Channel Polarization

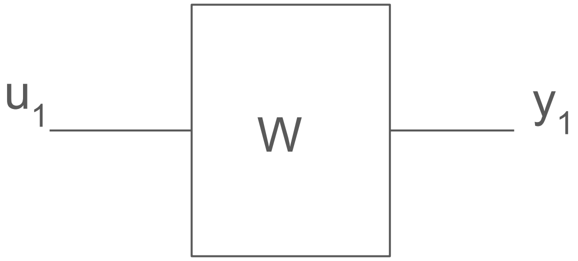
Erdal Arikan

Topics

- Motivation and Intuition
  - (Slides 3-11) - **Ameya**
- Empirical Analysis for BECs
  - (Slides 12-28, Conclusion) - **Evan**
- Mathematical Analysis and Proof Sketches
  - (Slides 29-47) - **Qiaobo**

Motivation

Noisy Channel



- Let $\mathbf{U_1}$ be an input, and W be a noisy channel through which $\mathbf{U_1}$ is passed. Let $\mathbf{Y_1}$ be the corresponding output for $\mathbf{U_1}$
- Now**,** since W is noisy, the resulting output $\mathbf{Y_1}$ might not be equal to $\mathbf{U_1}$**.** Let the error probability be **ε.**
- In such a case, how can we ensure that we get the correct output with a high probability?

Naive Method: Redundant "Encoding"

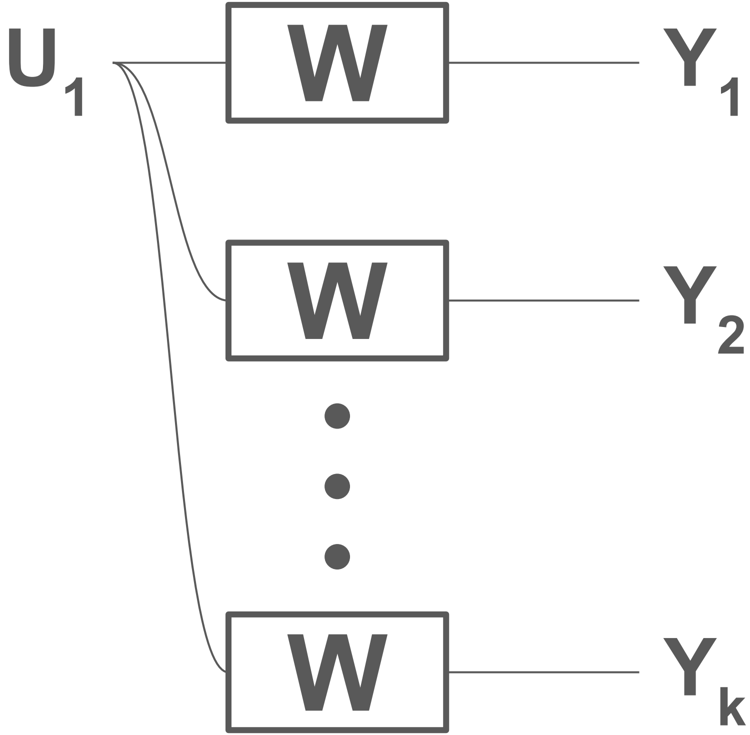
Let's suppose we're using erasure channels.

With this method, we can reconstruct $\mathbf{U_1}$ as long as one channel succeeds.

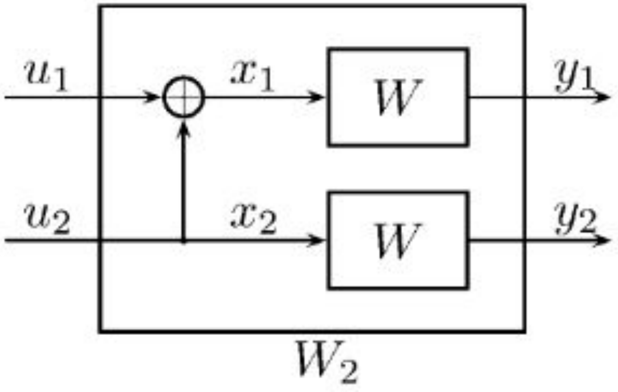
The probability of **k** independent channels all failing is $\boldsymbol{\varepsilon}^k$, which converges to 0 geometrically fast.

Is this the perfect channel?

- No, because we're using **k** channels to send 1 bit.

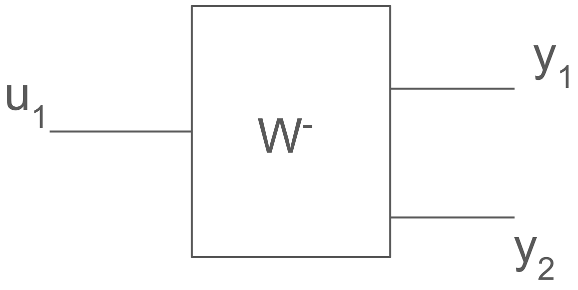


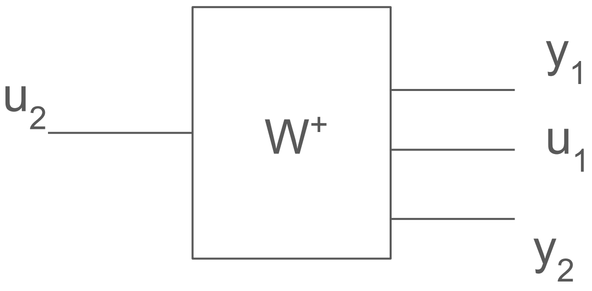
Channel Polarization



We follow the below steps for decoding.

1. Use y_1 and y_2 to decode u_1
2. Assume u_1 is decoded correctly, use u_1, y_1, y_2 to decode u_2

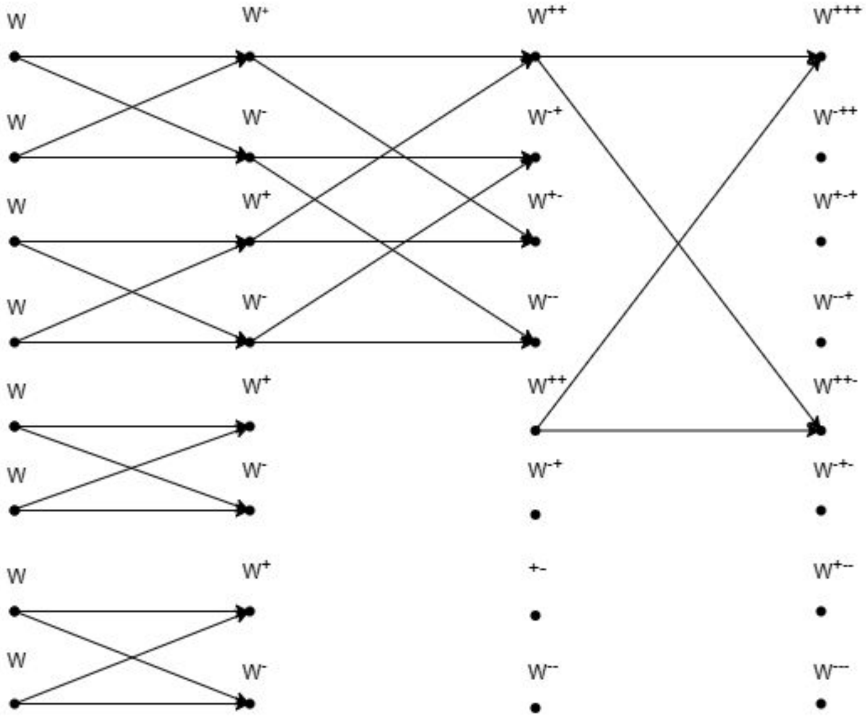




1. **W⁻**: With probability $(1-\varepsilon)^2$ receive $u_1 \oplus u_2$ and u_2. In all other cases, u_1 is lost.
2. Therefore **W⁻** is a BEC$(1-(1-\varepsilon)^2)$
3. **W⁺**: With probability ε^2, u_2 is lost. Therefore, **W⁺** is a BEC(ε^2)

Therefore, we can see that there is some level of polarization with **W⁺** and **W⁻** showing different error probabilities.

Visual Interpretation of the Polarized Channels



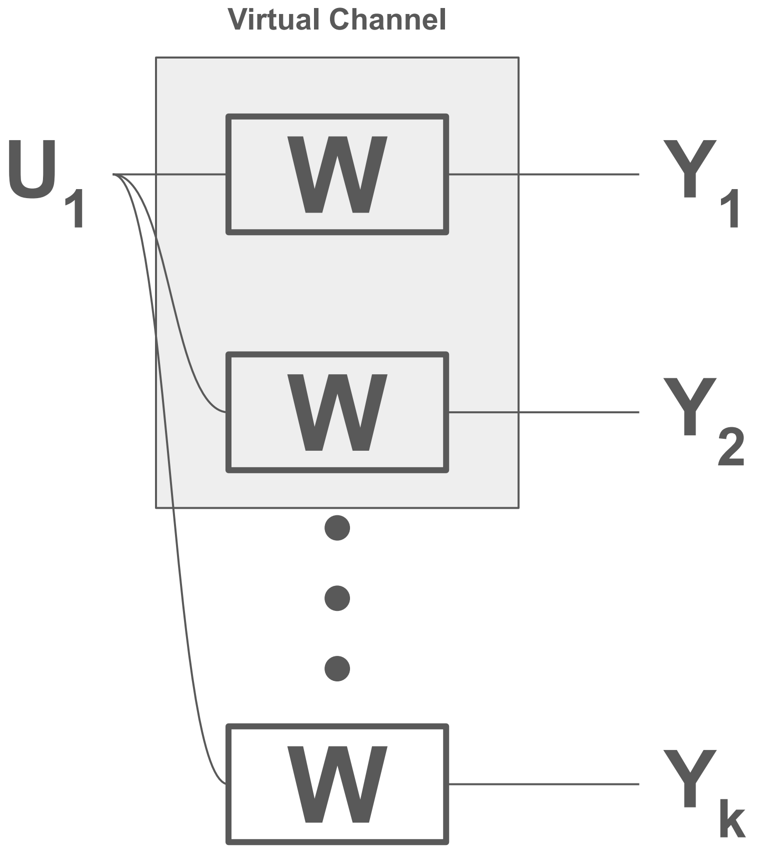
Channel Polarization

What is it?

- Combining a variety of memoryless binary symmetric channels into new **virtual channels**, which can be described in terms of inputs and outputs instead of a physical design.

What is useful about these virtual channels?

- We can construct them such that their channel capacities asymptotically approach 0 or 1 i.e. they are **polarized**



Why Would This Be Useful?

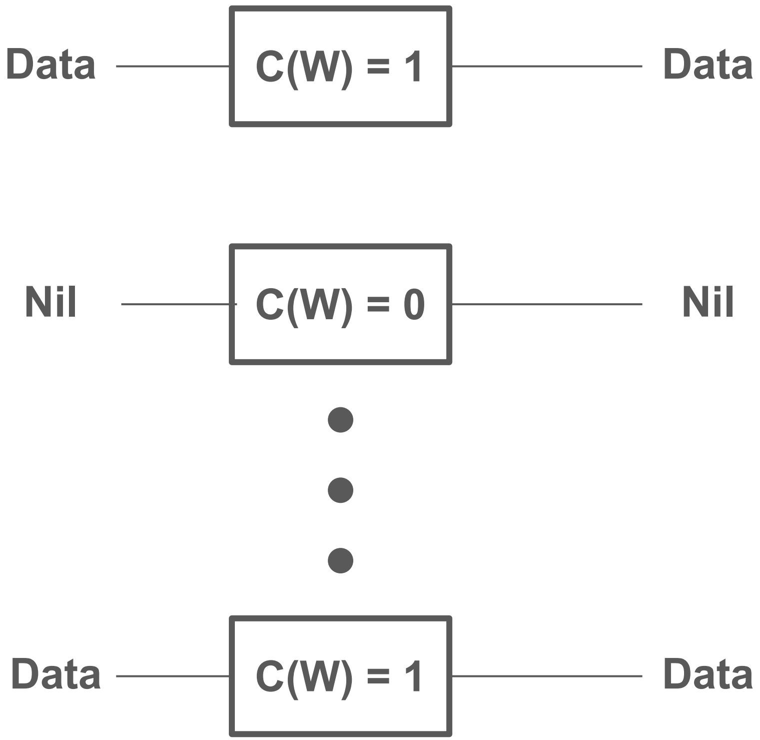
What if we could use lossy channels to make some **perfect channels** and some **useless channels**?

Perfect Channel:

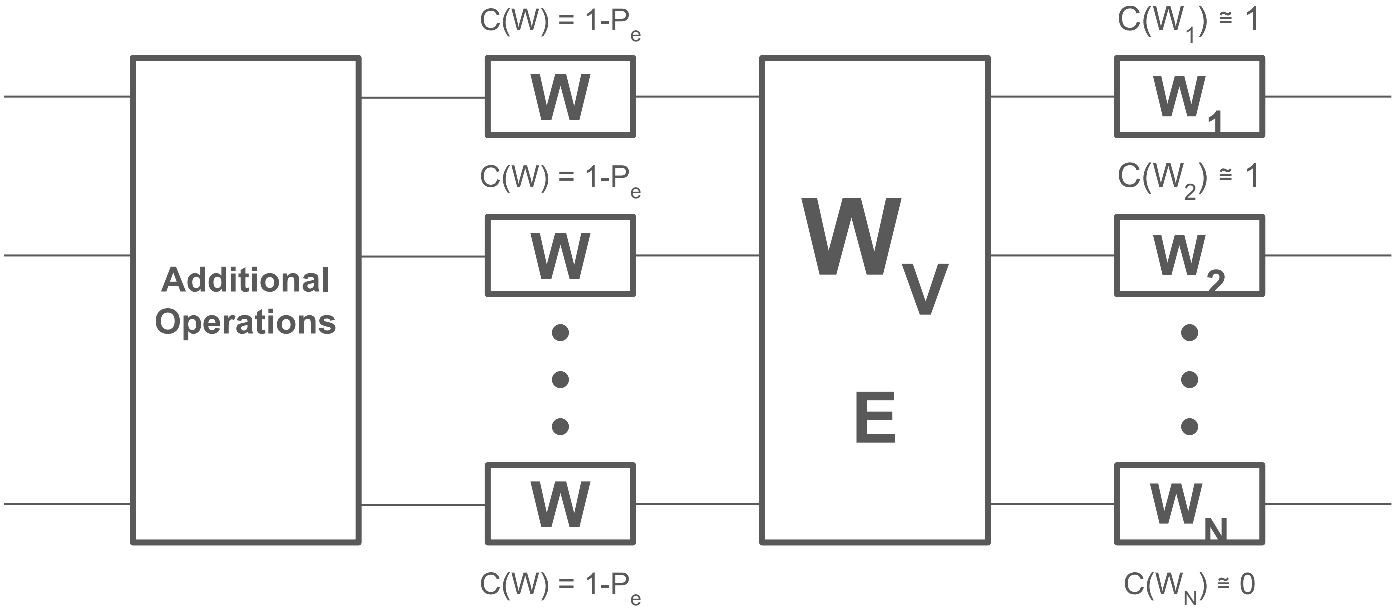
- Send data **without encoding**.

Useless Channel:

- Any data will be lost, so agree with the decoder to **never send data through this channel**.



Big Idea - Combine and Split Channels



Reducing Error While Maintaining Rate

How do we Combine Channels?

We will use the properties of three techniques:

- **Addition modulo 2**
- **Permutation**
- **Recursion**

Why These?

These properties relate Polar Codes to a broader class of channel codes called **block codes**, which we see in the textbook as **(M, n) codes**.

- We know that there exists **some (M, n) code** where **R ≅ C**.
- Can we find that code with a **tractable transformation** of our index set

Why These?

For Polar Codes, we use a **linear, invertible transformation** of the input index set.

- Addition modulo 2 is **always linear**, and **invertible in GF(2)**.
- If you express a set as a **vector**, permutation is a **matrix**
- Recursion can be captured through **Kronecker products**

The paper itself mentions that polar codes resemble **Reed-Muller codes**, that make **Plotkin construction** more flexible.

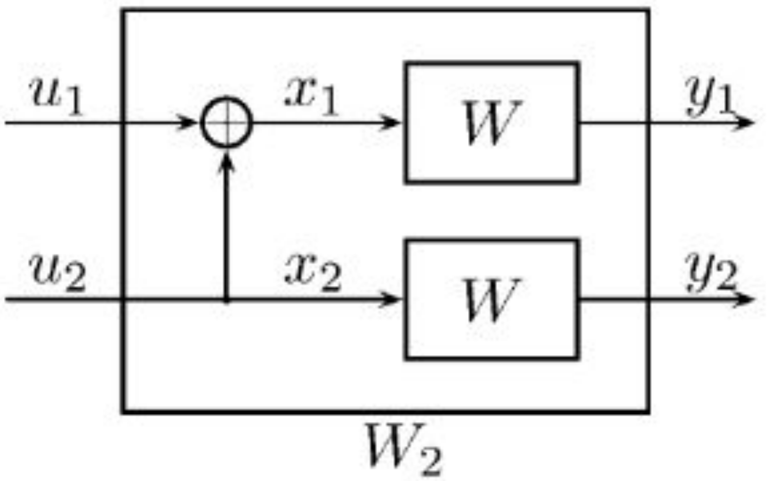
6 years after the paper was published, in a 2015 lecture, Arikan identifies this **computationally tractable transformation O(N log N)**.

How do we Split Channels?

We will use the **Chain Rule for Mutual Information**.

The W_2 Channel

Let's try the naive approach again, but increase the number of bits we send on our two channels. Assume $\mathbf{U_1}$ and $\mathbf{U_2}$ are independent i.i.d. uniform Bernoulli random variables that generate $\mathbf{u_1}$ and $\mathbf{u_2}$.



$$\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\begin{pmatrix}u_1\\ u_2\end{pmatrix} = \begin{pmatrix}x_1\\ x_2\end{pmatrix}$$

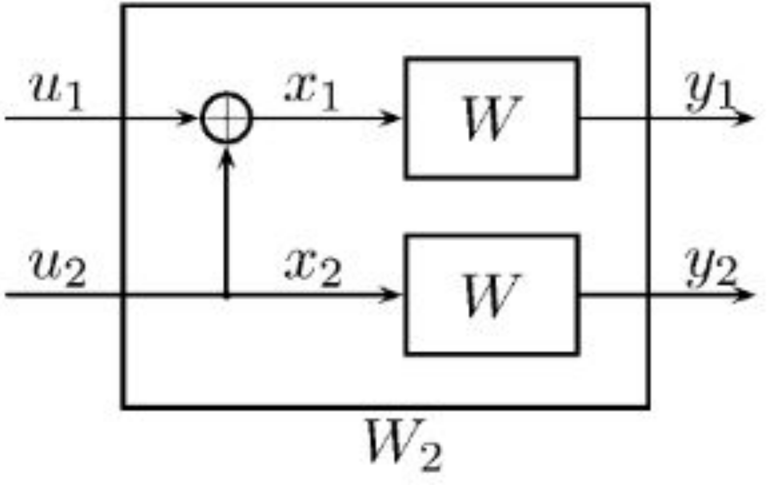
$$\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\begin{pmatrix}u_1\\ u_2\end{pmatrix} = \begin{pmatrix}u_1\\ u_2\end{pmatrix}$$

Transforming Combinations of Binary Input Channels

Remember that we will use the chain rule to split channels. To that end, we can relate $\mathbf{U_1}$ and $\mathbf{U_2}$ causally:

- $\mathbf{u_1}$ and $\mathbf{u_2}$ are sent.
- The decoder receives $\mathbf{y_1}$ and $\mathbf{y_2}$ and uses them to estimate $\mathbf{u_1}$, assuming that $\mathbf{u_2}$ is just noise.
- Then, using that estimate for $\mathbf{u_1}$, the decoder estimates $\mathbf{u_2}$.

We'll discuss the rate later.



Simple Computations—Chain Rule

Remember that the channel capacity we're interested in is the max of $\mathbf{I(U^{(N)};Y^{(N)})}$.
In our case, this is:

$$
\begin{aligned}
I(U^{(2)};Y^{(2)}) &= I(U_1;Y^{(2)}) + I(U_2;Y^{(2)}|U_1) \\
&= I(U_1;Y^{(2)}) + I(U_2;Y^{(2)},U_1)
\end{aligned}
$$

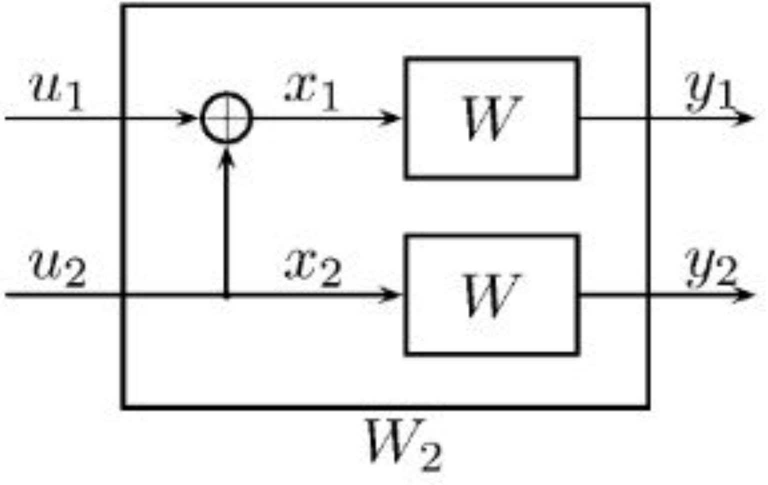
Where the second equality follows from the independence of $\mathbf{U_1}$ and $\mathbf{U_2}$.

$I(U_1;Y^{(2)})$

Suppose that we possess no information about U2, U_2 is an independent Ber(½) random variable. We can treat it as noise in our calculations.

For simplicity, assume **W** is a symmetric binary erasure channel.

What is the capacity of the virtual channel with "input" U_1 and "outputs" Y_1, Y_2?



$I(U_1;Y^{(2)})$

Suppose that $\mathbf{U_2}$ is an independent uniform Bernoulli RV we can treat as noise. Note that, if any channel is erased, then $\mathbf{U_1}$ cannot be reconstructed.

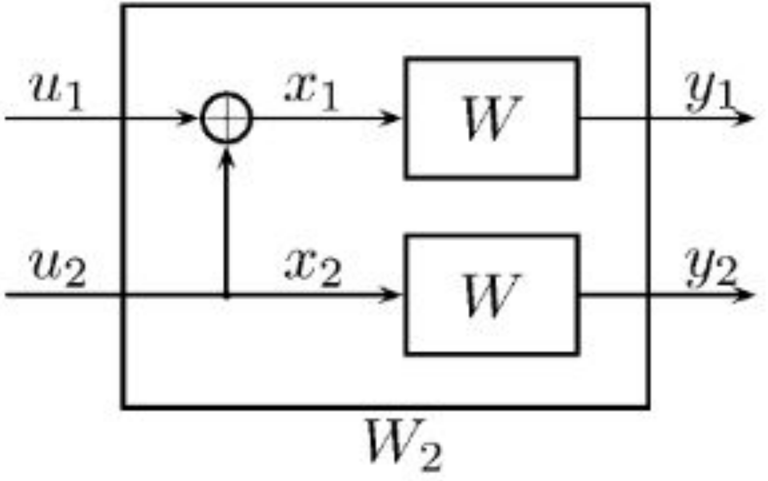
Let $\mathbf{E_i}$ be the event where $\mathbf{Y_i}$ is erased, and let $\mathbf{P(E_i) = \varepsilon}$.

$$
\begin{aligned}
I(U_1;Y^{(2)}) &= H(U_1) - H(U_1|Y_1,Y_2) \\
&= H(U_1) - H(U_1|E_1,E_2)(\epsilon^2) - H(U_1|E_1^c,E_2)(\epsilon-\epsilon^2) - H(U_1|E_1,E_2^c)(\epsilon-\epsilon^2) - H(U_1|E_1^c,E_2^c)(1-\epsilon)^2 \\
&= H(U_1)(1-2\epsilon+\epsilon^2)
\end{aligned}
$$

$I(U_2;Y^{(2)},U_1)$

Now, assume that we have estimated $\mathbf{u}_1$, and have also estimated it **correctly**.

What is the capacity of the virtual channel with “input” $\mathbf{U}_2$ and “outputs” $\mathbf{Y}_1$, $\mathbf{Y}_2$, given $\mathbf{U}_1$?



$I(U_2;Y^{(2)},U_1)$

This time, both channels must be erased in order to fail to reconstruct $\mathbf{U_2}$. If $\mathbf{Y_2}$ is not erased, then $\mathbf{U_2 = Y_2}$. If $\mathbf{Y_1}$ is not erased, then $\mathbf{U_2 = Y_1 \oplus U_1}$. Following a similar reduction as in the previous slide, we find that:

$$I(U_2;Y^{(2)},U_1) = H(U_2)(1-\epsilon^2)$$

Capacity Preservation

Note that our **capacities are preserved** across our combine-and-split operation.

$$(1-\epsilon^2)+(1-2\epsilon+\epsilon^2)=2-2\epsilon=(1-\epsilon)+(1+\epsilon)$$

Also, note that

$$1-2\epsilon+\epsilon^2\leq 1-\epsilon\leq 1-\epsilon^2$$

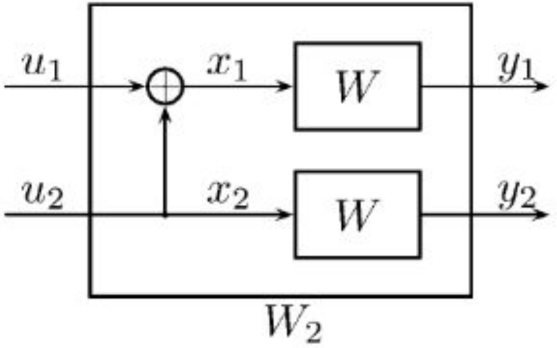
This is the first hint of our polarization—splitting and combining channels has created one virtual channel with **greater capacity** (W^+) and one virtual channel with **lower capacity** (W^-).

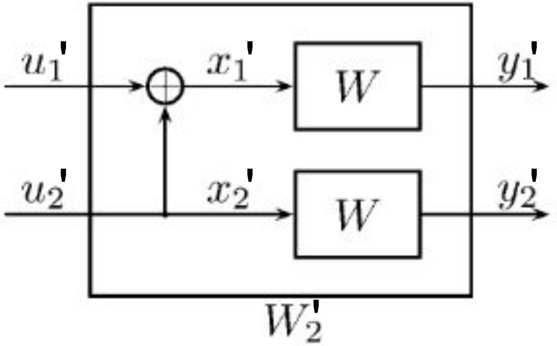
Extending W_2 to W_4

One convenient analytic property of binary erasure channels is that the resulting virtual channels can be **physically modeled as binary erasure channels.**

We want to group the two worst virtual channels, ($\mathbf{u_1;y_1,y_2}$) and ($\mathbf{u'_1;y'_1,y'_2}$), together.

We also want to group the two best channels ($\mathbf{u_2;u_1,y_1,y_2}$) and ($\mathbf{u'_2;u'_1,y'_1,y'_2}$) together.





Extending W_2 to W_4

We construct our outputs to attain the following channels, which we will denote by their respective mutual informations.

- **$C(W^{--}) = \max I(U_1 ; Y^{(4)})$**
- **$C(W^{+-}) = \max I(U_2 ; Y^{(4)}, U_1)$**
- **$C(W^{-+}) = \max I(U_3 ; Y^{(4)}, U_1, U_2)$**
- **$C(W^{++}) = \max I(U_4 ; Y^{(4)}, U_1, U_2, U_3)$**

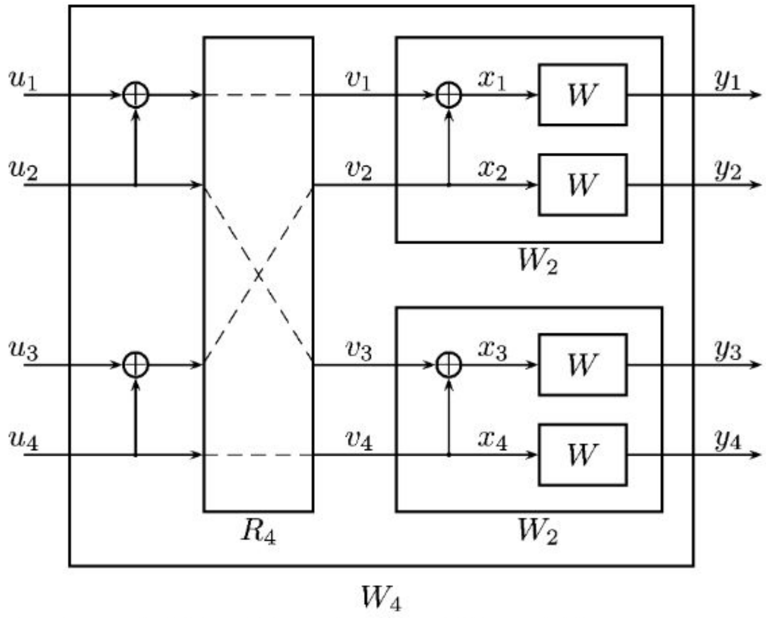


Fig. 2. The channel W_4 and its relation to W_2 and W.

Resulting Channel Capacities

Because we can treat our virtual channels as binary erasure channels, some recursive calculation gives us the following channel capacities:

- **$C(W^{--}) = 1 - 2(2\varepsilon - \varepsilon^2) + (2\varepsilon - \varepsilon^2)^2$**
- **$C(W^{+-}) = 1 - (2\varepsilon - \varepsilon^2)^2$**
- **$C(W^{-+}) = 1 - 2\varepsilon^2 + \varepsilon^4$**
- **$C(W^{++}) = 1 - \varepsilon^4$**

With some algebra, it becomes clear that these channels also conserve the sum of all channel capacities: **$4 - 4\varepsilon$**.

We can also see that

$C(W^{--}) \leq C(W^{+-}) \leq 1 - \varepsilon \leq C(W^{-+}) \leq C(W^{++})$

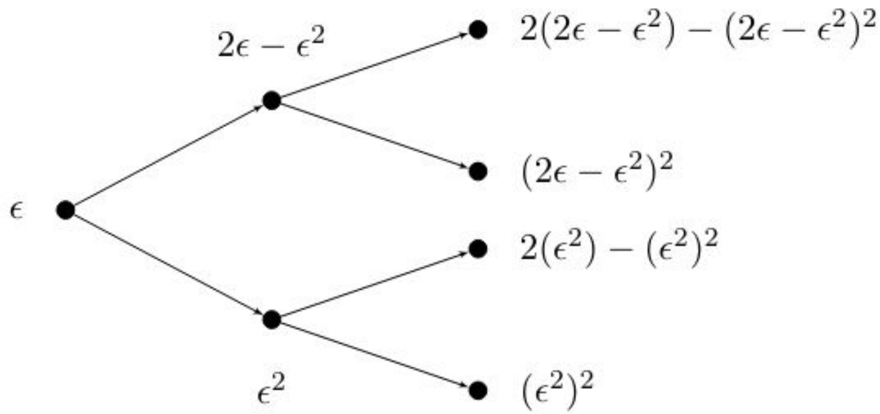
Polar Code Channels as a Bounded Martingale

Let **W'** denote a "parent channel," and let **W^-** and **W^+** denote its "child" virtual channels where **$C(W^-) \leq C(W') \leq C(W^+)$**. We can show that, for BECs,

$C(W^+) = 2C(W') - C(W')^2$

$C(W^-) = C(W')^2$

Assume that we are taking a uniform random walk through our "tree" of channels. What can we say about that process?



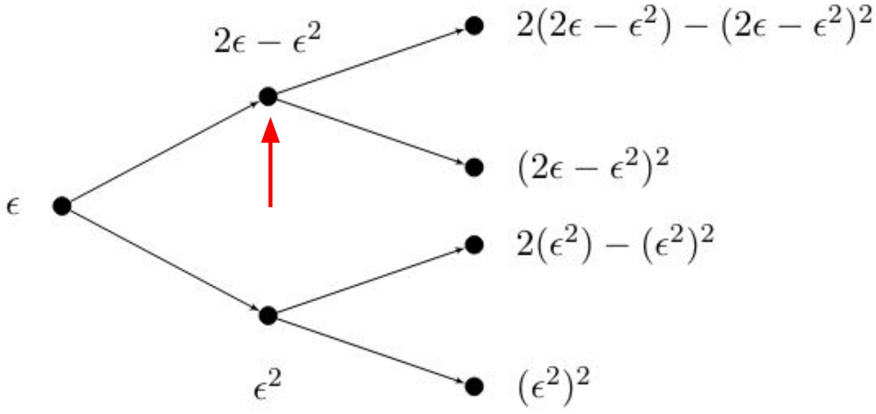
https://web.stanford.edu/class/ee376a/files/polarcodes.pdf

Random Walk through Polar Code Channels

Let **E[C(W)|C(W')]** denote the expected channel capacity of the next **step** we take in our walk.

E[C(W)|C(W')] = ½ C(W⁺) + ½ C(W⁻) = C(W')

This realization points to why a "process" of channel capacities appears as a martingale, a detail we will elaborate on later.



https://web.stanford.edu/class/ee376a/files/polarcodes.pdf

Mathematical Analysis

Symmetric Capacity

W: $\mathscr{X} \rightarrow \mathscr{Y}$ is our channel, a generic B-DMC with input alphabet $\mathscr{X}$, output alphabet $\mathscr{Y}$, and transition probabilities **W(y|x)**, x∈$\mathscr{X}$, y∈$\mathscr{Y}$. The input alphabet will always be {0,1}, the output alphabet and the transition probabilities may be arbitrary.

The symmetric capacity is defined as

$$I(W) \triangleq \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \frac{1}{2} W(y|x) \log \frac{W(y|x)}{\frac{1}{2}W(y|0) + \frac{1}{2}W(y|1)}$$

It is used as our measure of **capacity**, since it is the highest **rate** at which reliable communication is possible across using **inputs of with equal frequency.**

I(W) becomes the **Shannon capacity** under the assumption the distribution of errors is the same regardless of whether 0 or 1 is the input.

Symmetric Capacity and Shannon Capacity

With LOTP, $P_Y(y) = \frac{1}{2}W(y|0) + \frac{1}{2}W(y|1)$

Then,

$$
\begin{aligned}
I(W) &= \sum_{y \in Y} \sum_{x \in X} \frac{1}{2} W(y|x) \log_2\left(\frac{W(y|x)}{P_Y(y)}\right) \\
&= \sum_{y \in Y} \sum_{x \in X} P_X(x) W(y|x) \log_2\left(\frac{W(y|x) P_X(x)}{P_Y(y) P_X(x)}\right) \\
&= \sum_{y \in Y} \sum_{x \in X} P_{X,Y}(x,y) \log_2\left(\frac{P_{X,Y}(x,y)}{P_Y(y) P_X(x)}\right) \\
&= I(X;Y)
\end{aligned}
$$

Example: For a Binary Symmetric Channel (BSC), **P(y=0)=P(y=1)=½.** Then **I(W)=1-H(p)**, which is the familiar form of the Shannon capacity for a binary symmetric channel.

Bhattacharyya Parameter

The Bhattacharyya parameter is defined as

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

It is used as our measure of **reliability**. It is an **upper bound** on the probability of a maximum-likelihood decision error when **W** is used only once to transmit a 0 or 1.

Bhattacharyya Parameter Intuition

It is a measure of reliability because it quantifies how much uncertainty or confusion exists in **distinguishing between different input symbols** based on the channel's output.

For a perfect channel, the output distributions for x=0 and x=1 are disjoint, leading to **Z(W)=0**, which indicates that the channel is completely reliable. For a noisy channel, the overlap between the distributions increases, leading to **Z(W) approaching 1**.

Bounds and Relationships

From the definition,

- **I(W)** can be interpreted as an average Kullback-Leibler (KL) divergence between the conditional distributions and the marginal distribution, so it is in [0,1].
- **Z(W)** is non-negative, and is upper bounded by 1 by the AM-GM inequality, so it is also in [0,1].

They also have the following relationship (we will not prove this here):

Proposition 1: For any B-DMC W, we have

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

which suggests that I**(W)≈0 iff Z(W)≈1, I(W)≈1 iff Z(W)≈0.**

Transforming Rate and Reliability

We now first investigate how the rate and reliability parameters change through a local (single-step) transformation.

I(W) has the following property (Proposition 4):

$$I(W^-) + I(W^+) = 2I(W)$$
$$I(W^-) \leq I(W^+)$$
with equality iff $I(W)$ equals 0 or 1.

In other words, if **W** is neither perfect nor completely noisy, the single-step transform moves the symmetric capacity away from the center, i.e. $\mathbf{I(W^-)<I(W)<I(W^+)}$, thus helping polarization.

Reliability

Reliability **Z(W)** has the following property (Proposition 5):

$$Z(W^+) = Z(W)^2$$
$$Z(W^-) \leq 2Z(W) - Z(W)^2$$
$$Z(W^-) \geq Z(W) \geq Z(W^+)$$

We have $Z(W^-) = Z(W^+)$ if $Z(W)$ equals 0 or 1.

Equality holds in the second equation if W is a BEC.

Rate and Reliability

With these two propositions prepared for each single-step transformation, we can directly see the following relationships:

Let **$I(W_{2N}^{(2i-1)})$** be analogous to our single-step **$I(W^{+})$** channel.

Let **$I(W_{2N}^{(2)})$** be analogous to our single-step **$I(W^{-})$** channel.

$$I\left(W_{2N}^{(2i-1)}\right) + I\left(W_{2N}^{(2i)}\right) = 2I\left(W_{N}^{(i)}\right), Z\left(W_{2N}^{(2i-1)}\right) + Z\left(W_{2N}^{(2i)}\right) \leq 2Z\left(W_{N}^{(i)}\right)$$

$$I\left(W_{2N}^{(2i-1)}\right) \leq I\left(W_{N}^{(i)}\right) \leq I\left(W_{2N}^{(2i)}\right), Z\left(W_{2N}^{(2i-1)}\right) \geq Z\left(W_{N}^{(i)}\right) \geq Z\left(W_{2N}^{(2i)}\right)$$

$$Z\left(W_{2N}^{(2i-1)}\right) \leq 2Z\left(W_{N}^{(i)}\right) - Z\left(W_{N}^{(i)}\right)^{2}, Z\left(W_{2N}^{(2i)}\right) = Z\left(W_{N}^{(i)}\right)^{2}$$

Rate and Reliability

As a result, our cumulative rate and reliability satisfy:

$$\sum_{i=1}^{N} I\left(W_N^{(i)}\right) = NI(W), \sum_{i=1}^{N} Z\left(W_N^{(i)}\right) \leq NZ(W)$$

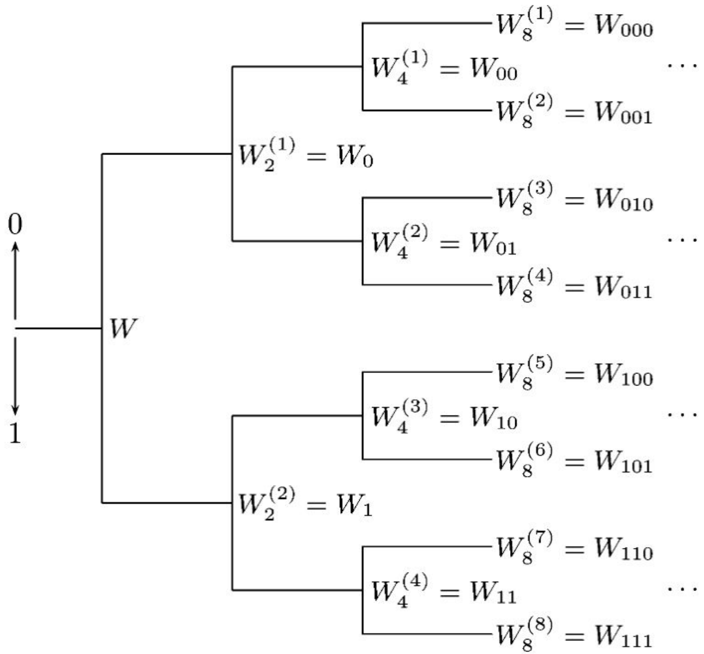
Channel Polarization

Now that we know the trend of **I(W)** and **Z(W)** during polarization, the asymptotic behavior can be derived as follows.

Theorem 1: For any B-DMC W, the channels $\{W_N^{(i)}\}$ *polarize* in the sense that, for any fixed $\delta \in (0,1)$, as N goes to infinity through powers of two, the fraction of indices $i \in \{1,\ldots,N\}$ for which $I(W_N^{(i)}) \in (1-\delta, 1]$ goes to $I(W)$ and the fraction for which $I(W_N^{(i)}) \in [0,\delta)$ goes to $1 - I(W)$.

Channel Polarization

Similarly to our empirical work, we define a random process. We denote each step as $\mathbf{K_n}$, and define $\mathbf{I_n = I(K_n)}$, $\mathbf{Z_n = Z(K_n)}$.



Proof Sketch

1. From our earlier proposition for reliability, we know that at each step $\mathbf{Z(W^-) + Z(W^+) \leq 2Z(W)}$. Following similar reasoning from our empirical work, we reason that $\mathbf{Z_n}$ is a **supermartingale**.
2. By Doob's Martingale Convergence Theorem, $\mathbf{Z_n}$ converges in $\mathbf{L_1}$ and **almost surely** to some random variable $\mathbf{Z_\infty}$.
3. Then,

$$E[|Z_{n+1} - Z_n|] = E[|Z_{n+1} - Z_\infty - Z_n + Z_\infty|] \leq E[|Z_{n+1} - Z_\infty|] + E[|Z_n - Z_\infty|] \to 0$$

Proof Sketch

1. Half of the time, $\mathbf{Z_{n+1}}$ will be $\mathbf{Z_n^2}$ from the proposition of reliability. As a result, we can expect

$$\mathbf{E[|Z_{n+1} - Z_n|] >= ½\ E[|Z_n^2 - Z_n|] = ½\ E[Z_n(1 - Z_n)] \rightarrow 0}$$

2. $\mathbf{Z_\infty}$, the limit of $\mathbf{Z_n}$ can only be 1 or 0 **a.s.**
3. By relating $\mathbf{Z_\infty}$ to $\mathbf{I_\infty}$, we can find that $\mathbf{P(I_\infty=1)=I_0}$ and $\mathbf{P(I_\infty=0)=1-I_0}$.

The symmetric capacity clusters around 0 and 1, except for a vanishing fraction, which implies the conclusion of Theorem 1.

Rate of Polarization

For any achievable rate **R**, there exists a subset of channels in the **Nth** level of channel polarization, with cardinality greater than **NR**, where the Bhattacharyya parameter increasingly diminishes (i.e. reliability increases) according to a “rate of polarization.”

Formally,

Theorem 2: For any B-DMC W with $I(W) > 0$, and any fixed $R < I(W)$, there exists a sequence of sets $\mathcal{A}_N \subset \{1, \ldots, N\}$, $N \in \{1, 2, \ldots, 2^n, \ldots\}$, such that $|\mathcal{A}_N| \geq NR$ and $Z(W_N^{(i)}) \leq O(N^{-5/4})$ for all $i \in \mathcal{A}_N$.

Proof Sketch

1. Let ω be an outcome, and denote a indicator random variable $\mathbf{B_i(\omega)}$. From the proposition of reliability, $\mathbf{B_{i+1}(\omega)=1}$ if $\mathbf{Z_{i+1}(\omega) = Z_i^2(\omega)}$. Otherwise, $\mathbf{B_{i+1}(\omega)=0}$ and $\mathbf{Z_{i+1}(\omega) \leq 2Z_i(\omega)}$.
2. Define a function that captures the set of outcomes after a "step," **m**, where our reliability is bounded, i.e., $\mathcal{T}_m(\zeta) \triangleq \{\omega \in \Omega : Z_i(\omega) \leq \zeta \text{ for all } i \geq m\}$.
3. Now, for that set of outcomes,

$$
\begin{aligned}
Z_n &\leq \frac{Z_n}{Z_{n-1}} \cdot \frac{Z_{n-1}}{Z_{n-2}} \cdot \dots \cdot \frac{Z_{m+1}}{Z_m} \cdot Z_m \\
&= Z_m \cdot \prod_{i=m+1}^{n} \frac{Z_i}{Z_{i-1}} \\
&\leq Z_m \cdot \zeta^{|\{i|B_i(\omega)=1, m\leq i\leq n-1\}|} \cdot 2^{|\{i|B_i(\omega)=0, m\leq i\leq n-1\}|} \\
&\leq \zeta \cdot 2^{n-m} \cdot \prod_{i=m+1}^{n} \left(\frac{\zeta}{2}\right)^{B_i(\omega)}
\end{aligned}
$$

Proof Sketch

1. We know that $$Z_n(\omega) \le \zeta \cdot 2^{n-m} \cdot \prod_{i=m+1}^{n} (\zeta/2)^{B_i(\omega)}, \quad \omega \in T_m(\zeta), n > m.$$

2. further denote $$\mathcal{U}_{m,n}(\eta) \triangleq \{\omega \in \Omega : \sum_{i=m+1}^{n} B_i(\omega) > (1/2 - \eta)(n-m)\}.$$

   we have $$Z_n(\omega) \le \zeta \cdot \left[2^{\frac{1}{2}+\eta} \zeta^{\frac{1}{2}-\eta}\right]^{n-m}, \quad \omega \in \mathcal{T}_m(\zeta) \cap \mathcal{U}_{m,n}(\eta)$$

Proof Sketch

4. Using a lemma in the paper (not proven here), we can bound the probability of our sets of outcomes:

$$P\left[T_{m_0}(\zeta)\right] \geq I_0 - \delta/2, P\left[U_{m,n}(\eta)\right] \geq 1 - 2^{-(n-m)\left[1-H\left(\frac{1}{2}-\eta\right)\right]}$$

one can obtain that

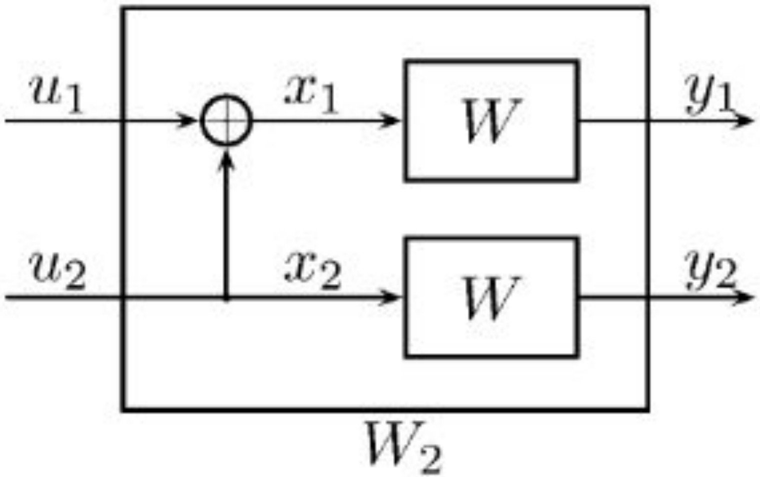
$$P\left[T_{m_1}(\zeta_0) \cap U_{m_1,n}(\eta_0)\right] \geq I_0 - \delta, \quad n \geq n_1$$

5. Combining these probability bounds, and our previous bounds on $\mathbf{Z_n}$, one can reach the conclusion of Theorem 2.

Rate Approaches Entropy of Input

For sufficiently large block size **N**, we find that:

$$
\begin{aligned}
NH(X) = H(X^N) &= H(U^N) \\
&= \sum_{i=1}^{N} H(U_i|U^{i-1}) \\
&\approx \sum_{H(U_i|U^{i-1})=1} 1 \\
&= |\{U_i | H(U_i|U^{i-1}) = 1 \forall 1 \leq i \leq N\}|
\end{aligned}
$$



It turns out that the fraction of channels that are perfect (i.e. channels we send data on) is around **H(X)**, meaning our rate is **H(X)**.

Recall that **I(X;Y) = H(X)** when **H(X|Y) = 0.** As a result, our capacity is maximized, as the output is a deterministic product of the input!

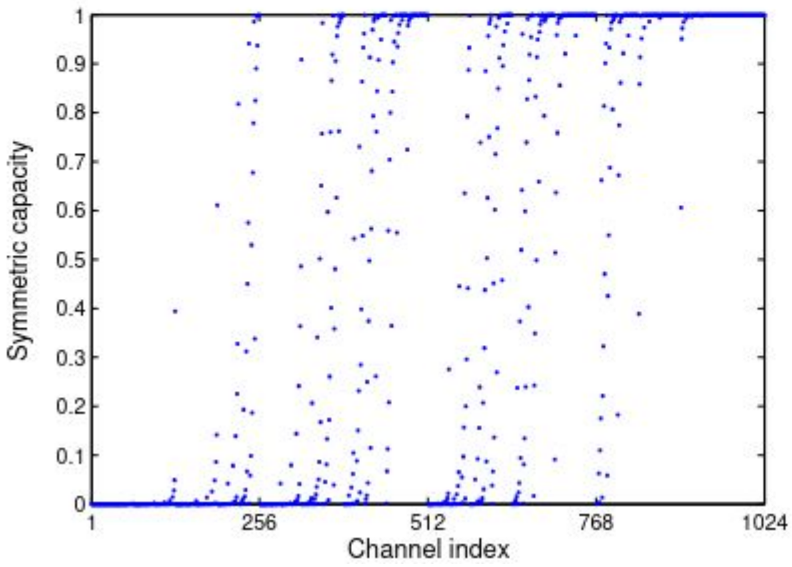
Issues

System Scale:

- The roughly-asymptotic results that make this a “good code” require a significant amount of recursion.

How do we identify these perfect channels?

- Other than for BECs, no algorithm known



Summary

Polar coding is a linear block coding technique that approaches channel capacity through a simple, recursive, invertible operation that is computationally feasible.