d_i = 0$, as they do not contribute anything to the LHS of (61). So from now on, we assume for simplicity that $d_i \ge 1$ for all $i = 1, 2, \ldots, m$.

Next, we split the interval $[1 : m]$ into two parts. In the first part the value of $d_i \cdot p_i$ is going to be small, and the sum of $d_i h(p_i)$ will also be small. For the second part, when $d_i \cdot p_i$ is large enough, we will be able to apply some concentration arguments. Denote:

$$F_1 \coloneqq \left\{i \,:\, p_i \le \frac{4 \log^2 \ell}{d_i}\right\},$$
$$F_2 \coloneqq \{1, 2, \ldots, m\} \setminus F_1.$$

Then

$$\sum_{i=1}^{m} d_i \left(h(p_i) - h\left(\frac{\chi_i}{d_i}\right)\right) \le \sum_{i\in F_1} d_i h(p_i) + \sum_{i\in F_2} d_i \left(h(p_i) - h\left(\frac{\chi_i}{d_i}\right)\right). \tag{62}$$

Let us deal with the summation over F_1 first. Split this set even further: $F_1^{(1)} = \{i \in F_1 : d_i \ge 8\log^2 \ell\}$, and $F_1^{(2)} = F_1 \setminus F_1^{(1)}$. Then for any $i \in F_1^{(1)}$ we use $h(p_i) \le 2p_i \log \frac{1}{p_i}$ from Proposition 4.2, since $p_i \le 1/2$. For any $i \in F_1^{(2)}$ we just use $h(p_i) \le 1$. Combining these, obtain

$$\begin{aligned}\sum_{i\in F_1} d_i h(p_i) &\le \sum_{i\in F_1^{(1)}} 2d_i p_i \log\frac{1}{p_i} + \sum_{i\in F_1^{(2)}} d_i \le \sum_{i\in F_1^{(1)}} 8\log^2\ell \cdot \log\left(\frac{d_i}{4\log^2\ell}\right) + \left|F_1^{(2)}\right| \cdot 8\log^2\ell \\ &\le 8\log^2\ell \cdot \sum_{i\in F_1^{(1)}} \log d_i + \left|F_1^{(2)}\right| \cdot 8\log^2\ell. \end{aligned}\tag{63}$$

For the second summand in the RHS above, we will just use $\left|F_1^{(2)}\right| \le m \le \ell^{1/2}$. For the first summand, we use Jensen's inequality, the fact that $\sum_{i=1}^{m} d_i = \ell$, and $\left|F_1^{(1)}\right| \le m \le \ell^{1/2}$ to derive

$$\sum_{i\in F_1^{(1)}} \log d_i \le \left|F_1^{(1)}\right| \cdot \log\left(\frac{\sum_{i\in F_1^{(1)}} d_i}{\left|F_1^{(1)}\right|}\right) \le \left|F_1^{(1)}\right| \cdot \log\left(\frac{\ell}{\left|F_1^{(1)}\right|}\right) \le \ell^{1/2}\log\left(\ell^{1/2}\right) = \frac{1}{2}\ell^{1/2}\log\ell,$$

where the last inequality uses that $x\log(\ell/x)$ is increasing while $x \le \ell/e$. Therefore, in (63) obtain

$$\sum_{i\in F_1} d_i h(p_i) \le 8\log^2\ell \cdot \sum_{i\in F_1^{(1)}} \log d_i + \left|F_1^{(2)}\right| \cdot 8\log^2\ell \le 5\ell^{1/2}\log^3\ell, \tag{64}$$

where we also used $8 \le \log\ell$.

Therefore, the first part of the RHS of (62) is always bounded by $5\ell^{1/2}\log^3\ell$. We will now deal with the remaining summations over $i \in F_2$.

For any $i \in F_2$, we know that $d_i p_i \ge 4\log^2\ell$. Now we apply the multiplicative Chernoff bound (24) for $\chi_i \sim \mathrm{Binom}(d_i, p_i)$ and $\delta = \frac{\log\ell}{\sqrt{d_i p_i}}$ to get

$$\mathop{\mathbb{P}}_{\chi_i}\left[|\chi_i - d_i p_i| \ge \sqrt{d_i p_i}\log\ell\right] \le 2e^{-\log^2\ell/3} \le \ell^{-(\log\ell)/3} \qquad \text{if } \log\ell \le \sqrt{d_i p_i}, \tag{65}$$

where the last inequality holds for $\log\ell > 3$ because the log in the exponent is to base 2. The condition $\log\ell \le \sqrt{d_i p_i}$ is needed in order to have $\delta \le 1$ for the multiplicative Chernoff bound (24) to hold.

Then, by the union bound, we derive

$$\mathop{\mathbb{P}}_{\chi\sim\Omega}\left[|\chi_i - d_i p_i| \ge \sqrt{d_i p_i}\log\ell \text{ for some } i \in F_2\right] \le |F_2| \cdot \ell^{-(\log\ell)/3} \le \ell^{-(\log\ell)/3+1/2}. \tag{66}$$

Define the sets $\mathcal{T}_1^{(i)}$ for all $i = 1, 2, \ldots, m$ as follows:

$$\begin{aligned}\mathcal{T}_1^{(i)} &:= \left\{s_i \in [0 : d_i] \,:\, |s_i - d_i p_i| \le \sqrt{d_i p_i}\log\ell\right\}, && \text{for } i \in F_2; \\ \mathcal{T}_1^{(i)} &:= [0 : d_i], && \text{for } i \notin F_2.\end{aligned}\tag{67}$$

and let

$$\theta_i \coloneqq \mathbb{P}[\chi_i \in \mathcal{T}_1^{(i)}]. \tag{68}$$

Then by (65) we have

$$\begin{aligned} &\theta_i \geq 1 - \ell^{-(\log \ell)/3}, && \text{for } i \in F_2; \\ &\theta_i = 1, && \text{for } i \notin F_2. \end{aligned}$$

Finally, define

$$\theta \coloneqq \prod_{i=1}^{m} \theta_i = \prod_{i \in F_2} \theta_i = \prod_{i \in F_2} \mathbb{P}[\chi_i \in \mathcal{T}_1^{(i)}] = \mathop{\mathbb{P}}_{\chi \sim \Omega}[\chi_i \in \mathcal{T}_1^{(i)} \text{ for all } i \in F_2] \geq 1 - \ell^{-(\log \ell)/3 + 1/2}, \tag{69}$$

where the last inequality is a direct implication of (66).

We will now define a set of new probability distributions $\mathcal{D}_i$ for all $i = 1, 2, \ldots, m$, as binomial distributions $\mathrm{Binom}(d_i, p_i)$ restricted to intervals $\mathcal{T}_1^{(i)}$. Formally, let us write

$$\mathop{\mathbb{P}}_{\eta_i \sim \mathcal{D}_i}[\eta_i = x] = \begin{cases} 0, & \text{if } x \notin \mathcal{T}_1^{(i)}; \\ \mathbb{P}_{\chi_i \sim \mathrm{Binom}(d_i, p_i)}[\chi_i = x] \cdot \theta_i^{-1}, & \text{if } x \in \mathcal{T}_1^{(i)}. \end{cases} \tag{70}$$

(So to get $\mathcal{D}_i$ we just took a distribution $\mathrm{Binom}(d_i, p_i)$, truncated it so it does not have any mass outside of $\mathcal{T}_1^{(i)}$, and rescaled appropriately.)

Next, define a product distribution $\mathcal{D} \coloneqq \bigtimes_{i=1}^{m} \mathcal{D}_i$ on the set $\mathcal{T}_1 \coloneqq \bigtimes_{i=1}^{m} \mathcal{T}_1^{(i)}$. Notice now that it is trivial that for any subset $\mathcal{R} \subseteq \mathcal{T}_1$ it holds

$$\mathop{\mathbb{P}}_{\chi \sim \Omega}[\chi \in \mathcal{R}] = \mathop{\mathbb{P}}_{\eta \sim \mathcal{D}}[\eta \in \mathcal{R}] \cdot \theta. \tag{71}$$

Since θ is very close to 1, it suffices to prove the claims for $\mathcal{D}$ instead of Ω.

Recall that our goal was to show that $\sum_{i \in F_2} d_i \left(h(p_i) - h\left(\frac{\chi_i}{d_i}\right)\right)$ (the second part from (62)) is bounded above by $O(\ell^{1/2} \log^3 \ell)$ with high probability, when $\chi \sim \Omega$. Instead now let us show that this summation is small with high probability when $\chi \sim \mathcal{D}$, and then use the arguments above to see that there is not much of a difference when $\chi \sim \Omega$.

Claim 7.13. *Let $i \in F_2$ and $\chi_i \sim \mathcal{D}_i$. Then*

$$\left| d_i \left(h(p_i) - h\left(\frac{\chi_i}{d_i}\right)\right)\right| \leq \sqrt{d_i p_i} \log^2 \ell. \tag{72}$$

Proof. First, $\left|\frac{\chi_i}{d_i} - p_i\right| \leq \sqrt{\frac{p_i}{d_i}} \log \ell$ for $\chi_i \sim \mathcal{D}_i$ by definition of the distribution $\mathcal{D}_i$. Now, for $i \in F_2$, $p_i \geq \frac{4 \log^2 \ell}{d_i}$, from which it follows that $\frac{p_i}{2} \geq \sqrt{\frac{p_i}{d_i}} \log \ell$, and therefore $\frac{p_i}{2} \leq \frac{\chi_i}{d_i} \leq \frac{3p_i}{2}$. We then use the concavity of the binary entropy function on $[0, 1]$. For a concave differentiable function f on an interval $[a, b]$, one has $|f(b) - f(a)| \leq |b - a| \cdot \max\{|f'(a)|, |f'(b)|\}$, which follows from a standard inequality $f(y) \leq f(x) + f'(x)(y - x)$ applied for (a, b) or (b, a), depending on which of $f(a)$ and $f(b)$ is larger. We apply this for the binary entropy function $h(\cdot)$ and one of the intervals $\left[\frac{\chi_i}{d_i}, p_i\right]$ and $\left[p_i, \frac{\chi_i}{d_i}\right]$, depending on which of $\frac{\chi_i}{d_i}$ and p_i is smaller:

$$\left| h\left(\frac{\chi_i}{d_i}\right) - h(p_i)\right| \leq \left|\frac{\chi_i}{d_i} - p_i\right| \cdot \max\left\{\left|\frac{dh}{dx}(p_i)\right|, \left|\frac{dh}{dx}\left(\frac{\chi_i}{d_i}\right)\right|\right\}.$$

Now, both p_i and $\frac{\chi_i}{d_i}$ lie in the interval $\left[\frac{p_i}{2}, \frac{3p_i}{2}\right]$, which is contained in $\left[\frac{p_i}{2}, 1-\frac{p_i}{2}\right]$, as $p_i < 1/2$. Out of symmetry of h around $1/2$, it follows that the maximal value of $\left|\frac{dh}{dx}(\cdot)\right|$ on the interval $\left[\frac{p_i}{2}, 1-\frac{p_i}{2}\right]$ is attained at $\frac{p_i}{2}$. Therefore, we have

$$\begin{aligned}\left|h\left(\frac{\chi_i}{d_i}\right) - h(p_i)\right| &\le \left|\frac{\chi_i}{d_i} - p_i\right| \cdot \max\left\{\left|\frac{dh}{dx}(p_i)\right|, \left|\frac{dh}{dx}\left(\frac{\chi_i}{d_i}\right)\right|\right\} \\ &\le \sqrt{\frac{p_i}{d_i}}\log\ell \cdot \left|\frac{dh}{dx}\left(\frac{p_i}{2}\right)\right| = \sqrt{\frac{p_i}{d_i}}\log\ell \cdot \log\frac{1-p_i/2}{p_i/2} \\ &\le \sqrt{\frac{p_i}{d_i}}\log\ell \cdot \log\frac{2}{p_i} \le \sqrt{\frac{p_i}{d_i}}\log\ell \cdot \log\left(\frac{d_i}{2\log^2\ell}\right) \le \sqrt{\frac{p_i}{d_i}}\log^2\ell,\end{aligned}$$

where the penultimate inequality follows from $p_i \ge \frac{4\log^2\ell}{d_i}$ for $i \in F_2$, and the last inequality uses $\frac{d_i}{2\log^2\ell} \le \ell$, as $\sum_{i=1}^m d_i = \ell$ and d_i's are nonnegative. Therefore, (72) follows. □

Let $\chi \sim \mathcal{D}$ here and further. Define for convenience new random variables $X_i = d_i\left(h(p_i) - h\left(\frac{\chi_i}{d_i}\right)\right)$ for all $i \in F_2$, and let also $X = \sum_{i\in F_2} X_i = \sum_{i\in F_2} d_i\left(h(p_i) - h\left(\frac{\chi_i}{d_i}\right)\right)$.

Claim 7.14. *With probability at least $1 - \ell^{-\log\ell}$ it holds that*

$$X - \mathbb{E}[X] \le \ell^{1/2}\log^3\ell$$

Proof. Obviously all the X_i's are independent, and also $X_i \in \left[-\sqrt{d_i p_i}\log^2\ell, \sqrt{d_i p_i}\log^2\ell\right]$ by Claim 7.13. Then we can apply Hoeffding's inequality for the sum of bounded independent random variables ([Hoe63, Theorem 2]), and obtain

$$\begin{aligned}\mathop{\mathbb{P}}_{\chi\sim\mathcal{D}}\left[X - \mathbb{E}[X] \ge \ell^{1/2}\log^3\ell\right] &\le \exp\left(-\frac{2\ell\log^6\ell}{\sum\limits_{i\in F_2}(2\sqrt{d_i p_i}\log^2\ell)^2}\right) \\ &= \exp\left(-\frac{2\ell\log^6\ell}{\log^4\ell\cdot\sum\limits_{i\in F_2}(4d_i p_i)}\right) \le \exp\left(-\frac{\ell\log^2\ell}{\sum\limits_{i\in F_2} d_i}\right) \\ &\le e^{-\log^2\ell} \le \ell^{-\log\ell},\end{aligned}$$

where we used $p_i \le 1/2$ and $\sum\limits_{i\in F_2} d_i \le \sum\limits_{i=1}^m d_i = \ell$ in the second and third inequalities, respectively. □

So by now we proved that $X = \sum_{i\in F_2} d_i\left(h(p_i) - h\left(\frac{\chi_i}{d_i}\right)\right)$ does not deviate much from its expectation. What we are left to show now is that $\mathbb{E}[X]$ is not very large by itself.

The following two claims show that the first moment and mean absolute deviation of the distribution $\mathcal{D}_i$ are close to those of Ω_i. This easily follows from the definition (70) of $\mathcal{D}_i$, and the proofs are deferred to Appendix B

Claim 7.15. *Let $i \in F_2$. Then $\left|\mathop{\mathbb{E}}_{\chi_i\sim\mathcal{D}_i}\left[\frac{\chi_i}{d_i}\right] - p_i\right| \le \frac{1}{d_i}$.*

Claim 7.16. *Let $\chi_i \sim \mathcal{D}_i$ and $\eta_i \sim \Omega_i$ for $i \in F_2$. Then $\mathbb{E}\left|\chi_i - \mathbb{E}[\chi_i]\right| \le \mathbb{E}\left|\eta_i - \mathbb{E}[\eta_i]\right| + 1$.*

These observations allow us we prove the following

Claim 7.17. *Let $i \in F_2$, and $\chi_i \sim \mathcal{D}_i$. Then $h\left(\mathbb{E}\left[\frac{\chi_i}{d_i}\right]\right) - \mathbb{E}\left[h\left(\frac{\chi_i}{d_i}\right)\right] \le \frac{5\log \ell}{d_i}$.*

Proof. Unfortunately, Jensen's inequality works in the opposite direction for us here. However, we use some form of converse Jensen's from [Dra11], which says the following:

Lemma 7.18 (Converse Jensen's inequality, [Dra11], Corollary 1.8)**.** *Let f be a concave differentiable function on an interval $[a,b]$, and let Z be a (discrete) random variable, taking values in $[a,b]$. Then*

$$0 \le f(\mathbb{E}[Z]) - \mathbb{E}[f(Z)] \le \frac{1}{2}\left(f'(a) - f'(b)\right) \cdot \mathbb{E}\left|Z - \mathbb{E}[Z]\right|.$$

We apply it here for the concave binary entropy function h, and random variable $Z = \frac{\chi_i}{d_i}$ for $\chi_i \sim \mathcal{D}_i$, which takes values in $[a,b] \coloneqq \left[p_i - \sqrt{\frac{p_i}{d_i}}\log \ell, p_i + \sqrt{\frac{p_i}{d_i}}\log \ell\right]$. Recall also that for $i \in F_2$, $p_i \ge \frac{4\log^2 \ell}{d_i}$ and then $\frac{p_i}{2} \ge \sqrt{\frac{p_i}{d_i}}\log \ell$, therefore $a = p_i - \sqrt{\frac{p_i}{d_i}}\log \ell \ge \frac{p_i}{2}$, and also $b = p_i + \sqrt{\frac{p_i}{d_i}}\log \ell \le \frac{3p_i}{2}$. Using the mean value theorem, for some $c \in [a,b] \subseteq \left[\frac{p_i}{2}, \frac{3p_i}{2}\right]$ we have

$$h'(a) - h'(b) = (b-a)\cdot(-h''(c)) \le 2\sqrt{\frac{p_i}{d_i}}\log \ell \cdot (-h''(c)).$$

Now we look at $(-h''(c)) = \frac{1}{c(1-c)\ln 2}$ for some $c \in \left[\frac{p_i}{2}, \frac{3p_i}{2}\right]$. As $p_i < 1/2$, it follows $\left[\frac{p_i}{2}, \frac{3p_i}{2}\right] \subseteq \left[\frac{p_i}{2}, 1 - \frac{p_i}{2}\right]$. Using the symmetry of a function $x(1-x)$ around $1/2$, we conclude that its minimal value over the interval $\left[\frac{p_i}{2}, \frac{3p_i}{2}\right]$ is attained at $p_i/2$. Thus derive $c(1-c) \ge \frac{p_i}{2}\left(1 - \frac{p_i}{2}\right) \ge \frac{3p_i}{8}$, since $p_i < 1/2$. And so $(-h''(c)) = \frac{1}{c(1-c)\ln 2} \le \frac{8}{p_i \cdot 3\ln 2} \le \frac{4}{p_i}$. Therefore

$$h'(a) - h'(b) \le \frac{8\log \ell}{\sqrt{d_i p_i}}.$$

Finally, Claim 7.16 gives $\mathbb{E}\left|Z - \mathbb{E}[Z]\right| \le \mathbb{E}\left|\frac{Z_2}{d_i} - \mathbb{E}\left[\frac{Z_2}{d_i}\right]\right| + \frac{1}{d_i}$ for $Z_2 \sim \mathrm{Binom}(d_i, p_i)$, and so

$$\mathbb{E}\left|Z - \mathbb{E}[Z]\right| \le \frac{1}{d_i}\mathbb{E}\left|Z_2 - \mathbb{E}[Z_2]\right| + \frac{1}{d_i} \le \frac{1}{d_i}\sqrt{\mathbb{E}[(Z_2 - \mathbb{E}[Z_2])^2]} + \frac{1}{d_i} = \sqrt{\frac{p_i(1-p_i)}{d_i}} + \frac{1}{d_i} \le \sqrt{\frac{p_i}{d_i}} + \frac{1}{d_i}.$$

Putting all this together, Lemma 7.18 gives us

$$0 \le h\left(\mathbb{E}\left[\frac{\chi_i}{d_i}\right]\right) - \mathbb{E}\left[h\left(\frac{\chi_i}{d_i}\right)\right] \le \frac{1}{2}\cdot\frac{8\log \ell}{\sqrt{d_i p_i}}\cdot\left(\sqrt{\frac{p_i}{d_i}} + \frac{1}{d_i}\right) = \frac{4\log \ell}{d_i} + \frac{4\log \ell}{d_i\sqrt{d_i p_i}} \le \frac{5\log \ell}{d_i},$$

where the last step uses $\sqrt{p_i d_i} \ge 2\log \ell$ for $i \in F_2$. □

We can now use the above claims and Proposition 4.1 to bound the expectation of X:

$$\begin{aligned}
\mathbb{E}[X] = \sum_{i\in F_2} d_i\left(h(p_i) - \mathbb{E}\left[h\left(\frac{\chi_i}{d_i}\right)\right]\right) &\le \sum_{i\in F_2} d_i\left(h(p_i) - h\left(\mathbb{E}\left[\frac{\chi_i}{d_i}\right]\right) + \frac{5\log \ell}{d_i}\right) \\
&\le \sum_{i\in F_2} d_i\left(h\left(\left|p_i - \mathbb{E}\left[\frac{\chi_i}{d_i}\right]\right|\right) + \frac{5\log \ell}{d_i}\right) \\
&\le \sum_{i\in F_2} d_i\left(h\left(\frac{1}{d_i}\right) + \frac{5\log \ell}{d_i}\right) \\
&\le \sum_{i\in F_2} d_i\left(\frac{2}{d_i}\log d_i + \frac{5\log \ell}{d_i}\right) \le 7\ell^{1/2}\log \ell \le \ell^{1/2}\log^3 \ell,
\end{aligned} \tag{73}$$

where the first inequality is from Claim 7.17, the second is by Proposition 4.1, the third one follows from Claim 7.15, the fourth inequality is from Proposition 4.2, and the next ones follow from $d_i \le \ell$, $|F_2| \le m \le \ell^{1/2}$, and $\log \ell > 8$ by the conditions for this Lemma 7.12.

So we showed in Claim 7.14 that X does not exceed its expectations by more than $\ell^{1/2}\log^3 \ell$ with high probability (over $\chi \sim \mathcal{D}$), and also that $E[X]$ is bounded by $\ell^{1/2}\log^3 \ell$ in (73), and therefore X does not exceed $2\ell^{1/2}\log^3 \ell$ with high probability. Specifically, it means that there exists $\mathcal{T} \subseteq \mathcal{T}_1$, such that $\mathbb{P}_{\chi\sim\mathcal{D}}[\chi \in \mathcal{T}] \ge 1 - \ell^{-\log \ell}$, and that for any $\boldsymbol{s} \in \mathcal{T}$ it holds $\sum\limits_{i\in F_2} d_i \left(h(p_i) - h\left(\frac{s_i}{d_i}\right)\right) \le 2\ell^{1/2}\log^3 \ell$. Recall that $\sum\limits_{i\in F_1} d_i h(p_i) \le 5\ell^{1/2}\log^3 \ell$ as we showed in (64). Thus, by summing these two inequalities, we conclude from (62) that $\sum_{i=1}^{m} d_i \left(h(p_i) - h\left(\frac{s_i}{d_i}\right)\right) \le 7\ell^{1/2}\log^3 \ell$ for any $\boldsymbol{s} \in \mathcal{T}$.

Finally, the last step is to return back from the product of "truncated binomials" $\mathcal{D}$ to the original product of binomials Ω. As we defined the set $\mathcal{T}$ above, we have $\mathbb{P}_{\chi\sim\mathcal{D}}[\chi \in \mathcal{T}] \ge 1 - \ell^{-\log \ell}$. But by (71) the distributions Ω and $\mathcal{D}$ are very close to each other, and therefore we obtain:

$$\mathop{\mathbb{P}}_{\chi\sim\Omega}[\chi \in \mathcal{T}] = \mathop{\mathbb{P}}_{\chi\sim\mathcal{D}}[\chi \in \mathcal{T}] \cdot \theta \ge \left(1 - \ell^{-\log \ell}\right)\left(1 - \ell^{-(\log \ell)/3+1/2}\right) \ge 1 - \ell^{-(\log \ell)/4},$$

where we used the bound (69) on θ for the first inequality and $\log \ell \ge 8$ for the second one. □

7.2 Arbitrary alphabet size

In this section we finish the proof of Theorem 5.8 for the general BMS channel using the results from the previous section.

For BMS channels with large output alphabet size we will use binning of the output, however we will do it in a way that *upgrades* the channel, rather then degrades it (recall Definition 4.3). Specifically, we will employ the following statement:

Proposition 7.19. *Let W be any BMS channel. Then there exists another BMS channel $\widetilde{W}$ with the following properties:*

(i) Output alphabet size of $\widetilde{W}$ is at most $2\sqrt{\ell}$;

(ii) $\widetilde{W}$ is upgraded with respect to W, i.e. $W \preceq \widetilde{W}$;

(iii) $H(\widetilde{W}) \ge H(W) - \dfrac{\log \ell}{\ell^{1/2}}$.

Before proving this proposition, we first show how we can finish a proof of Theorem 5.8 using it. So, consider any BMS channel W with output alphabet size larger than $2\sqrt{\ell}$, and consider the channel $\widetilde{W}$ which satisfies properties (i)-(iii) from Proposition 7.19 with respect to W. First of all, notice that $k \ge \ell(1 - H(W)) + 14\ell^{1/2}\log^3 \ell \ge \ell\left(1 - H(\widetilde{W}) - \frac{\log \ell}{\ell^{1/2}}\right) + 14\ell^{1/2}\log^3 \ell$, and thus $k \ge \ell(1 - H(\widetilde{W})) + 13\ell^{1/2}\log^3 \ell$. Taking the property (i) into consideration, it follows that the channel $\widetilde{W}$ satisfies all the conditions for the arguments in the Section 7.1 to be applied, i.e. the statement of Theorem 5.8 holds for $\widetilde{W}$. Therefore, we can argue that with probability at least $1 - \ell^{-(\log \ell)/20}$ over a random kernel G it holds $H(V_1 \mid \widetilde{\boldsymbol{Y}}) \ge 1 - \ell^{-(\log \ell)/20}$, where $\widetilde{\boldsymbol{Y}} = \widetilde{W}^{\ell}(\boldsymbol{V} \cdot G)$ is the output vector if one would use the channel $\widetilde{W}$ instead of W, for $\boldsymbol{V} \sim \{0,1\}^k$.

Now, let W_1 be the channel which "proves" that $\widetilde{W}$ is upgraded with respect to W, i.e. $W_1\left(\widetilde{W}(x)\right)$ and $W(x)$ are identically distributed for any $x \in \{0,1\}$. Trivially then, $W_1^{\ell}\left(\widetilde{W}^{\ell}(X)\right)$ and $W^{\ell}(X)$ are identically distributed for any random variable X supported on $\{0,1\}^{\ell}$.

Next, observe that the following forms a Markov chain

$$V_1 \to \boldsymbol{V} \to \boldsymbol{V} \cdot G \to \widetilde{W}^\ell(\boldsymbol{V}G) \to W_1^\ell\left(\widetilde{W}^\ell(\boldsymbol{V}G)\right),$$

where $\boldsymbol{V}$ is distributed uniformly over $\{0,1\}^k$. But then the data-processing inequality gives

$$I\left(V_1\,;\, W_1^\ell\left(\widetilde{W}^\ell(\boldsymbol{V}G)\right)\right) \le I\left(V_1\,;\, \widetilde{W}^\ell(\boldsymbol{V}G)\right).$$

However, as we discussed above, $W_1^\ell\left(\widetilde{W}^\ell(\boldsymbol{V}G)\right)$ and $W^\ell(\boldsymbol{V}G)$ are identically distributed, and so

$$I(V_1\,;\,\boldsymbol{Y}) = I\left(V_1\,;\, W^\ell(\boldsymbol{V}G)\right) = I\left(V_1\,;\, W_1^\ell\left(\widetilde{W}^\ell(\boldsymbol{V}G)\right)\right) \le I\left(V_1\,;\, \widetilde{W}^\ell(\boldsymbol{V}G)\right) = I(V_1\,;\,\widetilde{\boldsymbol{Y}}).$$

Therefore using $H(X|Y) = H(X) - I(X;Y)$ we derive that

$$H(V_1 \mid \boldsymbol{Y}) \ge H(V_1 \mid \widetilde{\boldsymbol{Y}}) \ge 1 - \ell^{-(\log \ell)/20}$$

with probability at least $1 - \ell^{-(\log \ell)/20}$. This concludes the proof of Theorem 5.8. □

Proof of Proposition 7.19. We are going to describe how to construct such an upgraded channel $\widetilde{W}$. We again are going to look at W as a convex combination of BSCs, as we discussed in Section 7.1: let W consist of m underlying BSC subchannels $W^{(1)}, W^{(2)} \dots, W^{(m)}$, each has probability q_j to be chosen. The subchannel $W^{(j)}$ has crossover probability p_j, and $0 \le p_1 \le \dots \le p_m \le \frac{1}{2}$. The subchannel $W^{(j)}$ can output $z_j^{(0)}$ or $z_j^{(1)}$, and the whole output alphabet is then $\mathcal{Y} = \{z_1^{(0)}, z_1^{(1)}, z_2^{(0)}, z_2^{(1)}, \dots, z_m^{(0)}, z_m^{(1)}\}$, $|\mathcal{Y}| = 2m$. It will be convenient to write the transmission probabilities of W explicitly: for any $k \in [m]$, $c, x \in \{0,1\}$:

$$W\left(z_k^{(c)} \,\middle|\, x\right) = \begin{cases} q_k \cdot (1-p_k), & x = c, \\ q_k \cdot p_k, & x \ne c. \end{cases} \tag{74}$$

The key ideas behind the construction of $\widetilde{W}$ are the following:

– decreasing a crossover probability in any BSC (sub)channel always upgrades the channel, i.e. $\mathrm{BSC}_{p_1} \preceq \mathrm{BSC}_{p_2}$ for any $0 \le p_2 \le p_1 \le \frac{1}{2}$ ([TV13, Lemma 9]). Indeed, one can simulate a flip of coin with bias p_1 by first flipping a coin with bias p_2, and then flipping the result one more time with probability $q = \frac{p_1 - p_2}{1 - 2p_2}$. In other words, $\mathrm{BSC}_{p_1}(x)$ and $\mathrm{BSC}_q\left(\mathrm{BSC}_{p_2}(x)\right)$ are identically distributed for $x \in \{0,1\}$.

– "binning" two BSC subchannels with the same crossover probability doesn't change the channel ([TV13, Corollary 10]).

Let us finally describe how to construct $\widetilde{W}$. Split the interval $[0, 1/2]$ into $\sqrt{\ell}$ parts evenly, i.e. let $\theta_j = \frac{j-1}{2\sqrt{\ell}}$ for $j = 1, 2, \dots, \sqrt{\ell} + 1$, and consider intevals $[\theta_j, \theta_{j+1})$ for $j = 1, 2, \dots, \sqrt{\ell}$ (include $1/2$ into the last interval). Now, to get $\widetilde{W}$, we first slightly decrease the crossover probabilities in all the BSC subchannels $W^{(1)}, W^{(2)} \dots, W^{(m)}$ so that they all become one of $\theta_1, \theta_2, \dots, \theta_{\sqrt{\ell}}$. After

that we bin together the subchannels with the same crossover probabilities and let the resulting channel be $\widetilde{W}$. Formally, we define

$$T_j \coloneqq \left\{ i \in [m] \ : \ p_i \in [\theta_j, \theta_{j+1}) \right\}, \qquad j = 1, 2, \ldots, \sqrt{\ell} - 1,$$
$$T_{\sqrt{\ell}} \coloneqq \left\{ i \in [m] \ : \ p_i \in [\theta_{\sqrt{\ell}}, \theta_{\sqrt{\ell}+1}] \right\}.$$

So, T_j is going to be the set of indices of subchannels of W for which we decrease the crossover probability to be equal to θ_j. Then the probability distribution over the new, binned, BSC subchannels $\widetilde{W^{(1)}}, \widetilde{W^{(2)}} \ldots, \widetilde{W^{(\sqrt{\ell})}}$ in the channel $\widetilde{W}$ is going to be $(\widetilde{q_1}, \widetilde{q_2}, \ldots, \widetilde{q_{\sqrt{\ell}}})$, where $\widetilde{q_j} \coloneqq \sum\limits_{i \in T_j} q_i$. The subchannel $\widetilde{W^{(j)}}$ has crossover probability θ_j, and it can output one of two new symbols $\widetilde{z_j^{(0)}}$ or $\widetilde{z_j^{(1)}}$. The whole output alphabet is then $\widetilde{\mathcal{Y}} = \{\widetilde{z_1^{(0)}}, \widetilde{z_1^{(1)}}, \widetilde{z_2^{(0)}}, \widetilde{z_2^{(1)}}, \ldots, \widetilde{z_{\sqrt{\ell}}^{(0)}}, \widetilde{z_{\sqrt{\ell}}^{(1)}}\}$, $|\widetilde{\mathcal{Y}}| = 2\sqrt{\ell}$. To be more specific, we describe $\widetilde{W} : \{0,1\} \to \widetilde{\mathcal{Y}}$, as follows: for any $j \in [\sqrt{\ell}]$ and any $b, x \in \{0,1\}$

$$\widetilde{W}\left(\widetilde{z_j^{(b)}} \,\middle|\, x \right) = \begin{cases} \sum\limits_{i \in T_j} q_i \cdot (1 - \theta_j), & x = b, \\ \sum\limits_{i \in T_j} q_i \cdot \theta_j, & x \neq b. \end{cases} \tag{75}$$

Property (i) on the output alphabet size for $\widetilde{W}$ then holds immediately. Let us verify (ii) by showing that $\widetilde{W}$ is indeed upgraded with respect to W.

One can imitate the usage of W using $\widetilde{W}$ as follows: on input $x \in \{0,1\}$, feed it through $\widetilde{W}$ to get output $\widetilde{z_j^{(b)}}$ for some $b \in \{0,1\}$ and $j \in [\sqrt{\ell}]$. We then know that the subchannel $\widetilde{W^{(j)}}$ was used, which by construction corresponds to the usage of a subchannel $W^{(i)}$ for some $i \in T_j$. Then we randomly choose an index k from T_j with probability of $i \in T_j$ being chosen equal to $\dfrac{q_i}{\widetilde{q_j}}$. This determines that we are going to use the subchannel $W^{(k)}$ while imitating the usage of W. By now we flipped the input with probability θ_j (since we used the subchannel $\widetilde{W^{(j)}}$), while we want it to be flipped with probability $p_k \geq \theta_j$ overall, since we decided to use $W^{(k)}$. So the only thing we need to do it to "flip" b to $(1-b)$ with probability $\frac{p_k - \theta_j}{1 - 2\theta_j}$, and then output $z_k^{(b)}$ or $z_k^{(1-b)}$ correspondingly.

Formally, we just describe the channel $W_1 : \widetilde{\mathcal{Y}} \to \mathcal{Y}$ which proves that $\widetilde{W}$ is upgraded with respect to W by all of its transmission probabilities: for all $k \in [m]$, $j \in [\sqrt{\ell}]$, $b, c \in \{0,1\}$ set

$$W_1\left(z_k^c \,\middle|\, \widetilde{z_j^{(b)}} \right) = \begin{cases} 0, & k \notin T_j \\ \dfrac{q_k}{\sum\limits_{i \in T_j} q_i} \cdot \left(1 - \dfrac{p_k - \theta_j}{1 - 2\theta_j} \right), & k \in T_j, \ b = c, \\ \dfrac{q_k}{\sum\limits_{i \in T_j} q_i} \cdot \left(\dfrac{p_k - \theta_j}{1 - 2\theta_j} \right), & k \in T_j, \ b \neq c. \end{cases} \tag{76}$$

It is easy to check that W_1 is a valid channel, and that it holds for any $k \in [m]$ and $c, x \in \{0,1\}$

$$\sum_{j \in [\sqrt{\ell}],\, b \in \{0,1\}} \widetilde{W}\left(\widetilde{z_j^{(b)}} \,\middle|\, x \right) W_1\left(z_k^{(c)} \,\middle|\, \widetilde{z_j^{(b)}} \right) = W\left(z_k^{(c)} \,\middle|\, x \right), \tag{77}$$

which proves that $\widetilde{W}$ is indeed upgraded to W. We prove the above equality in Appendix C.

It only remains to check that the property (iii) also holds, i.e. that the entropy did not decrease too much after we upgrade the channel W to $\widetilde{W}$. We have

$$H\left(\widetilde{W}\right) = \sum_{j\in[\sqrt{\ell}]} \widetilde{q}_j h(\theta_j) = \sum_{j\in[\sqrt{\ell}]} \left(\sum_{i\in T_j} q_i\right) h(\theta_j) = \sum_{k\in[m]} q_k h(\theta_{j_k}),$$

where we again denoted by j_k the index from $[\sqrt{\ell}]$ for which $k \in T_{j_k}$. Therefore

$$H(W) - H\left(\widetilde{W}\right) = \sum_{k\in[m]} q_k\big(h(p_k) - h(\theta_{j_k})\big) \le \sum_{k\in[m]} q_k\big(h(\theta_{j_k+1}) - h(\theta_{j_k})\big),$$

since $p_k \in [\theta_{j_k}, \theta_{j_k+1}]$ as $k \in T_{j_k}$. Finally, since $\theta_{j+1} - \theta_j = \frac{1}{2\sqrt{\ell}}$, Proposition 4.1 gives

$$H(W) - H\left(\widetilde{W}\right) \le \sum_{k\in[m]} q_k\big(h(\theta_{j_k+1}) - h(\theta_{j_k})\big) \le h\left(\frac{1}{2\sqrt{\ell}}\right) \le 2\cdot\frac{1}{2\sqrt{\ell}}\log\left(2\sqrt{\ell}\right) \le \frac{\log\ell}{\sqrt{\ell}}. \qquad \square$$

8 Suction at the ends

In this section we present the proof for Theorem 5.1 in the case the standard Arıkans kernel was chosen in Algorithm A – the so-called suction at the ends regime. Recall that, as we discussed in section 5.1, this regime applies when the entropy of the channel W falls into the interval $(\ell^{-4}, 1-\ell^{-4})$, and the algorithm directly takes a kernel $K = A_2^{\otimes\log\ell}$, where $A_2 = \left(\begin{smallmatrix}1 & 0\\ 1 & 1\end{smallmatrix}\right)$ is the kernel of Arıkan's original polarizing transform, instead of trying out all the possible matrices. Note that multiplying by such a kernel K is equivalent to just applying the Arıkan's 2×2 transform recursively $\log\ell$ times. Suppose we have a BMS channel W with $H(W)$ very close to 0 or 1. For Arıkan's basic transform, by working with the channel Bhattacharyya parameter $Z(W)$ instead of the entropy $H(W)$, it is well known that one of the two Arıkan bit-channels has Z value getting much closer (quadratically closer) to the boundary of the interval $(0,1)$ [Arı09, Kor09]. Using these ideas, we prove in this section that basic transform decreases the average of the potential function $g_\alpha(\cdot)$ of entropy at least by a factor of $\ell^{-1/2}$ after $\log\ell$ iterations for large enough ℓ.

The basic Arıkan's transform takes one channel W and splits it into a slightly worse channel W^- and a slightly better channel W^+. Then the transform is applied recursively to W^- and W^+, creating channels W^{--}, W^{-+}, W^{+-}, and W^{++}. One can think of the process as of a complete binary tree of depth $\log\ell$, with the root node W, and any node at the level i is of form W^{B_i} for some $B_i \in \{-,+\}^i$, with two children W^{B_i-} and W^{B_i+}. Denote $r = \log\ell$, then the channels at the leaves $\{W^{B_r}\}$, for all $B_r \in \{-,+\}^r$ are exactly the Arıkan's subchannels of W with respect to the kernel $K = A_2^{\otimes\log\ell}$. We are going to prove the following result

Lemma 8.1. *Let W be a BMS channel with $H(W) \notin (\ell^{-4}, 1-\ell^{-4})$, and $\alpha \in \left(0, \frac{1}{12}\right)$ be some constant. Let ℓ be a power of two and denote $r = \log\ell$. Then for ℓ large enough such that $r \ge \max\left\{\frac{1}{\alpha}, 128\right\}$*

$$\sum_{B\in\{-,+\}^r} g_\alpha\left(H\left(W^B\right)\right) \le \ell^{1/2} g_\alpha\left(H(W)\right), \tag{78}$$

where $g_\alpha(\cdot)$ is the potential function defined in (4).

Clearly, the above lemma will imply the suction at the end case of Theorem 5.1, as the inequality $\log \ell \geq \frac{1}{\alpha}$ holds by the conditions of this theorem.

For the analysis below, apart from the entropy of the channel, we will also use Bhattacharrya parameter $Z(W)$:

$$Z(W) = \sum_{y \in \mathcal{Y}} \sqrt{W(y \,|\, 0) W(y \,|\, 1)},$$

together with the inequalities which connect it to the entropy:

$$Z(W)^2 \leq H(W) \leq Z(W), \tag{79}$$

for any BMS channel W ([Kor09, Lemma 1.5], [Ari10, Proposition 2]). The reason we use this parameter is because of the following relations, which show how the Bhattacharrya parameter changes after the basic transform ([Arı09, Proposition 5] [RU08], [HAU14, eq (13)]):

$$Z(W^+) = Z(W)^2, \tag{80}$$

$$Z(W)\sqrt{2 - Z(W)^2} \leq Z(W^-) \leq 2Z(W). \tag{81}$$

We will also use the conservation of conditional entropy on application of Arıkan's transform

$$H(W^+) + H(W^-) = 2H(W). \tag{82}$$

Proof of Lemma 8.1. The proof is presented in the next two sections, as it is divided into two parts: the case when $H(W) \leq \ell^{-4}$ (suction at the lower end), and when $H(W) \geq 1 - \ell^{-4}$ (suction at the upper end).

8.1 Suction at the lower end

Suppose $H(W) \leq \ell^{-4}$ for this case, thus $Z(W) \leq \ell^{-2} = 2^{-2r}$.

First, recursive application of (82) gives

$$\sum_{B \in \{-,+\}^r} H\left(W^B\right) = 2^r H(W), \tag{83}$$

and since entropy is always nonnegative, this implies for any $B \in \{-,+\}^r$

$$H\left(W^B\right) \leq 2^r H(W). \tag{84}$$

Denote now $k = \left\lceil \log \frac{1}{\alpha} \right\rceil$, and notice that $\log r \geq k - 1$ since $r \geq \frac{1}{\alpha}$. For $B \in \{-,+\}^r$, define $wt_+(B)$ to be number of $+$'s in B. We will split the summation in (78) into two parts: the part with $wt_+(B) < k$, and when $wt_+(B) \geq k$.

First part. Out of (84) derive

$$\sum_{wt_+(B)<k} g_\alpha\left(H\left(W^B\right)\right) \leq \sum_{j=0}^{k-1} \binom{r}{j} g_\alpha\left(2^r H(W)\right) \leq \log r \cdot \binom{r}{\log r} \cdot 2^{r\alpha} H(W)^\alpha \leq 2^{\log^2 r + r\alpha} \cdot H(W)^\alpha, \tag{85}$$

where we used $\binom{r}{\log r} \leq \frac{r^{\log r}}{(\log r)!}$; the fact the g_α is increasing on $\left(0, \frac{1}{2}\right)$ together with $2^r H(W) \leq \ell^{-3} < \frac{1}{2}$, and that $g_\alpha(x) \leq x^\alpha$ for $x \in (0,1)$.

Second part. We are going to use the following observation, which was already established in [AT09, Lemma 1] and can be proved by induction based on (80) and (81):

Claim 8.2. *Let $B \in \{-,+\}^r$, such that number of $+$'s in B is equal to s. Then*

$$Z\left(W^B\right) \le \left(2^{r-s} \cdot Z(W)\right)^{2^s}.$$

This corresponds to first using the upper bound in (81) $(r-s)$ *times, and after that using* (80) s *times while walking* ***down*** *the recursive binary tree of channels.*

Then, using Claim 8.2 along with (79) and the fact that $Z(W) \le \ell^{-2} = 2^{-2r}$, we obtain the following for any $B \in \{-,+\}^r$ with $wt_+(B) = s \ge k$:

$$\begin{aligned} H\left(W^B\right) \le Z\left(W^B\right) \le \left(2^{r-s} \cdot Z(W)\right)^{2^s} &\le 2^{(r-s)2^s} \cdot Z(W)^{2^s-2} \cdot H(W) \\ &\le 2^{(r-s)2^s} \cdot 2^{-2r\cdot 2^s+4r} \cdot H(W) \\ &= 2^{-r2^s-s2^s+4r} \cdot H(W) \\ &\le 2^{-r2^k-k2^k+4r} \cdot H(W). \end{aligned}$$

Therefore

$$\sum_{wt_+(B)\ge k} g_\alpha\left(H\left(W^B\right)\right) \le \sum_{wt_+(B)\ge k} H\left(W^B\right)^\alpha \le 2^r \cdot 2^{\alpha(-r2^k-k2^k+4r)} \cdot H(W)^\alpha. \tag{86}$$

Observe now the following chain of inequalities

$$\frac{r}{2} + 4r\alpha + 2 \le r \le r \cdot 2^k\alpha \le r \cdot 2^k\alpha + k \cdot 2^k\alpha,$$

which trivially holds for $\alpha \le \dfrac{1}{12}$. Therefore

$$r + \alpha(-r2^k - k2^k + 4r) \le \frac{r}{2} - 2,$$

and thus in (86) obtain

$$\sum_{wt_+(B)\ge k} g_\alpha\left(H\left(W^B\right)\right) \le 2^{r/2-2} \cdot H(W)^\alpha. \tag{87}$$

Overall bound. Combining (85) and (87) we derive

$$\begin{aligned} \sum_{B\in\{-,+\}^r} g_\alpha\left(H\left(W^B\right)\right) &\le \left(2^{\log^2 r + r\alpha} + 2^{r/2-2}\right) \cdot H(W)^\alpha \\ &\le 2^{r/2} \cdot \frac{H(W)^\alpha}{2} \\ &\le \ell^{1/2} g_\alpha(H(W)), \end{aligned}$$

where we used $\log^2 r + r\alpha \le \frac{r}{2} - 2$ for $r \ge 128$, and $\frac{1}{2} \le (1-x)^\alpha$ for any $x \le \frac{1}{2}$. This proves Lemma 8.1 for the lower end case $H(W) \le \ell^{-4}$.

8.2 Suction at the upper end

Now consider the case $H(W) \ge 1 - \ell^{-4}$. The proof is quite similar to the previous case, but we are going to track the distance from $H(W)$ (and $Z(W)$) to 1 now. Specifically, denote

$$\begin{aligned} I(W) &= 1 - H(W), \\ S(W) &= 1 - Z(W), \end{aligned}$$

where $I(W)$ is actually the (symmetric) capacity of the channel, and $S(W)$ is just a notation we use in this proof. Notice that $g_\alpha(x) = g_\alpha(1-x)$, therefore it suffices to prove (78) with capacities of the channels instead of entropies in the inequality. Also notice that $I(W) \le \ell^{-4}$ for the current case of suction at the upper end.

Let us now derive the relations between $I(W)$, $S(W)$, as well as evolution of $S(\cdot)$ for W^+ and W^-, similar to (79), (80), (81), and (82). Inequalities in (79) imply

$$\begin{aligned} S(W) &= 1 - Z(W) \le 1 - H(W) = I(W), \\ I(W) &= 1 - H(W) \le 1 - Z(W)^2 \le 2(1 - Z(W)) = 2S(W), \end{aligned}$$

so let us combine this to write

$$S(W) \le I(W) \le 2S(W). \tag{88}$$

Next, (80) and (81) give

$$S(W^+) = 1 - Z(W)^2 \le 2(1 - Z(W)) \le 2S(W), \tag{89}$$

$$S(W^-) \le 1 - Z(W)\sqrt{2 - Z(W)^2} \le 2(1 - Z(W))^2 = 2S(W)^2, \tag{90}$$

where we used $1 - x\sqrt{2 - x^2} \le 2(1-x)^2$ for any $x \in (0,1)$, which can be proven easily by showing that equality holds at $x = 1$ and that the derivative of RHS minus LHS is negative on $(0,1)$.

Finally, it easily follows from (83) that

$$\sum_{B \in \{-,+\}^r} I\left(W^B\right) = 2^r I(W),$$

and since capacity is nonnegative as well, we also obtain for any $B \in \{-,+\}^r$

$$I\left(W^B\right) \le 2^r I(W). \tag{91}$$

We now proceed with a very similar approach to the suction at the lower end case in Section 8.1: denote $k = \left\lceil \log \frac{1}{\alpha} \right\rceil$, and notice that $\log r \ge k - 1$ since $r \ge \frac{1}{\alpha}$. For $B \in \{-,+\}^r$, define $wt_-(B)$ to be number of $-$'s in B. We will split the summation in (78) (but with capacities of channels instead of entropies) into two parts: the part with $wt_-(B) < k$, and when $wt_-(B) \ge k$.

First part. Out of (91) derive, similarly to (85)

$$\sum_{wt_-(B)<k} g_\alpha\left(I\left(W^B\right)\right) \le \sum_{j=0}^{k-1} \binom{r}{j} g_\alpha\left(2^r I(W)\right) \le \log r \cdot \binom{r}{\log r} \cdot 2^{r\alpha} I(W)^\alpha \le 2^{\log^2 r + r\alpha} \cdot I(W)^\alpha. \tag{92}$$

Second part. Similarly to Claim 8.2, one can show via induction using (89) and (90) the following

Claim 8.3. *Let $B \in \{-,+\}^r$, such that number of $-$'s in B is equal to s. Then*

$$S\left(W^B\right) \le 2^{2^s - 1}\left(2^{r-s} \cdot S(W)\right)^{2^s}.$$

This corresponds to first using equality (89) *$(r-s)$ times, and after that using bound* (90) *s times while walking* ***down*** *the recursive binary tree of channels.*

Using this claim with (88) and the fact that $S(W) \le Z(W) \le \ell^{-4} \le 2^{-4r}$ obtain for any $B \in \{-,+\}^r$ with $wt_-(B) = s \ge k$

$$\begin{aligned} I\left(W^B\right) \le 2S\left(W^B\right) &\le 2^{2^s} \cdot \left(2^{r-s} \cdot S(W)\right)^{2^s} && \le 2^{(r-s+1)2^s} \cdot S(W)^{2^s-1} \cdot I(W) \\ &\le 2^{(r-s+1)2^s - 4r2^s + 4r} \cdot I(W) && = 2^{-2^s(3r+s-1)+4r} \cdot I(W) \\ &\le 2^{-2^k(3r+k-1)+4r} \cdot I(W) && \le 2^{-r2^k} \cdot I(W), \end{aligned}$$

where the last inequality uses $4r \le 2^k(2t+k-1)$, which holds trivially for $k \ge 1$. Therefore

$$\sum_{wt_-(B)\ge k} g_\alpha\left(I\left(W^B\right)\right) \le \sum_{wt_-(B)\ge k} I\left(W^B\right)^\alpha \le 2^r \cdot 2^{-\alpha r 2^k} \cdot I(W)^\alpha \le I(W)^\alpha, \tag{93}$$

since $\alpha \cdot 2^k \ge 1$ by the choice of k.

Overall bound. The bounds (92) and (93) give us

$$\sum_{B\in\{-,+\}^r} g_\alpha\left(H\left(W^B\right)\right) = \sum_{B\in\{-,+\}^r} g_\alpha\left(I\left(W^B\right)\right) \le \left(2^{\log^2 r + r\alpha} + 1\right) \cdot I(W)^\alpha \le \ell^{1/2} g_\alpha(H(W))$$

for large enough r when $H(W) \ge 1 - \ell^{-4}$. This completes the proof of Lemma 8.1. □

9 Code construction, encoding and decoding procedures

Before presenting our code construction and encoding/decoding procedures, we first distinguish the difference between the code construction and the encoding procedure. The objectives of code construction for polar-type codes are two-fold: First, find the $N \times N$ encoding matrix; second, find the set of noiseless bits under the successive cancellation decoder, which will carry the message bits. On the other hand, by encoding we simply mean the procedure of obtaining the codeword $\boldsymbol{X}_{[1:N]}$ by multiplying the information vector $\boldsymbol{U}_{[1:N]}$ with the encoding matrix, where we only put information in the noiseless bits in $\boldsymbol{U}_{[1:N]}$ and set all the frozen bits to be 0. As we will see at the end of this section, while the code construction has complexity polynomial in N, the encoding procedure only has complexity $O_\ell(N \log N)$.

For polar codes with a fixed invertible kernel $K \in \{0,1\}^{\ell\times\ell}$, the polarization process works as follows: We start with some BMS channel W. After applying the polar transform to W using kernel K, we obtain ℓ bit-channels $\{W_i : i \in [\ell]\}$ as defined in (1). Next we apply the polar transform using kernel K to each of these ℓ bit-channels, and we write the polar transform of W_i as $\{W_{ij} : j \in [\ell]\}$. Then we apply the polar transform to each of the ℓ^2 bit channels $\{W_{i_1,i_2} : i_1, i_2 \in [\ell]\}$ and obtain $\{W_{i_1,i_2,i_3} : i_1, i_2, i_3 \in [\ell]\}$, so on and so forth. After t rounds of polar transforms, we obtain ℓ^t bit-channels $\{W_{i_1,\dots,i_t} : i_1, \dots, i_t \in [\ell]\}$, and one can show that these are the bit-channels seen by the successive cancellation decoder when decoding the corresponding polar codes constructed from kernel K.

For our purpose, we need to use polar codes with mixed kernels, and we need to search for a "good" kernel at each step of polarization. We will also introduce new notation for the bit-channels in order to indicate the usage of different kernels for different bit-channels. As mentioned in Sections 2.7 and 5.1, we need to use a binning algorithm (Algorithm B) to quantize all the bit-channels we obtain in the code construction procedure. As long as we choose the parameter Q in Algorithm B to be a large enough polynomial of N, the quantized channel can be used as a very

Algorithm B: Degraded binning algorithm

Input: $W : \{0,1\} \to \mathcal{Y}$, bound Q on the output alphabet size after binning
Output: $\widetilde{W} : \{0,1\} \to \widetilde{\mathcal{Y}}$, where $|\widetilde{\mathcal{Y}}| \le \mathsf{Q}$

Initialize the new channel $\widetilde{W}$ with output symbols $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{\mathsf{Q}}$ by setting $\widetilde{W}(\tilde{y}_i|x) = 0$ for all $i \in [\mathsf{Q}]$ and $x \in \{0,1\}$
for $y \in \mathcal{Y}$ **do**
  $p(0|y) \leftarrow \frac{W(y|0)}{W(y|0)+W(y|1)}$
  $i \leftarrow \lceil \mathsf{Q} \cdot p(0|y) \rceil$
  if $i = 0$ **then**
    $i \leftarrow 1$ // $i = 0$ if and only if $p(0|y) = 0$; we merge this single point into the next bin
  end
  $\widetilde{W}(\tilde{y}_i|0) \leftarrow \widetilde{W}(\tilde{y}_i|0) + W(y|0)$
  $\widetilde{W}(\tilde{y}_i|1) \leftarrow \widetilde{W}(\tilde{y}_i|1) + W(y|1)$
end
return $\widetilde{W}$

good approximation of the original channel. This is made precise by [GX15, Proposition 13]: For W and $\widetilde{W}$ in Algorithm B, we have[6]

$$H(W) \le H(\widetilde{W}) \le H(W) + \frac{2 \log \mathsf{Q}}{\mathsf{Q}}. \tag{94}$$

Given a BMS channel W, our code construction works as follows:

1. **Step 0:** We first use Algorithm B to quantize/bin the output alphabet of W such that the resulting (degraded) channel has at most N^3 outputs, i.e., we set $\mathsf{Q} = N^3$ in Algorithm B. Note that the parameter Q can be chosen as any polynomial of N. By changing the value of Q, we obtain a tradeoff between the decoding error probability and the gap to capacity; see Theorem 9.6 at the end of this section. Here we choose the special case of $\mathsf{Q} = N^3$ to give a concrete example of code construction. Next we use Algorithm A in Section 5 to find a good kernel[7] for the quantized channel and denote it as $K_1^{(0)}$. Recall from Section 2.4 that a kernel is good if all but a $\tilde{O}(\ell^{-1/2})$ fraction of the bit-channels obtained after polar transform by this kernel have entropy $\ell^{-\Omega(\log \ell)}$-close to either 0 or 1. The superscript (0) in $K_1^{(0)}$ indicates that this is the kernel used in Step 0 of polarization. In this case, we use $\{W_i(B, K_1^{(0)}) : i \in [\ell]\}$ to denote the ℓ bit-channels resulting from the polar transform of the quantized version of W using kernel $K_1^{(0)}$. Here B stands for the binning operation, and the arguments in the brackets are the operations to obtain the bit-channel $W_i(B, K_1^{(0)})$ from W: first bin the outputs of W and then perform the polar transform using kernel $K_1^{(0)}$. For each $i \in [\ell]$, we again use Algorithm B to quantize/bin the output alphabet of $W_i(B, K_1^{(0)})$ such that the resulting (degraded) bit-channel $W_i(B, K_1^{(0)}, B)$ has at most N^3 outputs.

[6] Note that the binning algorithm (Algorithm 2) in [GX15] has one minor difference from the binning algorithm (Algorithm B) in this paper: In [GX15], the binning algorithm outputs a channel with $\mathsf{Q}+1$ outputs in contrast to Q outputs in this paper. More precisely, line 5-7 in Algorithm B of this paper is not included in the algorithm in [GX15], but one can easily check that this minor difference does not affect the proof at all.

[7] We will prove in Proposition 9.3 that the error parameter Δ in Algorithm A can be chosen as $\Delta = \frac{6\ell \log N}{N^2}$ when we set $\mathsf{Q} = N^3$.

2. **Step 1:** For each $i_1 \in [\ell]$, we use Algorithm A to find a good kernel for the quantized bit-channel $W_{i_1}(B, K_1^{(0)}, B)$ and denote it as $K_{i_1}^{(1)}$. The ℓ bit-channels resulting from the polar transform of $W_{i_1}(B, K_1^{(0)}, B)$ using kernel $K_{i_1}^{(1)}$ are denoted as $\{W_{i_1,i_2}(B, K_1^{(0)}, B, K_{i_1}^{(1)}) : i_2 \in [\ell]\}$. In this step, we will obtain ℓ^2 bit-channels $\{W_{i_1,i_2}(B, K_1^{(0)}, B, K_{i_1}^{(1)}) : i_1, i_2 \in [\ell]\}$. For each of them, we use Algorithm B to quantize/bin its output alphabet such that the resulting (degraded) bit-channels $\{W_{i_1,i_2}(B, K_1^{(0)}, B, K_{i_1}^{(1)}, B) : i_1, i_2 \in [\ell]\}$ has at most N^3 outputs. See Fig. 2 for an illustration of this procedure for the special case of $\ell = 3$.

3. We repeat the polar transforms and binning operations at each step of the code construction. More precisely, at **Step** j we have ℓ^j bit-channels

$$\{W_{i_1,i_2,\ldots,i_j}(B, K_1^{(0)}, B, K_{i_1}^{(1)}, B, \ldots, K_{i_1,\ldots,i_{j-1}}^{(j-1)}, B) : i_1, i_2, \ldots, i_j \in [\ell]\}.$$

This notation is a bit messy, so we introduce some simplified notation for the bit-channels obtained with and without binning operations: We still use

$$W_{i_1,i_2,\ldots,i_j}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{j-1}}^{(j-1)})$$

to denote the bit-channel obtained without the binning operations at all, and we use

$$W_{i_1,i_2,\ldots,i_j}^{\text{bin}}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{j-1}}^{(j-1)})$$

to denote the bit-channel obtained with binning operations performed at every step from Step 0 to Step $j-1$, i.e.,

$$W_{i_1,i_2,\ldots,i_j}^{\text{bin}}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{j-1}}^{(j-1)}) := W_{i_1,i_2,\ldots,i_j}(B, K_1^{(0)}, B, K_{i_1}^{(1)}, B, \ldots, K_{i_1,\ldots,i_{j-1}}^{(j-1)}, B).$$

Moreover, we use $W_{i_1,i_2,\ldots,i_j}^{\text{bin}*}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{j-1}}^{(j-1)})$ to denote the bit-channel obtained with binning operations performed at every step except for the last step, i.e.,

$$W_{i_1,i_2,\ldots,i_j}^{\text{bin}*}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{j-1}}^{(j-1)}) := W_{i_1,i_2,\ldots,i_j}(B, K_1^{(0)}, B, K_{i_1}^{(1)}, B, \ldots, B, K_{i_1,\ldots,i_{j-1}}^{(j-1)}).$$

Next we use Algorithm A to find a good kernel for each of them and denote the kernel as $K_{i_1,\ldots,i_j}^{(j)}$. After applying polar transforms using these kernels, we obtain ℓ^{j+1} bit-channels

$$\{W_{i_1,\ldots,i_{j+1}}^{\text{bin}*}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_j}^{(j)}) : i_1, \ldots, i_{j+1} \in [\ell]\}.$$

Then we quantize/bin the output alphabets of these bit-channels using Algorithm B and obtain the following ℓ^{j+1} quantized bit-channels

$$\{W_{i_1,\ldots,i_{j+1}}^{\text{bin}}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_j}^{(j)}) : i_1, \ldots, i_{j+1} \in [\ell]\}.$$

4. After **step** $t-1$, we obtain $N = \ell^t$ quantized bit-channels

$$\{W_{i_1,\ldots,i_t}^{\text{bin}}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{t-1}}^{(t-1)}) : i_1, i_2, \ldots, i_j \in [\ell]\},$$

and we have also obtained all the kernels in each step of polarization. More precisely, we have ℓ^i kernels in step i, so from step 0 to step $t-1$, we have $1 + \ell + \cdots + \ell^{t-1} = \frac{N-1}{\ell-1}$ kernels in total.

5. Find the set of good (noiseless) indices. More precisely, we use the shorthand notation[8]

$$\begin{aligned} H_{i_1,\ldots,i_t}(W) &:= H(W_{i_1,\ldots,i_t}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{t-1}}^{(t-1)})) \\ H_{i_1,\ldots,i_t}^{\mathrm{bin}}(W) &:= H(W_{i_1,\ldots,i_t}^{\mathrm{bin}}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{t-1}}^{(t-1)})) \end{aligned} \tag{95}$$

and define the set of good indices as

$$\mathcal{S}_{\mathrm{good}} := \left\{(i_1, i_2, \ldots, i_t) \in [\ell]^t : H_{i_1,\ldots,i_t}^{\mathrm{bin}}(W) \le \frac{7\ell \log N}{N^2}\right\}. \tag{96}$$

6. Finally, we need to construct the encoding matrix from these $\frac{N-1}{\ell-1}$ kernels. The kernels we obtained in step j are

$$\{K_{i_1,\ldots,i_j}^{(j)} : i_1, \ldots, i_j \in [\ell]\}.$$

For an integer $i \in [\ell^j]$, we write the j-digit ℓ-ary expansion of $i-1$ as $(\tilde{i}_1, \tilde{i}_2, \ldots, \tilde{i}_j)$, where $\tilde{i}_j$ is the least significant digit and $\tilde{i}_1$ is the most significant digit, and each digit takes value in $\{0, 1, \ldots, \ell-1\}$. Let $(i_1, i_2, \ldots, i_j) := (\tilde{i}_1 + 1, \tilde{i}_2 + 1, \ldots, \tilde{i}_j + 1)$, and define the mapping $\tau_j : [\ell^j] \to [\ell]^j$ as

$$\tau_j(i) := (i_1, i_2, \ldots, i_j) \quad \text{for } i \in [\ell^j]. \tag{97}$$

This is a one-to-one mapping between $[\ell^j]$ and $[\ell]^j$, and we use the shorthand notation $K_i^{(j)}$ to denote $K_{\tau_j(i)}^{(j)}$ for $i \in [\ell^j]$. For each $j \in \{0, 1, \ldots, t-1\}$, we define the block diagonal matrices $\overline{D}^{(j)}$ with size $\ell^{j+1} \times \ell^{j+1}$ and $D^{(j)}$ with size $N \times N$ as

$$\overline{D}^{(j)} := \mathrm{Diag}(K_1^{(j)}, K_2^{(j)}, \ldots, K_{\ell^j}^{(j)}), \qquad D^{(j)} := \underbrace{\{\overline{D}^{(j)}, \overline{D}^{(j)}, \ldots, \overline{D}^{(j)}\}}_{\text{number of } \overline{D}^{(j)} \text{ is } \ell^{t-j-1}}. \tag{98}$$

For $i \in [\ell^t]$, we have $\tau_t(i) = (i_1, \ldots, i_t)$. For $j \in [t-1]$, we define the permutation $\pi^{(j)}$ on the set $[\ell^t]$ as

$$\pi^{(j)}(i) := \tau_t^{-1}(i_1, \ldots, i_{t-j-1}, i_t, i_{t-j}, i_{t-j+1}, \ldots, i_{t-1}) \quad \forall i \in [\ell^t]. \tag{99}$$

By this definition, $\pi^{(j)}$ simply keeps the first $t-j-1$ digits of i to be the same and performs a cyclic shift on the last $j+1$ digits. Here we give some concrete examples:

$$\begin{aligned} \pi^{(1)}(i) &= \tau_t^{-1}(i_1, \ldots, i_{t-2}, i_t, i_{t-1}), \\ \pi^{(2)}(i) &= \tau_t^{-1}(i_1, \ldots, i_{t-3}, i_t, i_{t-2}, i_{t-1}), \\ \pi^{(3)}(i) &= \tau_t^{-1}(i_1, \ldots, i_{t-4}, i_t, i_{t-3}, i_{t-2}, i_{t-1}), \\ \pi^{(t-1)}(i) &= \tau_t^{-1}(i_t, i_1, i_2, \ldots, i_{t-1}). \end{aligned}$$

For each $j \in [t-1]$, let $Q^{(j)}$ be the $\ell^t \times \ell^t$ permutation matrix corresponding to the permutation $\pi^{(j)}$, i.e., $Q^{(j)}$ is the permutation matrix such that

$$(U_1, U_2, \ldots, U_{\ell^t})Q^{(j)} = (U_{\pi^{(j)}(1)}, U_{\pi^{(j)}(2)}, \ldots, U_{\pi^{(j)}(\ell^t)}). \tag{100}$$

Finally, for each $j \in [t]$, we define the $N \times N$ matrix

$$M^{(j)} := D^{(j-1)}Q^{(j-1)}D^{(j-2)}Q^{(j-2)} \ldots D^{(1)}Q^{(1)}D^{(0)}. \tag{101}$$

[8]We omit the reference to the kernels in the notation $H_{i_1,\ldots,i_t}(W)$ and $H_{i_1,\ldots,i_t}^{\mathrm{bin}}(W)$.

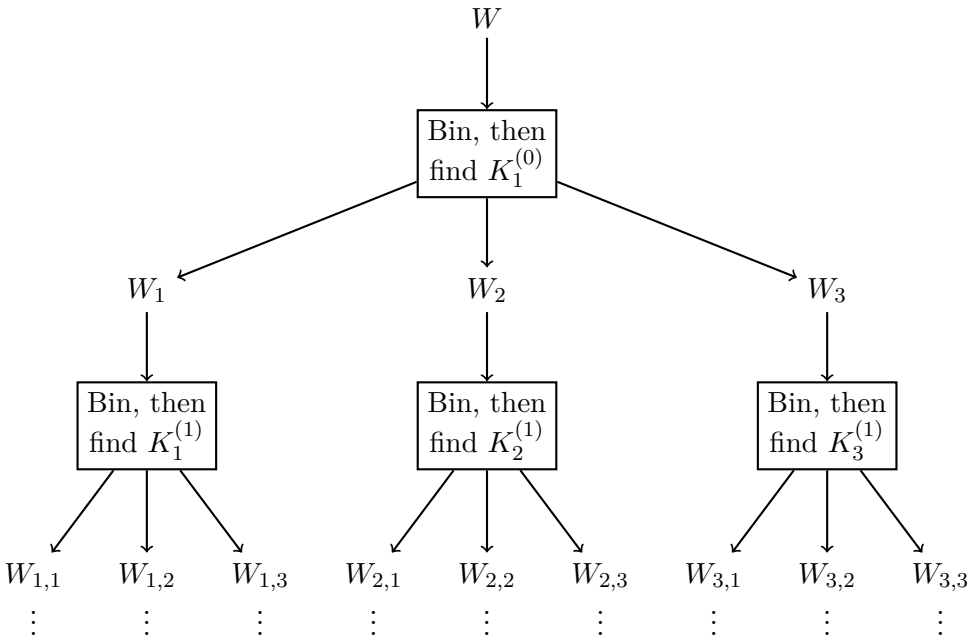

Figure 2: Illustration of code construction for the special case of $\ell = 3$.

Therefore, $M^{(j)}, j \in [t]$ satisfy the following recursive relation:

$$M^{(1)} = D^{(0)}, \qquad M^{(j+1)} = D^{(j)} Q^{(j)} M^{(j)}.$$

Our encoding matrix for code length $N = \ell^t$ is the submatrix of $M^{(t)}$ consisting of all the row vectors with indices belonging to the set $\mathcal{S}_{\text{good}}$ defined in (96); see the next paragraph for a detailed description of the encoding procedure.

Once we obtain the matrix $M^{(t)}$ and the set $\mathcal{S}_{\text{good}}$ in the code construction, the encoding procedure is standard; it is essentially the same as the original polar codes [Arı09]. Let $\boldsymbol{U}_{[1:N]}$ be a random vector consisting of N i.i.d. Bernoulli-1/2 random variables, and let $\boldsymbol{X}_{[1:N]} = \boldsymbol{U}_{[1:N]} M^{(t)}$. Recall that we use $\{W_i(M^{(t)}) : i \in [\ell^t]\}$ to denote the ℓ^t bit-channels resulting from the polar transform of W using matrix $M^{(t)}$. If we transmit the random vector $\boldsymbol{X}_{[1:N]}$ through N independent copies of W and denote the channel outputs as $\boldsymbol{Y}_{[1:N]}$, then by definition, the bit-channel mapping from U_i to $(\boldsymbol{U}_{[1:i-1]}, \boldsymbol{Y}_{[1:N]})$ is exactly $W_i(M^{(t)})$. Therefore, if we use a successive cancellation decoder to decode the input vector $\boldsymbol{U}_{[1:N]}$ bit by bit from all the channel outputs $\boldsymbol{Y}_{[1:N]}$ and all the previous input bits $\boldsymbol{U}_{[1:i-1]}$, then $W_i(M^{(t)})$ is the channel seen by the successive cancellation decoder when it decodes U_i. Clearly, $H(W_i(M^{(t)})) \approx 0$ means that the successive cancellation decoder can decode U_i correctly with high probability. For every $i \in \ell^t$, we write $\tau_t(i) = (i_1, i_2, \ldots, i_t)$. In Proposition 9.1 below, we will show that $H(W_i(M^{(t)})) = H_{i_1,\ldots,i_t}(W)$. Then in Proposition 9.3, we further show that $H_{i_1,\ldots,i_t}(W) \approx H^{\text{bin}}_{i_1,\ldots,i_t}(W)$. Therefore, $H(W_i(M^{(t)})) \approx H^{\text{bin}}_{i_1,\ldots,i_t}(W)$. By definition (96), the set $\mathcal{S}_{\text{good}}$ contains all the indices $(i_1, \ldots, i_t)$ for which $H^{\text{bin}}_{i_1,\ldots,i_t}(W) \approx 0$, so for all i such that $\tau_t(i) \in \mathcal{S}_{\text{good}}$, we also have $H(W_i(M^{(t)})) \approx 0$, meaning that the successive cancellation decoder can decode all the bits $\{U_i : \tau_t(i) \in \mathcal{S}_{\text{good}}\}$ correctly with high probability. In the encoding procedure, we put all the information in the set of good bits $\{U_i : \tau_t(i) \in \mathcal{S}_{\text{good}}\}$, and we set all the other bits to be some pre-determined value, e.g., set all of them to be 0. It is clear that the

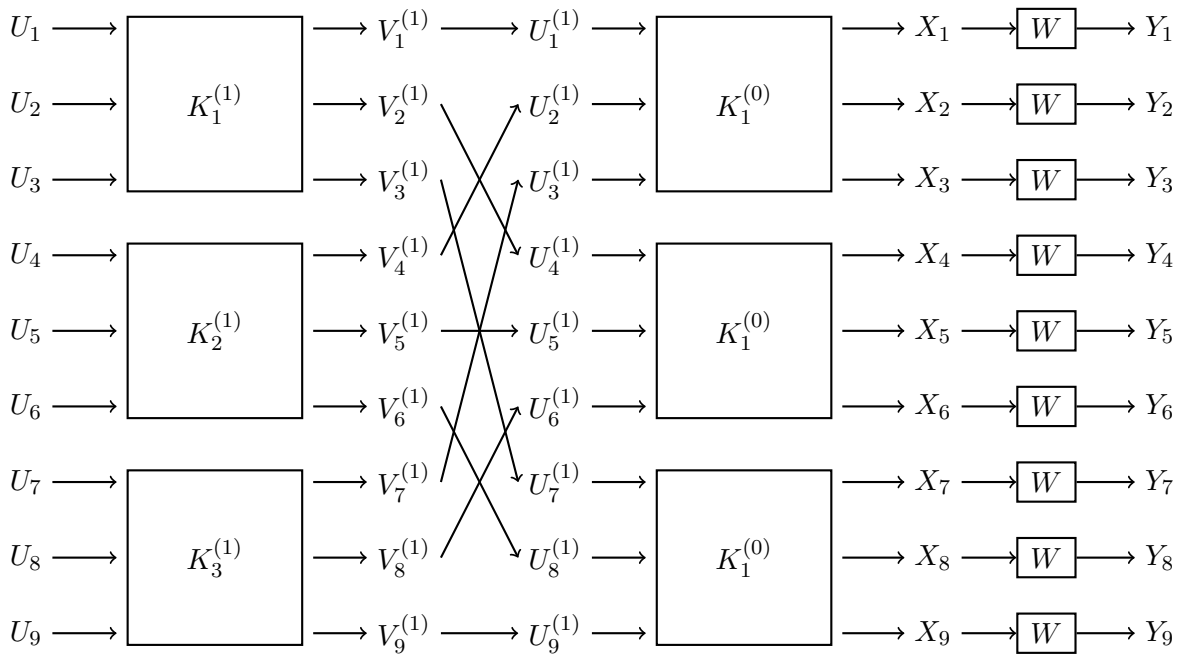

Figure 3: Illustration of the encoding process $\boldsymbol{X}_{[1:N]} = \boldsymbol{U}_{[1:N]} M^{(t)}$ for the special case of $\ell = 3$ and $t = 2$. Here $\boldsymbol{X}_{[1:N]}$ and $\boldsymbol{U}_{[1:N]}$ are row vectors. All four kernels in this figure $K_1^{(0)}, K_1^{(1)}, K_2^{(1)}, K_3^{(1)}$ have size 3×3, and the outputs of each kernel is obtained by multiplying the inputs with the kernel, e.g. $\boldsymbol{V}_{[1:3]}^{(1)} = \boldsymbol{U}_{[1:3]} K_1^{(1)}$.

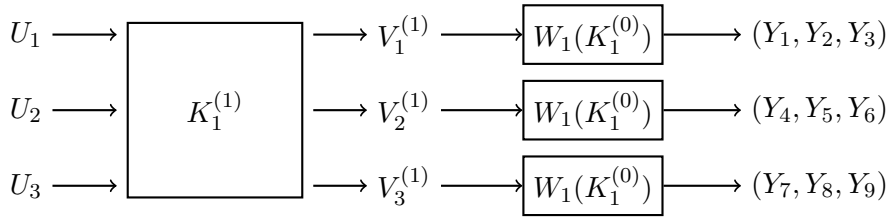

Figure 4: The (stochastic) mapping from $\boldsymbol{U}_{[1:3]}$ to $\boldsymbol{Y}_{[1:9]}$

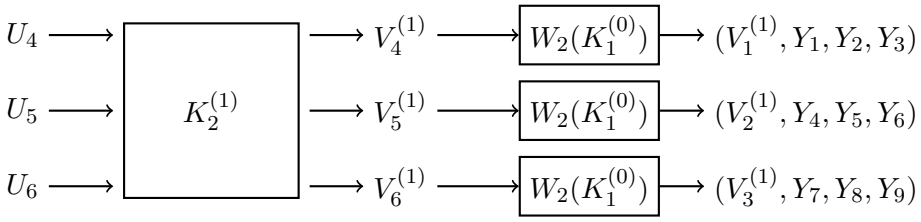

Figure 5: The (stochastic) mapping from $\boldsymbol{U}_{[4:6]}$ to $(\boldsymbol{V}_{[1:3]}^{(1)}, \boldsymbol{Y}_{[1:9]})$

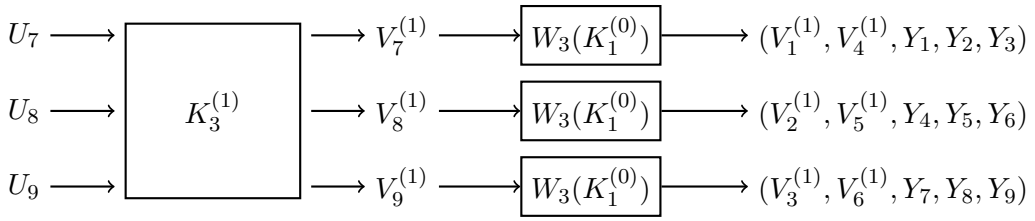

Figure 6: The (stochastic) mapping from $\boldsymbol{U}_{[7:9]}$ to $(\boldsymbol{V}_{[1:6]}^{(1)}, \boldsymbol{Y}_{[1:9]})$

generator matrix of this code is the submatrix of $M^{(t)}$ consisting of all the row vectors with indices belonging to the set $\mathcal{S}_{\text{good}}$.

9.1 Analysis of bit-channels

We say that two channels $W_1 : \{0,1\} \to \mathcal{Y}_1$ and $W_2 : \{0,1\} \to \mathcal{Y}_2$ are equivalent if there is a one-to-one mapping π between $\mathcal{Y}_1$ and $\mathcal{Y}_2$ such that $W_1(y_1|x) = W_2(\pi(y_1)|x)$ for all $y_1 \in \mathcal{Y}_1$ and $x \in \{0,1\}$. Denote this equivalence relation as $W_1 \equiv W_2$. Then we have the following result.

Proposition 9.1. *For every $i \in \ell^t$, we write $\tau_t(i) = (i_1, i_2, \ldots, i_t)$. Then we always have*

$$W_i(M^{(t)}) \equiv W_{i_1,\ldots,i_t}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{t-1}}^{(t-1)}).$$

Before formally proving this proposition, we first use the special case of $t = 2$ and $\ell = 3$ to illustrate the main idea behind the proof. In this case, we obtained one kernel $K_1^{(0)}$ in step 0 and three kernels $K_1^{(1)}, K_2^{(1)}, K_3^{(1)}$ in step 1. See Fig. 3 for an illustration of the encoding process $\boldsymbol{X}_{[1:9]} = \boldsymbol{U}_{[1:9]}M^{(2)}$. In particular, we can see that

$$\boldsymbol{V}_{[1:9]}^{(1)} = \boldsymbol{U}_{[1:9]}D^{(1)}, \qquad \boldsymbol{U}_{[1:9]}^{(1)} = \boldsymbol{V}_{[1:9]}^{(1)}Q^{(1)}, \qquad X_{[1:9]} = \boldsymbol{U}_{[1:9]}^{(1)}D^{(0)}.$$

Therefore, we indeed have $\boldsymbol{X}_{[1:9]} = \boldsymbol{U}_{[1:9]}D^{(1)}Q^{(1)}D^{(0)} = \boldsymbol{U}_{[1:9]}M^{(2)}$. Assume that $\boldsymbol{U}_{[1:9]}$ consists of 9 i.i.d. Bernoulli-1/2 random variables. Since $D^{(1)}, Q^{(1)}, D^{(0)}$ are all invertible matrices, the random vectors $\boldsymbol{V}_{[1:9]}^{(1)}, \boldsymbol{U}_{[1:9]}^{(1)}$ and $\boldsymbol{X}_{[1:9]}$ also consist of i.i.d. Bernoulli-1/2 random variables.

In order to analyze the bit-channels, we view Fig. 3 from the right side to the left side. First observe that the following three vectors

$$(U_1^{(1)}, U_2^{(1)}, U_3^{(1)}, Y_1, Y_2, Y_3), \qquad (U_4^{(1)}, U_5^{(1)}, U_6^{(1)}, Y_4, Y_5, Y_6), \qquad (U_7^{(1)}, U_8^{(1)}, U_9^{(1)}, Y_7, Y_8, Y_9)$$

are independent and identically distributed (i.i.d.).

Given a channel $W_1 : \mathcal{X} \to \mathcal{Y}$ and a pair of random variables (X, Y) that take values in $\mathcal{X}$ and $\mathcal{Y}$ respectively, we write

$$\mathbb{P}(X \to Y) \equiv W_1$$

if $\mathbb{P}(Y = y|X = x) = W(y|x)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, where $\mathbb{P}(X \to Y)$ means the channel that takes X as input and gives Y as output. By this definition, we have

$$\mathbb{P}(U_1^{(1)} \to \boldsymbol{Y}_{[1:3]}) \equiv \mathbb{P}(U_4^{(1)} \to \boldsymbol{Y}_{[4:6]}) \equiv \mathbb{P}(U_7^{(1)} \to \boldsymbol{Y}_{[7:9]}) \equiv W_1(K_1^{(0)}).$$

Since $V_1^{(1)} = U_1^{(1)}, V_2^{(1)} = U_4^{(1)}, V_3^{(1)} = U_7^{(1)}$, we also have

$$\mathbb{P}(V_1^{(1)} \to \boldsymbol{Y}_{[1:3]}) \equiv \mathbb{P}(V_2^{(1)} \to \boldsymbol{Y}_{[4:6]}) \equiv \mathbb{P}(V_3^{(1)} \to \boldsymbol{Y}_{[7:9]}) \equiv W_1(K_1^{(0)}).$$

Moreover, the following three vectors

$$(V_1^{(1)}, \mathbf{Y}_{[1:3]}), \qquad (V_2^{(1)}, \mathbf{Y}_{[4:6]}), \qquad (V_3^{(1)}, \mathbf{Y}_{[7:9]})$$

are independent. Therefore, the (stochastic) mapping from $U_{[1:3]}$ to $Y_{[1:9]}$ in Fig. 3 can be represented in a more compact form in Fig. 4. From Fig. 4, we can see that

$$\begin{aligned}
W_1(M^{(2)}) &\equiv \mathbb{P}(U_1 \to \mathbf{Y}_{[1:9]}) \equiv W_{1,1}(K_1^{(0)}, K_1^{(1)}),\\
W_2(M^{(2)}) &\equiv \mathbb{P}(U_2 \to (U_1, \mathbf{Y}_{[1:9]})) \equiv W_{1,2}(K_1^{(0)}, K_1^{(1)}),\\
W_3(M^{(2)}) &\equiv \mathbb{P}(U_3 \to (U_1, U_2, \mathbf{Y}_{[1:9]})) \equiv W_{1,3}(K_1^{(0)}, K_1^{(1)}).
\end{aligned}$$

Next we investigate $W_4(M^{(2)}), W_5(M^{(2)}), W_6(M^{(2)})$. Observe that

$$\mathbb{P}(U_2^{(1)} \to (U_1^{(1)}, \mathbf{Y}_{[1:3]})) \equiv \mathbb{P}(U_5^{(1)} \to (U_4^{(1)}, \mathbf{Y}_{[4:6]})) \equiv \mathbb{P}(U_8^{(1)} \to (U_7^{(1)}, \mathbf{Y}_{[7:9]})) \equiv W_2(K_1^{(0)}).$$

Therefore,

$$\mathbb{P}(V_4^{(1)} \to (V_1^{(1)}, \mathbf{Y}_{[1:3]})) \equiv \mathbb{P}(V_5^{(1)} \to (V_2^{(1)}, \mathbf{Y}_{[4:6]})) \equiv \mathbb{P}(V_6^{(1)} \to (V_3^{(1)}, \mathbf{Y}_{[7:9]})) \equiv W_2(K_1^{(0)}).$$

Moreover, since

$$(V_1^{(1)}, V_4^{(1)}, \mathbf{Y}_{[1:3]}), \qquad (V_2^{(1)}, V_5^{(1)}, \mathbf{Y}_{[4:6]}), \qquad (V_3^{(1)}, V_6^{(1)}, \mathbf{Y}_{[7:9]})$$

are independent, the (stochastic) mapping from $\boldsymbol{U}_{[4:6]}$ to $(\mathbf{V}_{[1:3]}^{(1)}, \mathbf{Y}_{[1:9]})$ in Fig. 3 can be represented in a more compact form in Fig. 5. Notice that there is a bijection between $\boldsymbol{U}_{[1:3]}$ and $\mathbf{V}_{[1:3]}^{(1)}$. Thus we can conclude from Fig. 5 that

$$\begin{aligned}
W_4(M^{(2)}) &\equiv \mathbb{P}(U_4 \to (\boldsymbol{U}_{[1:3]}, \mathbf{Y}_{[1:9]})) \equiv \mathbb{P}(U_4 \to (\mathbf{V}_{[1:3]}^{(1)}, \mathbf{Y}_{[1:9]})) \equiv W_{2,1}(K_1^{(0)}, K_2^{(1)}),\\
W_5(M^{(2)}) &\equiv \mathbb{P}(U_5 \to (\boldsymbol{U}_{[1:4]}, \mathbf{Y}_{[1:9]})) \equiv \mathbb{P}(U_5 \to (U_4, \mathbf{V}_{[1:3]}^{(1)}, \mathbf{Y}_{[1:9]})) \equiv W_{2,2}(K_1^{(0)}, K_2^{(1)}),\\
W_6(M^{(2)}) &\equiv \mathbb{P}(U_6 \to (\boldsymbol{U}_{[1:5]}, \mathbf{Y}_{[1:9]})) \equiv \mathbb{P}(U_6 \to (U_4, U_5, \mathbf{V}_{[1:3]}^{(1)}, \mathbf{Y}_{[1:9]})) \equiv W_{2,3}(K_1^{(0)}, K_2^{(1)}).
\end{aligned}$$

Finally, we can use the same method to show that

$$\begin{aligned}
&\mathbb{P}(V_7^{(1)} \to (V_1^{(1)}, V_4^{(1)}, \mathbf{Y}_{[1:3]})) \equiv \mathbb{P}(V_8^{(1)} \to (V_2^{(1)}, V_5^{(1)}, \mathbf{Y}_{[4:6]}))\\
\equiv &\mathbb{P}(V_9^{(1)} \to (V_3^{(1)}, V_6^{(1)}, \mathbf{Y}_{[7:9]})) \equiv W_3(K_1^{(0)}).
\end{aligned}$$

Therefore, the (stochastic) mapping from $\boldsymbol{U}_{[7:9]}$ to $(\mathbf{V}_{[1:6]}^{(1)}, \mathbf{Y}_{[1:9]})$ in Fig. 3 can be represented in a more compact form in Fig. 6. Notice that there is a bijection between $\boldsymbol{U}_{[1:6]}$ and $\mathbf{V}_{[1:6]}^{(1)}$. Thus we can conclude from Fig. 6 that

$$\begin{aligned}
W_7(M^{(2)}) &\equiv \mathbb{P}(U_7 \to (\boldsymbol{U}_{[1:6]}, \mathbf{Y}_{[1:9]})) \equiv \mathbb{P}(U_7 \to (\mathbf{V}_{[1:6]}^{(1)}, \mathbf{Y}_{[1:9]})) \equiv W_{3,1}(K_1^{(0)}, K_3^{(1)}),\\
W_8(M^{(2)}) &\equiv \mathbb{P}(U_8 \to (\boldsymbol{U}_{[1:7]}, \mathbf{Y}_{[1:9]})) \equiv \mathbb{P}(U_8 \to (U_7, \mathbf{V}_{[1:6]}^{(1)}, \mathbf{Y}_{[1:9]})) \equiv W_{3,2}(K_1^{(0)}, K_3^{(1)}),\\
W_9(M^{(2)}) &\equiv \mathbb{P}(U_9 \to (\boldsymbol{U}_{[1:8]}, \mathbf{Y}_{[1:9]})) \equiv \mathbb{P}(U_9 \to (U_7, U_8, \mathbf{V}_{[1:6]}^{(1)}, \mathbf{Y}_{[1:9]})) \equiv W_{3,3}(K_1^{(0)}, K_3^{(1)}).
\end{aligned}$$

Now we have proved Proposition 9.1 for the special case of $\ell = 3$ and $t = 2$. The proof for the general case follows the same idea, and we defer it to Appendix D.

9.2 Complexity of code construction, encoding and decoding

Proposition 9.2. *The code construction has $N^{O_\ell(1)}$ complexity. Both the encoding and successive decoding procedures have $O_\ell(N \log N)$ complexity.*

Proof. The key in our proof is that we consider ℓ as a (possibly very large) constant. We start with the code construction and we first show that both Algorithm A and Algorithm B have $\mathrm{poly}(N)$ time complexity. In the worst case, we need to check all 2^{ℓ^2} possible kernels in Algorithm A, and for each kernel we need to calculate the conditional entropy of the ℓ subchannels. Since we always work with the quantized channel with output size upper bounded by N^3, each subchannel of the quantized channels has no more than $2^\ell N^{3\ell}$ outputs. Therefore, the conditional entropy of these subchannels can be calculated in $\mathrm{poly}(N)$ time, so Algorithm A also has $\mathrm{poly}(N)$ complexity. After finding the good kernels, we need to use Algorithm B to quantize/bin the output alphabet of the subchannels produced by these good kernels. As mentioned above, the original alphabet size of these subchannels is no more than $2^\ell N^{3\ell}$. Therefore, Algorithm B also has $\mathrm{poly}(N)$ complexity. At Step i, we use Algorithm A ℓ^i times to find good kernels, and then we use Algorithm B ℓ^{i+1} times to quantize the bit-channels produced by these kernels, so in total we use Algorithm A $\frac{N-1}{\ell-1}$ times and we use Algorithm B $\frac{\ell(N-1)}{\ell-1}$ times. Finally, finding the set $\mathcal{S}_{\mathrm{good}}$ only requires calculating the conditional entropy of the bit-channels in the last step, so this can also be done in polynomial time. Thus we conclude that the code construction has $\mathrm{poly}(N)$ complexity, albeit the degree in $\mathrm{poly}(N)$ complexity depends on ℓ.

In the encoding procedure, we first form the vector $\boldsymbol{U}_{[1:N]}$ by putting all the information in the bits $\{U_i : \tau_t(i) \in \mathcal{S}_{\mathrm{good}}\}$ and setting all the other bits $\{U_i : \tau_t(i) \notin \mathcal{S}_{\mathrm{good}}\}$ to be 0. Then we multiply $\boldsymbol{U}_{[1:N]}$ with the encoding matrix $M^{(t)}$ and obtain the codeword $\boldsymbol{X}_{[1:N]} = \boldsymbol{U}_{[1:N]} M^{(t)}$. Since the matrix $M^{(t)}$ has size $N \times N$, a naive implementation of the encoding procedure would require $O(N^2)$ operations. Fortunately, we can use (101) to accelerate the encoding procedure. Namely, we first multiply $\boldsymbol{U}_{[1:N]}$ with $D^{(t-1)}$, then multiply the result with $Q^{(t-1)}$, then multiply by $D^{(t-2)}$, so on and so forth. As mentioned above, for $j = 0, 1, \ldots, t-1$, each $D^{(j)}$ is a block diagonal matrix with N/ℓ blocks on the diagonal, where each block has size $\ell \times \ell$. Therefore, multiplication with $D^{(j)}$ only requires $N\ell$ operations. By definition, $Q^{(j)}, j \in [t-1]$ are permutation matrices, so multiplication with them only requires N operations. In total, we multiply with $2t - 1 = 2\log_\ell N - 1$ matrices. Therefore, the encoding procedure can be computed in $O_\ell(N \log N)$ time, where O_ℓ means that the constant in big-O depends on ℓ.

The decoding algorithm uses exactly the same idea as the algorithm in Arıkan's original paper [Arı09, Section VIII-B]. Here we only use the special case of $\ell = 3$ and $t = 2$ in Fig. 3 to explain how Arıkan's decoding algorithm works for large (and mixed) kernels, and we omit the proof for general parameters. We start with the decoding of U_1, U_2, U_3 in Fig. 3. It is clear that decoding U_1, U_2, U_3 is equivalent to decoding $U_1^{(1)}, U_4^{(1)}, U_7^{(1)}$. Then the log-likelihood ratio (LLR) of each of these three bits can be calculated locally from only three output symbols. More precisely, the LLR of $U_1^{(1)}$ can be computed from $\boldsymbol{Y}_{[1:3]}$, the LLR of $U_4^{(1)}$ can be computed from $\boldsymbol{Y}_{[4:6]}$, and the LLR of $U_7^{(1)}$ can be computed from $\boldsymbol{Y}_{[7:9]}$. Therefore, the complexity of calculating each LLR only depends on the value of ℓ. Since ℓ is considered as a constant, the calculation of each LLR also has constant time complexity (although the complexity is exponential in ℓ). The next step is to decode $\boldsymbol{U}_{[4:6]}$ from $\boldsymbol{Y}_{[1:9]}$ together with $\boldsymbol{U}_{[1:3]}$. This is equivalent to calculating the LLRs of $U_2^{(1)}, U_5^{(1)}, U_8^{(1)}$ given $\boldsymbol{Y}_{[1:9]}$ and $U_1^{(1)}, U_4^{(1)}, U_7^{(1)}$. This again can be done locally: To compute the LLR of $U_2^{(1)}$, we only need the values of $\boldsymbol{Y}_{[1:3]}$ and $U_1^{(1)}$; to compute the LLR of $U_5^{(1)}$, we only need the values of $\boldsymbol{Y}_{[4:6]}$ and

$U_4^{(1)}$; to compute the LLR of $U_8^{(1)}$, we only need the values of $\mathbf{Y}_{[7:9]}$ and $U_7^{(1)}$. Finally, the decoding of $\boldsymbol{U}_{[7:9]}$ from $\mathbf{Y}_{[1:9]}$ and $\boldsymbol{U}_{[1:6]}$ can be decomposed into local computations in a similar way. Using this idea, one can show that for general values of ℓ and t, the decoding can also be decomposed into $t = \log_\ell N$ stages, and in each stage, the decoding can further be decomposed into N/ℓ local tasks, each of which has constant time complexity (although the complexity is exponential in ℓ). Therefore, the decoding complexity at each stage is $O_\ell(N)$ and the overall decoding complexity is $O_\ell(N \log N)$. As a final remark, we mention that after calculating the LLRs of all U_i's, we will only use the LLRs of the bits $\{U_i : \tau_t(i) \in \mathcal{S}_{\text{good}}\}$. For these bits, we decode U_i as 0 if its LLR is larger than 0 and decode it 1 otherwise. Recall that in the encoding procedure, we have set all the other bits $\{U_i : \tau_t(i) \notin \mathcal{S}_{\text{good}}\}$ to be 0, so for these bits we simply decode them as 0. □

9.3 Code rate and decoding error probability

In (95), we have defined the conditional entropy for all the bit-channels obtained in the last step (Step $t-1$). Here we also define the conditional entropy for the bit-channels obtained in the previous steps. More precisely, for every $j \in [t]$ and every $(i_1, i_2, \ldots, i_j) \in [\ell]^j$, we use the following short-hand notation:

$$\begin{aligned}
H_{i_1,\ldots,i_j}(W) &:= H(W_{i_1,\ldots,i_j}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{j-1}}^{(j-1)})) \\
H_{i_1,\ldots,i_j}^{\text{bin}}(W) &:= H(W_{i_1,\ldots,i_j}^{\text{bin}}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{j-1}}^{(j-1)})) \\
H_{i_1,\ldots,i_j}^{\text{bin}*}(W) &:= H(W_{i_1,\ldots,i_j}^{\text{bin}*}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{j-1}}^{(j-1)})).
\end{aligned}$$

According to (94), we have

$$H_{i_1,\ldots,i_j}^{\text{bin}*}(W) \le H_{i_1,\ldots,i_j}^{\text{bin}}(W) \le H_{i_1,\ldots,i_j}^{\text{bin}*}(W) + \frac{6 \log N}{N^3} \tag{102}$$

for every $j \in [t]$ and every $(i_1, i_2, \ldots, i_j) \in [\ell]^j$.

Proposition 9.3. *For every $j \in [t]$ and $(i_1, i_2, \ldots, i_j) \in [\ell]^j$, the conditional entropy $H_{i_1,\ldots,i_j}(W)$ and $H_{i_1,\ldots,i_j}^{\text{bin}}(W)$ satisfy the following inequality*

$$H_{i_1,\ldots,i_j}(W) \le H_{i_1,\ldots,i_j}^{\text{bin}}(W) \le H_{i_1,\ldots,i_j}(W) + \frac{6\ell \log N}{N^2} \tag{103}$$

Proof. Since the binning algorithm (Algorithm B) always produces a channel that is degraded with respect to the original channel, the first inequality in (103) follows immediately by applying Proposition 4.5 recursively in our t-step code construction.

Now we prove the second inequality in (103). We will prove the following inequality by induction on j:

$$H_{i_1,\ldots,i_j}^{\text{bin}}(W) \le H_{i_1,\ldots,i_j}(W) + \frac{6 \log N}{N^3}(1 + \ell + \ell^2 + \cdots + \ell^j) \qquad \forall (i_1, i_2, \ldots, i_j) \in [\ell]^j. \tag{104}$$

The base case of $j = 0$ is trivial. Now assume that this inequality holds for j and we prove it for $j+1$. By chain rule, we know that

$$\sum_{i_{j+1}=1}^{\ell} H_{i_1,\ldots,i_j,i_{j+1}}^{\text{bin}*}(W) = \ell H_{i_1,\ldots,i_j}^{\text{bin}}(W), \qquad \sum_{i_{j+1}=1}^{\ell} H_{i_1,\ldots,i_j,i_{j+1}}(W) = \ell H_{i_1,\ldots,i_j}(W).$$

Therefore,

$$\sum_{i_{j+1}=1}^{\ell}\Big(H^{\mathrm{bin}\,*}_{i_1,\dots,i_j,i_{j+1}}(W)-H_{i_1,\dots,i_j,i_{j+1}}(W)\Big)=\ell\Big(H^{\mathrm{bin}}_{i_1,\dots,i_j}(W)-H_{i_1,\dots,i_j}(W)\Big).$$

Since every summand on the left-hand side is non-negative, we have

$$H^{\mathrm{bin}\,*}_{i_1,\dots,i_j,i_{j+1}}(W)-H_{i_1,\dots,i_j,i_{j+1}}(W)\le\ell\Big(H^{\mathrm{bin}}_{i_1,\dots,i_j}(W)-H_{i_1,\dots,i_j}(W)\Big)\le\frac{6\log N}{N^3}(\ell+\ell^2+\cdots+\ell^{j+1}),$$

where the second inequality follows from the induction hypothesis. Combining this with (102), we obtain that

$$H^{\mathrm{bin}}_{i_1,\dots,i_j,i_{j+1}}(W)\le H_{i_1,\dots,i_j,i_{j+1}}(W)+\frac{6\log N}{N^3}(1+\ell+\ell^2+\cdots+\ell^{j+1}).$$

This establishes the inductive step and completes the proof of (104). The inequality (103) then follows directly from (104) by using the fact that $1+\ell+\cdots+\ell^j<\ell N$ for all $j\le t$. □

Recall that in Remark 5.2 we denoted by $\ell\ge\exp(\Omega(\alpha^{-1.01}))$ the conditions on ℓ to be large enough so that $\log\ell\ge\frac{11}{\alpha}$ and $\frac{\log\ell}{\log\log\ell+2}\ge\frac{3}{\alpha}$. In the theorems below, even though the statements hold for any $\alpha\in(0,1/12)$, we modify the intervals of α so that the rate appears positive in the formulations. This is also why in the formulation of the Theorem 1.1 we take α from $(0,1/36)$.

We now can formulate

Theorem 9.4. *For arbitrarily small $\alpha\in\left(0,\frac{1}{14}\right)$, if we choose a large enough constant $\ell\ge\exp(\Omega(\alpha^{-1.01}))$ to be a power of 2 and let $t=\log_\ell N$ grow, then the codes constructed from the above procedure have decoding error probability $O_\alpha(\log N/N)$ under successive decoding and code rate $I(W)-N^{-1/2+7\alpha}$, where $N=\ell^t$ is the code length.*

Proof. By (103) and the definition of $\mathcal{S}_{\mathrm{good}}$ in (96), we know that for every $(i_1,\dots,i_t)\in\mathcal{S}_{\mathrm{good}}$, we have $H_{i_1,\dots,i_t}(W)\le H^{\mathrm{bin}}_{i_1,\dots,i_t}(W)\le\frac{7\ell\log N}{N^2}$. Then by Lemma 2.2 in [BGN+18], we know that the ML decoding error probability of the bit-channel $W_{i_1,\dots,i_t}(K_1^{(0)},K_{i_1}^{(1)},\dots,K_{i_1,\dots,i_{t-1}}^{(t-1)})$ is also upper bounded by $\frac{7\ell\log N}{N^2}$. Since the cardinality of $\mathcal{S}_{\mathrm{good}}$ is at most N, we can conclude that the overall decoding error probability under the successive cancellation decoder is $O_\alpha(\log N/N)$ using the union bound.

Notice that $|\mathcal{S}_{\mathrm{good}}|$ is the code dimension. Therefore, we only need to lower bound $|\mathcal{S}_{\mathrm{good}}|$ in order to get the lower bound on the code rate. Define another set

$$\mathcal{S}'_{\mathrm{good}}:=\left\{(i_1,i_2,\dots,i_t)\in[\ell]^t:H_{i_1,\dots,i_t}(W)\le\frac{\ell\log N}{N^2}\right\}.\tag{105}$$

According to (103), if $H_{i_1,\dots,i_t}(W)<\frac{\ell\log N}{N^2}$, then $H^{\mathrm{bin}}_{i_1,\dots,i_t}(W)\le\frac{7\ell\log N}{N^2}$. Therefore, $\mathcal{S}'_{\mathrm{good}}\subseteq\mathcal{S}_{\mathrm{good}}$, so $|\mathcal{S}_{\mathrm{good}}|\ge|\mathcal{S}'_{\mathrm{good}}|$. In Lemma 9.5 below, we will prove that $|\mathcal{S}'_{\mathrm{good}}|\ge N(I(W)-N^{-1/2+7\alpha})$. Therefore, $|\mathcal{S}_{\mathrm{good}}|\ge N(I(W)-N^{-1/2+7\alpha})$. This completes the proof of the theorem. □

Lemma 9.5. *If $\alpha\in\left(0,\frac{1}{14}\right)$ and ℓ is large enough so that $\log\ell\ge\frac{11}{\alpha}$ and $\frac{\log\ell}{\log\log\ell+2}\ge\frac{3}{\alpha}$, then the set $\mathcal{S}'_{\mathrm{good}}$ defined in (105) satisfies the following inequality*

$$\left|\mathcal{S}'_{\mathrm{good}}\right|\ge N\left(I(W)-N^{-\frac{1}{2}+7\alpha}\right).$$

Proof. The proof is the same as in [BGN+18, Claim A.2]. Recall that we proved in (5)–(8)

$$\mathbb{P}\left[H^{(t)} \in \left(\frac{\ell \log N}{N^2}, 1 - \frac{\ell \log N}{N^2}\right)\right] \leq 2\frac{N^{2\alpha}}{(\ell \log N)^{\alpha}} \cdot \lambda_{\alpha}^{t},$$

where $H^{(t)}$ is (marginally) the entropy of the random channel at the last level of construction, i.e. $H^{(t)}$ is uniformly distributed over $H_{i_1,\ldots,i_t}(W)$ for all possible $(i_1, i_2, \ldots, i_t) \in [\ell]^t$, and λ_{α} is such that (6) holds for any channel W' throughout the construction. By Proposition 9.3, we can choose the error parameter Δ in Algorithm A to be $\Delta = \frac{6\ell \log N}{N^2}$, which satisfies the condition $\Delta \leq \ell^{-\log \ell}$ in Theorem 5.1. Then Theorem 5.1 and Remark 5.2 tell us that as long as the conditions on ℓ and α specified in this lemma hold, Algorithm A allows us to choose kernels such that $\lambda_{\alpha} \leq \ell^{-1/2+5\alpha}$, which gives

$$\mathbb{P}\left[H^{(t)} \in \left(\frac{\ell \log N}{N^2}, 1 - \frac{\ell \log N}{N^2}\right)\right] \leq \frac{2N^{-1/2+7\alpha}}{(\ell \log N)^{\alpha}}. \tag{106}$$

On the other hand, conservation of entropy throughout the process implies $E\left[H^{(t)}\right] = H(W)$, therefore by Markov's inequality

$$\mathbb{P}\left[H^{(t)} \geq 1 - \frac{\ell \log N}{N^2}\right] \leq \frac{H(W)}{1 - \frac{\ell \log N}{N^2}} \leq H(W) + \frac{2\ell \log N}{N^2}.$$

Since $H(W) = 1 - I(W)$ for symmetric channels and $\left|\mathcal{S}'_{\text{good}}\right| = N \cdot \mathbb{P}\left[H^{(t)} \leq \frac{\ell \log N}{N^2}\right]$, we have

$$\begin{aligned}
\left|\mathcal{S}'_{\text{good}}\right| &\geq N\left(1 - \frac{2N^{-1/2+7\alpha}}{(\ell \log N)^{\alpha}} - H(W) - \frac{2\ell \log N}{N^2}\right) \\
&\geq N\left(I(W) - \frac{3N^{-1/2+7\alpha}}{(\ell \log N)^{\alpha}}\right) \\
&\geq N\Big(I(W) - N^{-1/2+7\alpha}\Big).
\end{aligned}$$

□

9.4 Main theorem: Putting everything together

As we mentioned at the beginning of this section, the code construction presented above only takes the special case of $\mathsf{Q} = N^3$ as a concrete example, where Q is the upper bound on the output alphabet size after binning; see Algorithm B. In fact, we can change the value of Q to be any polynomial of N, and this will allow us to obtain a trade-off between the decoding error probability and the gap to capacity while maintaining the polynomial-time code construction as well as the $O_{\alpha}(N \log N)$ encoding and decoding complexity. More precisely, we have the following theorem.

Theorem 9.6. *For any BMS channel W, any $c > 0$ and arbitrarily small $\alpha \in \left(0, \frac{1}{12+2c}\right)$, if we choose a large constant ℓ to be a power of 2 which satisfies $\log \ell \geq \frac{11}{\alpha}$ and $\frac{\log \ell}{\log \log \ell + 2} \geq \frac{3}{\alpha}$, and set $\mathsf{Q} = N^{c+2}$ in the above code construction procedure, then we can construct a code $\mathcal{C}$ with code length $N = \ell^t$ such that the following four properties hold when t grows: (1) the code construction has $N^{O_{\alpha}(1)}$ complexity; (2) both encoding and decoding have $O_{\alpha}(N \log N)$ complexity; (3) rate of $\mathcal{C}$ is $I(W) - O(N^{-1/2+(c+6)\alpha})$; (4) decoding error probability of $\mathcal{C}$ is $O_{\alpha}(\log N / N^c)$ under successive decoding when $\mathcal{C}$ is used for channel coding over W.*

Proof. The proof of properties (1) and (2) is exactly the same as Proposition 9.2. Here we only briefly explain how to adjust the proof of Theorem 9.4 to show properties (3) and (4). First, we change the definitions of $\mathcal{S}_{\text{good}}$ and $\mathcal{S}'_{\text{good}}$ to

$$\mathcal{S}_{\text{good}} := \left\{(i_1, i_2, \dots, i_t) \in [\ell]^t : H^{\text{bin}}_{i_1,\dots,i_t}(W) \leq \frac{(2c+3)\ell \log N}{N^{c+1}}\right\},$$
$$\mathcal{S}'_{\text{good}} := \left\{(i_1, i_2, \dots, i_t) \in [\ell]^t : H_{i_1,\dots,i_t}(W) \leq \frac{\ell \log N}{N^{c+1}}\right\}.$$

The definition of $\mathcal{S}_{\text{good}}$ immediately implies property (4). Next we prove property (3). Since we change Q from N^3 to N^{c+2}, inequality (102) becomes

$$H^{\text{bin}\,*}_{i_1,\dots,i_j}(W) \leq H^{\text{bin}}_{i_1,\dots,i_j}(W) \leq H^{\text{bin}\,*}_{i_1,\dots,i_j}(W) + \frac{2(c+2)\log N}{N^{c+2}}.$$

As a consequence, inequality (103) in Proposition 9.3 becomes

$$H_{i_1,\dots,i_j}(W) \leq H^{\text{bin}}_{i_1,\dots,i_j}(W) \leq H_{i_1,\dots,i_j}(W) + \frac{2(c+2)\ell \log N}{N^{c+1}}.$$

This inequality tells us that $\mathcal{S}'_{\text{good}} \subseteq \mathcal{S}_{\text{good}}$, so $|\mathcal{S}_{\text{good}}| \geq |\mathcal{S}'_{\text{good}}|$. Then we follow Lemma 9.5 to lower bound $|\mathcal{S}'_{\text{good}}|$. Inequality (106) now becomes

$$\mathbb{P}\left[H^{(t)} \in \left(\frac{\ell \log N}{N^{c+1}}, 1 - \frac{\ell \log N}{N^{c+1}}\right)\right] \leq \frac{2N^{-1/2+(c+6)\alpha}}{(\ell \log N)^{\alpha}}.$$

Therefore, we obtain that

$$|\mathcal{S}_{\text{good}}| \geq |\mathcal{S}'_{\text{good}}| \geq N\Big(I(W) - N^{-1/2+(c+6)\alpha}\Big).$$

This completes the proof of the theorem. □

10 Inverse sub-exponential decoding error probability

In this section we finish proving our main result (Theorem 1.1), by showing how to obtain inverse sub-exponential $\exp(-N^{\alpha})$ probability of error decoding within our construction of polar codes, while still having $\text{poly}(N)$ time complexity of construction. Note that up to this point we only claimed inverse polynomial decoding error probability in Theorem 9.6. This restriction came from the fact that we need to approximate the channels we see in the tree during the construction phase (recall the discussion at the beginning of Sections 5.1 and 9), and to get a polynomial-time construction we need the binning parameter Q to be $\text{poly}(N)$ itself. But this means that we are only able to track the parameters (entropies, for instance) of the bit-channels approximately, with an additive error which is inverse polynomial in N, see (102). Since the decoding error probability relates directly to the upper bound on the entropies of the "good" bit-channels we choose, this leads to only being able to claim inverse polynomial decoding error probability.

It was proved in a recent work [WD19] that it is possible to achieve a fast scaling of polar codes (good scaling exponent) and good decoding error probability (inverse sub-exponential instead of inverse polynomial in N) simultaneously, using the idea of multiple (dynamic) kernels in the construction. Specifically, for any constants $\pi, \mu > 0$ such that $\pi + 2\mu < 1$, it is shown that one can construct a polar code with rate $N^{-\mu}$ close to capacity of the channel (which corresponds to

scaling exponent μ) and decoding error probability $\exp(-N^{\pi})$, as $N \to \infty$. Moreover, it is shown that this is an optimal scaling of these two parameters one can obtain for *any* (not just polar) codes. However, the construction phase in [WD19] tracked the *true* bit-channels that are obtained in the ℓ-ary tree of channels, which makes the construction intractable. This is because (most of) the true bit-channels cannot even be described in a tractable way, since they have exponential size of output alphabet.

In what follows we combine our approach of using Arıkan's kernels for polarized bit-channels with a stronger analysis of polarization from [WD19] to overcome this issue of intractable construction. Specifically, we show that even though we only track *approximations* (binned versions) of the bit-channels in the tree, if we use Arıkan's channels for suction at the end regime, then we are still able to prove very strong polarization, as in [WD19]. This comes from the fact that we know very well how Arıkan's basic 2×2 kernel evolves the parameters of the bit-channels. This allows us to get very strong bounds on the parameters of the *true* bit-channels (which leads to good decoding error probability), while still only tracking their *approximations* (which keeps the construction time polynomial). Somewhat surprisingly, the phase of the construction where the local kernels are chosen is exactly the same as it was before in Section 9, and the difference lies in a much tighter analysis of how to choose a set of "good" indices to actually construct a polar code.

Notations

We fix a small positive parameter $\alpha > 0$ from the statement of Theorem 1.1, which corresponds to how close the scaling exponent will be to $1/2$. Specifically, we will have the scaling exponent $\mu = 2 + O(\alpha)$. As before, the size of the kernel is denoted by $\ell = 2^s$, where ℓ is large enough in terms of α (specifically, the bounds from the statement of the Theorem 5.1 must hold).

We are going to work with the complete ℓ-ary tree of bit-channels, as described in Section 2.4. Let t be the depth of this tree, then there are $N = \ell^t$ bit-channels at the last level, denoted as W_i for $i \in [\ell^t]$ (these notations depend on the depth t of the tree at which we are looking, but it will always be clear from the context). Throughout this section we will denote such a tree of depth t as $\mathcal{T}_t$.

We will again have a random process of going down the tree, starting from the root, and picking a random child of a current bit-channel at each step. To be more precise, the random process W_i is defined as follows: $\mathsf{W}_0 = W$ (the initial channel, i.e. the root of the tree), and $\mathsf{W}_{j+1} = (\mathsf{W}_j)_k$, where $k \sim [\ell]$, and $(\mathsf{W}_j)_k$ is the k^{th} Arıkan's bit-channel of W_j with respect to the corresponding kernel in the tree. This indeed is equivalent to a random walk down the tree. Then we also define the random processes $\mathsf{Z}_j = Z(\mathsf{W}_j)$ and $\mathsf{H}_j = H(\mathsf{W}_j)$. Note that W_t marginally is distributed as W_i for $i \sim [N]$, where $N = \ell^t$, i.e. W_t is just a random bit-channel at the level t of the tree. Further, we will also look at random processes $\mathsf{W}_j^{\text{bin}}, \mathsf{H}_j^{\text{bin}}, \mathsf{Z}_j^{\text{bin}}$, which mean that we also do the binning procedure as described in the construction phase in Section 9. Note that $\mathsf{W}_j^{\text{bin}}$ are the channels that we actually track during the construction of the code, while W_j are the *true* bit-channels in the tree.

Finally, by $\exp(\bullet)$ we will denote $2^{\bullet}$ in this section, and we denote by $x^+ = \max\{x, 0\}$ the positive part of x.

Plan

First, notice that building the tree $\mathcal{T}_t$ of bit-channels is itself a part of construction of our polar codes. This includes tracking the binned versions of the bit-channels, and picking the kernels using

Algorithm A. This part will stay exactly the same as it is described in Section 9, with the binning parameter $\mathsf{Q} = N^3$, and the same threshold of ℓ^{-4} in the Algorithm A. The only part of the construction that is going to change is how we pick the set of good indices which we use to transmit information.

We will closely follow the analysis from [WD19, Appendices B, C] (also appearing in [WD18]), modified for our purposes. Specifically, we will prove the needed polarization of the construction presented in Section 9 in three steps (recall that $s = \log_2 \ell$):

1) $\mathbb{P}\left[\mathsf{Z}_t \leq \exp(-2st)\right] \geq I(W) - \ell^{-(1/2-10\alpha)t}$,

2) $\mathbb{P}\left[\mathsf{Z}_t \leq \exp\left(-2^{t^{1/3}}\right)\right] \geq I(W) - \ell^{-(1/2-11\alpha)t+\sqrt{t}}$,

3) $\mathbb{P}\left[\mathsf{Z}_t \leq \exp\left(-st \cdot \ell^{\alpha \cdot t}\right)\right] \geq I(W) - \ell^{-(1/2-16\alpha)t+2\sqrt{t}}$ for $t = \Omega(\log^6 s)$.

Moreover, for each step, we prove that the polarization at each step is *poly-time constructible*:

Definition 10.1. *We call the polarization $\mathbb{P}[\mathsf{Z}_t \leq p(t)] \geq R(t)$ to be* poly-time constructible *if one can find at least $N \cdot R(t)$ indexes $i \in [N]$ such that $Z(W_i) \leq p(t)$, where $N = \ell^t$, in time polynomial in N.*

Notice that if polarization $\mathbb{P}[\mathsf{Z}_t \leq p(t)] \geq R(t)$ is poly-time constructible, then by choosing these $N \cdot R(t)$ indexes as information bits of the code, a standard argument implies that one obtains a polar code of rate $R(t)$ and decoding error probability at most $N \cdot p(t)$. Moreover, since the indexes of the information bits were found in $\mathrm{poly}(N)$ time, this makes the whole code construction complexity polynomial in N.

The polarization behavior from Step 3 with $t \geq \frac{1}{\alpha^2}$ will then correspond to polar codes with rate $I(W) - N^{-1/2+18\alpha}$ (i.e. codes with scaling exponent $(2 + O(\alpha))$ and sub-exponentially small decoding error probability $N \cdot \exp\left(-st \cdot \ell^{\alpha \cdot t}\right) = \exp(-N^{\alpha})$, with $\mathrm{poly}(N)$ construction time, which finishes the proof of the main result of this paper.

10.1 Step 1

Lemma 10.2. $\mathbb{P}\left[\mathsf{Z}_t \leq \exp(-2st)\right] \geq I(W) - \ell^{-(1/2-10\alpha)t}$. *Moreover, this polarization is poly-time constructible.*

Proof. This follows from the analysis of the construction we already have in the previous sections. Fix some t and let $N = \ell^t$. Then the following is implied from Section 9.4 if one takes $\mathsf{Q} = N^3$, i.e. $c = 3$:

$$\mathop{\mathbb{P}}_{i \sim [N]}\left[H(W_i^{\mathrm{bin}}) \leq \frac{1}{N^4}\right] \geq I(W) - N^{-(1/2-10\alpha)}.$$

Note here that $H(W_i^{\mathrm{bin}})$ are the entropies of the binned bit-channels that we are actually tracking during the construction phase, so they are computable in polynomial time. This means that there is $\mathrm{poly}(N)$-time procedure which returns all the indices i for which $H(W_i^{\mathrm{bin}}) \leq \frac{1}{N^4}$. Then $Z(W_i^{\mathrm{bin}}) < \sqrt{H(W_i^{\mathrm{bin}})} \leq \frac{1}{N^2}$ for these indices, so we have for the random process $\mathsf{Z}_t^{\mathrm{bin}}$:

$$\mathbb{P}\left[\mathsf{Z}_t^{\mathrm{bin}} \leq \ell^{-2t}\right] = \mathbb{P}\left[\mathsf{Z}_t^{\mathrm{bin}} \leq 2^{-2st}\right] = \mathbb{P}\left[\mathsf{Z}_t^{\mathrm{bin}} \leq \exp\left(-2st\right)\right] \geq I(W) - N^{-(1/2-10\alpha)},$$

and moreover, one can find at least $N(I(W) - N^{-(1/2-10\alpha)})$ indexes within $i \in [N]$ for which the inequality $Z(W_i^{\text{bin}}) \leq \exp(-2st)$ holds in poly(N) time (just by returning the indices for which $H(W_i^{\text{bin}}) \leq \frac{1}{N^4}$). Since it always holds $\mathsf{Z}_t \leq \mathsf{Z}_t^{\text{bin}}$, the statement of the lemma follows. □

10.2 Step 2

Next, we are going to strengthen the polarization of the construction, using the result of Lemma 10.2. Specifically, we prove

Lemma 10.3. $\mathbb{P}\left[\mathsf{Z}_n \leq \exp\left(-2^{n^{1/3}}\right)\right] \geq I(W) - \ell^{-(1/2-11\alpha)n+\sqrt{n}}$. *Moreover, this polarization is poly-time constructible.*

Proof. For this lemma, we fix n to be the total depth of the tree (instead of t), and we want to prove the speed of polarization at level n. To do this, we will divide the tree into $\sqrt{n}$ stages, each of depth $\sqrt{n}$, and apply the polarization we obtained at Step 1 at each stage. So, we look at m being $\sqrt{n}$, $2\sqrt{n}, \ldots, n - \sqrt{n}$. Define the following events, starting with $E_0^{(0)} = \emptyset$ (again, closely following [WD19]):

$$
\begin{aligned}
A_m &= \left\{\mathsf{Z}_m^{\text{bin}} < \exp(-2sm)\right\} \setminus E_0^{(m-\sqrt{n})} \\
B_m &= A_m \bigcap \left\{\sum_{i=1}^{s\sqrt{n}} g_{sm+i} \leq \beta \cdot s\sqrt{n}\right\} \\
E_m &= A_m \setminus B_m \\
E_0^{(m)} &= E_0^{(m-\sqrt{n})} \cup E_m,
\end{aligned}
$$

where for now one can think of g_j's as of independent Bern(1/2) random variables for all $j \in [s \cdot n]$. In the following several paragraphs we explain what these events are going to correspond to. First of all, the actual random variable we are tracking here is W_n, and its realizations are ℓ^n bit-channels W_i for $i \in [\ell^n]$ at the last level of the tree. We can then think of events and subsets of bit-channels at level n interchangeably.

Notice that each bit-channel W_i for $i \in [\ell^n]$ corresponds to a unique path in the tree $\mathcal{T}_n$ from the root W (the initial channel) to the leaf W_i on the n^{th} level. We will be interested in the bit-channels on these path, their binned versions, and the parameters of both versions (true and binned) of these channels during the ensuing arguments. We denote this path of true bit-channels as $W_i^{(0)} = W, W_i^{(1)}, \ldots, W_i^{(n-1)}, W_i^{(n)} = W_i$. Clearly, this path is just a realization of a random walk $\mathsf{W}_0, \mathsf{W}_1, \ldots, \mathsf{W}_n$, when W_n ends up being W_i. In the same way, we will denote by $W_i^{(k),\text{bin}}$, for $k = 0, 1, \ldots, n$ the binned version of the bit-channel along this path, and by $H_i^{(k)}$, $H_i^{(k),\text{bin}}$, $Z_i^{(k)}$, and $Z_i^{(k),\text{bin}}$ the corresponding parameters of these channels.

We are going to construct a set of "good" bit-channels $E_0^{(n-\sqrt{n})}$ incrementally, by inspecting the tree from top to bottom. We start with the set $E_0^{(0)} = \emptyset$. Then, at each stage $m = \sqrt{n}$, $2\sqrt{n}, \ldots, n-\sqrt{n}$, we find a set E_m of bit-channels which we mark to be "good" at level m. Precisely, the channel W_i, for some $i \in [\ell^n]$, is going to be in E_m, if: a) it is not marked as good before that (i.e. it is not in $E_0^{(m-\sqrt{n})}$); b) the Bhattacharyya parameter $Z_i^{(m),\text{bin}}$ is small, specifically smaller then $\exp(-2sm)$; and c) a certain condition holds for how the branches are chosen in the path for W_i between levels m and $m + \sqrt{n}$ in the tree (more details on this later). Here conditions a) and

b) correspond together to the event A_m, while condition c) further defines the event B_m. Then the set $E_0^{(m)}$ will be the set of all bit-channels that we marked to be good up to the level m in the tree, and in the end, by collecting all the bit-channels that we marked as good at the stages $m = \sqrt{n}, 2\sqrt{n}, \dots, n - \sqrt{n}$, we obtain the final set $E_0^{(n-\sqrt{n})}$.

Denote by corresponding lowercase letters the probabilities of the events described before, i.e. $a_m \coloneqq \mathbb{P}[A_m]$, etc.. Finally, let $q_m = I(W) - e_0^{(m)}$, i.e. q_m is the gap between the capacity and the fraction of the channels which we marked as "good" up to level m.

To begin the formal analysis, let us first consider what happens in case of the event A_m. First, it means that $\mathsf{Z}_m^{\text{bin}} < \exp(-2sm)$. But then we know that we are going to apply Arıkan's kernel $A_2^{\otimes s}$ to this bit-channel at level m, since the threshold for picking Arıkan's kernel in Algorithm A, which we use in the construction phaze, is $\ell^{-4} = \exp(-4s)$. This means that, conditioned on A_m, we have $\mathsf{Z}_{m+1} \le \mathsf{Z}_m \cdot 2^s \le \mathsf{Z}_m^{\text{bin}} \cdot 2^s < 2^s \cdot \exp(-2sm)$, where the first inequality follows from that we know how Bhattacharrya parameter evolves when we use basic Arıkan's transforms. Precisely, using the kernel $A_2^{\otimes s}$ is equivalent to using the basic 2×2 kernel A_2 for s times, and the kernel A_2 in the worst case doubles the Bhattacharyya parameter. Thus s applications of A_2 can increase the Bhattacharyya parameter by at most a factor of 2^s.

Then it is easy to see that even after we apply Arıkan's kernel $A_2^{\otimes s}$ a total of $\sqrt{n}$ times, the Bhattacharyya parameter will still be below the threshold ℓ^{-4}: conditioned on A_m, one has $\mathsf{Z}_{m+\sqrt{n}} \le \mathsf{Z}_m \cdot (2^s)^{\sqrt{n}} < \exp(-2sm) \cdot \exp(s\sqrt{n}) < \exp(-sm) < \ell^{-4}$, as $m \ge \sqrt{n}$. It is easy to verify, using Proposition 9.3 and the relation (79) between the entropy and Bhattacharyya parameter of the bit-channel, that the binned parameter $\mathsf{H}_{m+j}^{\text{bin}}$ will also be below ℓ^{-4} for $j = 1, 2, \dots, \sqrt{n}$. This means that indeed for these $\sqrt{n}$ levels, the Arıkan's kernel was taken in the construction phase. Therefore, we know that only the kernel $A_2^{\otimes s}$ was applied at levels between m and $m + \sqrt{n}$, which can also be viewed as applying the basic 2×2 kernel A_2 for $s\sqrt{n}$ levels in the tree. Further this can be viewed as taking $s\sqrt{n}$ "good" or "bad" branches while going down the tree, where the good branch corresponds to squaring the Bhattachryya parameter, and the bad branch at most doubles it. Denote then by bits $g_{sm+i} \in \{0, 1\}$, for $i \in [s\sqrt{n}]$, the indicators of these branches being good or bad, where $g_{sm+i} = 0$ means the branch is bad, and $g_{sm+i} = 1$ means the branch is good. It is clear then that since we consider the random process of going down the tree choosing the next child randomly, then all g_{sm+i}'s are independent Bern(1/2) random variables. These are exactly the random variables appearing in the definition of B_m.

Notice then that

$$\frac{b_m}{a_m} = \mathbb{P}\left[\sum_{i=1}^{s\sqrt{n}} g_{sm+i} \le \beta \cdot s\sqrt{n}\right] \le 2^{-s\sqrt{n}(1-h_2(\beta))} \le 2^{-\gamma s\sqrt{n}} ,$$

where we can take, for instance, $\beta = 1/20$ and $\gamma = 0.85$. The inequality follows from entropic bound on the sum of binomial coefficients (one could also just use the Chernoff bound).

Recall that we defined $q_m = I(W) - e_0^{(m)}$. We then can write $q_{m-\sqrt{n}} - a_m = I(W) - (e_0^{(m-\sqrt{n})} + a_m)$. But note that by definition, the event $\left\{\mathsf{Z}_m^{\text{bin}} < \exp(-2sm)\right\}$ is a subevent of $A_m \cup E_0^{(m-\sqrt{n})}$, and thus using the bound from Lemma 10.2 (applied for the depth m) we know that

$$(e_0^{(m-\sqrt{n})} + a_m) \ge \mathbb{P}[A_m \cup E_0^{(m-\sqrt{n})}] \ge \mathbb{P}[\mathsf{Z}_m^{\text{bin}} < \exp(-2sm)] \ge I(W) - 2^{(-1/2+10\alpha)sm} .$$

Therefore we conclude

$$(q_{m-\sqrt{n}} - a_m)^+ \le 2^{(-1/2+10\alpha)sm}.$$

We can then derive

$$\begin{aligned} q_m &= I(W) - e_0^{(m)} = I(W) - (e_0^{(m-\sqrt{n})} + e_m) = q_{m-\sqrt{n}} - e_m \\ &= q_{m-\sqrt{n}} \left(1 - \frac{e_m}{a_m}\right) + \frac{e_m}{a_m}(q_{m-\sqrt{n}} - a_m) \\ &\leq q^+_{m-\sqrt{n}} \cdot \frac{b_m}{a_m} + (q_{m-\sqrt{n}} - a_m)^+ \\ &\leq q^+_{m-\sqrt{n}} \cdot 2^{-\gamma s\sqrt{n}} + 2^{(-1/2+10\alpha)sm}. \end{aligned}$$

Thus we end up we the following recurrence on q_m^+ (recall that $\ell = 2^s$):

$$\begin{aligned} q^+_{\sqrt{n}} &\leq 1 \\ q^+_m &\leq q^+_{m-\sqrt{n}} \cdot \ell^{-\gamma\sqrt{n}} + \ell^{-\frac{m}{2}+10\alpha m}. \end{aligned}$$

Solving this recurrence gives us $q^+_{n-\sqrt{n}} \leq \ell^{-\frac{n}{2}+11\alpha n+\sqrt{n}}$, since $\gamma > 1/2$. Therefore we can conclude

$$e_0^{(n-\sqrt{n})} \geq I(W) - \ell^{-\frac{n}{2}+11\alpha n+\sqrt{n}}. \tag{107}$$

Next, let us look at an arbitrary bit-channel (realization of Z_n) for which the event $E_0^{(n-\sqrt{n})}$ happens, and prove that such a bit-channel is indeed "good." Since $E_0^{(n-\sqrt{n})}$ happened, it means that E_m happened at some stage, thus $\mathsf{Z}_m^{\text{bin}} < \exp(-2sm)$ and $\sum_{i=1}^{s\sqrt{n}} g_{sm+i} \geq \beta \cdot s\sqrt{n}$, where g_{sm+i} for $i \in [s\sqrt{n}]$ correspond to taking bad or good branches in the basic 2×2 Arıkan's kernel. Similarly to Claim 8.2, we then can bound

$$\mathsf{Z}_{m+\sqrt{n}} < \left(2^{s\sqrt{n}}\mathsf{Z}_m\right)^{2^{\beta\cdot s\sqrt{n}}} < (2^{sm}\exp(-2sm))^{2^{\beta\cdot s\sqrt{n}}} \leq \exp\left(-sm \cdot 2^{\beta\cdot s\sqrt{n}}\right).$$

Then for the remaining $(n - m - \sqrt{n})$ levels of the tree, it is easy to see that the Bhattacharyaa parameter will also not ever be above the threshold of picking Arıkan's kernel in Algorithm A, thus, similarly as before, we can argue that the Bhattacharyya parameter increases by at most a factor of 2^s at each level. Therefore, we derive

$$\mathsf{Z}_n < 2^{s(n-m-\sqrt{n})} Z_{m+\sqrt{n}} \leq 2^{sn} \exp\left(-sm \cdot 2^{\beta\cdot s\sqrt{n}}\right) < \exp\left(-2^{n^{1/3}}\right),$$

where the last inequality follows from $m \geq \sqrt{n}$, $\beta = \frac{1}{20}$, and the condition $s \geq \frac{11}{\alpha}$ from Theorem 5.1 combined with the fact that α is small.

Since we proved that the event $E_0^{(n-\sqrt{n})}$ implies $\mathsf{Z}_n < \exp(-2^{n^{1/3}})$, we conclude, using (107):

$$\mathbb{P}[\mathsf{Z}_n < \exp(-2^{n^{1/3}})] \geq e_0^{(n-\sqrt{n})} \geq I(W) - \ell^{-\frac{n}{2}+11\alpha n+\sqrt{n}},$$

which precisely proves the polarization that was stated in the lemma.

The only thing left to prove then is that this polarization is poly-time constructible. To do this, we show that one can find the set $E_0^{(n-\sqrt{n})}$ of bit-channels in poly-time (recall here the equivalence between events and subsets of the bit-channel at the level n of the tree $\mathcal{T}_n$). But one can see that checking if a particular bit-channel W_i, for some $i \in [\ell^n]$, is easy. Indeed, to check if W_i is in $E_0^{(n-\sqrt{n})}$, it suffices to check if W_i is in E_m for any $m = \sqrt{n}, 2\sqrt{n}, \ldots, n - \sqrt{n}$. But this corresponds to looking at a Bhattacharyya parameter $Z_i^{(m),\text{bin}}$ and checking if it is smaller than $\exp(-2sm)$, and,

if this is the case, also looking at how many "good" branches (in the basic 2×2 Arıkan's transforms) there were within the next stage ($\sqrt{n}$ levels) in the tree $\mathcal{T}_n$. The latter can be done easily, since this information is essentially given by the index i of the bit-channel W_i (by its binary representation, to be precise). The former is actually also straightforward, since $Z_i^{(m),\mathrm{bin}}$ is the parameter of the binned bit-channel $W_i^{(m),\mathrm{bin}}$ that we are *actually tracking* during the construction phase, so we have this channel written down explicitly, and thus calculating its Bhattacharyya parameter is simple. Therefore all this can be done in time, polynomial in ℓ^n, and then the whole set $E_0^{(n-\sqrt{n})}$ can be found in poly-time (we can also say that the event $E_0^{(n-\sqrt{n})}$ is poly-time checkable). This finishes the proof of this lemma. □

For the following step, we will use the event $E_0^{(n-\sqrt{n})}$ as was defined in the proof of the above lemma. For convenience, we denote it as $R_n = E_0^{(n-\sqrt{n})}$, for any integer n. What we will use is that $\mathbb{P}[R_n] \geq I(W) - \ell^{-\frac{n}{2}+11\alpha n+\sqrt{n}}$; if R_n happens, then $\mathsf{Z}_n < \exp\left(-2^{n^{1/3}}\right)$; and that for any bit-channel it can be checked in poly-time if R_n happened, all of which is proven in Lemma 10.3.

10.3 Step 3

Here we will finally prove the polarization that implies the main result of this paper:

Lemma 10.4. $\mathbb{P}\left[\mathsf{Z}_t \leq \exp\left(-st\cdot\ell^{\alpha\cdot t}\right)\right] \geq I(W) - \ell^{-(1/2-16\alpha)t+2\sqrt{t}}$ *for* $t \geq C\cdot\log^6 s$, *where* C *is an absolute constant. Moreover, this polarization is poly-time constructible.*

Proof. We will again closely follow the approach from [WD19], though we are going to change the indexing notations to avoid any confusion with the previous step. We return to having the total depth of the tree to be t, and we will have $\sqrt{t}$ stages in the tree, each of length $\sqrt{t}$, similarly to the previous step. As before, we will define several events, starting with $C_0^{(0)} = \emptyset$ and $Q_0^{(0)} = \emptyset$. Then, for n being $\sqrt{t}, 2\sqrt{t}, \ldots, t-\sqrt{t}$, we define:

$$\begin{aligned}
C_n &= R_n \setminus C_0^{(n-\sqrt{t})}\\
C_0^{(n)} &= C_0^{(n-\sqrt{t})} \cup C_n\\
D_n &= C_n \bigcap \left\{ \sum_{i=1}^{s(t-n)} g_i \leq \alpha\cdot s\cdot t \right\}\\
Q_n &= C_n \setminus D_n\\
Q_0^{(n)} &= Q_0^{(n-\sqrt{t})} \cup Q_n,
\end{aligned}$$

where R_n is defined at the end of previous step, and g_i's can again be thought of as independent Bern(1/2) random variables. The intuition behind what these events correspond to is almost the same as in Step 2, but the bit-channels in D_n have conditions on branching from level n down to the bottom level t (instead of levels between n and $n+\sqrt{t}$). Here, the channels in $Q_0^{(n)}$ are the channels that we mark as "good" up to level n in the tree, and we will be interested in the final set $Q_0^{(t-\sqrt{t})}$ of "good" channels in the end. We again denote by corresponding lowercase letters the probabilities of these events. Define also

$$f_n = I(W) - c_0^{(n)} \quad \text{and} \quad p_n = I(W) - q_0^{(n)}\ .$$

First, consider event C_n happening. It means that R_n happens, so $\mathsf{Z}_n < \exp\left(-2^{n^{1/3}}\right)$. Then at least for some time, we are going to pick Arıkan's kernel in the construction phase, since the Bhattacharyya parameter is small enough. But assuming that we take Arıkan's kernels all the way down to the bottom of the tree, one can see

$$\mathsf{Z}_t < \ell^{t-n} \cdot \mathsf{Z}_n < \ell^t \cdot \exp\left(-2^{n^{1/3}}\right) \leq 2^{st} \cdot \exp\left(-2^{t^{1/6}}\right) < 2^{-4s} = \ell^{-4}$$

for $t \geq C \log^6 s$, where C is large enough. Again, by using Proposition 9.3 and (79) it is easy to show that the entropy of the binned version of the bit-channel will also always be below the threshold ℓ^{-4}. It means that we cannot in $(t-n)$ levels go over the threshold of choosing Arıkan's kernel, thus we indeed take Arıkan's kernel all the way down in the tree for the path for which R_n happens. Thus, similarly to the proof of Lemma 10.3 in the Step 2, we can think of it as taking the basic 2×2 Arıkan's kernels $s \cdot (t-n)$ times, starting at level n. Therefore if R_n happens, the branching down from level n can be viewed as taking "good" or "bad" branches in the A_2 kernels, so we again define indicator random variables g_i, for $i \in [s(t-n)]$, to denote these branches. It is clear that these random variables are going to be independent Bern(1/2). These are exactly the random variables g_i, for $i \in [s(t-n)]$, appearing in the definition of D_n.

We have

$$\frac{d_n}{c_n} = \mathbb{P}\left[\sum_{i=1}^{s(t-n)} g_i \leq \alpha s t\right] \leq 2^{-s(t-n)(1-h_2(\delta))},$$

where we denote $\delta \coloneqq \min\left\{\frac{\alpha t}{t-n}, 1\right\}$. The inequality again follows from the entropic inequality on the sum of binomial coefficients.

Recall that we denoted $f_n = I(W) - c_0^{(n)}$. The event $C_0^{(n)}$ contains the event R_n, thus $f_n \leq \ell^{-\frac{n}{2}+11\alpha n+\sqrt{n}}$, which follows from the proof of Lemma 10.3. Same inequality holds for f_n^+.

We will obtain a recurrence on $p_n - f_n^+$ as follows:

$$\begin{aligned}
p_n - f_n^+ &= I(W) - q_0^{(n)} - (I(W) - c_0^{(n)})^+ \\
&= p_{n-\sqrt{t}} - q_n - (f_{n-\sqrt{t}} - c_n)^+ \\
&\leq p_{n-\sqrt{t}} - q_n - \frac{q_n}{c_n}(f_{n-\sqrt{t}} - c_n)^+ \\
&\leq p_{n-\sqrt{t}} - q_n - \frac{q_n}{c_n}(f^+_{n-\sqrt{t}} - c_n) \\
&\leq p_{n-\sqrt{t}} - f^+_{n-\sqrt{t}} + \left(1 - \frac{q_n}{c_n}\right) f^+_{n-\sqrt{t}} \\
&= p_{n-\sqrt{t}} - f^+_{n-\sqrt{t}} + \frac{d_n}{c_n} f^+_{n-\sqrt{t}} \\
&\leq p_{n-\sqrt{t}} - f^+_{n-\sqrt{t}} + \ell^{-(1/2-11\alpha)(n-\sqrt{t})+\sqrt{n}} \cdot 2^{-s(t-n)(1-h_2(\delta))},
\end{aligned}$$

where recall that $\delta = \min\left\{\frac{\alpha t}{t-n}, 1\right\}$. We want to obtain an upper bound on the additive term in the inequality above. Consider the following two cases:

i) $\delta > \frac{1}{10}$, i.e. $10\alpha t > t - n$, thus $n > (1 - 10\alpha)t$. Then we give up on the term $2^{-s(t-n)(1-h_2(\delta))}$ completely, and we can write

$$\ell^{-(1/2-11\alpha)(n-\sqrt{t})+\sqrt{n}} \cdot 2^{-s(t-n)(1-h_2(\delta))} \leq \ell^{-(1/2-11\alpha)(1-10\alpha)t+\frac{3}{2}\sqrt{t}} \leq \ell^{-(1/2-16\alpha)t+\frac{3}{2}\sqrt{t}};$$

ii) $\delta \le \frac{1}{10}$, and then $h_2(\delta) < 1/2$. In this case we derive

$$\ell^{-(1/2-11\alpha)(n-\sqrt{t})+\sqrt{n}} \cdot 2^{-s(t-n)(1-h_2(\delta))} \le \ell^{-(1/2-11\alpha)n+\frac{3}{2}\sqrt{t}} \cdot \ell^{-1/2\cdot(t-n)} = \ell^{-1/2\cdot t+11\alpha n+\frac{3}{2}\sqrt{t}} < \ell^{-1/2\cdot t+11\alpha t+\frac{3}{2}\sqrt{t}}.$$

Putting the above together, we obtain

$$\begin{aligned} p_0 - f_0^+ &= 0 \\ p_n - f_n^+ &\le p_{n-\sqrt{t}} - f_{n-\sqrt{t}}^+ + \ell^{-(1/2-16\alpha)t+\frac{3}{2}\sqrt{t}}. \end{aligned}$$

Therefore $p_{t-\sqrt{t}} - f_{t-\sqrt{t}}^+ \le \sqrt{t} \cdot \ell^{-(1/2-16\alpha)t+\frac{3}{2}\sqrt{t}}$. Combining this with $f_{t-\sqrt{t}}^+ \le \ell^{-(1/2-11\alpha)(t-\sqrt{t})+\sqrt{t}}$, we obtain $p_{t-\sqrt{t}} \le \ell^{-(1/2-16\alpha)t+2\sqrt{t}}$, and thus

$$\mathbb{P}\left[Q_0^{(t-\sqrt{t})}\right] = q_0^{(t-\sqrt{t})} \ge I(W) - \ell^{-(1/2-16\alpha)t+2\sqrt{t}}. \tag{108}$$

Let us now check that the event $Q_0^{(t-\sqrt{t})}$ is actually "good" and allows us achieve the needed polarization. If $Q_0^{(t-\sqrt{t})}$ happens, then Q_n happened for some $n = k \cdot \sqrt{t}$. It means that C_n, and therefore R_n takes place, thus $\mathsf{Z}_n < \exp\left(-2^{n^{1/3}}\right)$. It also means that D_n does not happen, and thus there is at least αst "good" branches taken in the way down the tree, which corresponds to αst squarings of the Bhattacharyya parameter. Therefore

$$\mathsf{Z}_t \le \left(\ell^{t-n}\mathsf{Z}_n\right)^{2^{\alpha st}} < \left(2^{st}\exp\left(-2^{n^{1/3}}\right)\right)^{2^{\alpha st}} < \exp\left(-st\cdot 2^{\alpha st}\right) = \exp\left(-st\cdot\ell^{\alpha t}\right) = \frac{1}{N}\exp\left(-N^{\alpha}\right),$$

where the third inequality trivially follows from $n \ge \sqrt{t}$ and $t \ge C\log^6 s$ for large enough C. Combining this with (108), we obtain the desired polarization:

$$\mathbb{P}\left[\mathsf{Z}_t < \exp\left(-st\cdot 2^{\alpha st}\right)\right] \ge q_0^{(t-\sqrt{t})} \ge I(W) - \ell^{-(1/2-16\alpha)t+o(t)}.$$

It only remains to argue that this polarization is poly-time constructible. But this easily follows from the fact that the event R_n is poly-time checkable, which we proved in Step 2. Indeed, now for any bit-channel W_i, $i \in [\ell^t]$, we need to check if it is in $Q_0^{(t-\sqrt{t})}$. This means that one need to see if Q_n happened for some $n = k\sqrt{t}$. To do this, one checks in poly-time if C_n happened, which reduces to checking R_n (which can be done in poly-time). If R_n happened, then the only thing to check is how many "good" branches the remaining path to W_i has, which is easily (in poly-time) retrievable information from the index i. Therefore, the event $Q_0^{(t-\sqrt{t})}$ is indeed poly-time checkable, which finishes the proof of the lemma. □

Appendices

A Proofs of entropic lemmas for BMS channels

In the following two proofs we use the representation of BMS channel W as a convex combination of several BSC subchannels $W^{(1)}$, $W^{(2)}$, …, $W^{(m)}$, see the beginning of Section 7.1 for details. Each subchannel $W^{(j)}$ can output one of two symbols $z_j^{(0)}, z_j^{(1)}$, and

$W^{(j)}(z_j^{(0)}|0) = W^{(j)}(z_j^{(1)}|1)$, $W^{(j)}(z_j^{(1)}|0) = W^{(j)}(z_j^{(0)}|1)$. The output alphabet for W is thus $\mathcal{Y} = \{z_1^{(0)}, z_1^{(1)}, z_2^{(0)}, z_2^{(1)}, \ldots, z_m^{(0)}, z_m^{(1)}\}$. Define for these proofs the "flip" operator $\oplus : \mathcal{Y} \times \{0,1\} \to \mathcal{Y}$ as follows: $z_j^{(c)} \oplus b = z_j^{(b+c)}$, where $b, c \in \{0,1\}$, and $(b+c)$ is addition mod 2. In other words, $z_j^{(c)} \oplus 0$ doesn't change anything, and $z_j^{(c)} \oplus 1$ flips the output of the subchannel $W^{(j)}$ to the opposite symbol. Note then that $W^{(j)}(z_j^{(c)} \,|\, b) = W^{(j)}(z_j^{(c)} \oplus b \,|\, 0)$. Finally, we overload the operator to also work on $\mathcal{Y}^\ell \times \{0,1\}^\ell \to \mathcal{Y}^\ell$ by applying it coordinate-wise. It then easily follows that $W^\ell(\boldsymbol{y} \,|\, \boldsymbol{x}) = W^\ell(\boldsymbol{y} \oplus \boldsymbol{x} \,|\, \boldsymbol{0})$ for any $\boldsymbol{y} \in \mathcal{Y}^\ell$ and $\boldsymbol{x} \in \{0,1\}^\ell$.

Proof of Lemma 6.1. We can write

$$\begin{aligned}
\mathop{\mathbb{E}}_{g\sim G}[H^{(g)}(V_1|\boldsymbol{Y})] &= \sum_g \mathbb{P}(G=g)\left(\sum_{\boldsymbol{y}\in\mathcal{Y}^\ell} \mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}] H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right) \\
&= \sum_g \mathbb{P}(G=g)\left(\sum_{\boldsymbol{y}\in\mathcal{Y}^\ell}\left(\sum_{\boldsymbol{v}\in\{0,1\}^k} \mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}, \boldsymbol{V}=\boldsymbol{v}]\right) h\left(\frac{\mathbb{P}^{(g)}[V_1=0, \boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}]}\right)\right) \\
&= \frac{1}{2^k}\sum_{\boldsymbol{v}\in\{0,1\}^k}\sum_g \mathbb{P}(G=g)\sum_{\boldsymbol{y}\in\mathcal{Y}^\ell} \mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{v}] h\left(\frac{\mathbb{P}^{(g)}[V_1=0, \boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}]}\right) \qquad (109)
\end{aligned}$$

where $h(x) := -x\log_2 x - (1-x)\log_2(1-x)$ is the binary entropy function. Next, we show that for any fixed codebook g and any fixed $\boldsymbol{v} \in \{0,1\}^k$ it holds

$$\sum_{\boldsymbol{y}\in\mathcal{Y}^\ell} \mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{v}] h\left(\frac{\mathbb{P}^{(g)}[V_1=0, \boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}]}\right) = \sum_{\boldsymbol{y}\in\mathcal{Y}^\ell} \mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{0}] h\left(\frac{\mathbb{P}^{(g)}[V_1=0, \boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}]}\right), \qquad (110)$$

where $\boldsymbol{0}$ is the all-zero vector.

First of all, we know that

$$\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{v}] = W^\ell(\boldsymbol{y} \,|\, \boldsymbol{v}G) = W^\ell(\boldsymbol{y} \oplus \boldsymbol{v}G \,|\, \boldsymbol{0}) = \mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G \,|\, \boldsymbol{V}=\boldsymbol{0}], \qquad (111)$$

as was discussed at the beginning of this appendix. In the same way, it's easy to see

$$\begin{aligned}
\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}] = \frac{1}{2^k}\sum_{\boldsymbol{u}\in\{0,1\}^k} \mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{V}=\boldsymbol{u}] &= \frac{1}{2^k}\sum_{\boldsymbol{u}\in\{0,1\}^k} \mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G|\boldsymbol{V}=\boldsymbol{u}+\boldsymbol{v}] \\
&= \frac{1}{2^k}\sum_{\boldsymbol{u}+\boldsymbol{v}\in\{0,1\}^k} \mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G|\boldsymbol{V}=\boldsymbol{u}+\boldsymbol{v}] \\
&= \mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G]. \qquad (112)
\end{aligned}$$

The above equality uses the fact the we are considering linear codes, and $\boldsymbol{v}G$ is an arbitrary codeword. It follows from the symmetry of linear codes that "shifting" the output by a codeword does not change anything. Shifting here means the usual shifting for the BSC case, though for general BMS channel this is actually flipping the outputs or appropriate BSC subchannels, without changing which subchannel was actually used for which bit.

Denote now $\widetilde{\boldsymbol{V}} = \boldsymbol{V}_{>1}$, and recall that we are considering fixed $\boldsymbol{v}$ for now. Denote then also v_1 as the first coordinate of $\boldsymbol{v}$ and $\widetilde{\boldsymbol{v}} = \boldsymbol{v}_{>1}$. Then we derive similarly

$$\begin{aligned}\mathbb{P}^{(g)}[V_1=0,\boldsymbol{Y}=\boldsymbol{y}] &= \frac{1}{2^k}\sum_{\widetilde{\boldsymbol{u}}\in\{0,1\}^{k-1}}\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}|V_1=0,\widetilde{\boldsymbol{V}}=\widetilde{\boldsymbol{u}}]\\ &= \frac{1}{2^k}\sum_{\widetilde{\boldsymbol{u}}\in\{0,1\}^{k-1}}\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G\,|\,V_1=v_1,\widetilde{\boldsymbol{V}}=\widetilde{\boldsymbol{u}}+\widetilde{\boldsymbol{v}}]\\ &= \frac{1}{2^k}\sum_{\widetilde{\boldsymbol{u}}+\widetilde{\boldsymbol{v}}\in\{0,1\}^{k-1}}\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G\,|\,V_1=v_1,\widetilde{\boldsymbol{V}}=\widetilde{\boldsymbol{u}}+\widetilde{\boldsymbol{v}}]\\ &= \mathbb{P}^{(g)}[V_1=v_1,\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G].\end{aligned}\tag{113}$$

Notice that $\mathbb{P}^{(g)}[V_1=v_1,\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G]+\mathbb{P}^{(g)}[V_1=1-v_1,\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G]=\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G]$, and thus using the symmetry of the binary entropy function around $1/2$ obtain

$$h\left(\frac{\mathbb{P}^{(g)}[V_1=v_1,\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G]}\right)=h\left(\frac{\mathbb{P}^{(g)}[V_1=1-v_1,\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G]}\right).$$

Using this and (111)–(113) derive

$$\begin{aligned}&\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\big|\boldsymbol{V}=\boldsymbol{v}]h\left(\frac{\mathbb{P}^{(g)}[V_1=0,\boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}]}\right)\\ =&\,\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G\,|\,\boldsymbol{V}=\boldsymbol{0}]h\left(\frac{\mathbb{P}^{(g)}[V_1=0,\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{v}G]}\right).\end{aligned}$$

Finally, summing both parts over $\boldsymbol{y}\in\mathcal{Y}^\ell$ and noticing that $\boldsymbol{y}\oplus\boldsymbol{v}G$ will also range through all $\mathcal{Y}^\ell$ in this case, we establish (110). Then in (109) deduce

$$\begin{aligned}\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y})\right] &= \frac{1}{2^k}\sum_{\boldsymbol{v}\in\{0,1\}^k}\sum_g\mathbb{P}(G=g)\sum_{\boldsymbol{y}\in\mathcal{Y}^\ell}\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\big|\boldsymbol{V}=\boldsymbol{0}]h\left(\frac{\mathbb{P}^{(g)}[V_1=0,\boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}]}\right)\\ &= \sum_{\boldsymbol{y}\in\mathcal{Y}^\ell}\sum_g\mathbb{P}(G=g)\,\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\big|\boldsymbol{V}=\boldsymbol{0}]h\left(\frac{\mathbb{P}^{(g)}[V_1=0,\boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}]}\right)\\ &= \sum_{\boldsymbol{y}\in\mathcal{Y}^\ell}\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}\big|\boldsymbol{V}=\boldsymbol{0}]\mathop{\mathbb{E}}_{g\sim G}\left[H^{(g)}(V_1|\boldsymbol{Y}=\boldsymbol{y})\right],\end{aligned}$$

since $\mathbb{P}^{(g)}[\boldsymbol{Y}=\boldsymbol{y}\big|\boldsymbol{V}=\boldsymbol{0}]$ does not depend on the matrix g. □

Proof of Proposition 5.5. Let us unfold the conditioning in the LHS as follows

$$H\left(U_i\;\middle|\;W^\ell(\boldsymbol{U}\cdot K),\boldsymbol{U}_{<i}\right)=\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\{0,1\}^{i-1}}\left[H\left(U_i\;\middle|\;W^\ell(\boldsymbol{U}\cdot K),\boldsymbol{U}_{<i}=\boldsymbol{w}\right)\right].\tag{114}$$

We are going to show that the conditional entropy inside the expectation doesn't depend on the choice of $\boldsymbol{w}$, which will allow us to restrict to $\boldsymbol{w}=\boldsymbol{0}$.

Return now to the settings of the Proposition, and denote the (random) output $\boldsymbol{Y}=W^\ell(\boldsymbol{U}\cdot K)$. Let us now fix some $\boldsymbol{w}\in\{0,1\}^{i-1}$ and consider $H\left(U_i\;\middle|\;\boldsymbol{Y},\boldsymbol{U}_{<i}=\boldsymbol{w}\right)$. Unfolding the conditional entropy even more, derive

$$H\left(U_i\;\middle|\;\boldsymbol{Y},\boldsymbol{U}_{<i}=\boldsymbol{w}\right)=\sum_{\boldsymbol{y}\in\mathcal{Y}^\ell}\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}\,|\,\boldsymbol{U}_{<i}=\boldsymbol{w}]\cdot H\left(U_i\;\middle|\;\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{U}_{<i}=\boldsymbol{w}\right).\tag{115}$$

Denote now by B the first $(i-1)$ rows of K, and thus $\boldsymbol{Y} = W^\ell(\boldsymbol{U}\cdot K) = W^\ell(\boldsymbol{U}_{<i}\cdot B + \boldsymbol{U}_{\geq i}\cdot G)$. We then have

$$\begin{aligned}
\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}\,|\,\boldsymbol{U}_{<i}=\boldsymbol{w}] &= \sum_{\boldsymbol{v}\in\{0,1\}^k}\frac{1}{2^k}\,\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}\,|\,\boldsymbol{U}_{<i}=\boldsymbol{w},\boldsymbol{U}_{\geq i}=\boldsymbol{v}]\\
&= \sum_{\boldsymbol{v}\in\{0,1\}^k}\frac{1}{2^k}W^\ell\Big(\boldsymbol{y}\,\Big|\,\boldsymbol{w}\cdot B+\boldsymbol{v}\cdot G\Big)\\
&= \sum_{\boldsymbol{v}\in\{0,1\}^k}\frac{1}{2^k}W^\ell\Big(\boldsymbol{y}\oplus\boldsymbol{w}B\,\Big|\,\boldsymbol{v}\cdot G\Big)\\
&= \sum_{\boldsymbol{v}\in\{0,1\}^k}\frac{1}{2^k}\,\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{w}B\,|\,\boldsymbol{U}_{<i}=\boldsymbol{0},\boldsymbol{U}_{\geq i}=\boldsymbol{v}]\\
&= \mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{w}B\,|\,\boldsymbol{U}_{<i}=\boldsymbol{0}]. \qquad (116)
\end{aligned}$$

For the entropy i the RHS of (115), observe

$$H\Big(U_i\,\Big|\,\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{U}_{<i}=\boldsymbol{w}\Big) = h\Big(\,\mathbb{P}\,[U_i=0\,|\,\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{U}_{<i}=\boldsymbol{w}]\,\Big),$$

where $h(\cdot)$ is a binary entropy function. Out of the definition of conditional probability, obtain

$$\begin{aligned}
\mathbb{P}\,[U_i=0\,|\,\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{U}_{<i}=\boldsymbol{w}] &= \frac{\mathbb{P}[U_i=0,\boldsymbol{Y}=\boldsymbol{y}\,|\,\boldsymbol{U}_{<i}=\boldsymbol{w}]}{\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}\,|\,\boldsymbol{U}_{<i}=\boldsymbol{w}]}\\
&= \frac{\mathbb{P}[U_i=0,\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{w}B\,|\,\boldsymbol{U}_{<i}=\boldsymbol{0}]}{\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{w}B\,|\,\boldsymbol{U}_{<i}=\boldsymbol{0}]}\\
&= \mathbb{P}\,[U_i=0\,|\,\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{w}B,\boldsymbol{U}_{<i}=\boldsymbol{0}]\,,
\end{aligned}$$

where the second equality also uses (116) (and similar equality with $U_i=0$ inside the probability, which is completely analogical to (116)). Therefore, deduce in (115)

$$\begin{aligned}
H\Big(U_i\,\Big|\,\boldsymbol{Y},\boldsymbol{U}_{<i}=\boldsymbol{w}\Big) &= \sum_{\boldsymbol{y}\in\mathcal{Y}^\ell}\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{w}B\,|\,\boldsymbol{U}_{<i}=\boldsymbol{0}]\cdot H\Big(U_i\,\Big|\,\boldsymbol{Y}=\boldsymbol{y}\oplus\boldsymbol{w}B,\boldsymbol{U}_{<i}=\boldsymbol{0}\Big)\\
&= \sum_{\boldsymbol{z}\in\mathcal{Y}^\ell}\mathbb{P}[\boldsymbol{Y}=\boldsymbol{z}\,|\,\boldsymbol{U}_{<i}=\boldsymbol{0}]\cdot H\Big(U_i\,\Big|\,\boldsymbol{Y}=\boldsymbol{z},\boldsymbol{U}_{<i}=\boldsymbol{0}\Big)\\
&= H\Big(U_i\,\Big|\,\boldsymbol{Y},\boldsymbol{U}_{<i}=\boldsymbol{0}\Big)\,,
\end{aligned}$$

since $\boldsymbol{z}=\boldsymbol{y}\oplus\boldsymbol{w}B$ ranges over all $\mathcal{Y}^\ell$ for $\boldsymbol{y}\in\mathcal{Y}^\ell$. Therefore, in (114) there is no actual dependence on $\boldsymbol{w}$ under the expectation in the RHS, and thus

$$H\Big(U_i\,\Big|\,W^\ell(\boldsymbol{U}\cdot K),\boldsymbol{U}_{<i}\Big) = H\Big(U_i\,\Big|\,W^\ell(\boldsymbol{U}\cdot K),\boldsymbol{U}_{<i}=\boldsymbol{0}\Big)\,.$$

Finally, note that we can take $\boldsymbol{V}=\boldsymbol{U}_{\geq i}$, since it is uniformly distributed over $\{0,1\}^k$, and then $V_1=U_i$. Since $\boldsymbol{U}\cdot K=\boldsymbol{U}_{\geq i}\cdot G=\boldsymbol{V}\cdot G$ when $\boldsymbol{U}_{<i}=\boldsymbol{0}$, we indeed obtain

$$H\Big(U_i\,\Big|\,W^\ell(\boldsymbol{U}\cdot K),\boldsymbol{U}_{<i}\Big) = H\Big(U_i\,\Big|\,W^\ell(\boldsymbol{U}\cdot K),\boldsymbol{U}_{<i}=\boldsymbol{0}\Big) = H\Big(V_1\,\Big|\,W^\ell(\boldsymbol{V}\cdot G)\Big)\,. \qquad \square$$

B Proofs in Section 7.1.4

Proof of Claim 7.15. Denote for convenience the distribution $\Omega_i \coloneqq \mathrm{Binom}(d_i, p_i)$. Note that $\mathbb{E}_{\chi_i \sim \Omega_i}\left[\frac{\chi_i}{d_i}\right] = p_i$. Then we derive

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{\chi_i\sim\mathcal{D}_i}\left[\frac{\chi_i}{d_i}\right] - p_i\right| &= \left|\mathop{\mathbb{E}}_{\chi_i\sim\mathcal{D}_i}\left[\frac{\chi_i}{d_i}\right] - \mathop{\mathbb{E}}_{\chi_i\sim\Omega_i}\left[\frac{\chi_i}{d_i}\right]\right| \\
&= \left|\sum_{s\in[0:d_i]} \frac{s}{d_i} \mathop{\mathbb{P}}_{\chi_i\sim\mathcal{D}_i}[\chi_i = s] - \sum_{s\in[0:d_i]} \frac{s}{d_i} \mathop{\mathbb{P}}_{\chi_i\sim\Omega_i}[\chi_i = s]\right| \\
&\overset{(70)}{=} \left|\sum_{s\in\mathcal{T}_1^{(i)}} \frac{s}{d_i} \mathop{\mathbb{P}}_{\chi_i\sim\Omega_i}[\chi_i = s]\cdot\theta_i^{-1} - \sum_{s\in[0:d_i]} \frac{s}{d_i} \mathop{\mathbb{P}}_{\chi_i\sim\Omega_i}[\chi_i = s]\right| \\
&= \left|\sum_{s\in\mathcal{T}_1^{(i)}} \frac{s}{d_i} \mathop{\mathbb{P}}_{\chi_i\sim\Omega_i}[\chi_i = s]\cdot\left(\theta_i^{-1} - 1\right) - \sum_{s\notin\mathcal{T}_1^{(i)}} \frac{s}{d_i} \mathop{\mathbb{P}}_{\chi_i\sim\Omega_i}[\chi_i = s]\right| \\
&\le \sum_{s\in\mathcal{T}_1^{(i)}} \frac{s}{d_i} \mathop{\mathbb{P}}_{\chi_i\sim\Omega_i}[\chi_i = s]\cdot\left(\theta_i^{-1} - 1\right) + \sum_{s\notin\mathcal{T}_1^{(i)}} \frac{s}{d_i} \mathop{\mathbb{P}}_{\chi_i\sim\Omega_i}[\chi_i = s].
\end{aligned}$$

We have $\sum\limits_{s\notin\mathcal{T}_1^{(i)}} \frac{s}{d_i}\,\mathbb{P}_{\chi_i\sim\Omega_i}[\chi_i = s] \le \sum\limits_{s\notin\mathcal{T}_1^{(i)}} \mathbb{P}_{\chi_i\sim\Omega_i}[\chi_i = s] \overset{(68)}{=} (1-\theta_i) \overset{(65)}{\le} 2\ell^{-(\log\ell)/3}$.
Next, $\sum_{s\in\mathcal{T}_1^{(i)}} \frac{s}{d_i}\,\mathbb{P}_{\chi_i\sim\Omega_i}[\chi_i = s] \le \mathbb{E}_{\chi_i\sim\Omega_i}\left[\frac{\chi_i}{d_i}\right] \le 1$, and $\theta_i^{-1} - 1 = \frac{1-\theta_i}{\theta_i} \le 2(1-\theta_i) \le 4\ell^{-(\log\ell)/3}$.
Combining the above together, conclude $\left|\mathbb{E}\left[\frac{\chi_i}{d_i}\right] - p_i\right| \le 6\ell^{-(\log\ell)/3} \le \frac{1}{\ell} \le \frac{1}{d_i}$. □

Proof of Claim 7.16. Using the result of Claim 7.15 derive

$$\mathbb{E}\left|\chi_i - \mathbb{E}[\chi_i]\right| \le \mathbb{E}\left|\chi_i - p_i d_i\right| + \mathbb{E}\left|p_i d_i - \mathbb{E}[\chi_i]\right| \le \mathbb{E}\left|\chi_i - p_i d_i\right| + 1. \tag{117}$$

From (68), (71), and definition (67) of $\mathcal{T}_1^{(i)}$ for $i \in F_2$ observe also the following:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\chi_i\sim\mathcal{D}_i}\left|\chi_i - p_i d_i\right| &= \sum_{s\in\mathcal{T}_1^{(i)}} \left|s - p_i d_i\right| \cdot \mathop{\mathbb{P}}_{\eta_i\sim\Omega_i}[s]\cdot\theta_i^{-1} \\
&= \sum_{s\in\mathcal{T}_1^{(i)}} \left|s - p_i d_i\right| \cdot \mathop{\mathbb{P}}_{\eta_i\sim\Omega_i}[s] + \sum_{s\in\mathcal{T}_1^{(i)}} \left|s - p_i d_i\right| \cdot \mathop{\mathbb{P}}_{\eta_i\sim\Omega_i}[s]\cdot(\theta_i^{-1} - 1) \\
&\le \sum_{s\in\mathcal{T}_1^{(i)}} \left|s - p_i d_i\right| \cdot \mathop{\mathbb{P}}_{\eta_i\sim\Omega_i}[s] + \sqrt{d_i p_i}\log\ell \cdot \underbrace{\sum_{s\in\mathcal{T}_1^{(i)}} \mathop{\mathbb{P}}_{\eta_i\sim\Omega_i}[s]}_{\theta_i} \cdot \left(\frac{1-\theta_i}{\theta_i}\right) \\
&= \sum_{s\in\mathcal{T}_1^{(i)}} \left|s - p_i d_i\right| \cdot \mathop{\mathbb{P}}_{\eta_i\sim\Omega_i}[s] + \sqrt{d_i p_i}\log\ell \cdot (1-\theta_i) \\
&= \sum_{s\in\mathcal{T}_1^{(i)}} \left|s - p_i d_i\right| \cdot \mathop{\mathbb{P}}_{\eta_i\sim\Omega_i}[s] + \sum_{s\notin\mathcal{T}_1^{(i)}} \sqrt{d_i p_i}\log\ell \cdot \mathop{\mathbb{P}}_{\eta_i\sim\Omega_i}[s]
\end{aligned}$$

$$\leq \sum_{s\in\mathcal{T}_1^{(i)}} \left|s-p_id_i\right| \cdot \mathop{\mathbb{P}}_{\eta_i\sim\Omega_i}[s] + \sum_{s\notin\mathcal{T}_1^{(i)}} \left|s-p_id_i\right| \cdot \mathop{\mathbb{P}}_{\eta_i\sim\Omega_i}[s] = \mathop{\mathbb{E}}_{\eta_i\sim\Omega_i} \left|\eta_i - p_id_i\right|.$$

Combining this with (117), obtain the needed. □

C Proof in Section 7.2

Here we formally show that the channel $\widetilde{W}$ we constructed in Section 7.2 is indeed upgraded with respect to W. Recall that W, $\widetilde{W}$, and W_1 are defined in (74), (75), and (76) correspondingly, and our goal is to prove (77). First, to check that W_1 is a valid channel, observe

$$\sum_{k\in[m],\, c\in\{0,1\}} W_1\left(z_k^{(c)} \,\middle|\, \widetilde{z_j^{(b)}}\right) = \sum_{k\in T_j}\left(W_1\left(z_k^0 \,\middle|\, \widetilde{z_j^{(b)}}\right) + W_1\left(z_k^1 \,\middle|\, \widetilde{z_j^{(b)}}\right)\right) = \sum_{k\in T_j} \frac{q_k}{\sum\limits_{i\in T_j} q_i} = 1.$$

Finally, for any $k\in[m],\ c\in\{0,1\}$, let j_k be such that $k\in T_{j_k}$. Then we have for any $x\in\{0,1\}$

$$\sum_{j\in[\sqrt{\ell}],\, b\in\{0,1\}} \widetilde{W}\left(\widetilde{z_j^{(b)}} \,\middle|\, x\right) W_1\left(z_k^{(c)} \,\middle|\, \widetilde{z_j^{(b)}}\right) = \sum_{b\in\{0,1\}} \widetilde{W}\left(\widetilde{z_{j_k}^{(b)}} \,\middle|\, x\right) W_1\left(z_k^{(c)} \,\middle|\, \widetilde{z_{j_k}^{(b)}}\right).$$

Now, if $x=c$, we derive

$$\begin{aligned}
&\sum_{b\in\{0,1\}} \widetilde{W}\left(\widetilde{z_{j_k}^{(b)}} \,\middle|\, x\right) W_1\left(z_k^{(c)} \,\middle|\, \widetilde{z_{j_k}^{(b)}}\right) \\
&\qquad = \widetilde{W}\left(\widetilde{z_{j_k}^{(x)}} \,\middle|\, x\right) W_1\left(z_k^{(x)} \,\middle|\, \widetilde{z_{j_k}^{(x)}}\right) + \widetilde{W}\left(\widetilde{z_{j_k}^{(1-x)}} \,\middle|\, x\right) W_1\left(z_k^{(x)} \,\middle|\, \widetilde{z_{j_k}^{(1-x)}}\right) \\
&\qquad = \sum_{i\in T_{j_k}} q_i\cdot(1-\theta_{j_k})\cdot\frac{q_k}{\sum\limits_{i\in T_{j_k}} q_i}\cdot\left(1-\frac{p_k-\theta_{j_k}}{1-2\theta_{j_k}}\right) + \sum_{i\in T_{j_k}} q_i\cdot\theta_{j_k}\cdot\frac{q_k}{\sum\limits_{i\in T_{j_k}} q_i}\cdot\left(\frac{p_k-\theta_{j_k}}{1-2\theta_{j_k}}\right) \\
&\qquad = q_k\left(1-\theta_{j_k}-(1-\theta_{j_k})\cdot\left(\frac{p_k-\theta_{j_k}}{1-2\theta_{j_k}}\right)+\theta_{k_j}\cdot\left(\frac{p_k-\theta_{j_k}}{1-2\theta_{j_k}}\right)\right) \\
&\qquad = q_k\left(1-\theta_{j_k}-(1-2\theta_{j_k})\cdot\left(\frac{p_k-\theta_{j_k}}{1-2\theta_{j_k}}\right)\right) = q_k\cdot(1-p_k).
\end{aligned}$$

Otherwise, then $x=1-c$, obtain

$$\begin{aligned}
&\sum_{b\in\{0,1\}} \widetilde{W}\left(\widetilde{z_{j_k}^{(b)}} \,\middle|\, x\right) W_1\left(z_k^{(c)} \,\middle|\, \widetilde{z_{j_k}^{(b)}}\right) \\
&\qquad = \widetilde{W}\left(\widetilde{z_{j_k}^{(x)}} \,\middle|\, x\right) W_1\left(z_k^{(1-x)} \,\middle|\, \widetilde{z_{j_k}^{(x)}}\right) + \widetilde{W}\left(\widetilde{z_{j_k}^{(1-x)}} \,\middle|\, x\right) W_1\left(z_k^{(1-x)} \,\middle|\, \widetilde{z_{j_k}^{(1-x)}}\right) \\
&\qquad = \sum_{i\in T_{j_k}} q_i\cdot(1-\theta_{j_k})\cdot\frac{q_k}{\sum\limits_{i\in T_{j_k}} q_i}\cdot\left(\frac{p_k-\theta_{j_k}}{1-2\theta_{j_k}}\right) + \sum_{i\in T_{j_k}} q_i\cdot\theta_{j_k}\cdot\frac{q_k}{\sum\limits_{i\in T_{j_k}} q_i}\cdot\left(1-\frac{p_k-\theta_{j_k}}{1-2\theta_{j_k}}\right) \\
&\qquad = q_k\left((1-\theta_{j_k})\cdot\left(\frac{p_k-\theta_{j_k}}{1-2\theta_{j_k}}\right)+\theta_{j_k}-\theta_{j_k}\cdot\left(\frac{p_k-\theta_{j_k}}{1-2\theta_{j_k}}\right)\right)
\end{aligned}$$

$$= q_k \left((1 - 2\theta_{j_k}) \cdot \left(\frac{p_k - \theta_{j_k}}{1 - 2\theta_{j_k}} \right) + \theta_{j_k} \right) = q_k \cdot p_k.$$

Therefore, for any $k \in [m]$ and $c, x \in \{0,1\}$ it holds

$$\sum_{j \in [\sqrt{\ell}],\, b \in \{0,1\}} \widetilde{W}\left(\widetilde{z_j^{(b)}} \,\middle|\, x \right) W_1\left(z_k^{(c)} \,\middle|\, \widetilde{z_j^{(b)}} \right) = W\left(z_k^{(c)} \,\middle|\, x \right).$$

D Proof of Proposition 9.1

We still use $\boldsymbol{U}_{[1:N]}$ to denote the information vector and use $\boldsymbol{X}_{[1:N]} = \boldsymbol{U}_{[1:N]} M^{(t)}$ to denote the encoded vector. Assume that $\boldsymbol{U}_{[1:N]}$ consists of N i.i.d. Bernoulli-1/2 random variables. Similarly to the example in Section 9.1, we define the random vectors $\boldsymbol{V}_{[1:N]}^{(j)}, \boldsymbol{U}_{[1:N]}^{(j)}$ for $j = t-1, t-2, \ldots, 1$ recursively

$$\begin{aligned}
\boldsymbol{V}_{[1:N]}^{(t-1)} &= \boldsymbol{U}_{[1:N]} D^{(t-1)}, \\
\boldsymbol{U}_{[1:N]}^{(j)} &= \boldsymbol{V}_{[1:N]}^{(j)} Q^{(j)} \text{ for } j = t-1, t-2, \ldots, 1, \\
\boldsymbol{V}_{[1:N]}^{(j)} &= \boldsymbol{U}_{[1:N]}^{(j+1)} D^{(j)} \text{ for } j = t-2, t-3, \ldots, 1, \\
\boldsymbol{X}_{[1:N]} &= \boldsymbol{U}_{[1:N]}^{(1)} D^{(0)}.
\end{aligned} \tag{118}$$

Moreover, let $\boldsymbol{U}_{[1:N]}^{(t)} := \boldsymbol{U}_{[1:N]}$. We will prove the following two claims:

1. For every $a = 1, 2, \ldots, t$, the following ℓ^{t-a} random vectors

$$(\boldsymbol{U}_{[h\ell^a+1:h\ell^a+\ell^a]}^{(a)}, \boldsymbol{Y}_{[h\ell^a+1:h\ell^a+\ell^a]}), \quad h = 0, 1, \ldots, \ell^{t-a} - 1$$

   are i.i.d.

2. For every $a = 1, 2, \ldots, t$ and every $i \in [\ell^a]$, we write $\tau_a(i) = (i_1, i_2, \ldots, i_a)$, where τ_a is the a-digit expansion function defined in (97). Then for every $h = 0, 1, \ldots, \ell^{t-a} - 1$ and every $i \in [\ell^a]$, we have

$$\mathbb{P}(U_{h\ell^a+i}^{(a)} \to (\boldsymbol{U}_{[h\ell^a+1:h\ell^a+i-1]}^{(a)}, \boldsymbol{Y}_{[h\ell^a+1:h\ell^a+\ell^a]})) \equiv W_{i_1,\ldots,i_a}(K_1^{(0)}, K_{i_1}^{(1)}, \ldots, K_{i_1,\ldots,i_{a-1}}^{(a-1)}). \tag{119}$$

Note that Proposition 9.1 follows immediately from taking $a = t$ in (119). Therefore, we only need to prove these two claims.

We start with the first claim. By (98), for every $j = 0, 1, \ldots, t-1$, the matrix $D^{(j)}$ is a block diagonal matrix with ℓ^{t-j-1} blocks on the diagonal, where each block has size $\ell^{j+1} \times \ell^{j+1}$, and all the ℓ^{t-j-1} blocks are the same. According to (99)–(100), the permutation matrix $Q^{(j)}$ keeps the first $t-j-1$ digits of the ℓ-ary expansion to be the same and performs a cyclic shift on the last $j+1$ digits. Therefore, for every $j = 1, \ldots, t-1$, the permutation matrix $Q^{(j)}$ is also a block diagonal matrix with ℓ^{t-j-1} blocks on the diagonal, where each block has size $\ell^{j+1} \times \ell^{j+1}$, and all the ℓ^{t-j-1} blocks are the same. Therefore, for every $j \in [t]$, the matrix $M^{(j)}$ defined in (101) can be written in the following block diagonal form

$$M^{(j)} := \underbrace{\{\overline{M}^{(j)}, \overline{M}^{(j)}, \ldots, \overline{M}^{(j)}\}}_{\text{number of } \overline{M}^{(j)} \text{ is } \ell^{t-j}}, \tag{120}$$

where the size of $\overline{M}^{(j)}$ is $\ell^j \times \ell^j$. By the recursive definition (118), one can show that for every $j \in [t]$, we have

$$\boldsymbol{X}_{[1:N]} = \boldsymbol{U}^{(j)}_{[1:N]} M^{(j)}.$$

Combining this with (120), we obtain that for every $a \in [t]$ and every $h = 0, 1, \ldots, \ell^{t-a} - 1$,

$$\boldsymbol{X}_{[h\ell^a+1:h\ell^a+\ell^a]} = \boldsymbol{U}^{(a)}_{[h\ell^a+1:h\ell^a+\ell^a]} \overline{M}^{(a)}. \tag{121}$$

Since $\boldsymbol{X}_{[1:N]}$ consists of N i.i.d. Bernoulli-1/2 random variables, the following ℓ^{t-a} random vectors

$$(\boldsymbol{X}_{[h\ell^a+1:h\ell^a+\ell^a]}, \boldsymbol{Y}_{[h\ell^a+1:h\ell^a+\ell^a]}), \quad h = 0, 1, \ldots, \ell^{t-a} - 1$$

are i.i.d. Combining this with (121), we conclude that the random vectors

$$(\boldsymbol{U}^{(a)}_{[h\ell^a+1:h\ell^a+\ell^a]}, \boldsymbol{Y}_{[h\ell^a+1:h\ell^a+\ell^a]}), \quad h = 0, 1, \ldots, \ell^{t-a} - 1$$

are also i.i.d. This proves claim 1.

Next we prove claim 2 by induction. The case of $a = 1$ is trivial. Now we assume that (119) holds for a and prove it for $a + 1$. In light of claim 1, we only need to prove (119) for the special case of $h = 0$ because the distributions for different values of h are identical, i.e. we only need to prove that

$$\mathbb{P}(U^{(a+1)}_i \to (\boldsymbol{U}^{(a+1)}_{[1:i-1]}, \boldsymbol{Y}_{[1:\ell^{a+1}]})) \equiv W_{i_1,\ldots,i_{a+1}}(K^{(0)}_1, K^{(1)}_{i_1}, \ldots, K^{(a)}_{i_1,\ldots,i_a}) \qquad \forall i \in [\ell^{a+1}]. \tag{122}$$

For a given $i \in [\ell^{a+1}]$, we write its $(a+1)$-digit expansion as $\tau_{a+1}(i) = (i_1, i_2, \ldots, i_{a+1})$. By (118), we know that $\boldsymbol{V}^{(a)}_{[1:N]} = \boldsymbol{U}^{(a+1)}_{[1:N]} D^{(a)}$. By (98), the matrix $D^{(a)}$ is a block diagonal matrix with ℓ^{t-1} blocks on the diagonal, where each block has size $\ell \times \ell$. (Note that these ℓ^{t-1} blocks are not all the same unless $a = 0$.) Therefore, for every $h = 0, 1, \ldots, \ell^{t-1} - 1$, there is a bijection between the two vectors $\boldsymbol{V}^{(a)}_{[h\ell+1:h\ell+\ell]}$ and $\boldsymbol{U}^{(a+1)}_{[h\ell+1:h\ell+\ell]}$. Consequently, there is a bijection between the two vectors $\boldsymbol{U}^{(a+1)}_{[1:i-i_{a+1}]}$ and $\boldsymbol{V}^{(a)}_{[1:i-i_{a+1}]}$, so we have

$$\mathbb{P}(U^{(a+1)}_i \to (\boldsymbol{U}^{(a+1)}_{[1:i-1]}, \boldsymbol{Y}_{[1:\ell^{a+1}]})) \equiv \mathbb{P}(U^{(a+1)}_i \to (\boldsymbol{U}^{(a+1)}_{[i-i_{a+1}+1:i-1]}, \boldsymbol{V}^{(a)}_{[1:i-i_{a+1}]}, \boldsymbol{Y}_{[1:\ell^{a+1}]})). \tag{123}$$

By (98), we also have that

$$\boldsymbol{V}^{(a)}_{[i-i_{a+1}+1:i-i_{a+1}+\ell]} = \boldsymbol{U}^{(a+1)}_{[i-i_{a+1}+1:i-i_{a+1}+\ell]} K^{(a)}_{i_1,i_2,\ldots,i_a}. \tag{124}$$

Let $\hat{i} := (i - i_{a+1})/\ell$, so $\tau_a(\hat{i}) = (i_1, i_2, \ldots, i_a)$. According to the induction hypothesis,

$$\mathbb{P}(U^{(a)}_{\hat{i}} \to (\boldsymbol{U}^{(a)}_{[1:\hat{i}-1]}, \boldsymbol{Y}_{[1:\ell^a]})) \equiv W_{i_1,\ldots,i_a}(K^{(0)}_1, K^{(1)}_{i_1}, \ldots, K^{(a-1)}_{i_1,\ldots,i_{a-1}}).$$

Combining this with the relation $\boldsymbol{U}^{(a)}_{[1:N]} = \boldsymbol{V}^{(a)}_{[1:N]} Q^{(a)}$ and (123)–(124), we can prove (122) with the ideas illustrated in Fig. 4–6. This completes the proof of claim 2 as well as Proposition 9.1.

Acknowledgment

The authors are grateful to Hamed Hassani for useful discussions and sharing his insights on random coding theorems during the initial stages of this work. They also thank the anonymous reviewers for their careful reading and valuable suggestion and corrections to the final paper.

References

[Arı09] Erdal Arıkan. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on Information Theory*, pages 3051–3073, July 2009.

[Ari10] Erdal Arikan. Source polarization. *2010 IEEE International Symposium on Information Theory*, pages 899–903, 2010.

[AT09] Erdal Arıkan and Emre Telatar. On the rate of channel polarization. In *Proceedings of 2009 IEEE International Symposium on Information Theory*, pages 1493–1495, 2009.

[BBGL17] M. Benammar, V. Bioglio, F. Gabry, and I. Land. Multi-kernel polar codes: Proof of polarization and error exponents. In *2017 IEEE Information Theory Workshop (ITW)*, pages 101–105. IEEE, 2017.

[BF02] Alexander Barg and G. David Forney. Random codes: minimum distances and error exponents. *IEEE Transactions on Information Theory*, 48(9):2568–2573, Sep. 2002.

[BGN+18] Jaroslaw Blasiok, Venkatesan Guruswami, Preetum Nakkiran, Atri Rudra, and Madhu Sudan. General strong polarization. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 485–492. ACM, 2018.

[DR96] Devdatt Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *BRICS Report Series*, 3(25), Jan. 1996.

[Dra11] S. Dragomir. A refinement and a divided difference reverse of jensen's inequality with applications. *Revista Colombiana de Matemáticas*, 50, 2011.

[DZF16] Y. Domb, R. Zamir, and M. Feder. The random coding bound is tight for the average linear code or lattice. *IEEE Transactions on Information Theory*, 62(1):121–130, Jan 2016.

[FHMV17] Arman Fazeli, S. Hamed Hassani, Marco Mondelli, and Alexander Vardy. Binary Linear Codes with Optimal Scaling and Quasi-Linear Complexity. *ArXiv e-prints*, November 2017.

[For67] G. David Forney. *Concatenated codes*. PhD thesis, Massachusetts Institute of Technology, 1967.

[For05] G. David Forney. On exponential error bounds for random codes on the BSC. *Lecture notes*, 2005. Available at http://web.mit.edu/6.441/spring05/reading/Forney_ExpEBBSC.pdf.

[FV14] Arman Fazeli and Alexander Vardy. On the scaling exponent of binary polarization kernels. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 797–804, Sep. 2014.

[Gal65] R. Gallager. A simple derivation of the coding theorem and some applications. *IEEE Transactions on Information Theory*, 11(1):3–18, January 1965.

[Gal68] Robert G Gallager. *Information theory and reliable communication*, volume 2. Springer, 1968.

[GB14] Dina Goldin and David Burshtein. Improved bounds on the finite length scaling of polar codes. *IEEE Trans. Information Theory*, 60(11):6966–6978, 2014.

[GBLB17] F. Gabry, V. Bioglio, I. Land, and J. Belfiore. Multi-kernel construction of polar codes. In *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 761–765. IEEE, 2017.

[GRY20] Venkatesan Guruswami, Andrii Riazanov, and Min Ye. Arikan meets shannon: Polar codes with near-optimal convergence to channel capacity. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 552–564, New York, NY, USA, 2020. Association for Computing Machinery.

[GX15] Venkatesan Guruswami and Patrick Xia. Polar codes: Speed of polarization and polynomial gap to capacity. *IEEE Trans. Information Theory*, 61(1):3–16, 2015. Preliminary version in Proc. of FOCS 2013.

[HAU14] S. H. Hassani, K. Alishahi, and R. L. Urbanke. Finite-length scaling for polar codes. *IEEE Transactions on Information Theory*, 60(10):5875–5898, Oct 2014.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[iFLM11] A. G. i. Fàbregas, I. Land, and A. Martinez. Extremes of random coding error exponents. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2896–2898, July 2011.

[JDP83] Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *The Annals of Statistics*, 11(1):286–295, 1983.

[Kor09] Satish Babu Korada. *Polar Codes for Channel and Source Coding*. PhD thesis, École Polytechnique Fédérale De Lausanne, 2009.

[KSU10] Satish Babu Korada, Eren Sasoglu, and Rüdiger L. Urbanke. Polar codes: Characterization of exponent, bounds, and constructions. *IEEE Transactions on Information Theory*, 56(12):6253–6264, 2010.

[KU10] Satish Babu Korada and Rüdiger L. Urbanke. Polar codes are optimal for lossy source coding. *IEEE Transactions on Information Theory*, 56(4):1751–1768, 2010.

[LH06] Ingmar Land and Johannes Huber. Information combining. *Foundations and Trends in Communications and Information Theory*, 3(3):227–330, 2006.

[MHU16] Marco Mondelli, S. Hamed Hassani, and Rüdiger L. Urbanke. Unified scaling of polar codes: Error exponent, scaling exponent, moderate deviations, and error floors. *IEEE Trans. Information Theory*, 62(12):6698–6712, 2016.

[MS77] Florence Jessie MacWilliams and Neil James Alexander Sloane. *The theory of error-correcting codes*, volume 16. Elsevier, 1977.

[MT12] Vera Miloslavskaya and Peter Trifonov. Design of binary polar codes with arbitrary kernel. *2012 IEEE Information Theory Workshop*, pages 119–123, 2012.

[MT14] Ryuhei Mori and Toshiyuki Tanaka. Source and channel polarization over finite fields and reed-solomon matrices. *IEEE Trans. Information Theory*, 60(5):2720–2736, 2014.

[PPV10] Yury Polyanskiy, H Vincent Poor, and Sergio Verdú. Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307, 2010.

[PSL15] Noam Presman, Ofer Shapira, and Simon Litsyn. Mixed-kernels constructions of polar codes. *IEEE Journal on Selected Areas in Communications*, 34(2):239–253, 2015.

[PU16] Henry D. Pfister and Rüdiger L. Urbanke. Near-optimal finite-length scaling for polar codes over large alphabets. In *IEEE International Symposium on Information Theory, ISIT*, pages 215–219, 2016.

[RU08] Thomas Richardson and Rudiger Urbanke. *Modern Coding Theory.* Cambridge University Press, 2008.

[Sha48] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[Str62] Volker Strassen. Asymptotische Abschatzungen in Shannon's Informationstheories. In *Trans. 3rd Prague Conf. Info. Theory*, pages 689–723, 1962.

[Str09] Volker Strassen. Asymptotic estimates in Shannon's information theory. In *Proc. Trans. 3rd Prague Conf. Inf. Theory*, pages 689–723, 2009.

[Top01] Flemming Topsøe. Bounds for entropy and divergence for distributions over a two-element set. *JIPAM. Journal of Inequalities in Pure & Applied Mathematics [electronic only]*, 2(2):Paper No. 25, 13 p.–Paper No. 25, 13 p., 2001.

[TV13] Ido Tal and Alexander Vardy. How to construct polar codes. *IEEE Transactions on Information Theory*, 59(10):6562–6582, Oct 2013.

[WD18] Hsin-Po Wang and Iwan Duursma. Polar-like codes and asymptotic tradeoff among block length, code rate, and error probability. *arXiv:1812.08112*, 2018.

[WD19] Hsin-Po Wang and Iwan M. Duursma. Polar codes' simplicity, random codes' durability. *ArXiv*, abs/1912.08995, 2019.

[Wol57] Jacob Wolfowitz. The coding of messages subject to chance errors. *Illinois J. Math.*, 1:591–606, 1957.

[YB15] Min Ye and Alexander Barg. Polar codes using dynamic kernels. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 231–235. IEEE, 2015.

[YFV19] Hanwen Yao, Arman Fazeli, and Alexander Vardy. Explicit polar codes with small scaling exponent. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1757–1761, July 2019.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il