INFORMATION THEORY AND APPLICATIONS IN COMPUTER SCIENCE

Polar codes for discrete memoryless channels

A self-contained tutorial

Independent further reading after Lectures 1–5

1 Overview

Random coding shows that good codes exist. Polar coding organizes the communication problem so that a fast recursive decoder can exploit it. An invertible transform creates synthetic channels of unequal quality; we send information only through the reliable ones and fix the remaining inputs in advance. The main difficulty is proving that nearly a capacity-sized fraction become reliable enough for *whole-block* decoding.

We work first with a fixed binary-input discrete memoryless channel $W : \{0,1\} \to \mathcal{Y}$, where $\mathcal{Y}$ is finite. Its N-use transition distribution is

$$W^N(y^N \mid x^N) = \prod_{j=1}^{N} W(y_j \mid x_j).$$

There is no feedback. Encoding and decoding are deterministic once the code is chosen, and the decoder knows the channel and the code. Unless we explicitly give a stronger guarantee, error probability averages over a uniform message and the channel noise.

The coding objective. A length-N block code sends a message of k bits through N channel uses. Its encoder maps $\{0,1\}^k$ into channel-input strings, and its decoder maps channel-output strings back to messages. Its rate is k/N bits per channel use. A block error means that the decoded message differs in at least one bit from the transmitted message. Thus the average error is

$$P_e = 2^{-k} \sum_{m \in \{0,1\}^k} \Pr\{\mathsf{dec}(Y^N) \neq m \mid X^N = \mathsf{enc}(m)\}.$$

The probability inside the sum is over the channel noise. These are the code and error conventions of Lecture 5; no later coding construction is required here.

Write

$$I(W) = I(X;Y) \quad \text{for } X \sim \text{Bernoulli}(1/2), \qquad C(W) = \max_{P_X} I(X;Y).$$

$I(W)$ is the *uniform-input mutual information*, often called symmetric capacity. It need not equal $C(W)$ for an asymmetric channel.

A binary-input channel is *symmetric* here if its output alphabet has a permutation π with $\pi(\pi(y)) = y$ and $W(\pi(y) \mid b \oplus 1) = W(y \mid b)$. Relabeling the outputs can then exchange the two inputs without changing their transition probabilities.

Theorem 1.1 (What is proved in this tutorial)**.** *For every finite-output binary-input DMC W and every $R < I(W)$ with $R \geq 0$, there are deterministic polar codes of lengths $N = 2^m$ and rates tending to R with average block error tending to zero. Their encoders use $O(N \log N)$ bit operations and their successive-cancellation decoders use $O(N \log N)$ arithmetic operations, given their information sets and frozen bits. More precisely, a fraction tending to $I(W)$ of the bit-channels have reliability parameter at most $2^{-N^{1/8}}$, enough for block error at most $N2^{-N^{1/8}}$.*

For a binary-input symmetric channel, the frozen bits can all be zero, and the same block-error bound holds for every message. For arbitrary finite input and output alphabets, a finite-layer reduction achieves every rate below $C(W)$ with deterministic codes and vanishing average error. That reduction is proved in Section 8.

The theorem separates fast *use* of a chosen code from efficient *design* of the information set and frozen vector. Our elementary proof does not bound the required length polynomially in the inverse gap to capacity, or supply a fast general-purpose design algorithm. Section 9 identifies stronger results that address those questions. The exponent $1/8$ is deliberately sufficient rather than sharp.

Prerequisites. The course background is the core material of Lectures 1–5. Lecture 1 supplies entropy, concavity, and mutual-information chain rules; Lectures 2–3 supply nonnegativity of relative entropy and optimal fair-bit testing. The probability estimates and DMC coding objective connect to Lectures 4–5. No optional sections of those lectures are required. We use elementary Cauchy–Schwarz, Markov, and Chebyshev bounds, recalling their relevant forms below. Logs are base two; h_2 is binary entropy.

2 Two uses: combine, then split

For independent fair bits A, B, send

$$X_1 = A \oplus B, \qquad X_2 = B$$

through two independent copies of W. The transformation is invertible, and (X_1, X_2) is again a pair of independent fair bits. Decode A first from (Y_1, Y_2), and then B from (Y_1, Y_2, A). The corresponding binary-input channels are

$$W^-(y_1, y_2 \mid a) = \frac{1}{2} \sum_{b \in \{0,1\}} W(y_1 \mid a \oplus b) W(y_2 \mid b), \tag{2.1}$$

$$W^+(y_1, y_2, a \mid b) = \frac{1}{2} W(y_1 \mid a \oplus b) W(y_2 \mid b). \tag{2.2}$$

The minus channel averages over the still-unknown fair bit B. For the plus channel, the fair bit A is included in the output; this accounts for its factor $1/2$ without a sum over a. This is true side information in the mathematical channel definition. The actual decoder will supply its earlier decision instead; the first-error argument below justifies that replacement. These synthetic channels describe two inference tasks for the same physical observations.

Proposition 2.1 (Conservation and ordering)**.** *For every binary-input DMC,*

$$I(W^-) + I(W^+) = 2I(W), \qquad I(W^-) \le I(W) \le I(W^+).$$

Proof. Invertibility and the chain rule give

$$I(A; Y_1, Y_2) + I(B; Y_1, Y_2 \mid A) = I(X_1, X_2; Y_1, Y_2) = 2I(W).$$

The second term equals $I(B; Y_1, Y_2, A)$ because A and B are independent. For the ordering, the plus task includes the observation Y_2 from a direct transmission of B. The chain rule gives

$$I(W^+) = I(B; Y_2) + I(B; Y_1, A \mid Y_2) \ge I(B; Y_2) = I(W).$$

Conservation now gives $I(W^-) \le I(W)$. □

Information has been redistributed. Repeating this operation will concentrate it in some bit-channels, provided the splitting does not stall at intermediate reliability. The next parameter makes that claim quantitative.

3 A reliability parameter and its splitting rules

Define the Bhattacharyya parameter

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

It measures the overlap between the two conditional output distributions. Identical distributions give $Z = 1$; disjoint supports give $Z = 0$. The finite-sum Cauchy–Schwarz inequality $(\sum_y a_y b_y)^2 \le (\sum_y a_y^2)(\sum_y b_y^2)$, applied to $a_y = \sqrt{W(y \mid 0)}$ and $b_y = \sqrt{W(y \mid 1)}$, gives $0 \le Z(W) \le 1$. Under fair input, the optimal single-bit decision error is

$$P_{\text{bit}}(W) = \frac{1}{2}\sum_y \min\{W(y \mid 0), W(y \mid 1)\} \le \frac{1}{2}Z(W). \tag{3.1}$$

The minimum is the smaller contribution at each output; it is bounded by the geometric mean. Small Z therefore certifies small bit error.

Posterior bias. Under fair input the output distribution is $Q(y) = (W(y \mid 0) + W(y \mid 1))/2$. Disregard outputs of Q-mass zero. Bayes' rule gives

$$p_y = \Pr\{X = 0 \mid Y = y\} = \frac{W(y \mid 0)}{2Q(y)}, \qquad \theta(y) = 2p_y - 1, \qquad \Delta(y) = |\theta(y)|.$$

Thus θ is the signed posterior bias, while $\Delta \in [0, 1]$ measures how certain the bit is, regardless of which value is favored. Since $W(y \mid 0) = Q(y)(1 + \theta(y))$ and $W(y \mid 1) = Q(y)(1 - \theta(y))$, and binary entropy is unchanged by exchanging its two probabilities,

$$Z(W) = \mathbb{E}_Q\sqrt{1 - \Delta(Y)^2}, \qquad I(W) = 1 - \mathbb{E}_Q h_2((1 - \Delta(Y))/2).$$

We also need the bias of an XOR. For independent bits with probabilities p_1, p_2 of being zero, that probability for their XOR is $p_1p_2 + (1-p_1)(1-p_2)$. Its signed bias is therefore $(2p_1 - 1)(2p_2 - 1)$. Under independent fair inputs and separate channel uses, the posterior distribution of (X_1, X_2) given (Y_1, Y_2) factorizes by Bayes' rule. Consequently the absolute posterior bias of $A = X_1 \oplus X_2$ is $\Delta(Y_1)\Delta(Y_2)$.

Lemma 3.1 (The endpoints of reliability and information agree)**.** *For every binary-input DMC,*

$$1 - \log(1 + Z(W)) \le I(W) \le \sqrt{1 - Z(W)^2}.$$

Proof. For $p \in [0, 1]$, put $S = \sqrt{p} + \sqrt{1-p}$, and write $P = (p, 1-p)$ and $Q_p = (\sqrt{p}/S, \sqrt{1-p}/S)$. The direction of relative entropy matters: expanding $D(P\|Q_p)$ gives

$$0 \le D(P\|Q_p) = \log S - \frac{1}{2}h_2(p), \qquad h_2(p) \le \log\bigl(1 + 2\sqrt{p(1-p)}\bigr).$$

Endpoint values follow by continuity. Apply this inequality at each posterior and use concavity of log to obtain $1 - I(W) \le \log(1 + Z(W))$.

Concavity of binary entropy also gives $h_2(p) \ge 2\min\{p, 1-p\}$, by its chords on the two half intervals. Hence $I(W) \le \mathbb{E}_Q\Delta$. Finally,

$$(\mathbb{E}\Delta)^2 \le \mathbb{E}\Delta^2, \qquad Z(W)^2 \le \mathbb{E}(1 - \Delta^2),$$

so $I(W)^2 \le 1 - Z(W)^2$. □

Lemma 3.2 (Reliability splitting)**.** *The channels in* (2.1)–(2.2) *satisfy*

$$\boxed{Z(W^+) = Z(W)^2, \qquad Z(W^-) \le 2Z(W) - Z(W)^2.}$$

In particular, $(Z(W^-) + Z(W^+))/2 \le Z(W)$.

Proof. Using (2.2) directly,

$$\begin{aligned} Z(W^+) &= \frac{1}{2} \sum_{a,y_1,y_2} \sqrt{W(y_1 \mid a)W(y_1 \mid a \oplus 1)W(y_2 \mid 0)W(y_2 \mid 1)} \\ &= Z(W)^2. \end{aligned}$$

For the minus channel, Y_1, Y_2 are independent with distribution Q under fair A. Its posterior bias is $\Delta_1\Delta_2$. Write $a = \sqrt{1-\Delta_1^2}$ and $b = \sqrt{1-\Delta_2^2}$. Then

$$\sqrt{1-\Delta_1^2\Delta_2^2} = \sqrt{a^2+b^2-a^2b^2} \le a + b - ab.$$

Both sides are nonnegative, and the difference of the squared right and left sides is $2ab(1-a)(1-b) \ge 0$. Taking expectations and using independence gives the claimed upper bound. □

The plus step squares a small error parameter. The minus step can increase it by at most a factor of two. A typical long branch has many squaring steps; the proof must control what the intervening steps cost.

4 The recursive transform and the actual decoder

For bit strings, $\oplus$ denotes coordinatewise XOR. Write $[N] = \{1, \dots, N\}$ and $u_i^j = (u_i, \dots, u_j)$; if $i > j$, this is the empty string. A subscript set selects those coordinates in increasing order.

Let $N = 2^m$. Define a linear transform recursively by

$$\mathcal{E}_1(u) = u, \qquad \mathcal{E}_{2h}(a,b) = \big(\mathcal{E}_h(a) \oplus \mathcal{E}_h(b), \mathcal{E}_h(b)\big),$$

where a, b each have length h. The right half recovers $\mathcal{E}_h(b)$, and XOR of the two halves recovers $\mathcal{E}_h(a)$, so induction proves invertibility. Induction also gives the XOR identity

$$\mathcal{E}_N(u \oplus a) = \mathcal{E}_N(u) \oplus \mathcal{E}_N(a),$$

since every step uses only XORs. This is the precise linearity property used later, and requires no finite-field or linear-code background. The encoder uses the recurrence $T(N) = 2T(N/2) + O(N)$. Each of the $\log_2 N$ recursion levels processes $O(N)$ coordinates in total, giving $O(N \log N)$ XOR operations.

For independent uniform U^N, set $X^N = \mathcal{E}_N(U^N)$ and pass it through W^N. The ith bit-channel has input U_i and output (Y^N, U_1^{i-1}):

$$W_N^{(i)}(y^N, u_1^{i-1} \mid b) = 2^{-(N-1)} \sum_{v \in \{0,1\}^{N-i}} W^N\big(y^N \mid \mathcal{E}_N(u_1^{i-1}, b, v)\big). \tag{4.1}$$

The factor is $2^{-(i-1)}2^{-(N-i)}$: the first factor is the probability of the fair prefix in the output, and the second averages over the unknown future bits. The sum over an empty future has one term. The decoder will use these likelihoods even when some future bits are frozen. They are a deliberately chosen decoding metric; they need not be the actual posterior probabilities of that fixed code. Averaging over frozen vectors will recover the uniform model for the error analysis, after which a good deterministic vector can be fixed.

Relabeling outputs does not change a channel's quality. Call V and V' *equivalent*, written $V \simeq V'$, if an output bijection ϕ satisfies $V'(\phi(y) \mid b) = V(y \mid b)$ for both inputs. Reindexing the defining sums leaves I, Z, and optimal bit error unchanged. Their minus and plus children are also equivalent: apply ϕ to both physical outputs and retain the plus side bit. This qualification matters because a synthetic output may record the same side information in a different coordinate system.

Lemma 4.1 (The bit-channel recursion)**.** *For $N = 2h$,*

$$W_N^{(i)} \simeq (W^-)_h^{(i)} \quad (1 \le i \le h), \qquad W_N^{(h+k)} \simeq (W^+)_h^{(k)} \quad (1 \le k \le h).$$

Thus the bit-channels are, up to output relabeling, the leaves obtained by replacing each V with (V^-, V^+) for m levels. Minus descendants come before plus descendants in bit-index order.

Proof. Write $u = (a, b)$, $\alpha = \mathcal{E}_h(a)$, $\beta = \mathcal{E}_h(b)$, and pair the physical outputs as $z_j = (y_j, y_{j+h})$. The channel likelihood is

$$W^N(y \mid \mathcal{E}_N(a, b)) = \prod_{j=1}^{h} W(y_j \mid \alpha_j \oplus \beta_j) W(y_{j+h} \mid \beta_j).$$

For a first-half bit, all of b is unknown. Since $b \mapsto \beta$ is a bijection, its sum factorizes:

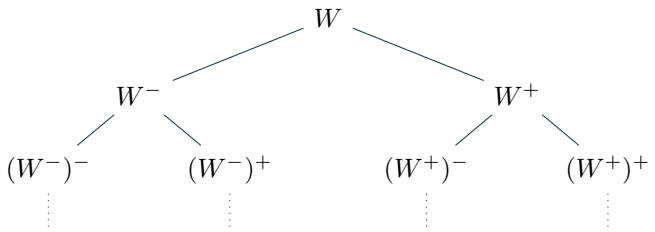
$$\sum_b W^N(y \mid \mathcal{E}_N(a, b)) = \prod_{j=1}^{h} \sum_{\beta_j \in \{0,1\}} W(y_j \mid \alpha_j \oplus \beta_j) W(y_{j+h} \mid \beta_j) = 2^h \prod_{j=1}^{h} W^-(z_j \mid \alpha_j).$$

Multiplying by $2^{-(2h-1)}$ from (4.1) leaves $2^{-(h-1)}$. Summing over the unknown suffix of a is therefore exactly the smaller bit-channel $(W^-)_h^{(i)}$, with outputs paired.

For a second-half bit, the full a is already in the output, and

$$W^N(y \mid \mathcal{E}_N(a, b)) = 2^h \prod_{j=1}^{h} W^+(z_j, \alpha_j \mid \beta_j).$$

Again the normalization becomes $2^{-(h-1)}$, and the remaining sum is over the unknown suffix of b. Replace the known output a by $\alpha = \mathcal{E}_h(a)$, an invertible relabeling. This gives $(W^+)_h^{(k)}$. Applying the same argument to each smaller transform proves the leaf statement by induction. □



Continue to depth m. A path chooses one minus or plus task at each level; the leaves are the bit-channels, up to relabeling.

Information conservation across all leaves can also be read directly from the chain rule. Independence of U_i from its prefix and the invertibility of the transform give

$$\begin{aligned} \sum_{i=1}^{N} I(W_N^{(i)}) &= \sum_{i=1}^{N} I(U_i; Y^N \mid U_1^{i-1}) \\ &= I(U^N; Y^N) = I(X^N; Y^N) = NI(W). \end{aligned} \tag{4.2}$$

The last equality uses the independent uniform X_i and memorylessness.

4.1 Frozen bits and successive cancellation

Choose an information set $A \subseteq [N]$, put $A^c = [N] \setminus A$, and choose a frozen vector $f \in \{0,1\}^{A^c}$. Put the message in u_A, set $u_{A^c} = f$, and transmit $\mathcal{E}_N(u)$. Invertibility makes this a code with $2^{|A|}$ different words, length N, and rate $|A|/N$.

Decode in increasing order of i. At a frozen position insert the known bit. At an information position choose the $b \in \{0,1\}$ maximizing $W_N^{(i)}(y^N, \widehat{u}_1^{i-1} \mid b)$, with a fixed tie rule. The decoder receives no true previous information bits; its previous decisions supply them. A *genie-aided* decision means the same rule with the true prefix substituted, solely for analysis.

Proposition 4.2 (The block-error bound and fixed frozen bits)**.** *For any W and information set A, there is a deterministic frozen vector f for which*

$$P_e \le \frac{1}{2} \sum_{i \in A} Z(W_N^{(i)}).$$

If W is binary-input symmetric, every frozen vector satisfies the slightly weaker bound $P_e \le \sum_{i \in A} Z(W_N^{(i)})$, for every message, so in particular $f = 0$ works.

Proof. Temporarily choose the frozen vector uniformly, independently of a uniform message. Then all U^N are independent fair bits, and the definition of the bit-channel gives the joint distribution

$$\Pr\{U_i = b, Y^N = y, U_1^{i-1} = c\} = \tfrac{1}{2} W_N^{(i)}(y, c \mid b).$$

Thus the genie faces exactly the fair-bit test analyzed in (3.1). Let E_i be the event that the first decoder error is at i. Frozen positions cannot be the first error. The events E_i, for $i \in A$, are disjoint and their union is the block-error event. On E_i the prefix is still true, so the genie-aided decision at i is wrong. Its probability is $P_{\text{bit}}(W_N^{(i)}) \le Z(W_N^{(i)})/2$. Consequently

$$\mathbb{E}_f P_e(f) \le \frac{1}{2} \sum_{i \in A} Z(W_N^{(i)}),$$

so some fixed f obeys the same bound. No fresh frozen randomness is needed during communication.

For the stronger symmetry assertion, use the output permutation π in the definition of a symmetric channel. For $a \in \{0,1\}^N$, transform the output coordinate j by $\pi^{\mathcal{E}_N(a)_j}$, and call this transformation T_a. The XOR identity and the change of future variable $v \mapsto v \oplus a_{i+1}^N$ in (4.1) give

$$W_N^{(i)}(T_a y, u_1^{i-1} \oplus a_1^{i-1} \mid b \oplus a_i) = W_N^{(i)}(y, u_1^{i-1} \mid b).$$

Thus simultaneously shifting the full input u to $u \oplus a$ preserves the true-versus-alternative likelihood comparison. It follows that the probability of the genie event

$$W_N^{(i)}(Y, U_1^{i-1} \mid U_i \oplus 1) \ge W_N^{(i)}(Y, U_1^{i-1} \mid U_i)$$

is the same for every full input vector u. This event includes ties, so it contains any erroneous genie decision regardless of the tie rule. Under uniform U, write $V = W_N^{(i)}$ and $z = (y, u_1^{i-1})$. The event's probability is

$$\frac{1}{2} \sum_{b \in \{0,1\}} \sum_{z \colon V(z|1-b) \ge V(z|b)} V(z \mid b) \le \frac{1}{2} \sum_b \sum_z \sqrt{V(z \mid 0) V(z \mid 1)} = Z(V).$$

Here the smaller likelihood is at most the geometric mean. Its probability for a fixed u has the same bound by symmetry. Union-bound the first error as before. □

The prefix in the analysis is true only until the first mistake. That is exactly why a first-error union bound is valid without assuming independent decoder errors. A fixed cutoff $Z \le \varepsilon$ would give only $P_e \le N\varepsilon$; this does not vanish as N grows. We will prove a shrinking cutoff that is sufficiently strong.

4.2 Why successive cancellation is fast

At a pair of channel outputs, let the incoming likelihood pairs be (a_0, a_1) and (b_0, b_1). The likelihood pairs for the two decisions, up to common positive factors, are

$$(a_0b_0 + a_1b_1,\ a_1b_0 + a_0b_1), \qquad (a_vb_0,\ a_{v\oplus 1}b_1),$$

where v is the already decoded bit for the plus decision. Likelihood pairs handle zero probabilities without dividing by zero. A zero pair can be assigned an arbitrary fixed decision; it represents a prefix of zero probability under the bit-channel model.

At blocklength $2h$, compute the h minus pairs, decode a recursively, compute $\mathcal{E}_h(a)$ as the plus side information, and decode b recursively from the plus pairs. Maintaining the partial sums $\mathcal{E}_h(a)$ in the same recursive computation costs $O(N)$ additional XORs at each tree level: each node combines its two returned encoded halves once. Thus both likelihood computation and partial-sum maintenance have total cost $O(N \log N)$. Re-encoding every prefix separately would be unnecessary extra work.

This is an arithmetic-operation count with access to channel likelihoods. Bit complexity also depends on their representation and the precision used. We make no unit-cost claim for exact arithmetic on arbitrary real channel probabilities.

5 Weak polarization without a convergence theorem

The proof has two tasks. First, show that an intermediate reliability cannot persist: a plus step would keep making a definite change, but the total expected squared change has a finite budget. Then use conservation of information to identify how much mass goes to the reliable endpoint. This second task makes the achievable rate equal to $I(W)$ rather than merely positive.

Choose a path through the synthetic-channel tree by independent fair signs. Let $W_0 = W$, and at step $j + 1$ replace W_j by W_j^- or W_j^+ according to that sign. Set

$$I_j = I(W_j), \qquad Z_j = Z(W_j).$$

At depth m, the path is uniform among the $N = 2^m$ bit-channels. Probabilities about (I_m, Z_m) therefore count fractions of leaves. Conservation and reliability splitting give

$$\mathbb{E}[I_{j+1} \mid \mathcal{F}_j] = I_j, \qquad \mathbb{E}[Z_{j+1} \mid \mathcal{F}_j] \le Z_j, \tag{5.1}$$

where $\mathcal{F}_j$ records the first j signs. In particular, $\mathbb{E}I_j = I(W)$ at every depth. We need only these finite conditional expectations, not a theorem about almost-sure convergence.

Concretely, conditioning on a sign history fixes a node, and the next conditional expectation averages its two children. For a finite-valued history H, any function g, and any random variable V,

$$\mathbb{E}[g(H)V] = \sum_h \Pr(H = h)g(h)\mathbb{E}[V \mid H = h] = \mathbb{E}[g(H)\mathbb{E}(V \mid H)].$$

This elementary identity is the only conditioning rule needed below. We write $\mathbf{1}_A$ for the indicator of event A.

Theorem 5.1 (Weak polarization)**.** *For every fixed* $0 < \eta < 1/2$*,*

$$\Pr\{\eta \le Z_m \le 1 - \eta\} \longrightarrow 0.$$

For every fixed $0 < \varepsilon < 1$*,*

$$\Pr\{Z_m \le \varepsilon\} \longrightarrow I(W).$$

Proof. The idea is to bound the total expected squared movement. A plus step must move an intermediate Z_j by $Z_j(1 - Z_j)$; too many such movements would exhaust that bound.

Let $D_j = Z_j - \mathbb{E}[Z_{j+1} \mid \mathcal{F}_j] \geq 0$. Telescoping gives

$$\sum_{j=0}^{m-1} \mathbb{E}D_j = Z_0 - \mathbb{E}Z_m \leq Z_0 \leq 1.$$

The conditioning identity gives $\mathbb{E}[Z_j Z_{j+1}] = \mathbb{E}[Z_j(Z_j - D_j)]$. Expand $(Z_{j+1} - Z_j)^2$, substitute this identity, and sum; the successive second moments cancel. This yields

$$\begin{aligned}\sum_{j=0}^{m-1} \mathbb{E}(Z_{j+1} - Z_j)^2 &= \mathbb{E}Z_m^2 - Z_0^2 + 2\sum_{j=0}^{m-1} \mathbb{E}[Z_j D_j] \\ &\leq 1 + 2\sum_{j=0}^{m-1} \mathbb{E}D_j \leq 3.\end{aligned}$$

On the plus branch, which has conditional probability $1/2$, $Z_{j+1} = Z_j^2$. Hence for every depth m,

$$\sum_{j=0}^{m-1} \mathbb{E}[Z_j^2(1 - Z_j)^2] \leq 6.$$

Each summand depends only on the first j signs, so extending the finite tree does not change it. The nonnegative partial sums are bounded, so the summands tend to zero: infinitely many terms bounded below by a fixed positive number would make the partial sums unbounded. On the middle interval the integrand is at least $\eta^2(1 - \eta)^2$, proving that its probability vanishes.

To identify the endpoint masses, fix ε and take $0 < \eta < \min\{\varepsilon, 1 - \varepsilon\}$. Lemma 3.1 gives

$$\left| I_m - \mathbf{1}_{\{Z_m \leq \varepsilon\}} \right| \leq \begin{cases} \log(1 + \eta), & Z_m \leq \eta, \\ 1, & \eta < Z_m < 1 - \eta, \\ \sqrt{2\eta - \eta^2}, & Z_m \geq 1 - \eta. \end{cases}$$

Take expectations. First let $m \to \infty$, using the vanishing middle mass and $\mathbb{E}I_m = I(W)$, and then let $\eta \downarrow 0$. This proves the second assertion. □

For any fixed reliability target, nearly an $I(W)$ fraction eventually meet it. This statement does not permit replacing the fixed target by $1/N^2$ or any other length-dependent target. The next proof supplies the missing quantitative step.

6 From moderately reliable to sufficiently reliable

Weak polarization reaches any fixed small target, but block decoding needs a target that shrinks with N. We use the first half of the tree depth to reach a fixed small parameter. During the remaining steps, squaring can reduce it very rapidly. The proof normalizes out those squarings so that the intervening losses have bounded expectation.

Theorem 6.1 (A sufficient rate of polarization)**.** *For $N = 2^m$, put*

$$b_m = 2^{\lceil m/8 \rceil}, \qquad \tau_N = 2^{-b_m}.$$

Then

$$\frac{1}{N}\left| \{i : Z(W_N^{(i)}) \leq \tau_N\} \right| = \Pr\{Z_m \leq \tau_N\} \longrightarrow I(W), \qquad \tau_N \leq 2^{-N^{1/8}}.$$

Proof. Put $L = -\log Z$ on paths with positive parameters. If a path reaches $Z = 0$, both children have parameter zero by Lemma 3.2; it remains zero and automatically meets every later target. For the remaining paths, a plus step doubles L and a minus step decreases it by at most one, since $Z^- \le 2Z$.

Consider r further signs from a state with $L_0 \ge a > 1$. Let B_i indicate a plus sign and $S_j = \sum_{i=1}^{j} B_i$, with $S_0 = 0$. For the normalized quantity $\Lambda_i = L_i/2^{S_i}$, a plus step leaves $\Lambda_i = \Lambda_{i-1}$, whereas a minus step gives $\Lambda_i \ge \Lambda_{i-1} - 2^{-S_{i-1}}$. Summing these one-step inequalities gives

$$\frac{L_r}{2^{S_r}} \ge L_0 - T_r, \qquad T_r = \sum_{i=1}^{r} (1 - B_i) 2^{-S_{i-1}}. \tag{6.1}$$

A plus step preserves the normalized quantity; a minus step subtracts at most the displayed weight. Define T_r on all fair future sign strings, including those on which the parameter reaches zero. Its expectation is over all these strings, without conditioning on avoiding zero. Independence gives

$$\mathbb{E}2^{-S_{i-1}} = \prod_{k=1}^{i-1} \mathbb{E}2^{-B_k} = \left(\frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}\right)^{i-1} = (3/4)^{i-1}.$$

The current sign is independent of the earlier ones, so

$$\mathbb{E}T_r = \frac{1}{2} \sum_{i=1}^{r} (3/4)^{i-1} \le 2, \qquad \Pr\{T_r > a - 1\} \le \frac{2}{a-1}.$$

Outside this exceptional event, $L_r \ge 2^{S_r}$. The probability bound is Markov's inequality: a nonnegative variable T satisfies $\mathbb{E}T \ge u \Pr\{T > u\}$ for every $u > 0$.

Split the total depth at $t = \lfloor m/2 \rfloor$, leaving $r = m - t$ independent signs. By weak polarization,

$$\Pr\{Z_t \le 2^{-a}\} \longrightarrow I(W)$$

for each fixed $a > 1$. Meanwhile S_r has binomial distribution with mean $r/2 \ge m/4$ and variance $r/4 \le m/4$. Falling below $m/8$ therefore requires a deviation of at least $m/8$ from the mean. Applying Markov to the squared deviation, which is Chebyshev's inequality, gives

$$\Pr\{S_r < m/8\} \le \frac{16}{m}.$$

If the initial small-parameter event holds and neither tail exception occurs, $L_m \ge 2^{S_r} \ge 2^{\lceil m/8 \rceil}$, since S_r is an integer. Therefore

$$\Pr\{Z_m \le \tau_N\} \ge \Pr\{Z_t \le 2^{-a}\} - \frac{2}{a-1} - \frac{16}{m}.$$

This subtracts the two tail probabilities by a union bound; their exceptional events need not be independent. Paths reaching zero are already counted as good. First let $m \to \infty$ and then $a \to \infty$. The lower limit is at least $I(W)$.

For the upper limit, Lemma 3.1 and conservation give

$$I(W) = \mathbb{E}I_m \ge (1 - \log(1 + \tau_N)) \Pr\{Z_m \le \tau_N\}.$$

The factor tends to one, so the upper limit is at most $I(W)$. Finally $b_m \ge 2^{m/8} = N^{1/8}$, proving the last inequality. □

7 Assembling the binary-input coding theorem

Let

$$A_N = \{i : Z(W_N^{(i)}) \le \tau_N\}.$$

The two preceding theorems give $|A_N|/N \to I(W)$. Choose a fixed frozen vector supplied by Proposition 4.2. Then

$$P_e \le \frac{1}{2}|A_N|\tau_N \le N2^{-N^{1/8}} \longrightarrow 0,$$

because $\log N - N^{1/8} \to -\infty$. For any fixed $0 < R < I(W)$, all sufficiently large N have at least $\lfloor RN \rfloor$ eligible positions. Retain that many and fix all the others, choosing a frozen vector for the smaller set by the same averaging argument. The rate tends to R and the error bound remains valid. Rate zero uses one message. This proves the first part of Theorem 1.1, including its algorithmic bounds from Section 4.

For a binary-input symmetric channel, all frozen bits can be zero by Proposition 4.2, and its stronger guarantee holds for every message. Its Shannon capacity equals $I(W)$: exchanging the input labels and permuting the output leaves mutual information unchanged. To justify the optimization, for any input distribution P,

$$I(P, W) = H(PW) - \sum_x P(x)H(W(\cdot \mid x)).$$

The output distribution PW is linear in P, so the first term is concave in P by entropy concavity; the second term is linear. Hence $I(P, W)$ is concave. Averaging P with its input-exchanged distribution gives uniform input, whose mutual information is therefore at least as large. This proves $C(W) = I(W)$.

The construction targets a *specified memoryless channel*. Its reliability theorem averages over channel noise, even when it holds for each message. It is not a guarantee against every adversarial error pattern of a given weight. Such a guarantee requires a separate distance and decoding analysis.

8 From uniform inputs to arbitrary finite DMC capacity

We now prove the general-DMC assertion without treating uniform-input capacity as the unrestricted capacity. Let $W : \mathcal{X} \to \mathcal{Y}$ have finite alphabets. The construction uses two ideas: approximate a desired input distribution with uniformly chosen binary labels, then decode the label bits in successive layers.

8.1 A deterministic mapper shapes the input distribution

Choose an integer $s \ge 1$ and a map $g : \{0,1\}^s \to \mathcal{X}$. For a uniform label $B = (B_1, \dots, B_s)$, transmit $X = g(B)$. If exactly n_x labels map to x, then $\Pr\{X = x\} = n_x/2^s$. Every distribution on a finite alphabet can be approximated by such dyadic distributions: round its masses to multiples of 2^{-s} and distribute the remaining integer counts so that the sum is 2^s. Each rounded mass differs by at most 2^{-s}, so the sum of their absolute differences is at most $|\mathcal{X}|2^{-s}$.

For each fixed s, g, define a binary-input channel for layer j with input B_j and output (Y, B_1^{j-1}):

$$V_j(y, b_1^{j-1} \mid b_j) = 2^{-(s-1)} \sum_{b_{j+1}^s} W(y \mid g(b_1^s)). \tag{8.1}$$

Here $2^{-(s-1)} = 2^{-(j-1)}2^{-(s-j)}$: the prefix is uniform output side information, while future label bits are averaged uniformly.

Its transition probabilities sum to one. By the chain rule and the fact that Y depends on B only through $g(B)$,

$$\sum_{j=1}^{s} I(V_j) = I(B;Y) = I(X;Y). \tag{8.2}$$

Each V_j may be asymmetric. Our binary-input theorem allows that.

8.2 Parallel polar encoders and successive layer decoding

At layer j, use a length-N polar encoder with an information set $A_{j,N}$ of bit-channels of V_j satisfying the same threshold τ_N. Let its encoded row be b_j^N. At physical coordinate i, send

$$x_i = g(b_{1,i}, \ldots, b_{s,i}).$$

The physical channel is used N times, while the total message has $\sum_j |A_{j,N}|$ bits. Decode layer 1, then layer 2, and so on. Within layer j, successive cancellation with the likelihoods of V_j estimates the pre-transform row $\widehat{u}_j^N$. Re-encode it to form

$$\widehat{b}_j^N = \mathcal{E}_N(\widehat{u}_j^N).$$

Supply these *encoded* rows, rather than the pre-transform rows, as the side information in later layers.

We must justify that fixed frozen bits do not invalidate the uniform-label model in (8.1). Temporarily average over independent fair frozen bits in *every* layer and a uniform message. Then every pre-transform row is uniform, so invertibility makes all the encoded rows independent uniform bit strings. The columns of the label array are consequently independent uniform labels. The following calculation verifies the channel seen by a layer.

Lemma 8.1 (The product channel for one layer)**.** *Under this uniform-array model, conditional on $B_j^N = b^N$, the output $(Y^N, B_1^N, \ldots, B_{j-1}^N)$ has transition distribution V_j^N.*

Proof. Write $c_{<j,i} = (c_{1,i}, \ldots, c_{j-1,i})$ for the earlier bits in column i. Summing out all later rows factorizes over columns, giving

$$\begin{aligned}
&\Pr\{B_j^N = b^N, Y^N = y^N, B_{<j}^N = c_{<j}^N\} \\
&\quad = \prod_{i=1}^{N} \left[2^{-s} \sum_{d \in \{0,1\}^{s-j}} W(y_i \mid g(c_{<j,i}, b_i, d)) \right] \\
&\quad = 2^{-N} \prod_{i=1}^{N} V_j(y_i, c_{<j,i} \mid b_i).
\end{aligned}$$

Each bracket is $V_j/2$ by (8.1). Divide by $\Pr\{B_j^N = b^N\} = 2^{-N}$ to obtain the asserted conditional distribution. □

With a genie supplying all true earlier encoded layers, the current pre-transform row therefore has exactly the polar bit-channels of V_j. With its true bit prefix supplied as well, a decision at information position i has error at most $Z((V_j)_N^{(i)})/2$.

If the actual decoder makes its first error at a given bit in a given layer, every required earlier layer and bit prefix is still correct. Union-bound over all such positions, and then average over the frozen vectors, to obtain

$$\mathbb{E}_f P_e(f) \le \frac{1}{2} \sum_{j=1}^{s} \sum_{i \in A_{j,N}} Z((V_j)_N^{(i)}) \le \frac{sN}{2} \tau_N.$$

Some deterministic tuple of frozen vectors attains this bound. Fix it once for that blocklength. No shared or encoder randomness is required during transmission. For fixed s,

$$P_e \longrightarrow 0, \qquad \frac{1}{N}\sum_{j=1}^{s}|A_{j,N}| \longrightarrow \sum_{j=1}^{s} I(V_j) = I(X;Y).$$

If the many-to-one mapper causes any message collisions, they can be removed with negligible rate loss. When $P_e < 1/4$, retain a set of at least half the messages whose individual errors are at most $2P_e$. Their success probabilities are strictly above $1/2$. Two retained messages cannot have the same transmitted word: their channel-output distributions would be identical, while disjoint decoder decisions cannot both have probability above $1/2$. The retained encoder is therefore injective. If the original message has $K \geq 1$ bits, keep exactly 2^{K-1} of these messages and label them by $\{0,1\}^{K-1}$; the rate loss is $1/N$. A one-message code is already injective.

8.3 Approaching the optimizing input distribution

On a finite alphabet,

$$I(P_X, W) = H(P_X W) - \sum_x P_X(x) H(W(\cdot \mid x))$$

is continuous in P_X, including zero probabilities, since $-p\log p$ has continuous value zero at $p = 0$. Given $0 \leq R < C(W)$, choose an input distribution P with $I(P, W) > R$, as permitted by the definition of capacity. Choose a sufficiently fine dyadic approximation P_s to P that $I(P_s, W) > R$. Fix its finite s and mapper g, and then let $N \to \infty$ in the layered construction. Its limiting rate exceeds R and its error vanishes. Restricting to exactly $2^{\lfloor RN \rfloor}$ messages and labeling them by $\{0,1\}^{\lfloor RN \rfloor}$ gives rates tending to R; use messages with the smallest individual error probabilities so that the restriction does not increase average error. Keep the original decoder's decisions on retained messages, and redirect every discarded-message output to one fixed retained message. This cannot increase any retained message's error. This proves the full finite-DMC capacity assertion of Theorem 1.1.

For fixed s and a fixed channel, the unrestricted layered encoder and decoder have $O(sN \log N)$ recursive operations, with constants that include the layer-channel likelihood computations and the mapper. Arbitrary message restriction or expurgation need not supply efficient indexing of the retained messages. Finding a suitable frozen tuple by this averaging proof need not be efficient. The order of limits matters: choose a finite shaping resolution for a target rate below capacity, then increase blocklength. We have not proved a uniform polynomial dependence on an arbitrarily small rate gap.

9 History, stronger results, and what the proof establishes

Arıkan introduced channel polarization and polar codes in his 2009 paper *Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels*. The essential change from the random coding proof of Lecture 5 is to organize the channel uses through a recursive invertible transform, so that individual input positions have different reliabilities and can be decoded in a fixed successive order.

Arıkan and Telatar's 2009 *On the rate of channel polarization* proves the stronger reliability scale with every fixed exponent $\beta < 1/2$. Our elementary two-stage proof establishes only the sufficient exponent $1/8$ and does not need that stronger result.

Guruswami and Xia's *Polar Codes: Speed of polarization and polynomial gap to capacity* shows, for binary-input symmetric memoryless channels, that the blocklength and the code construction complexity can be bounded polynomially in the inverse rate gap. Their analysis also bounds decoding complexity polynomially in that gap. This is a sharper question than proving that sufficiently large powers of two eventually work, and is not established by the proof above.

For general alphabet polarization and input-distribution issues, see Şaşoğlu, Telatar, and Arıkan, *Polarization for arbitrary discrete memoryless channels* (2009). The dyadic mapper and binary layer argument here give a direct reduction for the finite-alphabet, average-error capacity statement; they do not assert efficient construction of all frozen vectors or the strongest nonbinary polar coding theorem.