Polar Coding

Part 1: The method

Erdal Arıkan

Electrical-Electronics Engineering Department,
Bilkent University, Ankara, Turkey

2016 JTG / IEEE Information Theory Society Summer School,
Department of Electrical Communication Engineering,
Indian Institute of Science,
Bangalore, India
27 June - 1 July 2016

Table of Contents

1.1 Information theory review

1.2 Channel polarization

1.3 Polar coding

1.4 Performance

Information theory review

- Objective
  - Establish notation
  - Review the channel coding theorem
- Reference for this part: T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed., Wiley: 2006.

Information theory review

- Objective
  - Establish notation
  - Review the channel coding theorem
- Reference for this part: T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed., Wiley: 2006.

Information theory review

- Objective
  - Establish notation
  - Review the channel coding theorem
- Reference for this part: T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed., Wiley: 2006.

Information theory review

- Objective
  - Establish notation
  - Review the channel coding theorem
- Reference for this part: T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed., Wiley: 2006.

Notation - I

- Upper case letters $X, U, Y, \ldots$ denote random variables
- Lower case letters $x, u, y, \ldots$ denote realization values
- Script letters $\mathcal{X}, \mathcal{Y}, \cdots$ denote alphabets
- $X^N = (X_1, \ldots, X_N)$ denotes a vector of random variables
- $X_i^j = (X_i, \ldots, X_j)$ denotes a sub-vector of X^N
- Similar notation applies to realizations: x^N and x_i^j

Notation - I

- Upper case letters $X, U, Y, \ldots$ denote random variables
- Lower case letters $x, u, y, \ldots$ denote realization values
- Script letters $\mathcal{X}, \mathcal{Y}, \cdots$ denote alphabets
- $X^N = (X_1, \ldots, X_N)$ denotes a vector of random variables
- $X_i^j = (X_i, \ldots, X_j)$ denotes a sub-vector of X^N
- Similar notation applies to realizations: x^N and x_i^j

Notation - I

- Upper case letters $X, U, Y, \ldots$ denote random variables
- Lower case letters $x, u, y, \ldots$ denote realization values
- Script letters $\mathcal{X}, \mathcal{Y}, \cdots$ denote alphabets
- $X^N = (X_1, \ldots, X_N)$ denotes a vector of random variables
- $X_i^j = (X_i, \ldots, X_j)$ denotes a sub-vector of X^N
- Similar notation applies to realizations: x^N and x_i^j

Notation - I

- Upper case letters $X, U, Y, \ldots$ denote random variables
- Lower case letters $x, u, y, \ldots$ denote realization values
- Script letters $\mathcal{X}, \mathcal{Y}, \cdots$ denote alphabets
- $X^N = (X_1, \ldots, X_N)$ denotes a vector of random variables
- $X_i^j = (X_i, \ldots, X_j)$ denotes a sub-vector of X^N
- Similar notation applies to realizations: x^N and x_i^j

Notation - I

- Upper case letters $X, U, Y, \ldots$ denote random variables
- Lower case letters $x, u, y, \ldots$ denote realization values
- Script letters $\mathcal{X}, \mathcal{Y}, \cdots$ denote alphabets
- $X^N = (X_1, \ldots, X_N)$ denotes a vector of random variables
- $X_i^j = (X_i, \ldots, X_j)$ denotes a sub-vector of X^N
- Similar notation applies to realizations: x^N and x_i^j

Notation - I

- Upper case letters $X, U, Y, \ldots$ denote random variables
- Lower case letters $x, u, y, \ldots$ denote realization values
- Script letters $\mathcal{X}, \mathcal{Y}, \cdots$ denote alphabets
- $X^N = (X_1, \ldots, X_N)$ denotes a vector of random variables
- $X_i^j = (X_i, \ldots, X_j)$ denotes a sub-vector of X^N
- Similar notation applies to realizations: x^N and x_i^j

Notation - II

- $P_X(x)$ denotes the probability mass function (PMF) on a discrete rv X; we also write $X \sim P_X(x)$
- Likewise, we use the standard notation $P_{X,Y}(x, y)$, $P_{X|Y}(x|y)$ to denote the joint and conditional PMF on pairs of discrete rvs
- For simplicity, we drop the subscripts and write $P(x)$, $P(x, y)$, *etc.*, when there is no risk of ambiguity

Notation - II

- $P_X(x)$ denotes the probability mass function (PMF) on a discrete rv X; we also write $X \sim P_X(x)$
- Likewise, we use the standard notation $P_{X,Y}(x,y)$, $P_{X|Y}(x|y)$ to denote the joint and conditional PMF on pairs of discrete rvs
- For simplicity, we drop the subscripts and write $P(x)$, $P(x,y)$, *etc.*, when there is no risk of ambiguity

Notation - II

- $P_X(x)$ denotes the probability mass function (PMF) on a discrete rv X; we also write $X \sim P_X(x)$
- Likewise, we use the standard notation $P_{X,Y}(x, y)$, $P_{X|Y}(x|y)$ to denote the joint and conditional PMF on pairs of discrete rvs
- For simplicity, we drop the subscripts and write $P(x)$, $P(x, y)$, *etc.*, when there is no risk of ambiguity

Entropy

Entropy of $X \sim P(x)$ is defined as

$$H(X) = \mathbb{E}\left[\log \frac{1}{P(X)}\right] = \sum_{x \in \mathcal{X}} P(x) \log \frac{1}{P(x)}$$

- $H(X)$ is a non-negative convex function of the PMF P_X
- $H(X) = 0$ iff X is deterministic
- $H(X) \leq \log |\mathcal{X}|$ with equality iff P_X is uniform over $\mathcal{X}$

Entropy

Entropy of $X \sim P(x)$ is defined as

$$H(X) = \mathbb{E}\left[\log \frac{1}{P(X)}\right] = \sum_{x \in \mathcal{X}} P(x) \log \frac{1}{P(x)}$$

- $H(X)$ is a non-negative convex function of the PMF P_X
- $H(X) = 0$ iff X is deterministic
- $H(X) \leq \log |\mathcal{X}|$ with equality iff P_X is uniform over $\mathcal{X}$

Entropy

Entropy of $X \sim P(x)$ is defined as

$$H(X) = \mathbb{E}\left[\log \frac{1}{P(X)}\right] = \sum_{x \in \mathcal{X}} P(x) \log \frac{1}{P(x)}$$

- $H(X)$ is a non-negative convex function of the PMF P_X
- $H(X) = 0$ iff X is deterministic
- $H(X) \leq \log |\mathcal{X}|$ with equality iff P_X is uniform over $\mathcal{X}$

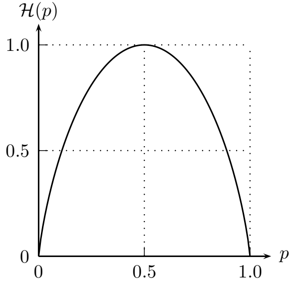
Binary entropy function

For $X \sim \text{Bern}(p)$, *i.e.*,

$$X = \begin{cases} 1, & \text{with prob. } p, \\ 0, & \text{with prob. } 1-p \end{cases}$$

entropy is given by

$$H(X) = \mathcal{H}(p)$$
$$\triangleq -p\log_2(p) - (1-p)\log_2(1-p)$$



Joint Entropy

- Joint entropy of $(X, Y) \sim P(x, y)$

$$H(X, Y) = \mathbb{E}\left[\log \frac{1}{P(X, Y)}\right] = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P(x, y) \log \frac{1}{P(x, y)}$$

- Conditional entropy of X given Y

$$H(X|Y) = H(X, Y) - H(Y)$$

- $H(X|Y) \geq 0$ with eq. iff X if a function of Y
- $H(X|Y) \leq H(X)$ with eq. iff X and Y are independent

Joint Entropy

- Joint entropy of $(X, Y) \sim P(x, y)$

$$H(X, Y) = \mathbb{E}\left[\log \frac{1}{P(X, Y)}\right] = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P(x, y) \log \frac{1}{P(x, y)}$$

- Conditional entropy of X given Y

$$H(X|Y) = H(X, Y) - H(Y)$$

- $H(X|Y) \geq 0$ with eq. iff X if a function of Y
- $H(X|Y) \leq H(X)$ with eq. iff X and Y are independent

Joint Entropy

- Joint entropy of $(X, Y) \sim P(x, y)$

$$H(X, Y) = \mathbb{E}\left[\log \frac{1}{P(X, Y)}\right] = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P(x, y) \log \frac{1}{P(x, y)}$$

- Conditional entropy of X given Y

$$H(X|Y) = H(X, Y) - H(Y)$$

- $H(X|Y) \geq 0$ with eq. iff X if a function of Y
- $H(X|Y) \leq H(X)$ with eq. iff X and Y are independent

Joint Entropy

- Joint entropy of $(X, Y) \sim P(x, y)$

$$H(X,Y) = \mathbb{E}\left[\log \frac{1}{P(X,Y)}\right] = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P(x,y)\log\frac{1}{P(x,y)}$$

- Conditional entropy of X given Y

$$H(X|Y) = H(X,Y) - H(Y)$$

- $H(X|Y) \geq 0$ with eq. iff X if a function of Y
- $H(X|Y) \leq H(X)$ with eq. iff X and Y are independent

Chain rule

- For any pair of rvs (X, Y),
  - $H(X, Y) = H(X) + H(Y|X)$
  - $H(X, Y) = H(Y) + H(X|Y)$
- $H(X, Y) \leq H(X) + H(Y)$ with equality iff X and Y are independent.

Chain rule

- For any pair of rvs (X, Y),
  - $H(X, Y) = H(X) + H(Y|X)$
  - $H(X, Y) = H(Y) + H(X|Y)$
- $H(X, Y) \leq H(X) + H(Y)$ with equality iff X and Y are independent.

Chain rule

- For any pair of rvs (X, Y),
  - $H(X, Y) = H(X) + H(Y|X)$
  - $H(X, Y) = H(Y) + H(X|Y)$
- $H(X, Y) \leq H(X) + H(Y)$ with equality iff X and Y are independent.

Chain rule

- For any pair of rvs (X, Y),
  - $H(X, Y) = H(X) + H(Y|X)$
  - $H(X, Y) = H(Y) + H(X|Y)$
- $H(X, Y) \leq H(X) + H(Y)$ with equality iff X and Y are independent.

Chain rule - II

For any random vector $X^N = (X_1, \ldots, X_N)$

$$\begin{aligned} H(X^N) &= H(X_1) + H(X_2|X_1) + \cdots + H(X_N|X^{N-1}) \\ &= \sum_{i=1}^{N} H(X_i|X^{i-1}) \\ &\leq \sum_{i=1}^{N} H(X_i) \end{aligned}$$

with equality iff $X_1, \ldots, X_N$ are independent.

Mutual information

- For any $(X, Y) \sim P(x, y)$, the mutual information between them is defined as

$$I(X;Y) = H(X) - H(X|Y) = \mathbb{E}\left[\log \frac{P(X|Y)}{P(X)}\right]$$

- Alternatively,

$$I(X;Y) = H(Y) - H(Y|Y)$$

or

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Mutual information

- For any $(X, Y) \sim P(x, y)$, the mutual information between them is defined as

$$I(X; Y) = H(X) - H(X|Y) = \mathbb{E}\left[\log \frac{P(X|Y)}{P(X)}\right]$$

- Alternatively,

$$I(X; Y) = H(Y) - H(Y|Y)$$

or

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

Conditional mutual information

- For any three-part ensemble $(X, Y, Z) \sim P(x, y, z)$, the mutual information between X and Y conditional on Z is defined as

$$I(X; Y|Z) = H(X|Z) - H(X|YZ)$$

- Examples exist for both

$$I(X; Y|Z) < I(X; Y) \quad \text{and} \quad I(X; Y|Z) > I(X; Y)$$

Conditional mutual information

- For any three-part ensemble $(X, Y, Z) \sim P(x, y, z)$, the mutual information between X and Y conditional on Z is defined as

$$I(X; Y|Z) = H(X|Z) - H(X|YZ)$$

- Examples exist for both

$$I(X; Y|Z) < I(X; Y) \quad \text{and} \quad I(X; Y|Z) > I(X; Y)$$

Conditional mutual information: a special case

- If $(X, Y, Z) \sim P(x)P(z)P(y|x, z)$ (i.e., if X and Z are independent, then

$$I(X; Y|Z) = I(X; Y, Z)$$

- Proof.

$$\begin{aligned} I(X; Y|Z) &= \mathbb{E}\left[\log \frac{P(X, Y|Z)}{P(X|Z)P(Y|Z)}\right] \\ &= \mathbb{E}\left[\log \frac{P(X, Y|Z)}{P(X)P(Y|Z)}\right] \\ &= \mathbb{E}\left[\log \frac{P(X|Y, Z)}{P(X)}\right] \\ &= I(X; Y, Z) \end{aligned}$$

Conditional mutual information: a special case

- If $(X, Y, Z) \sim P(x)P(z)P(y|x, z)$ (i.e., if X and Z are independent, then

$$I(X; Y|Z) = I(X; Y, Z)$$

- Proof.

$$\begin{aligned} I(X; Y|Z) &= \mathbb{E}\left[\log \frac{P(X, Y|Z)}{P(X|Z)P(Y|Z)}\right] \\ &= \mathbb{E}\left[\log \frac{P(X, Y|Z)}{P(X)P(Y|Z)}\right] \\ &= \mathbb{E}\left[\log \frac{P(X|Y, Z)}{P(X)}\right] \\ &= I(X; Y, Z) \end{aligned}$$

Chain rule of mutual information

For any ensemble $(X^N, Y) \sim P(x_1, \ldots, x_N, y)$, we have

$$
\begin{aligned}
I(X^N; Y) &= I(X_1; Y) + I(X_2; Y|X_1) + \cdots + I(X_N; Y|X^{N-1}) \\
&= \sum_{i=1}^{N} I(X_i; Y|X^{i-1})
\end{aligned}
$$

Chain rule of mutual information

For any ensemble $(X^N, Y) \sim P(x_1, \ldots, x_N, y)$, we have

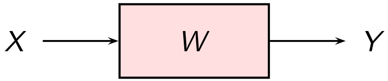
$$
\begin{aligned}
I(X^N; Y) &= I(X_1; Y) + I(X_2; Y|X_1) + \cdots + I(X_N; Y|X^{N-1}) \\
&= \sum_{i=1}^{N} I(X_i; Y|X^{i-1})
\end{aligned}
$$

If the components of X^N are statistically independent, then the chain rule can also be written as

$$
\begin{aligned}
I(X^N; Y) &= I(X_1; Y) + I(X_2; Y, X_1) + \cdots + I(X_N; Y, X^{N-1}) \\
&= \sum_{i=1}^{N} I(X_i; Y, X^{i-1})
\end{aligned}
$$

Discrete memoryless channels (DMC)

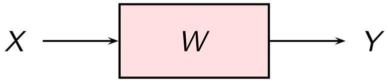
A DMC is a conditional probability assignment $\{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ for two discrete alphabets $\mathcal{X}$, $\mathcal{Y}$.



- We write $W : \mathcal{X} \to \mathcal{Y}$ or simply W to denote a DMC
- $\mathcal{X}$ is called the channel input alphabet
- $\mathcal{Y}$ is called the channel output alphabet
- W is called the channel transition probability matrix

Discrete memoryless channels (DMC)

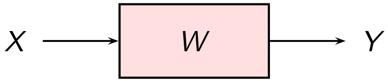
A DMC is a conditional probability assignment
$\{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ for two discrete alphabets $\mathcal{X}$, $\mathcal{Y}$.



- We write $W : \mathcal{X} \to \mathcal{Y}$ or simply W to denote a DMC
- $\mathcal{X}$ is called the channel input alphabet
- $\mathcal{Y}$ is called the channel output alphabet
- W is called the channel transition probability matrix

Discrete memoryless channels (DMC)

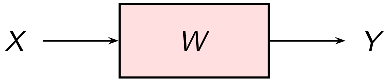
A DMC is a conditional probability assignment
$\{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ for two discrete alphabets $\mathcal{X}$, $\mathcal{Y}$.



- We write $W : \mathcal{X} \to \mathcal{Y}$ or simply W to denote a DMC
- $\mathcal{X}$ is called the channel input alphabet
- $\mathcal{Y}$ is called the channel output alphabet
- W is called the channel transition probability matrix

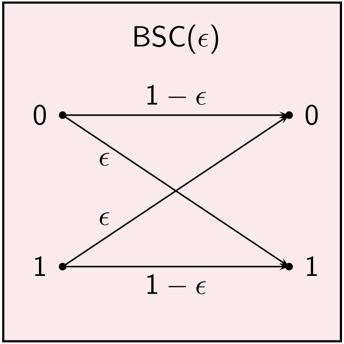
Discrete memoryless channels (DMC)

A DMC is a conditional probability assignment $\{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ for two discrete alphabets $\mathcal{X}$, $\mathcal{Y}$.



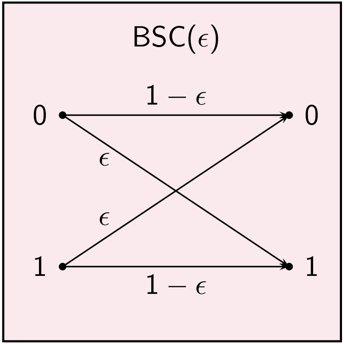
- We write $W : \mathcal{X} \to \mathcal{Y}$ or simply W to denote a DMC
- $\mathcal{X}$ is called the channel input alphabet
- $\mathcal{Y}$ is called the channel output alphabet
- W is called the channel transition probability matrix

An example: Binary Symmetric Channel



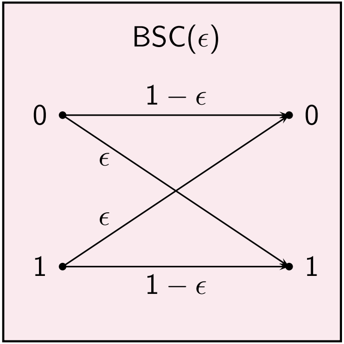
- Input alphabet $\mathcal{X} = \{0, 1\}$
- Output alphabet $\mathcal{Y} = \{0, 1\}$
- Transition probabilities $W(1|1) = W(0|0) = 1 - \epsilon$, $W(0|1) = W(1|0) = \epsilon$

An example: Binary Symmetric Channel



- Input alphabet $\mathcal{X} = \{0, 1\}$
- Output alphabet $\mathcal{Y} = \{0, 1\}$
- Transition probabilities $W(1|1) = W(0|0) = 1 - \epsilon$, $W(0|1) = W(1|0) = \epsilon$

An example: Binary Symmetric Channel



- Input alphabet $\mathcal{X} = \{0, 1\}$
- Output alphabet $\mathcal{Y} = \{0, 1\}$
- Transition probabilities $W(1|1) = W(0|0) = 1 - \epsilon$, $W(0|1) = W(1|0) = \epsilon$

Channel coding

Channel coding is an operation to achieve reliable communication over an unreliable channel. It has two parts.

- An encoder that maps messages to codewords
- A decoder that maps channel outputs back to messages

Channel coding

Channel coding is an operation to achieve reliable communication over an unreliable channel. It has two parts.

- An encoder that maps messages to codewords
- A decoder that maps channel outputs back to messages

Block code

Given a channel $W : \mathcal{X} \to \mathcal{Y}$, a block code with length N and rate R is such that

- the message set consists of integers $\{1, \ldots, M = 2^{NR}\}$
- the codeword for each message m is a sequence $x^N(m)$ of length N over $\mathcal{X}^N$
- the decoder operates on channel output blocks y^N over $\mathcal{Y}^N$ and produces estimates $\hat{m}$ of the transmitted message m.
- the performance is measured by the probability of frame (block) error, also called frame error rate (FER), which is defined as

$$P_e = \Pr(\hat{m} \neq m)$$

  where m is the transmitted message which is assumed equiprobable over the message set and $\hat{m}$ denotes the decoder output.

Block code

Given a channel $W : \mathcal{X} \to \mathcal{Y}$, a block code with length N and rate R is such that

- the message set consists of integers $\{1, \ldots, M = 2^{NR}\}$
- the codeword for each message m is a sequence $x^N(m)$ of length N over $\mathcal{X}^N$
- the decoder operates on channel output blocks y^N over $\mathcal{Y}^N$ and produces estimates $\hat{m}$ of the transmitted message m.
- the performance is measured by the probability of frame (block) error, also called frame error rate (FER), which is defined as

$$P_e = \Pr(\hat{m} \neq m)$$

  where m is the transmitted message which is assumed equiprobable over the message set and $\hat{m}$ denotes the decoder output.

Block code

Given a channel $W : \mathcal{X} \to \mathcal{Y}$, a block code with length N and rate R is such that

- the message set consists of integers $\{1, \ldots, M = 2^{NR}\}$
- the codeword for each message m is a sequence $x^N(m)$ of length N over $\mathcal{X}^N$
- the decoder operates on channel output blocks y^N over $\mathcal{Y}^N$ and produces estimates $\hat{m}$ of the transmitted message m.
- the performance is measured by the probability of frame (block) error, also called frame error rate (FER), which is defined as

$$P_e = \Pr(\hat{m} \neq m)$$

where m is the transmitted message which is assumed equiprobable over the message set and $\hat{m}$ denotes the decoder output.

Block code

Given a channel $W : \mathcal{X} \to \mathcal{Y}$, a block code with length N and rate R is such that

- the message set consists of integers $\{1, \ldots, M = 2^{NR}\}$
- the codeword for each message m is a sequence $x^N(m)$ of length N over $\mathcal{X}^N$
- the decoder operates on channel output blocks y^N over $\mathcal{Y}^N$ and produces estimates $\hat{m}$ of the transmitted message m.
- the performance is measured by the probability of frame (block) error, also called frame error rate (FER), which is defined as

  $$P_e = \Pr(\hat{m} \neq m)$$

  where m is the transmitted message which is assumed equiprobable over the message set and $\hat{m}$ denotes the decoder output.

Channel capacity

The capacity $C(W)$ of a DMC $W : \mathcal{X} \to \mathcal{Y}$ is defined as the maximum of $I(X; Y)$ over all probability assignments of the form

$$P_{X,Y}(x, y) = Q(x)W(y|x)$$

where Q is an arbitrary probability assignment over the channel input alphabet $\mathcal{X}$, or briefly,

$$C(W) = \max_{Q(x)} I(X; Y).$$

Lecture 2 – Channel polarization

- Objective: Explain channel polarization
- Topics:
  - Channel codes as polarizers of information
  - Low-complexity polarization by channel combining and splitting
  - The main polarization theorem
  - Rate of polarization

Lecture 2 – Channel polarization

- Objective: Explain channel polarization
- Topics:
  - Channel codes as polarizers of information
  - Low-complexity polarization by channel combining and splitting
  - The main polarization theorem
  - Rate of polarization

Lecture 2 – Channel polarization

- Objective: Explain channel polarization
- Topics:
  - Channel codes as polarizers of information
  - Low-complexity polarization by channel combining and splitting
  - The main polarization theorem
  - Rate of polarization

Lecture 2 – Channel polarization

- Objective: Explain channel polarization
- Topics:
  - Channel codes as polarizers of information
  - Low-complexity polarization by channel combining and splitting
  - The main polarization theorem
  - Rate of polarization

Lecture 2 – Channel polarization

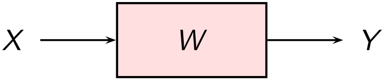
- Objective: Explain channel polarization
- Topics:
  - Channel codes as polarizers of information
  - Low-complexity polarization by channel combining and splitting
  - The main polarization theorem
  - Rate of polarization

Lecture 2 – Channel polarization

- Objective: Explain channel polarization
- Topics:
  - Channel codes as polarizers of information
  - Low-complexity polarization by channel combining and splitting
  - The main polarization theorem
  - Rate of polarization

The channel

Let $W : X \to Y$ be a binary-input discrete memoryless channel

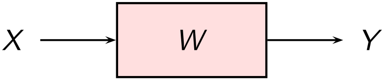


- input alphabet: $\mathcal{X} = \{0, 1\}$,
- output alphabet: $\mathcal{Y}$,
- transition probabilities:

$$W(y|x), \quad x \in \mathcal{X}, y \in \mathcal{Y}$$

The channel

Let $W : X \to Y$ be a binary-input discrete memoryless channel

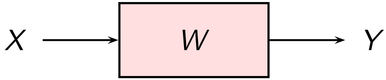


- input alphabet: $\mathcal{X} = \{0, 1\}$,
- output alphabet: $\mathcal{Y}$,
- transition probabilities:

$$W(y|x), \quad x \in \mathcal{X}, y \in \mathcal{Y}$$

The channel

Let $W : X \to Y$ be a binary-input discrete memoryless channel

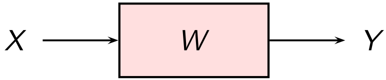


- input alphabet: $\mathcal{X} = \{0, 1\}$,
- output alphabet: $\mathcal{Y}$,
- transition probabilities:

$$W(y|x), \quad x \in \mathcal{X}, y \in \mathcal{Y}$$

The channel

Let $W: X \to Y$ be a binary-input discrete memoryless channel



- input alphabet: $\mathcal{X} = \{0, 1\}$,
- output alphabet: $\mathcal{Y}$,
- transition probabilities:

$$W(y|x), \quad x \in \mathcal{X}, y \in \mathcal{Y}$$

Channel capacity

Let W be an arbitrary binary-input DMC $W : \mathcal{X} = \{0,1\} \to \mathcal{Y}$.

- The capacity of W is defined as

$$C(W) = \max_Q I(X;Y), \qquad (X,Y) \sim Q(x)W(y|x).$$

- The capacity of W with uniform inputs (also called *symmetric capacity*) is defined as

$$I(W) = I(X;Y), \qquad (X,Y) \sim Q_{\text{unif}}(x)W(y|x) = \frac{1}{2}W(y|x).$$

- We use base-2 logarithms so that Use base-2 logarithms:

$$0 \le I(W) \le C(W) \le 1$$

Channel capacity

Let W be an arbitrary binary-input DMC $W : \mathcal{X} = \{0, 1\} \to \mathcal{Y}$.

- The capacity of W is defined as

$$C(W) = \max_Q I(X; Y), \qquad (X, Y) \sim Q(x)W(y|x).$$

- The capacity of W with uniform inputs (also called *symmetric capacity*) is defined as

$$I(W) = I(X; Y), \qquad (X, Y) \sim Q_{\text{unif}}(x)W(y|x) = \frac{1}{2}W(y|x).$$

- We use base-2 logarithms so that Use base-2 logarithms:

$$0 \leq I(W) \leq C(W) \leq 1$$

Channel capacity

Let W be an arbitrary binary-input DMC $W : \mathcal{X} = \{0, 1\} \to \mathcal{Y}$.

- The capacity of W is defined as

$$C(W) = \max_Q I(X; Y), \qquad (X, Y) \sim Q(x)W(y|x).$$

- The capacity of W with uniform inputs (also called *symmetric capacity*) is defined as

$$I(W) = I(X; Y), \qquad (X, Y) \sim Q_{\text{unif}}(x)W(y|x) = \frac{1}{2}W(y|x).$$

- We use base-2 logarithms so that Use base-2 logarithms:

$$0 \leq I(W) \leq C(W) \leq 1$$

Input-output symmetry for a binary-input channel

- A binary-input channel $W : \mathcal{X} = \{0, 1\} \to \mathcal{Y}$ is called *input-output symmetric* if there exists a permutation π of the output alphabet $\mathcal{Y}$ such that the following conditions are satisfied:
  - $\pi^{-1} = \pi$
  - $W(y|0) = W(\pi(y)|1)$ for all $y \in \mathcal{Y}$.
- Fact: If W is input-output symmetric, then $C(W) = I(W)$.
- Fact: $I(W)$ is the highest achievable rate by linear codes.

Input-output symmetry for a binary-input channel

- A binary-input channel $W : \mathcal{X} = \{0,1\} \to \mathcal{Y}$ is called *input-output symmetric* if there exists a permutation π of the output alphabet $\mathcal{Y}$ such that the following conditions are satisfied:
  - $\pi^{-1} = \pi$
  - $W(y|0) = W(\pi(y)|1)$ for all $y \in \mathcal{Y}$.
- Fact: If W is input-output symmetric, then $C(W) = I(W)$.
- Fact: $I(W)$ is the highest achievable rate by linear codes.

Input-output symmetry for a binary-input channel

- A binary-input channel $W : \mathcal{X} = \{0, 1\} \to \mathcal{Y}$ is called *input-output symmetric* if there exists a permutation π of the output alphabet $\mathcal{Y}$ such that the following conditions are satisfied:
  - $\pi^{-1} = \pi$
  - $W(y|0) = W(\pi(y)|1)$ for all $y \in \mathcal{Y}$.
- Fact: If W is input-output symmetric, then $C(W) = I(W)$.
- Fact: $I(W)$ is the highest achievable rate by linear codes.

Input-output symmetry for a binary-input channel

- A binary-input channel $W : \mathcal{X} = \{0,1\} \to \mathcal{Y}$ is called *input-output symmetric* if there exists a permutation π of the output alphabet $\mathcal{Y}$ such that the following conditions are satisfied:
  - $\pi^{-1} = \pi$
  - $W(y|0) = W(\pi(y)|1)$ for all $y \in \mathcal{Y}$.
- Fact: If W is input-output symmetric, then $C(W) = I(W)$.
- Fact: $I(W)$ is the highest achievable rate by linear codes.

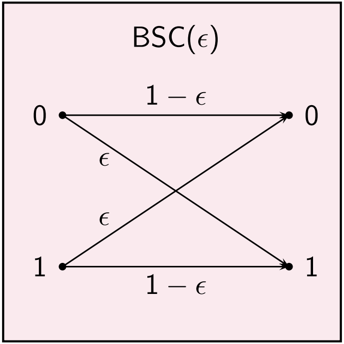
Input-output symmetry for a binary-input channel

- A binary-input channel $W : \mathcal{X} = \{0, 1\} \to \mathcal{Y}$ is called *input-output symmetric* if there exists a permutation π of the output alphabet $\mathcal{Y}$ such that the following conditions are satisfied:
  - $\pi^{-1} = \pi$
  - $W(y|0) = W(\pi(y)|1)$ for all $y \in \mathcal{Y}$.
- Fact: If W is input-output symmetric, then $C(W) = I(W)$.
- Fact: $I(W)$ is the highest achievable rate by linear codes.

Examples of input-output symmetric channels

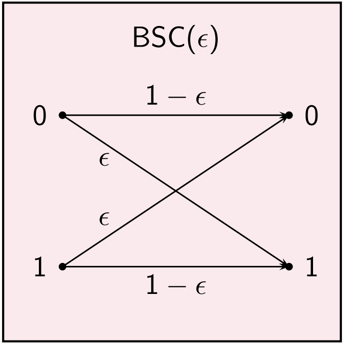
Examples of input-output symmetric channels

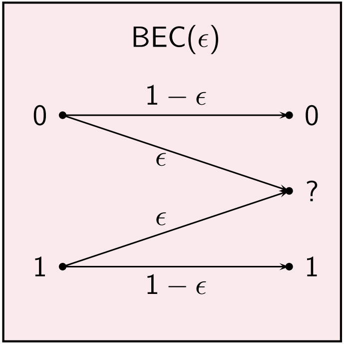
Examples:



Examples of input-output symmetric channels

Examples:





Assumption

In this presentation we will assume that the channel W under consideration is (input-output) symmetric.

- For a symmetric W, the capacity is given by

$$I(W) = H(X) - H(X|Y) = 1 - H(X|Y).$$

- The capacity of the BSC(ϵ):

$$I\left[\text{BSC}(\epsilon)\right] = 1 - \mathcal{H}(\epsilon)$$

- The capacity of the BEC(ϵ) is given by

$$I\left[\text{BEC}(\epsilon)\right] = 1 - \epsilon$$

Assumption

In this presentation we will assume that the channel W under consideration is (input-output) symmetric.

- For a symmetric W, the capacity is given by

$$I(W) = H(X) - H(X|Y) = 1 - H(X|Y).$$

- The capacity of the BSC(ϵ):

$$I\big[\text{BSC}(\epsilon)\big] = 1 - \mathcal{H}(\epsilon)$$

- The capacity of the BEC(ϵ) is given by

$$I\big[\text{BEC}(\epsilon)\big] = 1 - \epsilon$$

Assumption

In this presentation we will assume that the channel W under consideration is (input-output) symmetric.

- For a symmetric W, the capacity is given by

$$I(W) = H(X) - H(X|Y) = 1 - H(X|Y).$$

- The capacity of the BSC(ϵ):

$$I\big[\text{BSC}(\epsilon)\big] = 1 - \mathcal{H}(\epsilon)$$

- The capacity of the BEC(ϵ) is given by

$$I\big[\text{BEC}(\epsilon)\big] = 1 - \epsilon$$

The main idea

- Channel coding problem trivial for two types of channels
  - Perfect: $I(W) = 1$
  - Useless: $I(W) = 0$
- Transform ordinary W into such extreme channels

The main idea

- Channel coding problem trivial for two types of channels
  - Perfect: $I(W) = 1$
  - Useless: $I(W) = 0$
- Transform ordinary W into such extreme channels

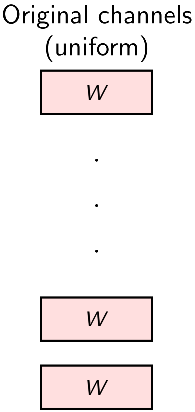
The main idea

- Channel coding problem trivial for two types of channels
  - Perfect: $I(W) = 1$
  - Useless: $I(W) = 0$
- Transform ordinary W into such extreme channels

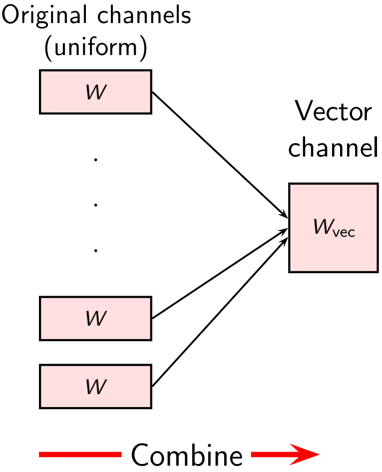
The main idea

- Channel coding problem trivial for two types of channels
  - Perfect: $I(W) = 1$
  - Useless: $I(W) = 0$
- Transform ordinary W into such extreme channels

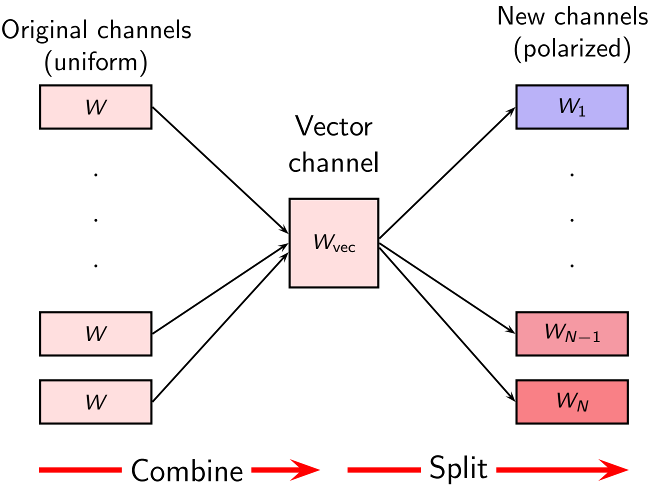
The method: aggregate and redistribute symmetric capacity



The method: aggregate and redistribute symmetric capacity



The method: aggregate and redistribute symmetric capacity



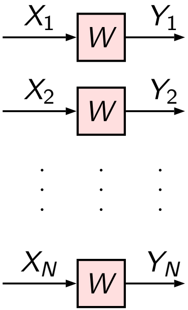
Combining

- Begin with N copies of W,
- use a 1-1 mapping

$$G_N : \{0,1\}^N \to \{0,1\}^N$$

- to create a vector channel

$$W_{\text{vec}} : U^N \to Y^N$$



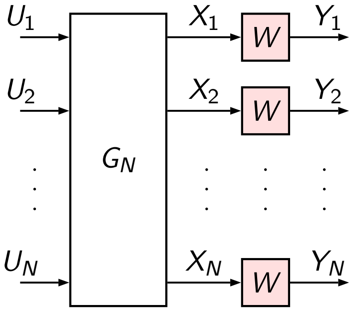
Combining

- Begin with N copies of W,
- use a 1-1 mapping

$$G_N : \{0,1\}^N \to \{0,1\}^N$$

- to create a vector channel

$$W_{\text{vec}} : U^N \to Y^N$$



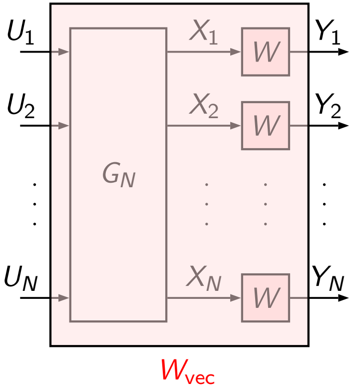
Combining

- Begin with N copies of W,
- use a 1-1 mapping

$$G_N : \{0,1\}^N \to \{0,1\}^N$$

- to create a vector channel

$$W_{\text{vec}} : U^N \to Y^N$$

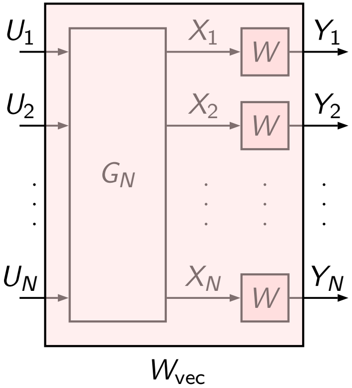


Conservation of symmetric capacity

Combining operation is lossless:

- Take $U_1, \ldots, U_N$ i.i.d. unif. $\{0,1\}$
- then, $X_1, \ldots, X_N$ i.i.d. unif. $\{0,1\}$
- and

$$\begin{aligned} I(W_{\text{vec}}) &= I(U^N; Y^N) \\ &= I(X^N; Y^N) \\ &= NI(W) \end{aligned}$$

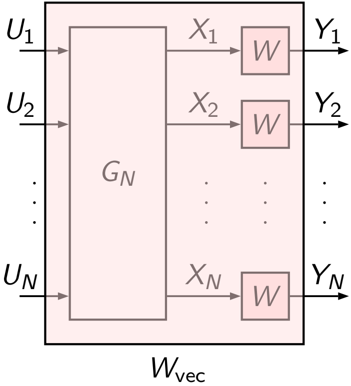


Conservation of symmetric capacity

Combining operation is lossless:

- Take $U_1, \ldots, U_N$ i.i.d. unif. $\{0,1\}$
- then, $X_1, \ldots, X_N$ i.i.d. unif. $\{0,1\}$
- and

$$\begin{aligned} I(W_{\text{vec}}) &= I(U^N; Y^N) \\ &= I(X^N; Y^N) \\ &= NI(W) \end{aligned}$$

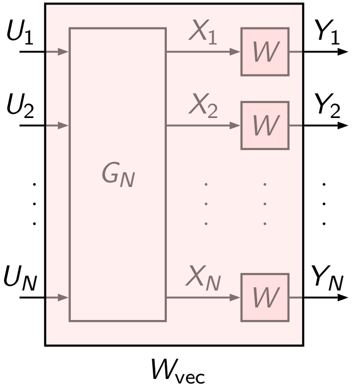


Conservation of symmetric capacity

Combining operation is lossless:

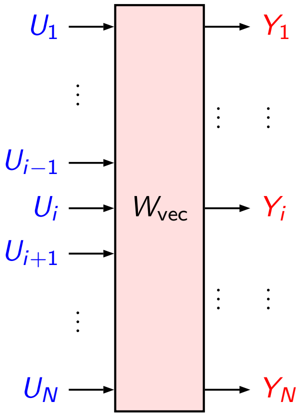
- Take $U_1, \dots, U_N$ i.i.d. unif. $\{0,1\}$
- then, $X_1, \dots, X_N$ i.i.d. unif. $\{0,1\}$
- and

$$
\begin{aligned}
I(W_{\text{vec}}) &= I(U^N; Y^N) \\
&= I(X^N; Y^N) \\
&= NI(W)
\end{aligned}
$$



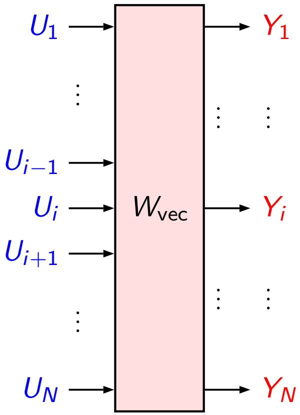
Splitting

$$I(W_{\text{vec}}) = I(U^N; Y^N)$$



Splitting

$$I(W_{\text{vec}}) = I(U^N; Y^N) = \sum_{i=1}^{N} I(U_i; Y^N, U^{i-1})$$

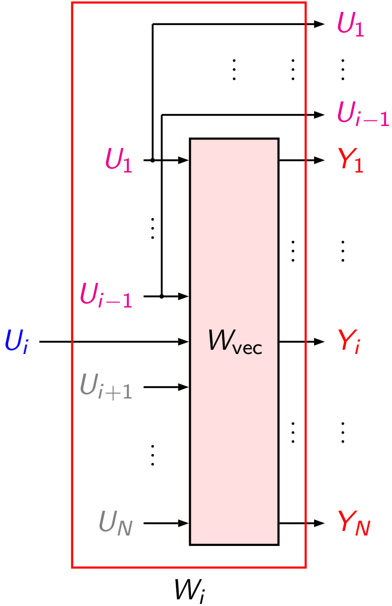


Splitting

$$I(W_{\text{vec}}) = I(U^N; Y^N)$$
$$= \sum_{i=1}^{N} I(U_i; Y^N, U^{i-1})$$

Define bit-channels

$$W_i : U_i \to (Y^N, U^{i-1})$$

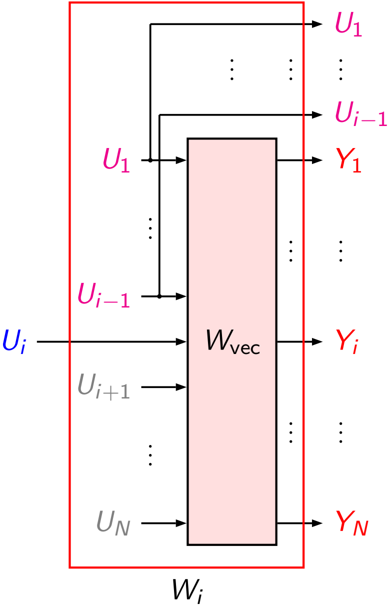


Splitting

$$I(W_{\text{vec}}) = I(U^N; Y^N)$$
$$= \sum_{i=1}^{N} I(U_i; Y^N, U^{i-1})$$
$$= \sum_{i=1}^{N} I(W_i)$$

Define bit-channels

$$W_i : U_i \to (Y^N, U^{i-1})$$

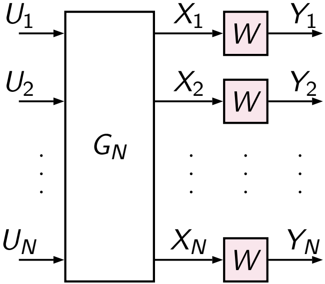


Polarization is commonplace

- Polarization is the rule not the exception
- A random permutation

$$G_N : \{0,1\}^N \to \{0,1\}^N$$

  is a good polarizer with high probability
- Equivalent to Shannon's random coding approach

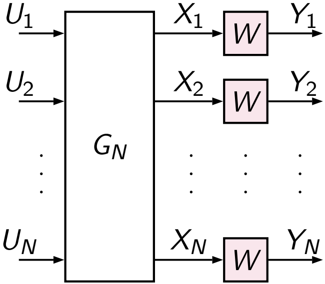


Polarization is commonplace

- Polarization is the rule not the exception
- A random permutation

$$G_N : \{0,1\}^N \to \{0,1\}^N$$

is a good polarizer with high probability
- Equivalent to Shannon's random coding approach



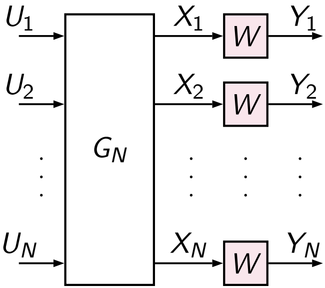
Polarization is commonplace

- Polarization is the rule not the exception
- A random permutation

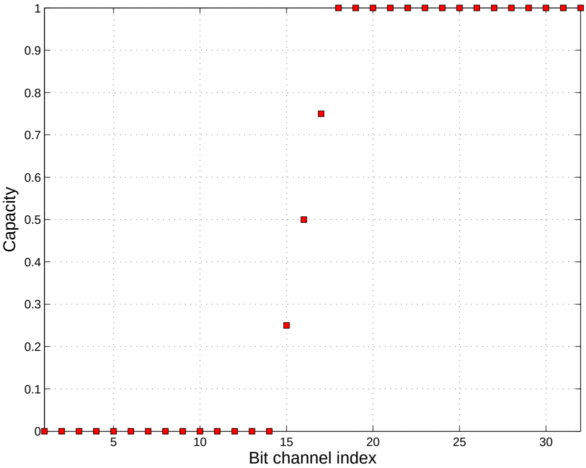
$$G_N : \{0,1\}^N \to \{0,1\}^N$$

is a good polarizer with high probability

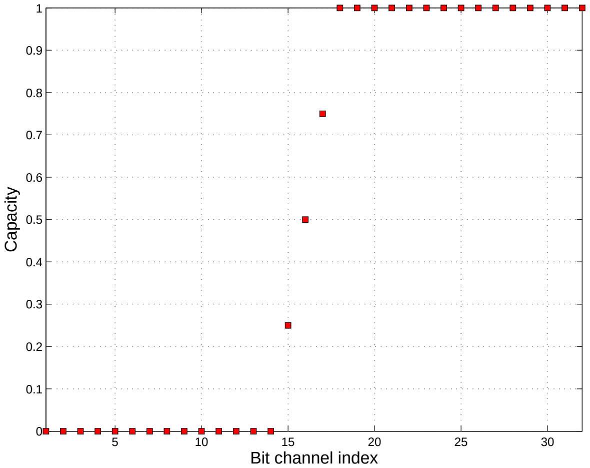
- Equivalent to Shannon's random coding approach



Random polarizers: stepwise, isotropic



Random polarizers: stepwise, isotropic



Isotropy: any redistribution order is as good as any other.

The complexity issue

- Random polarizers lack structure, too complex to implement
- Need a low-complexity polarizer
- May sacrifice stepwise, isotropic properties of random polarizers in return for less complexity

The complexity issue

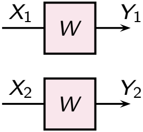
- Random polarizers lack structure, too complex to implement
- Need a low-complexity polarizer
- May sacrifice stepwise, isotropic properties of random polarizers in return for less complexity

The complexity issue

- Random polarizers lack structure, too complex to implement
- Need a low-complexity polarizer
- May sacrifice stepwise, isotropic properties of random polarizers in return for less complexity

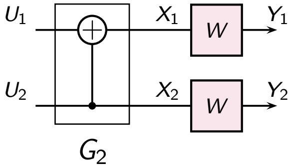
Basic module for a low-complexity scheme

Combine two copies of W



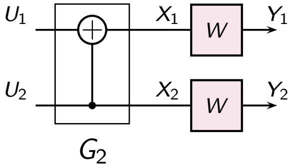
Basic module for a low-complexity scheme

Combine two copies of W



Basic module for a low-complexity scheme

Combine two copies of W

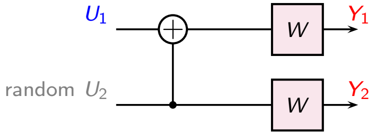


and split to create two bit-channels

$$W_1 : U_1 \to (Y_1, Y_2)$$
$$W_2 : U_2 \to (Y_1, Y_2, U_1)$$

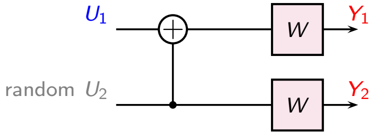
The first bit-channel W_1

$$W_1 : U_1 \to (Y_1, Y_2)$$



The first bit-channel W_1

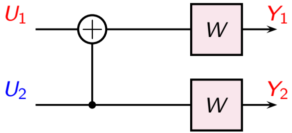
$$W_1 : U_1 \to (Y_1, Y_2)$$



$$I(W_1) = I(U_1; Y_1, Y_2)$$

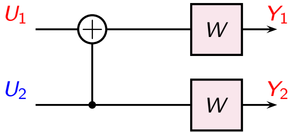
The second bit-channel W_2

$$W_2 : U_2 \to (Y_1, Y_2, U_1)$$



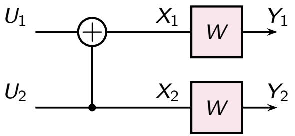
The second bit-channel W_2

$$W_2 : U_2 \to (Y_1, Y_2, U_1)$$



$$I(W_2) = I(U_2; Y_1, Y_2, U_1)$$

Symmetric capacity conserved but redistributed unevenly



- Conservation:

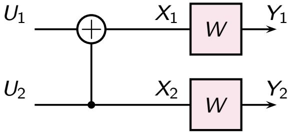
$$I(W_1) + I(W_2) = 2I(W)$$

- Extremization:

$$I(W_1) \leq I(W) \leq I(W_2)$$

with equality iff $I(W)$ equals 0 or 1.

Symmetric capacity conserved but redistributed unevenly



- Conservation:

$$I(W_1) + I(W_2) = 2I(W)$$

- Extremization:

$$I(W_1) \leq I(W) \leq I(W_2)$$

with equality iff $I(W)$ equals 0 or 1.

Recursive extension

- The basic polarization operation can be denoted as:

$$(W, W) \xrightarrow{\text{combine}} W_2 \xrightarrow{\text{split}} (W^-, W^+).$$

- The recursive extension will consist of the operations

$$(W^-, W^-) \longrightarrow \left(W^-\right)_2 \longrightarrow (W^{--}, W^{-+})$$

$$(W^+, W^+) \longrightarrow \left(W^+\right)_2 \longrightarrow (W^{+-}, W^{++})$$

where we wrote W^{--} for $\left(W^-\right)^-$, etc.

Recursive extension

- The basic polarization operation can be denoted as:

$$(W, W) \xrightarrow{\text{combine}} W_2 \xrightarrow{\text{split}} (W^-, W^+).$$

- The recursive extension will consist of the operations

$$\left(W^-, W^-\right) \longrightarrow \left(W^-\right)_2 \longrightarrow \left(W^{--}, W^{-+}\right)$$

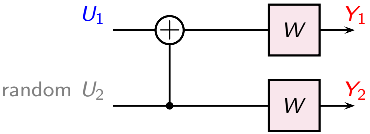
$$\left(W^+, W^+\right) \longrightarrow \left(W^+\right)_2 \longrightarrow \left(W^{+-}, W^{++}\right)$$

where we wrote W^{--} for $\left(W^-\right)^-$, etc.

Characterization of the *bad* channel W^-

The channel W^- is related to W by

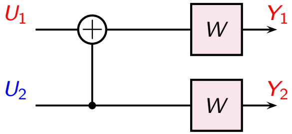
$$W^-(y_1, y_2|u_1) = \sum_{u_2} Q_{\text{unif}}(u_2) W_2(y_1, y_2|u_1, u_2)$$
$$= \sum_{u_2} \frac{1}{2} W(y_1|u_1 \oplus u_2) W(y_2|u_2)$$



Characterization of the *good* channel W^+

The channel W^+ is related to W by

$$\begin{aligned} W^+(y_1, y_2, u_1|u_2) &= P_{U_1|U_2}(u_1|u_2) W_2(y_1, y_2|u_1, u_2) \\ &= \frac{1}{2} W(y_1|u_1 \oplus u_2) W(y_2|u_2) \end{aligned}$$



Preservation of input-output symmetry

If W has input-output symmetry then W^- and W^+ each has input-output symmetry.

Specifically, if $W : \mathcal{X} \to \mathcal{Y}$ has symmetry with permutation $\pi : \mathcal{Y} \to \mathcal{Y}$, then

- $W^- : \mathcal{X} \to \mathcal{Y}^2$ is symmetric with

$$\pi^-(y_1, y_2) = \pi(y_1)\pi(y_2)$$

- $W^+ : \mathcal{X} \to \mathcal{Y}^2 \times \mathcal{X}$ is symmetric with

$$\pi^+(y_1, y_2, u_1) = \pi(y_1)\pi(y_2)(u_1 \oplus 1)$$

Preservation of input-output symmetry

If W has input-output symmetry then W^- and W^+ each has input-output symmetry.

Specifically, if $W : \mathcal{X} \to \mathcal{Y}$ has symmetry with permutation $\pi : \mathcal{Y} \to \mathcal{Y}$, then

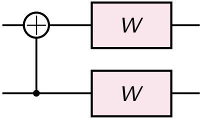
- $W^- : \mathcal{X} \to \mathcal{Y}^2$ is symmetric with

$$\pi^-(y_1, y_2) = \pi(y_1)\pi(y_2)$$

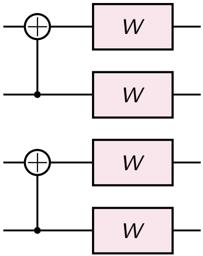
- $W^+ : \mathcal{X} \to \mathcal{Y}^2 \times \mathcal{X}$ is symmetric with

$$\pi^+(y_1, y_2, u_1) = \pi(y_1)\pi(y_2)(u_1 \oplus 1)$$

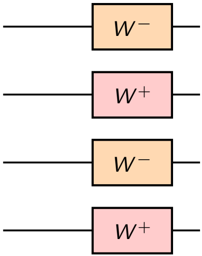
For the size-4 construction



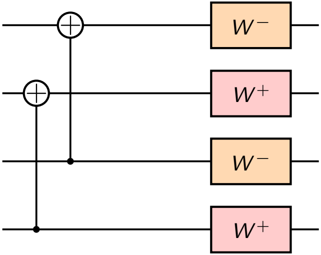
… duplicate the basic transform



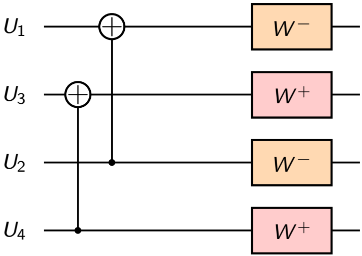
… obtain a pair of W^- and W^+ each



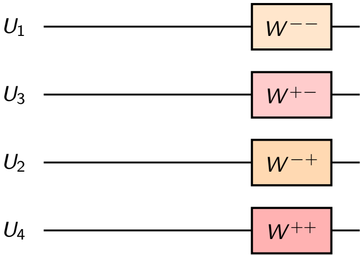
… apply basic transform on each pair



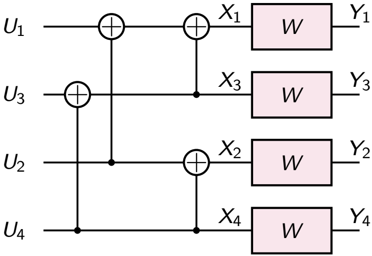
… decode in the indicated order



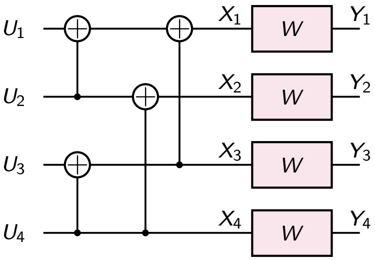
... obtain the four new bit-channels



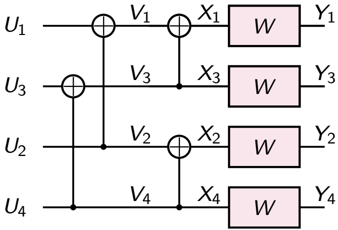
Overall size-4 construction



“Rewire” for standard-form size-4 construction



The first bit channel



Proposition

The first bit channel

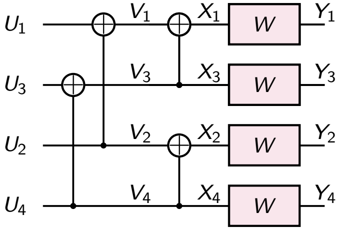
$$W_1 : U_1 \to Y_1^4$$

is equivalent to W^{--}.

Proof that $W_1 = W^{--}$

$$
\begin{aligned}
W_1(y_1^4|u_1) &= \sum_{u_2^4} P(y_1^4, u_2^4|u_1) = \sum_{u_2^4} P(u_2^4|u_1)P(y_1^4|u_1^4) \\
&= \sum_{u_2^4} \frac{1}{8} P(y_1^4|u_1 \oplus u_2, u_3 \oplus u_4, u_2, u_4) \\
&= \sum_{u_2, v_3, v_4} \frac{1}{8} P(y_1^4|u_1 \oplus u_2, v_3, u_2, v_4) \\
&= \sum_{u_2, v_3, v_4} \frac{1}{8} P(y_1, y_3|u_1 \oplus u_2, v_3) P(y_2, y_4|u_2, v_4) \\
&= \sum_{u_2} \frac{1}{2} \left(\sum_{v_3} \frac{1}{2} P(y_1, y_3|u_1 \oplus u_2, v_3) \right) \left(\sum_{v_4} \frac{1}{2} P(y_2, y_4|u_2, v_4) \right) \\
&= \sum_{u_2} \frac{1}{2} W^-(y_1, y_3|u_1 \oplus u_2) W^-(y_2, y_4|u_2) \\
&= \left(W^-\right)^-(y_1^4|u_1) = W^{--}(y_1^4|u_1).
\end{aligned}
$$

The second bit channel



Proposition

The second bit channel

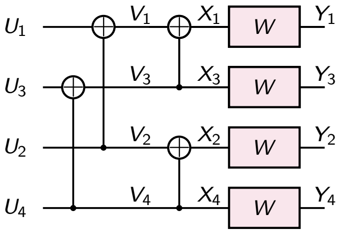
$$W_2 : U_2 \to (Y_1^4, U_1)$$

is equivalent to W^{-+}.

Proof that $W_2 = W^{-+}$

$$
\begin{aligned}
W_2(y_1^4, u_1|u_2) &= \sum_{u_3^4} P(y_1^4, u_1, u_3^4|u_2) = \sum_{u_3^4} \frac{1}{8} P(y_1^4|u_1^4) \\
&= \sum_{u_3^4} \frac{1}{8} P(y_1^4|u_1 \oplus u_2, u_3 \oplus u_4, u_2, u_4) \\
&= \sum_{v_3^4} \frac{1}{8} P(y_1^4|u_1 \oplus u_2, v_3, u_2, v_4) \\
&= \sum_{v_3^4} \frac{1}{8} P(y_1, y_3|u_1 \oplus u_2, v_3) P(y_2, y_4|u_2, v_4) \\
&= \frac{1}{2} \left(\sum_{v_3} \frac{1}{2} P(y_1, y_3|u_1 \oplus u_2, v_3) \right) \left(\sum_{v_4} \frac{1}{2} P(y_2, y_4|u_2, v_4) \right) \\
&= \frac{1}{2} W^{-}(y_1, y_3|u_1 \oplus u_2) W^{-}(y_2, y_4|u_2) \\
&= \left(W^{-}\right)^{+}(y_1^4, u_1|u_2) = W^{-+}(y_1^4, u_1|u_2).
\end{aligned}
$$

The third bit channel



Proposition

The third bit channel

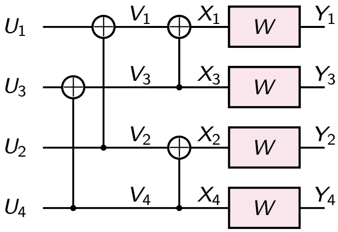
$$W_3 : U_3 \to (Y_1^4, U_1^2)$$

is equivalent to W^{+-}.

Proof that $W_3 = W^{+-}$

$$\begin{aligned}
W_3(y_1^4, u_1^2|u_3) &= \sum_{u_4} P(y_1^4, u_1^2, u_4|u_3) = \sum_{u_4} \frac{1}{8} P(y_1^4|u_1^4) \\
&= \sum_{u_4} \frac{1}{8} P(y_1^4|u_1 \oplus u_2, u_3 \oplus u_4, u_2, u_4) \\
&= \sum_{v_4} \frac{1}{8} P(y_1^4|v_1, v_3, v_2, v_4) \\
&= \sum_{v_4} \frac{1}{2} P(y_1, y_3, v_1|v_3) P(y_2, y_4, v_2|v_4) \\
&= \sum_{v_4} \frac{1}{2} W^+(y_1, y_3, v_1|v_3) W^+(y_2, y_4, v_2|v_4) \\
&= \sum_{u_4} \frac{1}{2} W^+(y_1, y_3, v_1|u_3 \oplus u_4) W^+(y_2, y_4, v_2|u_4) \\
&= \left(W^+\right)^-(y_1^4, v_1^2|u_3) = \left(W^+\right)^-(y_1^4, u_1^2|u_3) \\
&= W^{+-}(y_1^4, u_1^2|u_3).
\end{aligned}$$

The fourth bit channel



Proposition

The fourth bit channel

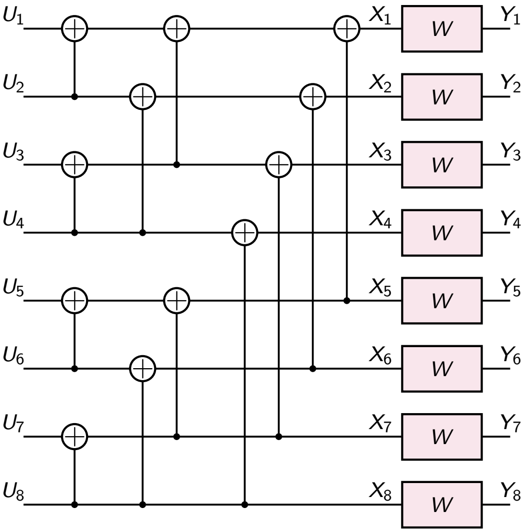
$$W_4 : U_4 \to (Y_1^4, U_1^3)$$

is equivalent to W^{++}.

Proof that $W_4 = W^{++}$

$$
\begin{aligned}
W_4(y_1^4, u_1^3 | u_4) &= \frac{1}{8} P(y_1^4 | u_4) \\
&= \frac{1}{8} P(y_1^4 | u_1 \oplus u_2, u_3 \oplus u_4, u_2, u_4) \\
&= \frac{1}{8} P(y_1^4 | v_1, v_3, v_2, v_4) \\
&= \frac{1}{2} P(y_1, y_3, v_1 | v_3) P(y_2, y_4, v_2 | v_4) \\
&= \frac{1}{2} W^+(y_1, y_3, v_1 | v_3) W^+(y_2, y_4, v_2 | v_4) \\
&= \frac{1}{2} W^+(y_1, y_3, v_1 | u_3 \oplus u_4) W^+(y_2, y_4, v_2 | u_4) \\
&= \left(W^+\right)^+(y_1^4, v_1^2, u_3 | u_4) \\
&= \left(W^+\right)^+(y_1^4, u_1^3 | u_3) \\
&= W^{++}(y_1^4, u_1^3 | u_4).
\end{aligned}
$$

Size-8 construction



Polarization of a BEC W

Polarization is easy to analyze when W is a BEC.

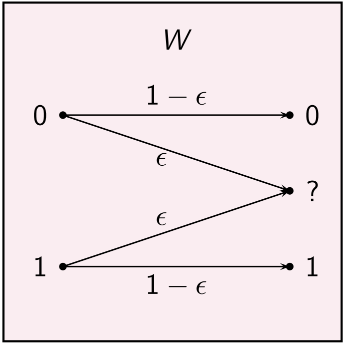
If W is a BEC(ϵ), then so are W^- and W^+, with erasure probabilities

$$\epsilon^- \triangleq 2\epsilon - \epsilon^2$$

and

$$\epsilon^+ \triangleq \epsilon^2$$

respectively.



The first bit channel W^-

The first bit channel W^- is a BEC.

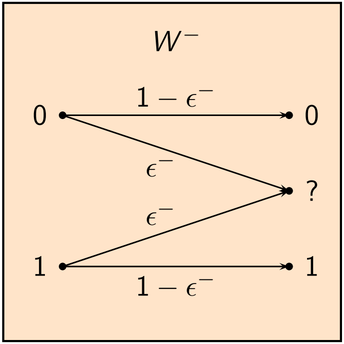
If W is a BEC(ϵ), then so are W^- and W^+, with erasure probabilities

$$\epsilon^- \triangleq 2\epsilon - \epsilon^2$$

and

$$\epsilon^+ \triangleq \epsilon^2$$

respectively.



The second bit channel W^+

The second bit channel W^+ is a BEC.

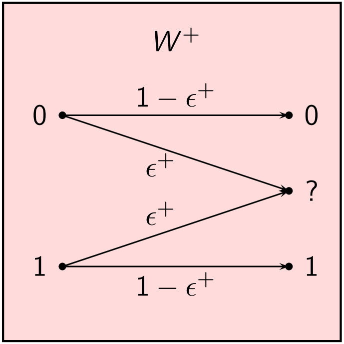
If W is a BEC(ϵ), then so are W^- and W^+, with erasure probabilities

$$\epsilon^- \triangleq 2\epsilon - \epsilon^2$$

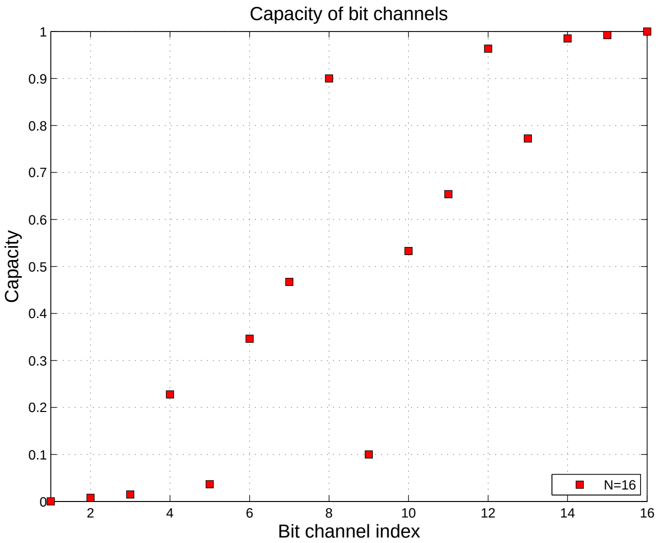
and

$$\epsilon^+ \triangleq \epsilon^2$$

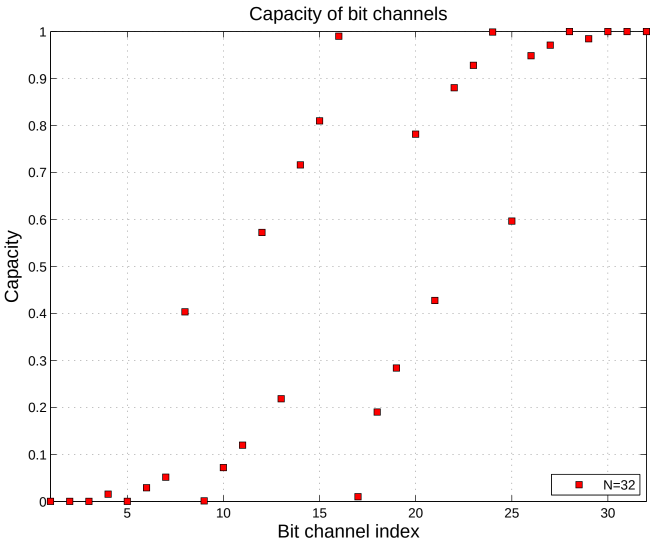
respectively.



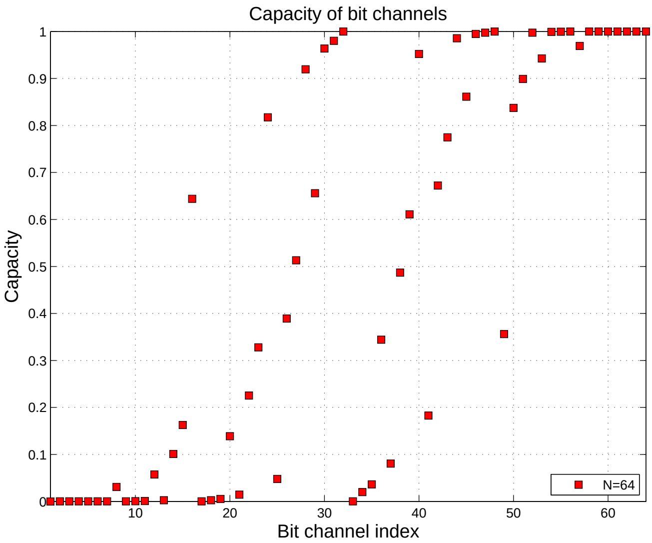
Polarization for BEC($\frac{1}{2}$): $N = 16$



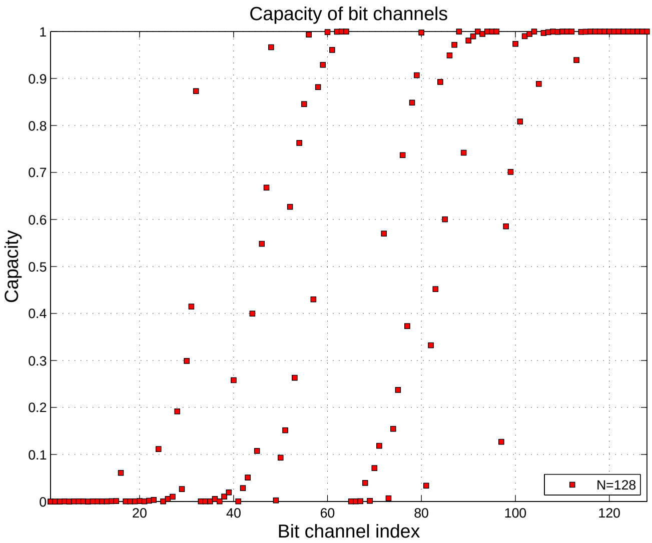
Polarization for BEC($\frac{1}{2}$): $N = 32$



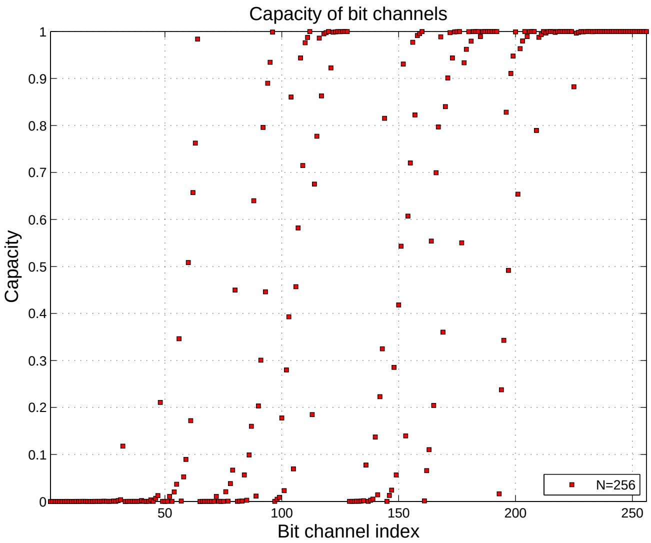
Polarization for BEC($\frac{1}{2}$): $N = 64$



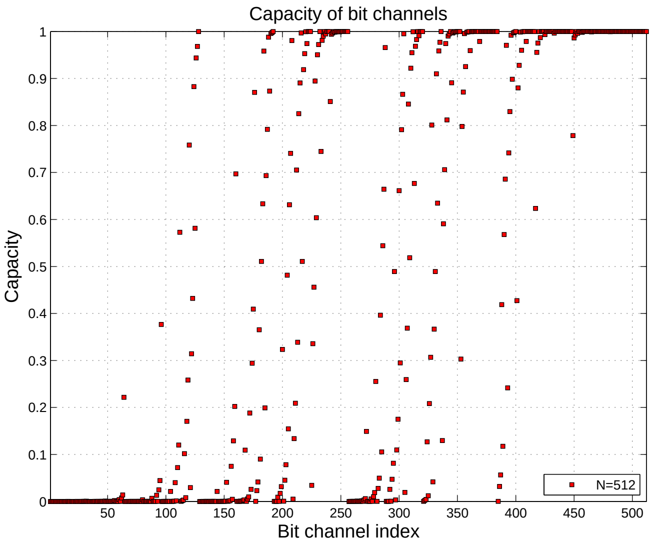
Polarization for BEC($\frac{1}{2}$): $N = 128$



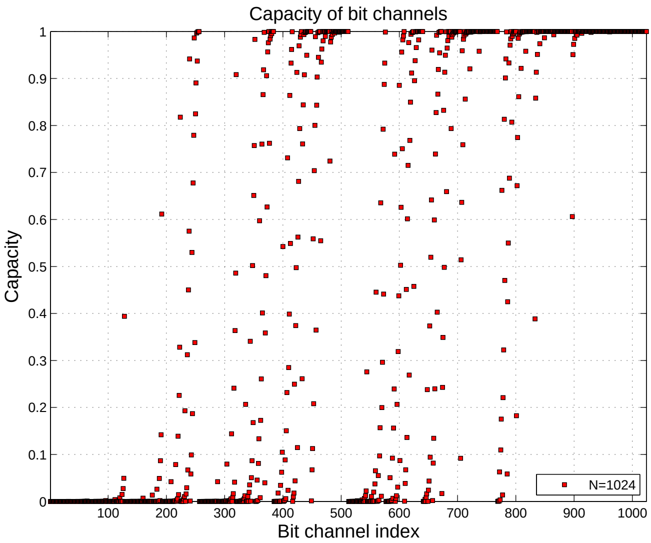
Polarization for BEC($\frac{1}{2}$): $N = 256$



Polarization for BEC($\frac{1}{2}$): $N = 512$

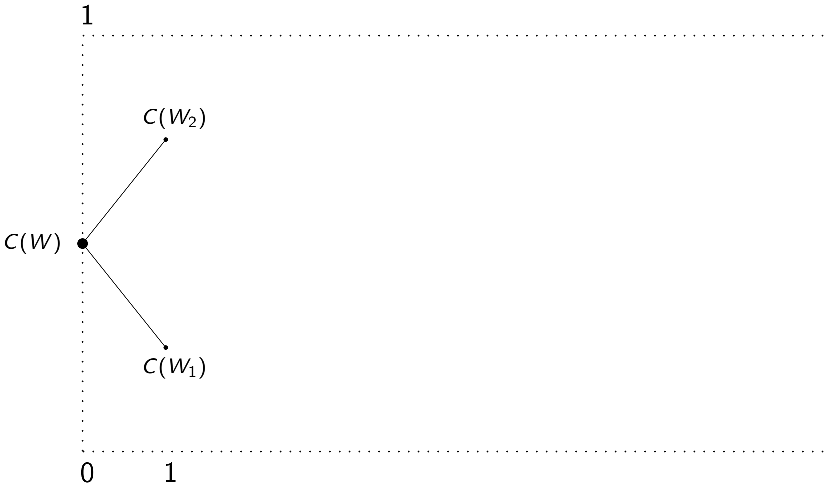


Polarization for BEC($\frac{1}{2}$): $N = 1024$

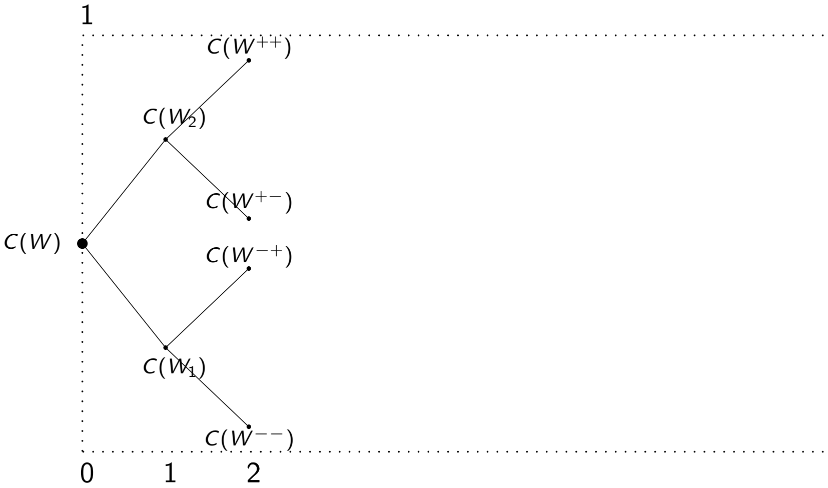


Polarization martingale for $W = \text{BEC}(\frac{1}{2})$

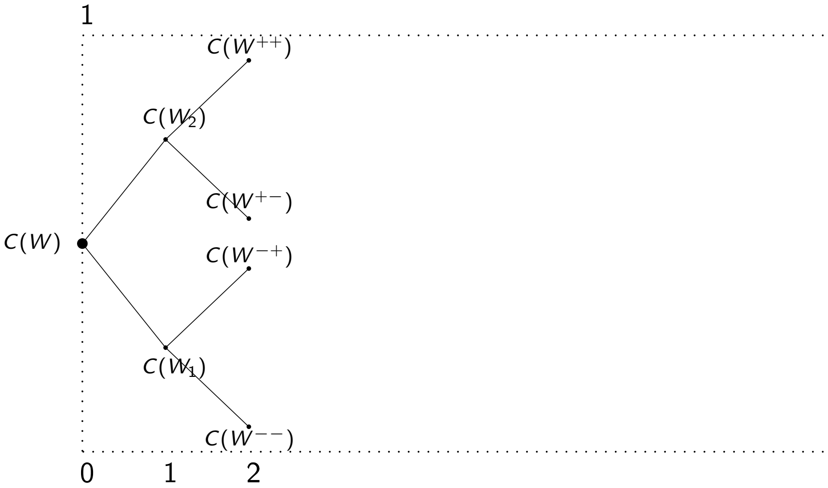
Polarization martingale for $W = \text{BEC}(\frac{1}{2})$



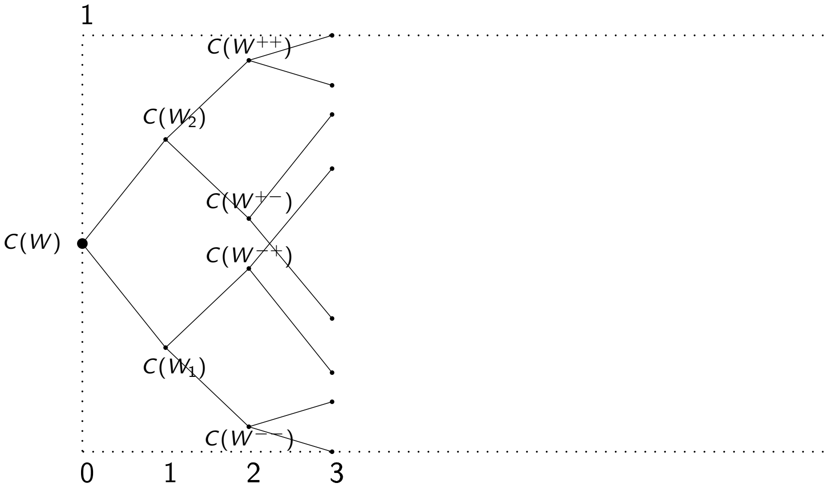
Polarization martingale for $W = \text{BEC}(\frac{1}{2})$



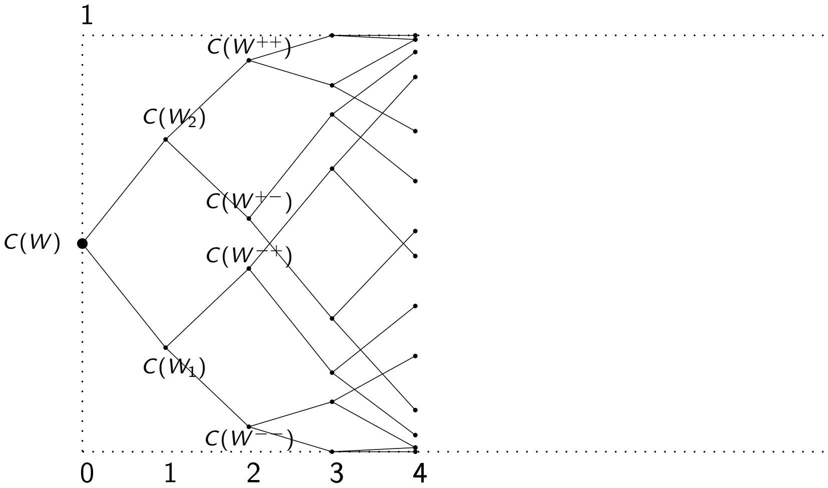
Polarization martingale for $W = \text{BEC}(\frac{1}{2})$



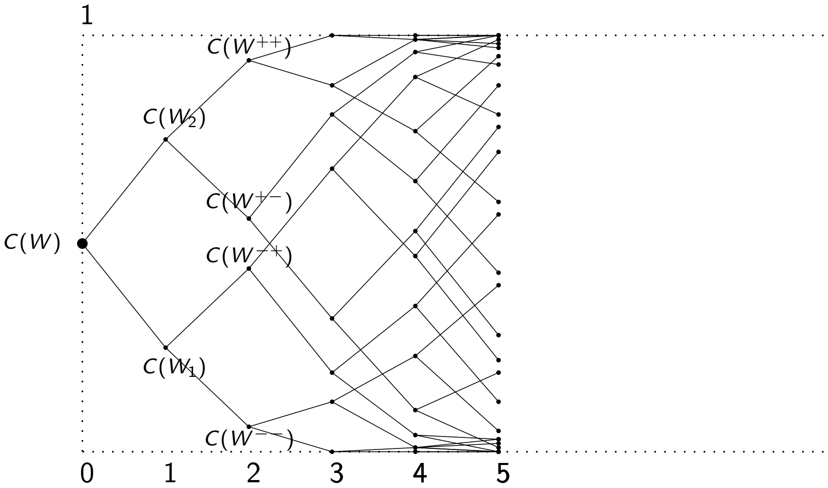
Polarization martingale for $W = \text{BEC}(\frac{1}{2})$



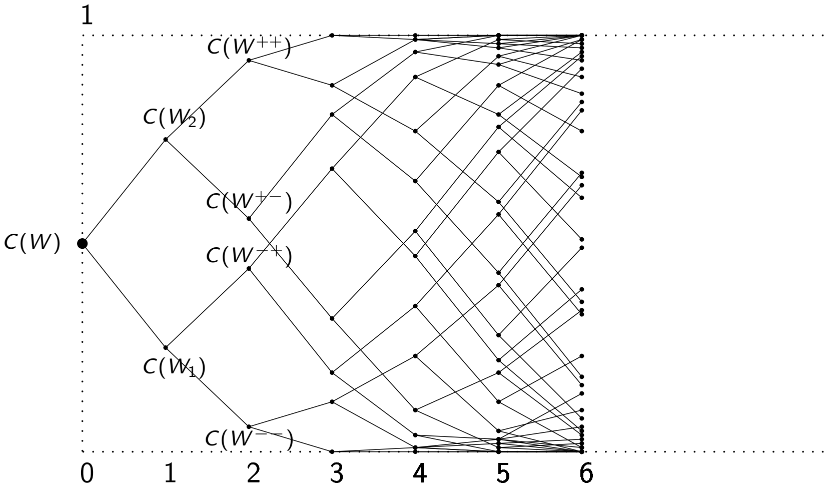
Polarization martingale for $W = \text{BEC}(\frac{1}{2})$



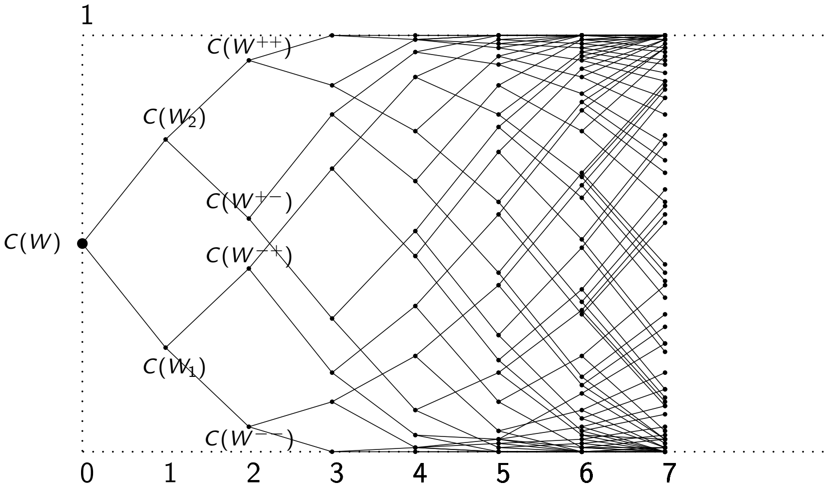
Polarization martingale for $W = \text{BEC}(\frac{1}{2})$



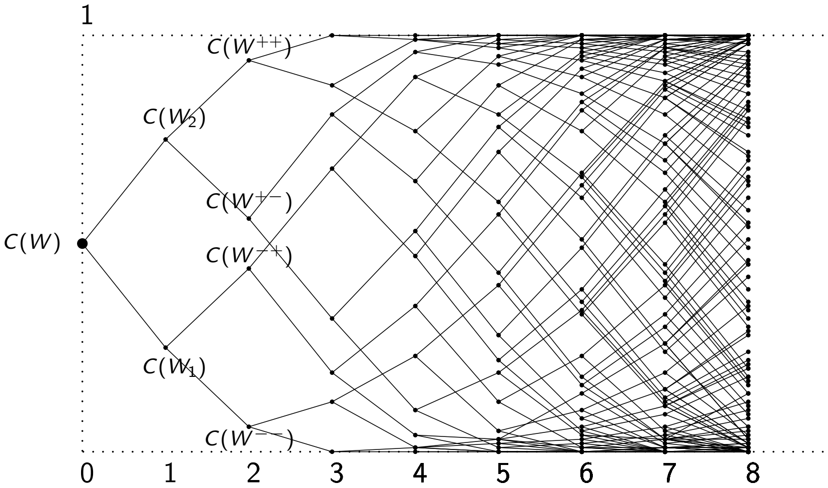
Polarization martingale for $W = \text{BEC}(\frac{1}{2})$



Polarization martingale for $W = \text{BEC}(\frac{1}{2})$



Polarization martingale for $W = \text{BEC}(\frac{1}{2})$



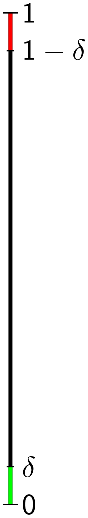
Theorem (Polarization, A. 2007)

The bit-channel capacities $\{I(W_i)\}$ *polarize: for any* $\delta \in (0,1)$*, as the construction size* N *grows*

$$\left[\frac{\textit{no. channels with } I(W_i) > 1-\delta}{N}\right] \longrightarrow I(W)$$

and

$$\left[\frac{\textit{no. channels with } I(W_i) < \delta}{N}\right] \longrightarrow 1 - I(W)$$



Theorem (Rate of polarization, A. and Telatar (2008))

Above theorem holds with $\delta = 2^{-N^{0.49}}$.

Theorem (Polarization, A. 2007)

The bit-channel capacities $\{I(W_i)\}$ *polarize: for any* $\delta \in (0,1)$*, as the construction size N grows*

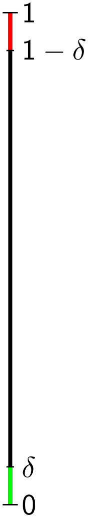
$$\left[\frac{\text{no. channels with } I(W_i) > 1-\delta}{N}\right] \longrightarrow I(W)$$

and

$$\left[\frac{\text{no. channels with } I(W_i) < \delta}{N}\right] \longrightarrow 1 - I(W)$$

Theorem (Rate of polarization, A. and Telatar (2008))

Above theorem holds with $\delta = 2^{-N^{0.49}}$*.*



1.1 Information theory review

1.2 Channel polarization

1.3 Polar coding

1.4 Performance

Section 1.3: Polar coding

- Objective: Introduce polar coding
- Topics
  - Code construction
  - Encoding
  - Decoding

Section 1.3: Polar coding

- Objective: Introduce polar coding
- Topics
  - Code construction
  - Encoding
  - Decoding

Section 1.3: Polar coding

- Objective: Introduce polar coding
- Topics
  - Code construction
  - Encoding
  - Decoding

Section 1.3: Polar coding

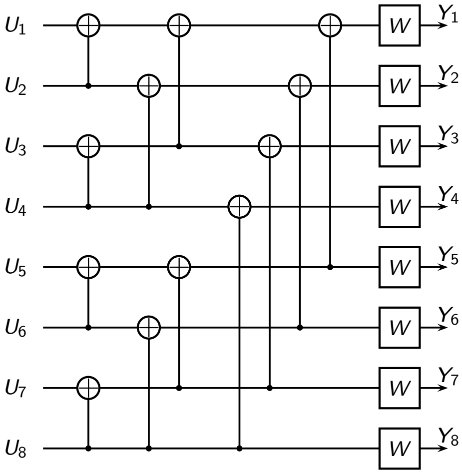
- Objective: Introduce polar coding
- Topics
  - Code construction
  - Encoding
  - Decoding

Section 1.3: Polar coding

- Objective: Introduce polar coding
- Topics
  - Code construction
  - Encoding
  - Decoding

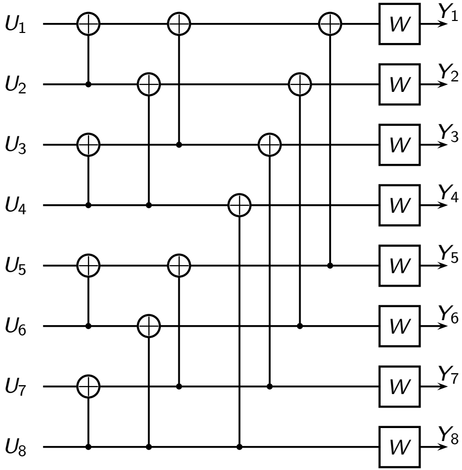
Polar code example: $W = \mathrm{BEC}(\frac{1}{2})$, $N = 8$, rate $1/2$

$I(W_i)$
0.0039
0.1211
0.1914
0.6836
0.3164
0.8086
0.8789
0.9961



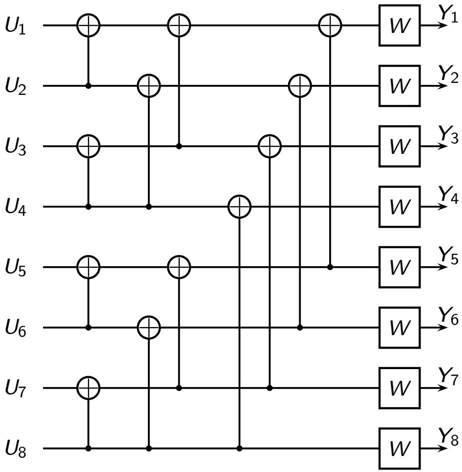
Polar code example: $W = \text{BEC}(\frac{1}{2})$, $N = 8$, rate $1/2$

$I(W_i)$	Rank
0.0039	8
0.1211	7
0.1914	6
0.6836	4
0.3164	5
0.8086	3
0.8789	2
0.9961	1



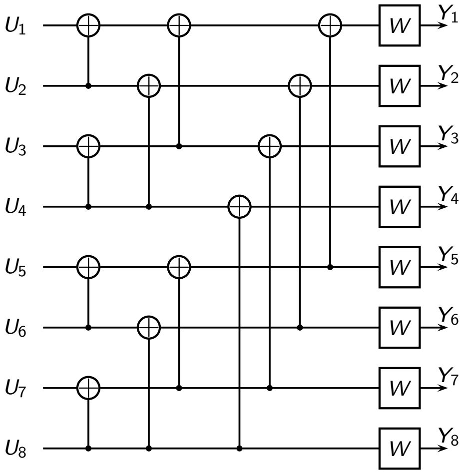
Polar code example: $W = \text{BEC}(\frac{1}{2})$, $N = 8$, rate $1/2$

$I(W_i)$	Rank	
0.0039	8	
0.1211	7	
0.1914	6	
0.6836	4	
0.3164	5	
0.8086	3	
0.8789	2	
0.9961	1	data



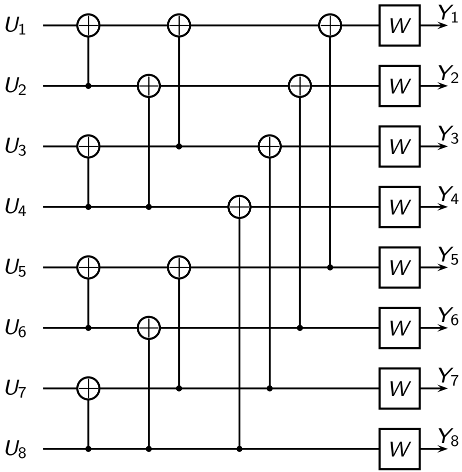
Polar code example: $W = \mathrm{BEC}(\frac{1}{2})$, $N = 8$, rate $1/2$

$I(W_i)$	Rank	
0.0039	8	
0.1211	7	
0.1914	6	
0.6836	4	
0.3164	5	
0.8086	3	
0.8789	2	data
0.9961	1	data



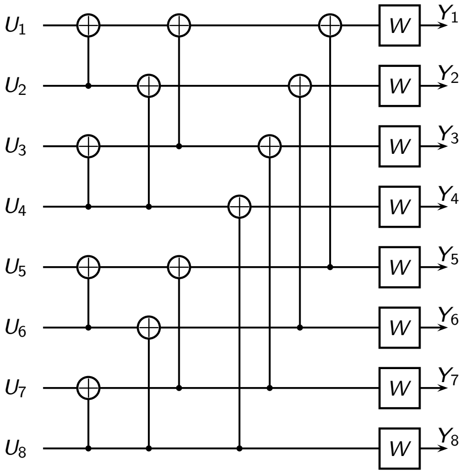
Polar code example: $W = \text{BEC}(\frac{1}{2})$, $N = 8$, rate $1/2$

$I(W_i)$	Rank	
0.0039	8	
0.1211	7	
0.1914	6	
0.6836	4	
0.3164	5	
0.8086	3	data
0.8789	2	data
0.9961	1	data



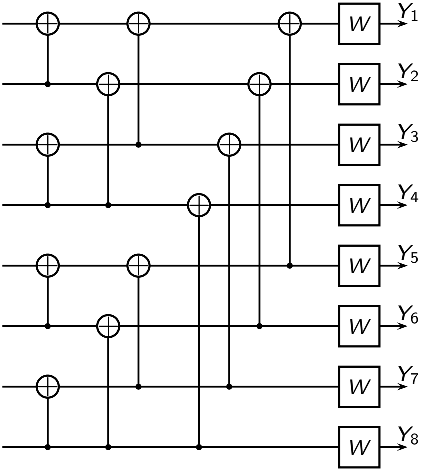
Polar code example: $W = \text{BEC}(\frac{1}{2})$, $N = 8$, rate $1/2$

$I(W_i)$	Rank	
0.0039	8	
0.1211	7	
0.1914	6	
0.6836	4	data
0.3164	5	
0.8086	3	data
0.8789	2	data
0.9961	1	data



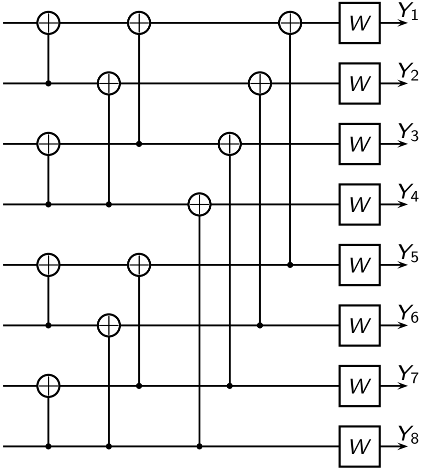
Polar code example: $W = \mathrm{BEC}(\frac{1}{2})$, $N = 8$, rate $1/2$

$I(W_i)$	Rank		
0.0039	8	frozen	U_1
0.1211	7	frozen	U_2
0.1914	6	frozen	U_3
0.6836	4	data	U_4
0.3164	5	frozen	U_5
0.8086	3	data	U_6
0.8789	2	data	U_7
0.9961	1	data	U_8



Polar code example: $W = \text{BEC}(\frac{1}{2})$, $N = 8$, rate $1/2$

$I(W_i)$	Rank		
0.0039	8	frozen	0
0.1211	7	frozen	0
0.1914	6	frozen	0
0.6836	4	data	U_4
0.3164	5	frozen	0
0.8086	3	data	U_6
0.8789	2	data	U_7
0.9961	1	data	U_8



Encoding complexity

Theorem

Encoding complexity for polar coding is $\mathcal{O}(N \log N)$.

Proof:

- Polar coding transform can be represented as a graph with $N[1 + \log(N)]$ variables.
- The graph has $(1 + \log(N))$ levels with N variables at each level.
- Computation begins at the source level and can be carried out level by level.
- Space complexity $O(N)$, time complexity $O(N \log N)$.

Encoding complexity

Theorem

Encoding complexity for polar coding is $\mathcal{O}(N \log N)$.

Proof:

- Polar coding transform can be represented as a graph with $N[1 + \log(N)]$ variables.
- The graph has $(1 + \log(N))$ levels with N variables at each level.
- Computation begins at the source level and can be carried out level by level.
- Space complexity $O(N)$, time complexity $O(N \log N)$.

Encoding complexity

Theorem

Encoding complexity for polar coding is $\mathcal{O}(N \log N)$.

Proof:

- Polar coding transform can be represented as a graph with $N[1 + \log(N)]$ variables.
- The graph has $(1 + \log(N))$ levels with N variables at each level.
- Computation begins at the source level and can be carried out level by level.
- Space complexity $O(N)$, time complexity $O(N \log N)$.

Encoding complexity

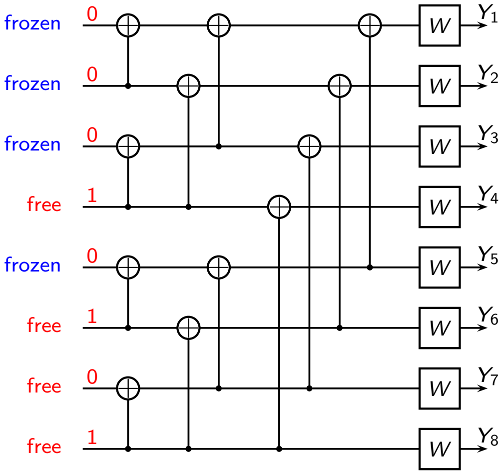
Theorem

Encoding complexity for polar coding is $\mathcal{O}(N \log N)$.

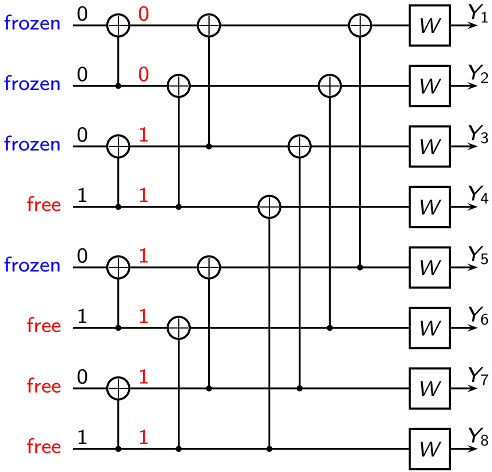
Proof:

- Polar coding transform can be represented as a graph with $N[1 + \log(N)]$ variables.
- The graph has $(1 + \log(N))$ levels with N variables at each level.
- Computation begins at the source level and can be carried out level by level.
- Space complexity $O(N)$, time complexity $O(N \log N)$.

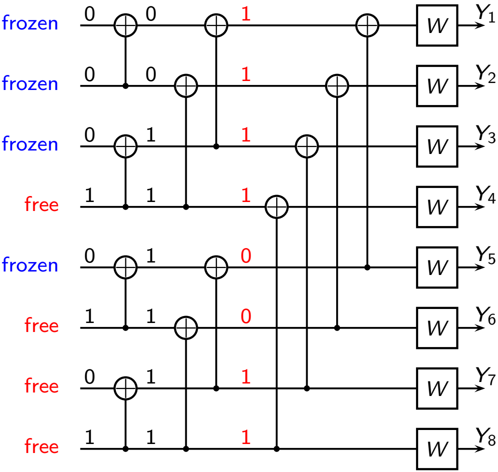
Encoding: an example



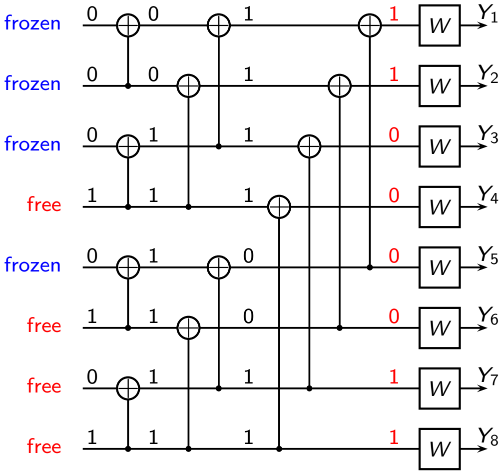
Encoding: an example



Encoding: an example



Encoding: an example



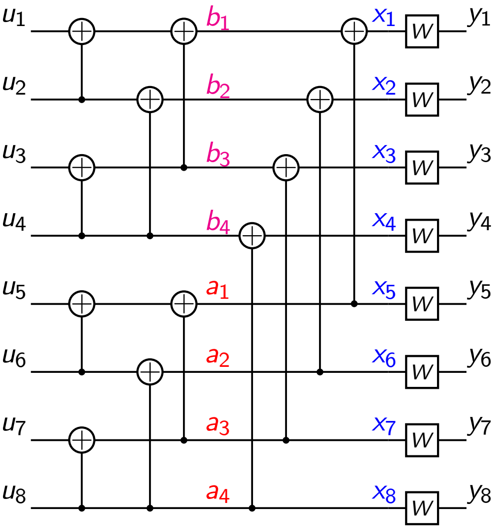
Successive Cancellation Decoding (SCD)

Theorem

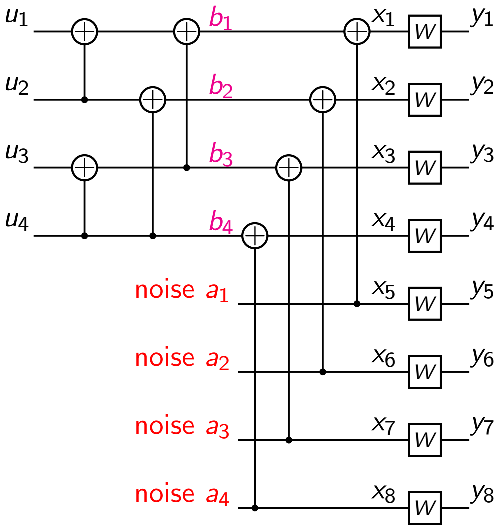
The complexity of successive cancellation decoding for polar codes is $\mathcal{O}(N \log N)$.

Proof: Given below.

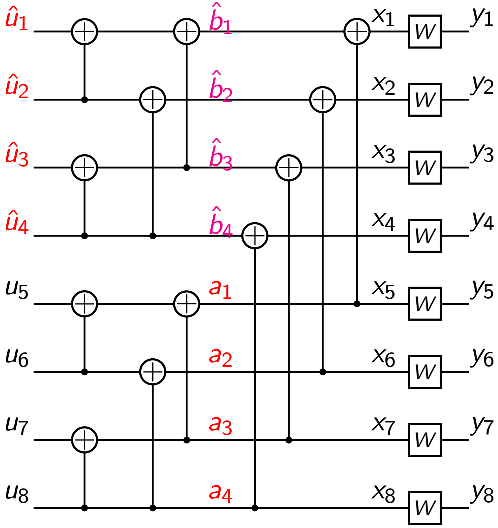
SCD: Exploit the $\mathbf{x} = |\mathbf{a}|\mathbf{a} + \mathbf{b}|$ structure



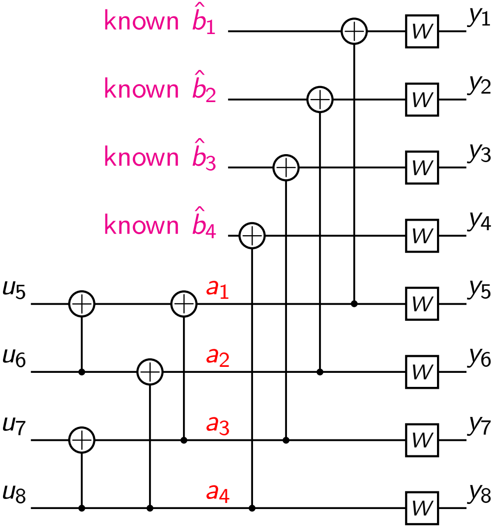
First phase: treat **a** as noise, decode (u_1, u_2, u_3, u_4)



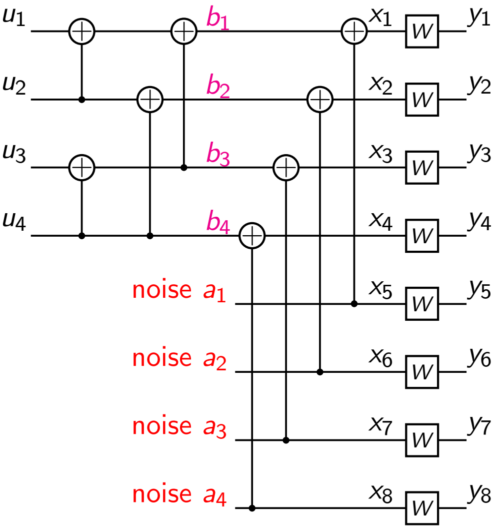
End of first phase



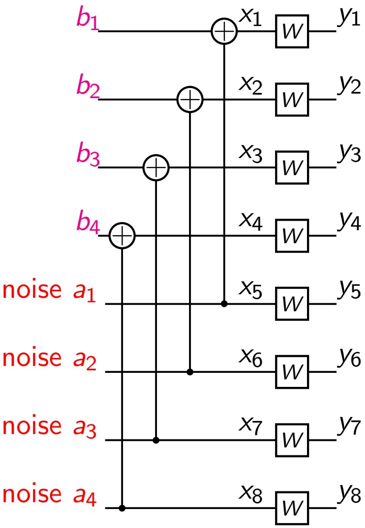
Second phase: Treat $\hat{\mathbf{b}}$ as known, decode (u_5, u_6, u_7, u_8)



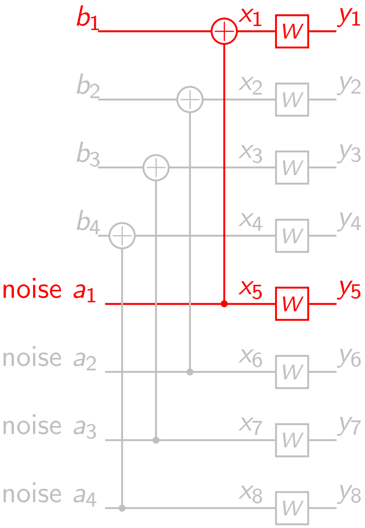
First phase in detail



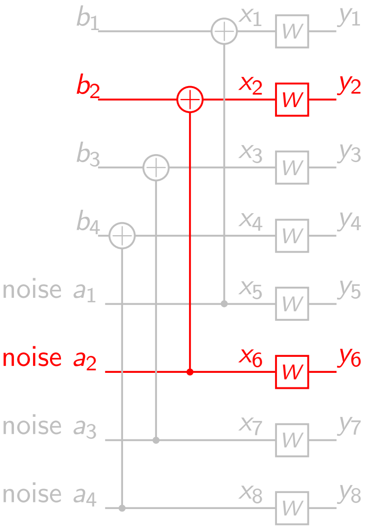
Equivalent channel model



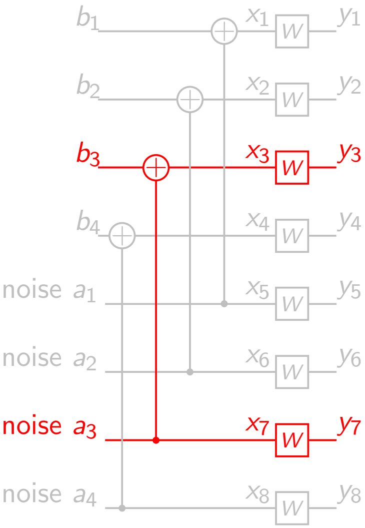
First copy of W^-



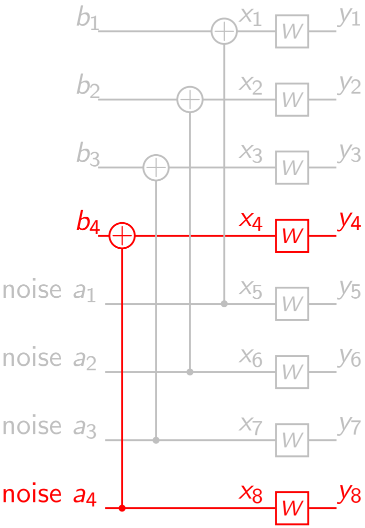
Second copy of W^-



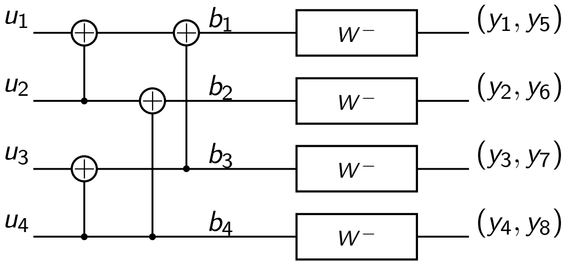
Third copy of W^-



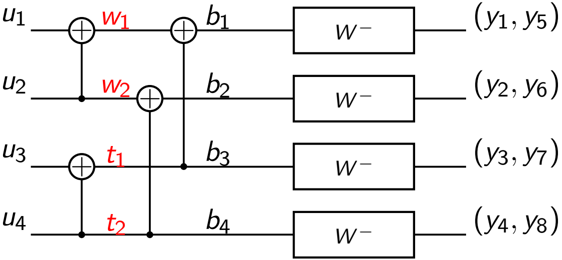
Fourth copy of W^-



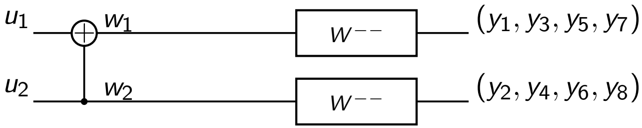
Decoding on W^-



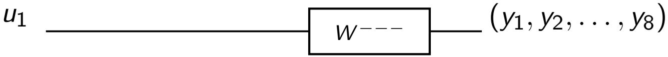
$\mathbf{b} = |\mathbf{t}|\mathbf{t} + \mathbf{w}|$



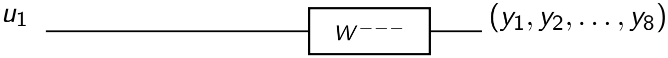
Decoding on W^{--}



Decoding on W^{---}



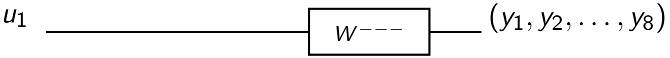
Decoding on W^{---}



Compute

$$L^{---} \triangleq \frac{W^{---}(y_1, \ldots, y_8 \mid u_1 = 0)}{W^{---}(y_1, \ldots, y_8 \mid u_1 = 1)}$$

Decoding on W^{---}



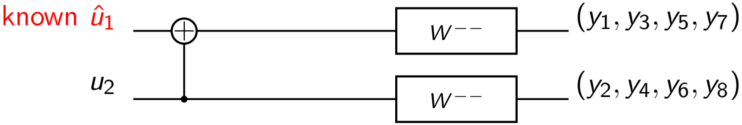
Compute

$$L^{---} \triangleq \frac{W^{---}(y_1, \ldots, y_8 \mid u_1 = 0)}{W^{---}(y_1, \ldots, y_8 \mid u_1 = 1)}$$

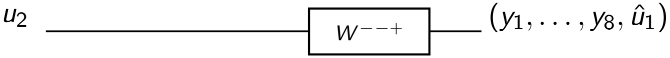
and set

$$\hat{u}_1 = \begin{cases} u_1 & \text{if } u_1 \text{ is frozen} \\ 0 & \text{else if } L^{---} > 0 \\ 1 & \text{else} \end{cases}$$

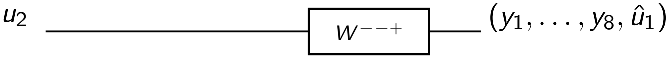
Decoding on W^{--+}



Decoding on W^{--+}



Decoding on W^{--+}



Compute

$$L^{--+} \triangleq \frac{W^{--+}(y_1, \ldots, y_8, \hat{u}_1 \mid u_2 = 0)}{W^{--+}(y_1, \ldots, y_8, \hat{u}_1 \mid u_2 = 1)}$$

and set

$$\hat{u}_2 = \begin{cases} u_2 & \text{if } u_2 \text{ is frozen} \\ 0 & \text{else if } L^{--+} > 0 \\ 1 & \text{else} \end{cases}$$

Complexity for successive cancelation decoding

- Let C_N be the complexity of decoding a code of length N
- Decoding problem of size N for W reduced to two decoding problems of size $N/2$ for W^- and W^+
- So
$$C_N = 2C_{N/2} + kN$$
for some constant k
- This gives $C_N = \mathcal{O}(N \log N)$

Complexity for successive cancelation decoding

- Let C_N be the complexity of decoding a code of length N
- Decoding problem of size N for W reduced to two decoding problems of size $N/2$ for W^- and W^+
- So

$$C_N = 2C_{N/2} + kN$$

  for some constant k
- This gives $C_N = \mathcal{O}(N \log N)$

Complexity for successive cancelation decoding

- Let C_N be the complexity of decoding a code of length N
- Decoding problem of size N for W reduced to two decoding problems of size $N/2$ for W^- and W^+
- So

$$C_N = 2C_{N/2} + kN$$

  for some constant k
- This gives $C_N = \mathcal{O}(N \log N)$

Complexity for successive cancelation decoding

- Let C_N be the complexity of decoding a code of length N
- Decoding problem of size N for W reduced to two decoding problems of size $N/2$ for W^- and W^+
- So

  $$C_N = 2C_{N/2} + kN$$

  for some constant k
- This gives $C_N = \mathcal{O}(N \log N)$

Performance of polar codes

Probability of Error (A. and Telatar (2008)

For any binary-input symmetric channel W, the probability of frame error for polar coding at rate $R < I(W)$ and using codes of length N is bounded as

$$P_e(N, R) \leq 2^{-N^{0.49}}$$

for sufficiently large N.

A more refined versions of this result has been given given by S. H. Hassani, R. Mori, T. Tanaka, and R. L. Urbanke (2011).

Construction complexity

Construction Complexity

Polar codes can be constructed in time $\mathcal{O}(N\text{poly}(log(N)))$.

This result has been developed in a sequence of papers by

- R. Mori and T. Tanaka (2009)
- I. Tal and A. Vardy (2011)
- R. Pedarsani, S. H. Hassani, I. Tal, and E. Telatar (2011)

Gaussian approximation

- Trifonov (2011) introduced a Gaussian approximation technique for constructing polar codes
- Dai *et al.* (2015) studied various refinements of Gaussian approximation for polar code construction
- These methods work extremely well although a satisfactory explanation of why they work is still missing

Gaussian approximation

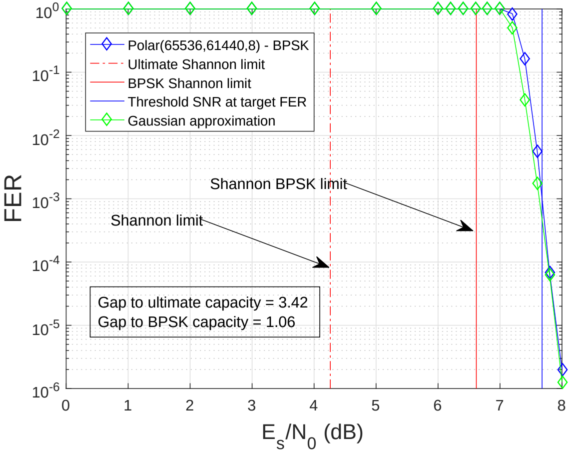
- Trifonov (2011) introduced a Gaussian approximation technique for constructing polar codes
- Dai *et al.* (2015) studied various refinements of Gaussian approximation for polar code construction
- These methods work extremely well although a satisfactory explanation of why they work is still missing

Gaussian approximation

- Trifonov (2011) introduced a Gaussian approximation technique for constructing polar codes
- Dai *et al.* (2015) studied various refinements of Gaussian approximation for polar code construction
- These methods work extremely well although a satisfactory explanation of why they work is still missing

Example of Gaussian approximation

Polar code construction and performance estimation by Gaussian approximation



Polar coding summary

Summary

Given W, $N = 2^n$, and $R < I(W)$, a polar code can be constructed such that it has

- construction complexity $\mathcal{O}(N\text{poly}(log(N)))$,
- encoding complexity $\approx N \log N$,
- successive-cancellation decoding complexity $\approx N \log N$,
- frame error probability $P_e(N, R) = \mathcal{O}(2^{-N^{0.49}})$.

Polar coding summary

Summary

Given W, $N = 2^n$, and $R < I(W)$, a polar code can be constructed such that it has

- construction complexity $\mathcal{O}(N\text{poly}(log(N)))$,
- encoding complexity $\approx N \log N$,
- successive-cancellation decoding complexity $\approx N \log N$,
- frame error probability $P_e(N, R) = \mathcal{O}(2^{-N^{0.49}})$.

Polar coding summary

Summary

Given W, $N = 2^n$, and $R < I(W)$, a polar code can be constructed such that it has

- construction complexity $\mathcal{O}(N\text{poly}(log(N)))$,
- encoding complexity $\approx N \log N$,
- successive-cancellation decoding complexity $\approx N \log N$,
- frame error probability $P_e(N, R) = \mathcal{O}(2^{-N^{0.49}})$.

Polar coding summary

Summary

Given W, $N = 2^n$, and $R < I(W)$, a polar code can be constructed such that it has

- construction complexity $\mathcal{O}(N\text{poly}(log(N)))$,
- encoding complexity $\approx N \log N$,
- successive-cancellation decoding complexity $\approx N \log N$,
- frame error probability $P_e(N, R) = \mathcal{O}(2^{-N^{0.49}})$.

Section 1.4: Polar coding performance

- Objective: Discuss the performance of polar coding and compare with state-of-the-art codes
- Topics
  - Performance of polar codes under various decoding algorithms
  - Comparisons with other codes
  - Implementation complexity
  - Concatenation schemes with polar codes

Section 1.4: Polar coding performance

- Objective: Discuss the performance of polar coding and compare with state-of-the-art codes
- Topics
  - Performance of polar codes under various decoding algorithms
  - Comparisons with other codes
  - Implementation complexity
  - Concatenation schemes with polar codes

Section 1.4: Polar coding performance

- Objective: Discuss the performance of polar coding and compare with state-of-the-art codes
- Topics
  - Performance of polar codes under various decoding algorithms
  - Comparisons with other codes
  - Implementation complexity
  - Concatenation schemes with polar codes

Section 1.4: Polar coding performance

- Objective: Discuss the performance of polar coding and compare with state-of-the-art codes
- Topics
  - Performance of polar codes under various decoding algorithms
  - Comparisons with other codes
  - Implementation complexity
  - Concatenation schemes with polar codes

Section 1.4: Polar coding performance

- Objective: Discuss the performance of polar coding and compare with state-of-the-art codes
- Topics
  - Performance of polar codes under various decoding algorithms
  - Comparisons with other codes
  - Implementation complexity
  - Concatenation schemes with polar codes

Section 1.4: Polar coding performance

- Objective: Discuss the performance of polar coding and compare with state-of-the-art codes
- Topics
  - Performance of polar codes under various decoding algorithms
  - Comparisons with other codes
  - Implementation complexity
  - Concatenation schemes with polar codes

Types of decoders for polar codes

- Maximum likelihood (ML)
- Successive cancellation (SC)
- Belief propagation (BP)
- List decoder
- List decoder with CRC
- Sphere-decoding

Types of decoders for polar codes

- Maximum likelihood (ML)
- Successive cancellation (SC)
- Belief propagation (BP)
- List decoder
- List decoder with CRC
- Sphere-decoding

Types of decoders for polar codes

- Maximum likelihood (ML)
- Successive cancellation (SC)
- Belief propagation (BP)
- List decoder
- List decoder with CRC
- Sphere-decoding

Types of decoders for polar codes

- Maximum likelihood (ML)
- Successive cancellation (SC)
- Belief propagation (BP)
- List decoder
- List decoder with CRC
- Sphere-decoding

Types of decoders for polar codes

- Maximum likelihood (ML)
- Successive cancellation (SC)
- Belief propagation (BP)
- List decoder
- List decoder with CRC
- Sphere-decoding

Types of decoders for polar codes

- Maximum likelihood (ML)
- Successive cancellation (SC)
- Belief propagation (BP)
- List decoder
- List decoder with CRC
- Sphere-decoding

Polar decoders as search heuristics

- Successive cancellation decoding: A depth-first search method with complexity roughly $N \log N$
  - Sufficient to achieve channel capacity
  - Not powerful enough to challenge LDPC and turbo codes in short to moderate lengths
- List decoding: A breadth-first search algorithm with limited branching (known as "beam search" in AI)
  - Introduced by Tal and Vardy (2011) based on a similar scheme for RM codes by Dumer and Shabunov (2000, 2002, 2006)
  - Sufficient to challenge the state-of-the-art at short to moderate lengths
  - Complexity grows as $O(LN \log N)$ for a list size L
- Sphere-decoding: "British Museum" search with branch and bound

Polar decoders as search heuristics

- Successive cancellation decoding: A depth-first search method with complexity roughly $N \log N$
  - Sufficient to achieve channel capacity
  - Not powerful enough to challenge LDPC and turbo codes in short to moderate lengths
- List decoding: A breadth-first search algorithm with limited branching (known as "beam search" in AI)
  - Introduced by Tal and Vardy (2011) based on a similar scheme for RM codes by Dumer and Shabunov (2000, 2002, 2006)
  - Sufficient to challenge the state-of-the-art at short to moderate lengths
  - Complexity grows as $O(LN \log N)$ for a list size L
- Sphere-decoding: "British Museum" search with branch and bound

Polar decoders as search heuristics

- Successive cancellation decoding: A depth-first search method with complexity roughly $N \log N$
  - Sufficient to achieve channel capacity
  - Not powerful enough to challenge LDPC and turbo codes in short to moderate lengths
- List decoding: A breadth-first search algorithm with limited branching (known as "beam search" in AI)
  - Introduced by Tal and Vardy (2011) based on a similar scheme for RM codes by Dumer and Shabunov (2000, 2002, 2006)
  - Sufficient to challenge the state-of-the-art at short to moderate lengths
  - Complexity grows as $O(LN \log N)$ for a list size L
- Sphere-decoding: "British Museum" search with branch and bound

Polar decoders as search heuristics

- Successive cancellation decoding: A depth-first search method with complexity roughly $N \log N$
  - Sufficient to achieve channel capacity
  - Not powerful enough to challenge LDPC and turbo codes in short to moderate lengths
- List decoding: A breadth-first search algorithm with limited branching (known as "beam search" in AI)
  - Introduced by Tal and Vardy (2011) based on a similar scheme for RM codes by Dumer and Shabunov (2000, 2002, 2006)
  - Sufficient to challenge the state-of-the-art at short to moderate lengths
  - Complexity grows as $O(LN \log N)$ for a list size L
- Sphere-decoding: "British Museum" search with branch and bound

Polar decoders as search heuristics

- Successive cancellation decoding: A depth-first search method with complexity roughly $N \log N$
  - Sufficient to achieve channel capacity
  - Not powerful enough to challenge LDPC and turbo codes in short to moderate lengths
- List decoding: A breadth-first search algorithm with limited branching (known as "beam search" in AI)
  - Introduced by Tal and Vardy (2011) based on a similar scheme for RM codes by Dumer and Shabunov (2000, 2002, 2006)
  - Sufficient to challenge the state-of-the-art at short to moderate lengths
  - Complexity grows as $O(LN \log N)$ for a list size L
- Sphere-decoding: "British Museum" search with branch and bound

Polar decoders as search heuristics

- Successive cancellation decoding: A depth-first search method with complexity roughly $N \log N$
  - Sufficient to achieve channel capacity
  - Not powerful enough to challenge LDPC and turbo codes in short to moderate lengths
- List decoding: A breadth-first search algorithm with limited branching (known as "beam search" in AI)
  - Introduced by Tal and Vardy (2011) based on a similar scheme for RM codes by Dumer and Shabunov (2000, 2002, 2006)
  - Sufficient to challenge the state-of-the-art at short to moderate lengths
  - Complexity grows as $O(LN \log N)$ for a list size L
- Sphere-decoding: "British Museum" search with branch and bound

Polar decoders as search heuristics

- Successive cancellation decoding: A depth-first search method with complexity roughly $N \log N$
  - Sufficient to achieve channel capacity
  - Not powerful enough to challenge LDPC and turbo codes in short to moderate lengths
- List decoding: A breadth-first search algorithm with limited branching (known as "beam search" in AI)
  - Introduced by Tal and Vardy (2011) based on a similar scheme for RM codes by Dumer and Shabunov (2000, 2002, 2006)
  - Sufficient to challenge the state-of-the-art at short to moderate lengths
  - Complexity grows as $O(LN \log N)$ for a list size L
- Sphere-decoding: "British Museum" search with branch and bound

Polar decoders as search heuristics

- Successive cancellation decoding: A depth-first search method with complexity roughly $N \log N$
  - Sufficient to achieve channel capacity
  - Not powerful enough to challenge LDPC and turbo codes in short to moderate lengths
- List decoding: A breadth-first search algorithm with limited branching (known as “beam search” in AI)
  - Introduced by Tal and Vardy (2011) based on a similar scheme for RM codes by Dumer and Shabunov (2000, 2002, 2006)
  - Sufficient to challenge the state-of-the-art at short to moderate lengths
  - Complexity grows as $O(LN \log N)$ for a list size L
- Sphere-decoding: “British Museum” search with branch and bound

List decoder for polar codes

- First produce L candidate decisions
- Pick the most likely word from the list
- In the CRC version, first discard the candidates that do not satisfy the CRC
- Complexity $\mathcal{O}(LN \log N)$

List decoder for polar codes

- First produce L candidate decisions
- Pick the most likely word from the list
- In the CRC version, first discard the candidates that do not satisfy the CRC
- Complexity $\mathcal{O}(LN \log N)$

List decoder for polar codes

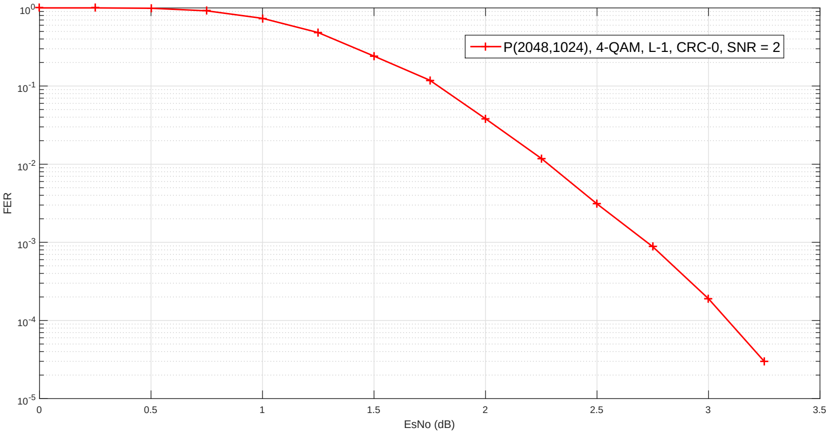
- First produce L candidate decisions
- Pick the most likely word from the list
- In the CRC version, first discard the candidates that do not satisfy the CRC
- Complexity $\mathcal{O}(LN \log N)$

List decoder for polar codes

- First produce L candidate decisions
- Pick the most likely word from the list
- In the CRC version, first discard the candidates that do not satisfy the CRC
- Complexity $\mathcal{O}(LN \log N)$

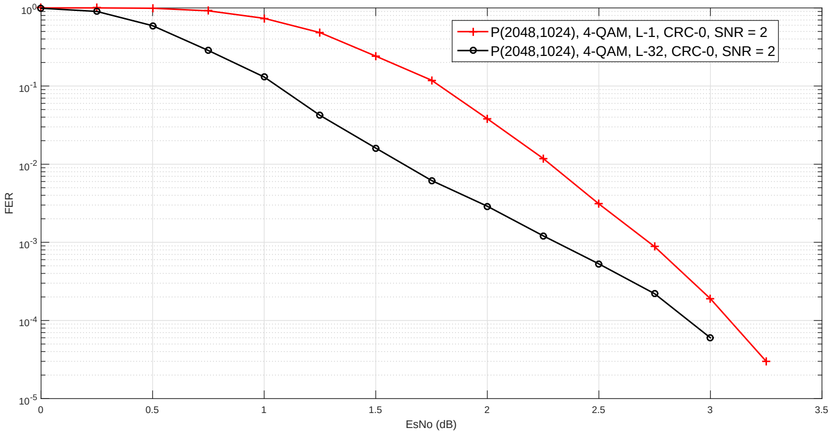
Polar code performance

Successive cancellation decoder



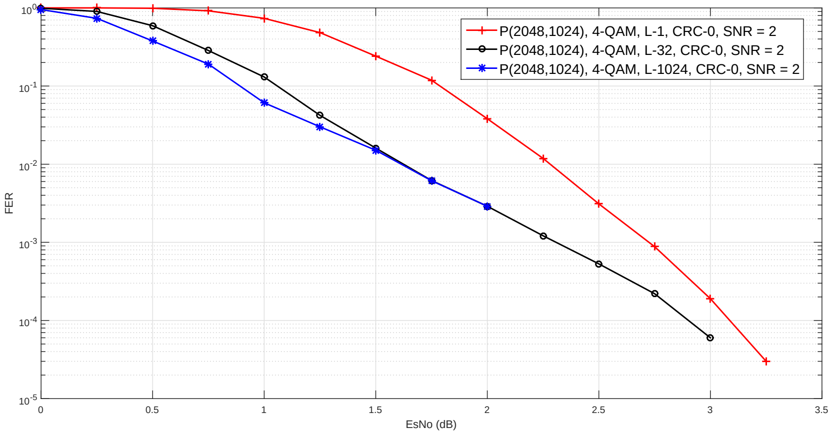
Polar code performance

Improvement by list-decoding: List-32



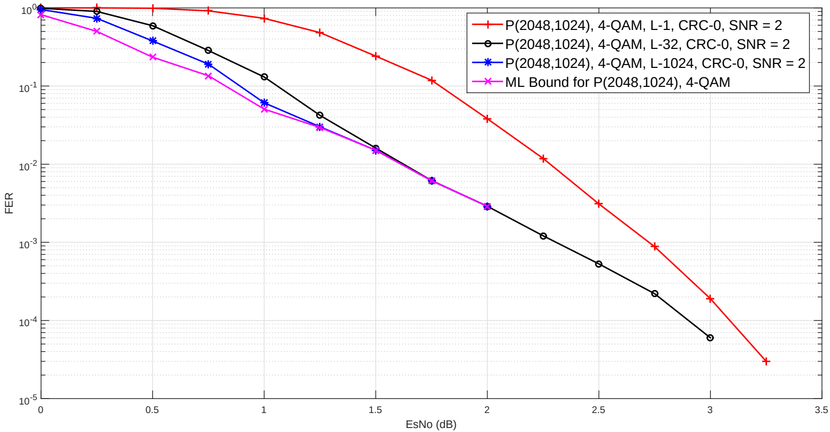
Polar code performance

Improvement by list-decoding: List-1024



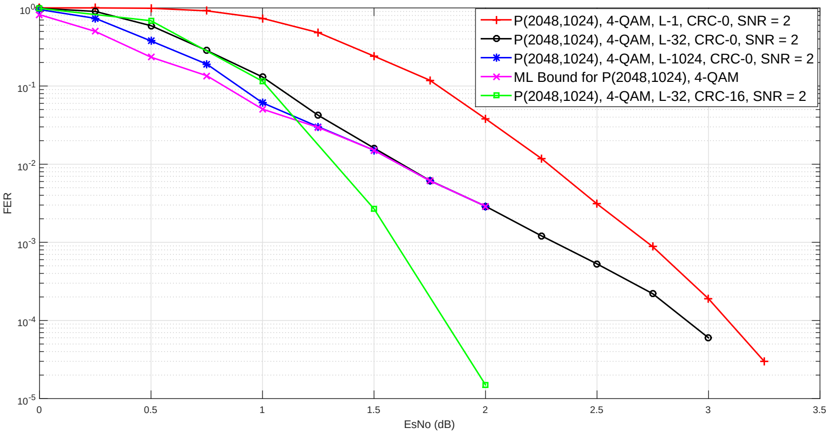
Polar code performance

Comparison with ML bound



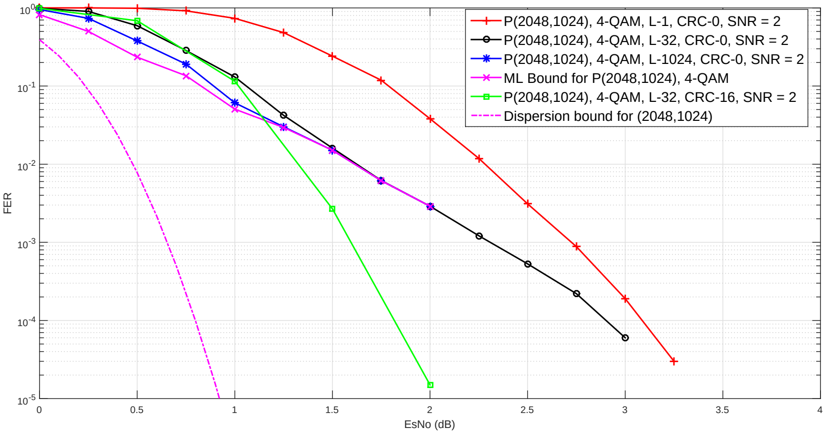
Polar code performance

Introducing CRC improves performance at high SNR



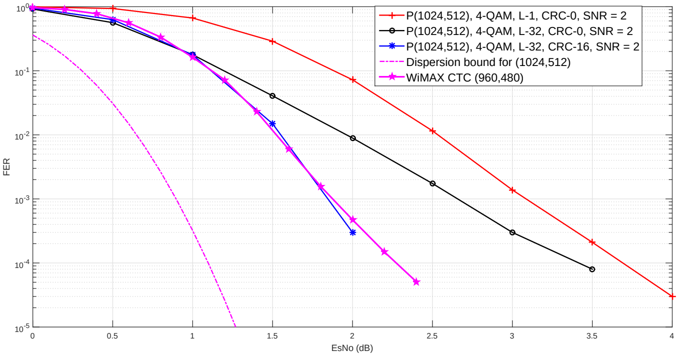
Polar code performance

Comparison with dispersion bound



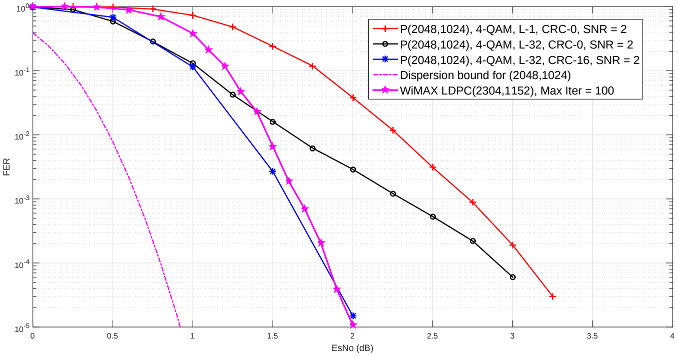
Polar codes vs WiMAX Turbo Codes

Comparable performance obtained with List-32 + CRC



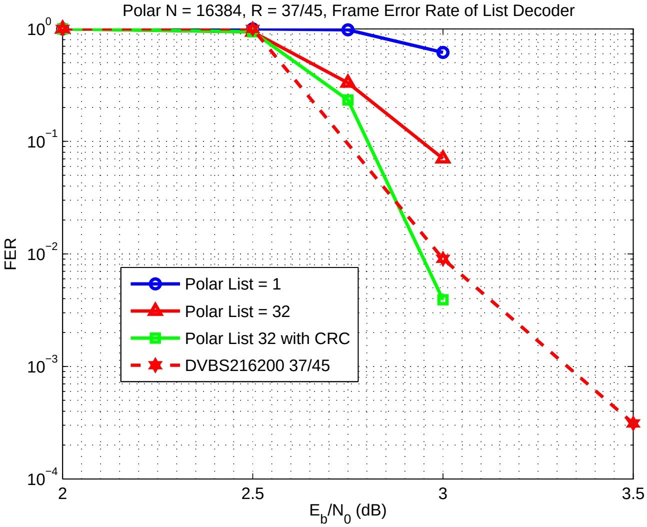
Polar codes vs WiMAX LDPC Codes

Better performance obtained with List-32 + CRC



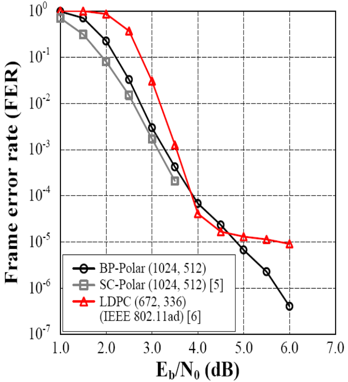
Polar Codes vs DVB-S2 LDPC Codes

LDPC (16200,13320), Polar (16384,13421). Rates = 0.82. BPSK-AWGN channel.



Polar codes vs IEEE 802.11ad LDPC codes

Park (2014) gives the following performance comparison.



(Park's result on LDPC conflicts with reference IEEE 802.11-10/0432r2. Whether there exists an error floor as shown needs to be confirmed independently.)

Source: Youn Sung Park, "Energy-Effcient Decoders of Near-Capacity Channel Codes," PhD Dissertation, The University of Michigan, 2014.

Summary of performance comparisons

- Successive cancellation decoder is simplest but inherently sequential which limits throughput
- BP decoder improves throughput and with careful design performance
- List decoder but significantly improves performance at low SNR
- Adding CRC to list decoding improves performance significantly at high SNR with little extra complexity
- Overall, polar codes under list-32 decoding with CRC offer performance comparable to codes used in present wireless standards

Summary of performance comparisons

- Successive cancellation decoder is simplest but inherently sequential which limits throughput
- BP decoder improves throughput and with careful design performance
- List decoder but significantly improves performance at low SNR
- Adding CRC to list decoding improves performance significantly at high SNR with little extra complexity
- Overall, polar codes under list-32 decoding with CRC offer performance comparable to codes used in present wireless standards

Summary of performance comparisons

- Successive cancellation decoder is simplest but inherently sequential which limits throughput
- BP decoder improves throughput and with careful design performance
- List decoder but significantly improves performance at low SNR
- Adding CRC to list decoding improves performance significantly at high SNR with little extra complexity
- Overall, polar codes under list-32 decoding with CRC offer performance comparable to codes used in present wireless standards

Summary of performance comparisons

- Successive cancellation decoder is simplest but inherently sequential which limits throughput
- BP decoder improves throughput and with careful design performance
- List decoder but significantly improves performance at low SNR
- Adding CRC to list decoding improves performance significantly at high SNR with little extra complexity
- Overall, polar codes under list-32 decoding with CRC offer performance comparable to codes used in present wireless standards

Summary of performance comparisons

- Successive cancellation decoder is simplest but inherently sequential which limits throughput
- BP decoder improves throughput and with careful design performance
- List decoder but significantly improves performance at low SNR
- Adding CRC to list decoding improves performance significantly at high SNR with little extra complexity
- Overall, polar codes under list-32 decoding with CRC offer performance comparable to codes used in present wireless standards

Implementation performance metrics

Implementation performance is measured by

- Chip area (mm2)
- Throughput (Mbits/sec)
- Energy efficiency (nJ/bit)
- Hardware efficiency (Mb/s/mm2)

Implementation performance metrics

Implementation performance is measured by

- Chip area (mm2)
- Throughput (Mbits/sec)
- Energy efficiency (nJ/bit)
- Hardware efficiency (Mb/s/mm2)

Implementation performance metrics

Implementation performance is measured by

- Chip area (mm2)
- Throughput (Mbits/sec)
- Energy efficiency (nJ/bit)
- Hardware efficiency (Mb/s/mm2)

Implementation performance metrics

Implementation performance is measured by

- Chip area (mm2)
- Throughput (Mbits/sec)
- Energy efficiency (nJ/bit)
- Hardware efficiency (Mb/s/mm2)

Successive cancellation decoder comparisons

	[1]	[2][1]	[3][2]	
Decoder Type	SC	SC	BP	
Block Length	1024	1024	1024	
Technology	90 nm	65 nm	65 nm	
Area [mm^2]	3.213	0.68	1.476	
Voltage [V]	1.0	1.2	1.0	0.475
Frequency [MHz]	2.79	1010	300	50
Power [mW]	32.75	-	477.5	18.6
Throughput [Mb/s]	2860	497	4676	779.3
Engy.-per-bit [pJ/b]	11.45	-	102.1	23.8
Hard. Eff. [Mb/s/mm^2]	890	730	3168	528

[1] O. Dizdar and E. Arıkan, arXiv:1412.3829, 2014.

[2] Y. Fan and C.-Y. Tsui, "An efficient partial-sum network architecture for semi-parallel polar codes decoder implementation," Signal Processing, IEEE Transactions on, vol. 62, no. 12, pp. 3165-3179, June 2014.

[3] C. Zhang, B. Yuan, and K. K. Parhi, "Reduced-latency SC polar decoder architectures," arxiv.org, 2011.

[1] Throughput 730 Mb/s calculated by technology conversion metrics

[2] Performance at 4 dB SNR with average no of iterations 6.57

BP decoder comparisons

Property	Unit	[1]	[2]	[3]	[3]	[4]	[4]
Decoding type and Scheduling		SCD with folded HPPSN	Specialized SC	BP Circular Unidirec-tional	BP Circular Unidirec-tional	BP All-ON, Fully Parallel	BP Circular Unidirec-tional, Reduced Complexity
Block length		1024	16384	1024	1024	1024	1024
Rate			0.9	0.5	0.5	0.5	0.5
Technology		CMOS	Altera Stratix 4	CMOS	CMOS	CMOS	CMOS
Process	nm	65	40	65	65	45	45
Core area	mm^2	0.068		1.48	1.48	12.46	1.65
Supply	V	1.2	1.35	1	0.475	1	1
Frequency	MHz	1010	106	300	50	606	555
Power	mW			477.5	18.6	2056.5	328.4
Iterations		1	1	15	15	15	15
Throughput*	Mb/s	497	1091	1024	171	2068	1960
Energy efficiency	pJ/b			102.1	23.8	110.5	19.3
Energy eff. per iter.	pJ/b/iter			15.54	3.63	7.36	1.28
Area efficiency	$Mb/s/mm^2$	7306.78		693.77	99.80	166.01	1187.71
Normalized to 45 nm according to ITRS roadmap							
Throughput*	Mb/s	613.4		1263.8	210.6	2068	1960
Energy efficiency	pJ/b			149.6	34.9	110.5	19.3
Area efficiency	$Mb/s/mm^2$	18036.5		1250.21	179.85	166.01	1187.71

∗ Throughput obtained by disabling the BP early-stopping rules for fair comparison.

[1] Y.-Z. Fan and C.-Y. Tsui, "An efficient partial-sum network architecture for semi-parallel polar codes decoder implementation," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3165–3179, June 2014.

[2] G. Sarkis, P. Giard, A. Vardy, C. Thibeault, and W. J. Gross, "Fast polar decoders: Algorithm and implementation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 5, pp. 946–957, May 2014.

[3] Y. S. Park, "Energy-efficient decoders of near-capacity channel codes," in *http://deepblue.lib.umich.edu/handle/2027.42/108731*, 23 October 2014 PhD.

[4] A. D. G. Biroli, G. Masera, E. Arıkan, "High-throughput belief propagation decoder architectures for polar codes," submitted 2015.

Concatenation

Method	Ref
Block turbo coding with polar constituents	AKMOP (2009)
Generalized concatenated coding with polar inner	AM (2009)
Reed-Solomon outer, polar inner	BJE (2010)
Polar outer, block inner	SH (2010)
Polar outer, LDPC inner	EP (ISIT'2011)

AKMOP: A., Kim, Markarian, Özgür, Poyraz

GCC: A., Markarian

BJE: Bakshi, Jaggi, and Effros

SH: Seidl and Huber

EP: Eslami and Pishro-Nik

Polar Coding

Applications

Erdal Arıkan

Electrical-Electronics Engineering Department,
Bilkent University, Ankara, Turkey

2016 JTG / IEEE Information Theory Society Summer School,
Department of Electrical Communication Engineering,
Indian Institute of Science,
Bangalore, India
27 June - 1 July 2016

Table of Contents

2.1 Polar coding for bandlimited channels

2.2 Polar codes for future applications

2.1 Polar Coding for bandlimited channels

- Objective: To discuss coding for bandlimited channels in general and with polar coding in particular
- Topics
  - Bit interleaved coded modulation (BICM)
  - Multi-level coding and modulation (MLCM)
  - Lattice coding
  - Direct polarization approach

2.1 Polar Coding for bandlimited channels

- Objective: To discuss coding for bandlimited channels in general and with polar coding in particular
- Topics
  - Bit interleaved coded modulation (BICM)
  - Multi-level coding and modulation (MLCM)
  - Lattice coding
  - Direct polarization approach

2.1 Polar Coding for bandlimited channels

- Objective: To discuss coding for bandlimited channels in general and with polar coding in particular
- Topics
  - Bit interleaved coded modulation (BICM)
  - Multi-level coding and modulation (MLCM)
  - Lattice coding
  - Direct polarization approach

2.1 Polar Coding for bandlimited channels

- Objective: To discuss coding for bandlimited channels in general and with polar coding in particular
- Topics
  - Bit interleaved coded modulation (BICM)
  - Multi-level coding and modulation (MLCM)
  - Lattice coding
  - Direct polarization approach

2.1 Polar Coding for bandlimited channels

- Objective: To discuss coding for bandlimited channels in general and with polar coding in particular
- Topics
  - Bit interleaved coded modulation (BICM)
  - Multi-level coding and modulation (MLCM)
  - Lattice coding
  - Direct polarization approach

2.1 Polar Coding for bandlimited channels

- Objective: To discuss coding for bandlimited channels in general and with polar coding in particular
- Topics
  - Bit interleaved coded modulation (BICM)
  - Multi-level coding and modulation (MLCM)
  - Lattice coding
  - Direct polarization approach

The AWGN Channel

The AWGN channel is a continuous-time channel

$$Y(t) = X(t) + N(t)$$

such that the input $X(t)$ is a random process bandlimited to W subject to a power constraint $\overline{X^2(t)} \leq P$, and $N(t)$ is white Gaussian noise with power spectral density $N_0/2$.

Capacity

Shannon's formula gives the capacity of the AWGN channel as

$$C_{[b/s]} = W \log_2(1 + P/WN_0) \quad (bits/s)$$

Discrete Time Model

An AWGN channel of bandwidth W gives rise to $2W$ independent discrete time channels per second with input-output mapping

$$Y = X + N$$

- X is a random variable with mean 0 and energy $E[X^2] \leq P/2W$
- N is Gaussian noise with 0-mean and energy $N_0/2$.
- It is customary to normalize the signal energies to joules per 2 dimensions and define

$$E_s = P/W \quad \text{Joules/2D}$$

  as signal energy (per two dimensions).
- One defines the the signal-to-noise ratio as E_s/N_0.

Discrete Time Model

An AWGN channel of bandwidth W gives rise to $2W$ independent discrete time channels per second with input-output mapping

$$Y = X + N$$

- X is a random variable with mean 0 and energy $E[X^2] \leq P/2W$
- N is Gaussian noise with 0-mean and energy $N_0/2$.
- It is customary to normalize the signal energies to joules per 2 dimensions and define

$$E_s = P/W \quad \text{Joules/2D}$$

  as signal energy (per two dimensions).
- One defines the the signal-to-noise ratio as E_s/N_0.

Discrete Time Model

An AWGN channel of bandwidth W gives rise to $2W$ independent discrete time channels per second with input-output mapping

$$Y = X + N$$

- X is a random variable with mean 0 and energy $E[X^2] \leq P/2W$
- N is Gaussian noise with 0-mean and energy $N_0/2$.
- It is customary to normalize the signal energies to joules per 2 dimensions and define

$$E_s = P/W \quad \text{Joules/2D}$$

  as signal energy (per two dimensions).
- One defines the the signal-to-noise ratio as E_s/N_0.

Discrete Time Model

An AWGN channel of bandwidth W gives rise to $2W$ independent discrete time channels per second with input-output mapping

$$Y = X + N$$

- X is a random variable with mean 0 and energy $E[X^2] \leq P/2W$
- N is Gaussian noise with 0-mean and energy $N_0/2$.
- It is customary to normalize the signal energies to joules per 2 dimensions and define

$$E_s = P/W \quad \text{Joules/2D}$$

as signal energy (per two dimensions).
- One defines the the signal-to-noise ratio as E_s/N_0.

Capacity

The capacity of the discrete-time AWGN channel is given by

$$C = \frac{1}{2}\log_2(1 + E_s/N_0), \quad (bits/D),$$

achieved by i.i.d. Gaussian inputs $X \sim N(0, E_s/2)$ per dimension.

Signal Design Problem

Now, we need a digital interface instead of real-valued inputs.

- Select a subset $\mathcal{A} \subset \mathcal{R}^n$ as the "signal set" or "modulation alphabet".
- Finding a signal set with good Euclidean distance properties and other desirable features is the "signal design" problem.
- Typically, the dimension n is 1 or 2, but can be higher.

Signal Design Problem

Now, we need a digital interface instead of real-valued inputs.

- Select a subset $\mathcal{A} \subset \mathcal{R}^n$ as the "signal set" or "modulation alphabet".
- Finding a signal set with good Euclidean distance properties and other desirable features is the "signal design" problem.
- Typically, the dimension n is 1 or 2, but can be higher.

Signal Design Problem

Now, we need a digital interface instead of real-valued inputs.

- Select a subset $\mathcal{A} \subset \mathcal{R}^n$ as the "signal set" or "modulation alphabet".
- Finding a signal set with good Euclidean distance properties and other desirable features is the "signal design" problem.
- Typically, the dimension n is 1 or 2, but can be higher.

Separation of coding and modulation

- Each constellation $\mathcal{A}$ has a capacity $C_{\mathcal{A}}$ (bits/D) which is a function of E_s/N_0.
- The spectral efficiency ρ (bits/D) has to satisfy

$$\rho < C_{\mathcal{A}}(E_s/N_0)$$

  at the operating E_s/N_0.
- The spectral efficiency is the product of two terms

$$\rho = R \times \frac{\log_2(|\mathcal{A}|)}{dim(\mathcal{A})}$$

  where R (dimensionless) is the rate of the FEC.
- For a given ρ, there any many choices w.r.t. R and $\mathcal{A}$.

Separation of coding and modulation

- Each constellation $\mathcal{A}$ has a capacity $C_{\mathcal{A}}$ (bits/D) which is a function of E_s/N_0.
- The spectral efficiency ρ (bits/D) has to satisfy

$$\rho < C_{\mathcal{A}}(E_s/N_0)$$

at the operating E_s/N_0.
- The spectral efficiency is the product of two terms

$$\rho = R \times \frac{\log_2(|\mathcal{A}|)}{dim(\mathcal{A})}$$

where R (dimensionless) is the rate of the FEC.
- For a given ρ, there any many choices w.r.t. R and $\mathcal{A}$.

Separation of coding and modulation

- Each constellation $\mathcal{A}$ has a capacity $C_{\mathcal{A}}$ (bits/D) which is a function of E_s/N_0.
- The spectral efficiency ρ (bits/D) has to satisfy

$$\rho < C_{\mathcal{A}}(E_s/N_0)$$

  at the operating E_s/N_0.
- The spectral efficiency is the product of two terms

$$\rho = R \times \frac{\log_2(|\mathcal{A}|)}{dim(\mathcal{A})}$$

  where R (dimensionless) is the rate of the FEC.
- For a given ρ, there any many choices w.r.t. R and $\mathcal{A}$.

Separation of coding and modulation

- Each constellation $\mathcal{A}$ has a capacity $C_{\mathcal{A}}$ (bits/D) which is a function of E_s/N_0.
- The spectral efficiency ρ (bits/D) has to satisfy

$$\rho < C_{\mathcal{A}}(E_s/N_0)$$

  at the operating E_s/N_0.
- The spectral efficiency is the product of two terms

$$\rho = R \times \frac{\log_2(|\mathcal{A}|)}{dim(\mathcal{A})}$$

  where R (dimensionless) is the rate of the FEC.
- For a given ρ, there any many choices w.r.t. R and $\mathcal{A}$.

M-ary Pulse Amplitude Modulation

A 1-D signal set with $\mathcal{A} = \{\pm\alpha, \pm 3\alpha, \ldots, \pm(M-1)\}$.

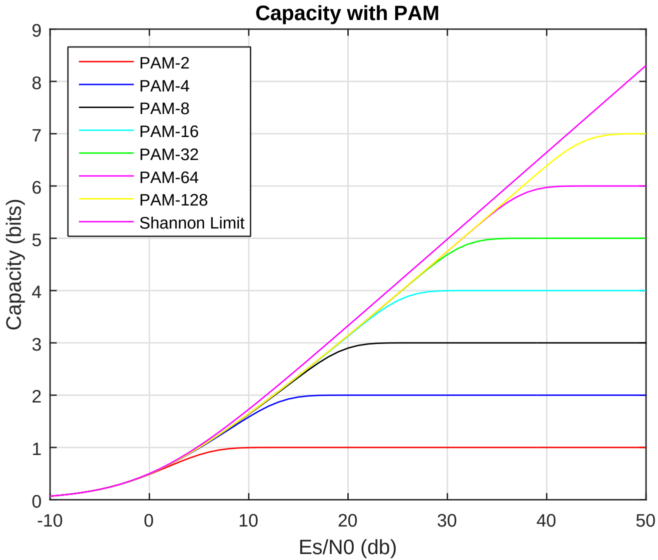
- Average energy: $E_s = 2\alpha^2(M^2 - 1)/3$ (J/2D)
- Consider the capacity, cutoff rate

M-ary Pulse Amplitude Modulation

A 1-D signal set with $\mathcal{A} = \{\pm\alpha, \pm3\alpha, \ldots, \pm(M-1)\}$.

- Average energy: $E_s = 2\alpha^2(M^2-1)/3$ (J/2D)
- Consider the capacity, cutoff rate

Capacity of *M*-PAM



M-PAM is good enough from a capacity viewpoint.

Conventional approach

Given a target spectral efficiency ρ and a target error rate P_e at a specific E_s/N_o,

- select M large enough so that M-PAM capacity is close enough to the Shannon capacity at the given E_s/N_o
- apply coding external to modulation to achieve the desired P_e

Conventional approach

Given a target spectral efficiency ρ and a target error rate P_e at a specific E_s/N_o,

- select M large enough so that M-PAM capacity is close enough to the Shannon capacity at the given E_s/N_o
- apply coding external to modulation to achieve the desired P_e

Conventional approach

Given a target spectral efficiency ρ and a target error rate P_e at a specific E_s/N_o,

- select M large enough so that M-PAM capacity is close enough to the Shannon capacity at the given E_s/N_o
- apply coding external to modulation to achieve the desired P_e

Such separation of coding and modulation was first challenged successfully by Ungerboeck (1981).

Conventional approach

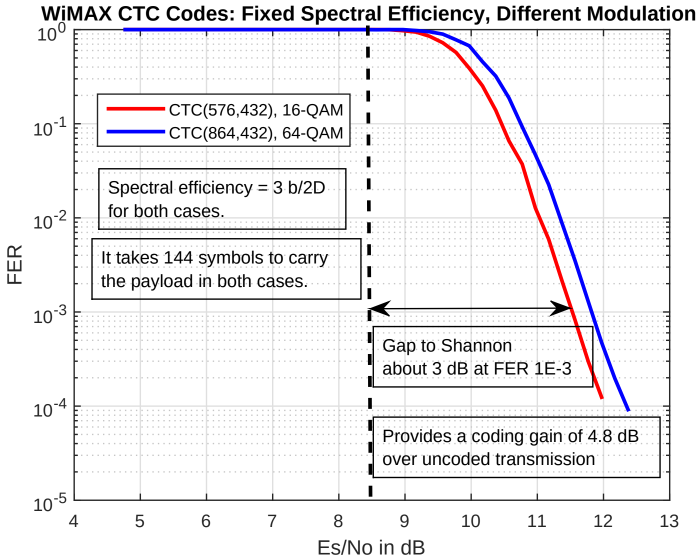
Given a target spectral efficiency ρ and a target error rate P_e at a specific E_s/N_o,

- select M large enough so that M-PAM capacity is close enough to the Shannon capacity at the given E_s/N_o
- apply coding external to modulation to achieve the desired P_e

Such separation of coding and modulation was first challenged successfully by Ungerboeck (1981).

However, with the advent of powerful codes at affordable complexity, there is a return to the conventional design methodology.

How does it work in practice?



Theory and practice don't match here!

Why change modulation instead of just the code rate?

- Suppose we fix the modulation as 64-QAM and wish to deliver data at spectral efficiencies 1, 2, 3, 4, 5 b/2D.
- We would need a coding scheme that works well at rates 1/6, 1/3, 1/2, 2/3, 5/6.
- The inability of delivering high quality coding over a wide range of rates forces one to change the order of modulation.
- The difficulty here is practical: it is a challenge to have a coding scheme that works well over all rates from 0 to 1.

Why change modulation instead of just the code rate?

- Suppose we fix the modulation as 64-QAM and wish to deliver data at spectral efficiencies 1, 2, 3, 4, 5 b/2D.
- We would need a coding scheme that works well at rates 1/6, 1/3, 1/2, 2/3, 5/6.
- The inability of delivering high quality coding over a wide range of rates forces one to change the order of modulation.
- The difficulty here is practical: it is a challenge to have a coding scheme that works well over all rates from 0 to 1.

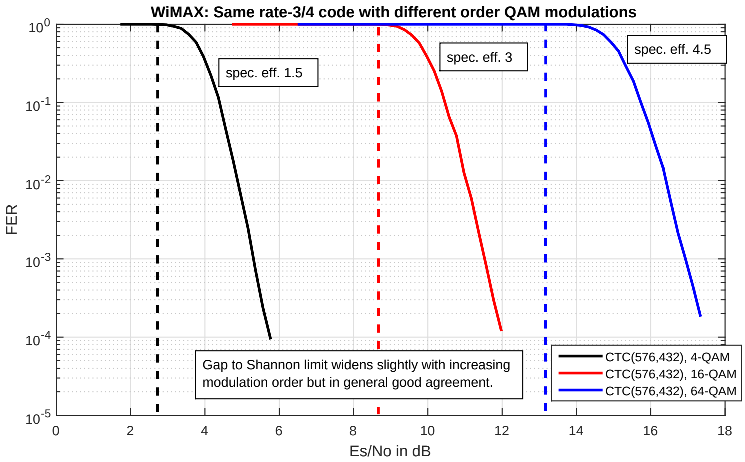
Why change modulation instead of just the code rate?

- Suppose we fix the modulation as 64-QAM and wish to deliver data at spectral efficiencies 1, 2, 3, 4, 5 b/2D.
- We would need a coding scheme that works well at rates 1/6, 1/3, 1/2, 2/3, 5/6.
- The inability of delivering high quality coding over a wide range of rates forces one to change the order of modulation.
- The difficulty here is practical: it is a challenge to have a coding scheme that works well over all rates from 0 to 1.

Why change modulation instead of just the code rate?

- Suppose we fix the modulation as 64-QAM and wish to deliver data at spectral efficiencies 1, 2, 3, 4, 5 b/2D.
- We would need a coding scheme that works well at rates 1/6, 1/3, 1/2, 2/3, 5/6.
- The inability of delivering high quality coding over a wide range of rates forces one to change the order of modulation.
- The difficulty here is practical: it is a challenge to have a coding scheme that works well over all rates from 0 to 1.

Alternative: Fixed code, variable modulation



Polar coding and modulation

Polar codes can be applied to modulation in at least three different ways.

- Direct polarization
- Multi-level techniques
- Polar lattices
- BICM

Polar coding and modulation

Polar codes can be applied to modulation in at least three different ways.

- Direct polarization
- Multi-level techniques
- Polar lattices
- BICM

Polar coding and modulation

Polar codes can be applied to modulation in at least three different ways.

- Direct polarization
- Multi-level techniques
- Polar lattices
- BICM

Polar coding and modulation

Polar codes can be applied to modulation in at least three different ways.

- Direct polarization
- Multi-level techniques
- Polar lattices
- BICM

Direct Method

- Idea: Given a system with q-ary modulation, treat it as an ordinary q-ary input memoryless channel and apply a suitable polarization transform.
- Theory of q-ary polarization exists:
  - Şaşoğlu, E., E. Telatar, and E. Arıkan. "Polarization for arbitrary discrete memoryless channels." IEEE ITW 2009.
  - Sahebi, A. G. and S. S. Pradhan, "Multilevel polarization of polar codes over arbitrary discrete memoryless channels." IEEE Allerton, 2011.
  - Park, W.-C. and A. Barg. "Polar codes for q-ary channels," IEEE Trans. Inform. Theory, 2013.
  - ...
- Deserves to be studied further.

Direct Method

- Idea: Given a system with q-ary modulation, treat it as an ordinary q-ary input memoryless channel and apply a suitable polarization transform.
- Theory of q-ary polarization exists:
  - Şasoğlu, E., E. Telatar, and E. Arıkan. "Polarization for arbitrary discrete memoryless channels." IEEE ITW 2009.
  - Sahebi, A. G. and S. S. Pradhan, "Multilevel polarization of polar codes over arbitrary discrete memoryless channels." IEEE Allerton, 2011.
  - Park, W.-C. and A. Barg. "Polar codes for q-ary channels," IEEE Trans. Inform. Theory, 2013.
  - ...
- Deserves to be studied further.

Direct Method

- Idea: Given a system with q-ary modulation, treat it as an ordinary q-ary input memoryless channel and apply a suitable polarization transform.
- Theory of q-ary polarization exists:
  - Şaşoğlu, E., E. Telatar, and E. Arıkan. "Polarization for arbitrary discrete memoryless channels." IEEE ITW 2009.
  - Sahebi, A. G. and S. S. Pradhan, "Multilevel polarization of polar codes over arbitrary discrete memoryless channels." IEEE Allerton, 2011.
  - Park, W.-C. and A. Barg. "Polar codes for q-ary channels," IEEE Trans. Inform. Theory, 2013.
  - ...
- Deserves to be studied further.

Direct Method

- Idea: Given a system with q-ary modulation, treat it as an ordinary q-ary input memoryless channel and apply a suitable polarization transform.
- Theory of q-ary polarization exists:
  - Şasoğlu, E., E. Telatar, and E. Arıkan. "Polarization for arbitrary discrete memoryless channels." IEEE ITW 2009.
  - Sahebi, A. G. and S. S. Pradhan, "Multilevel polarization of polar codes over arbitrary discrete memoryless channels." IEEE Allerton, 2011.
  - Park, W.-C. and A. Barg. "Polar codes for q-ary channels," IEEE Trans. Inform. Theory, 2013.
  - ...
- Deserves to be studied further.

Direct Method

- Idea: Given a system with q-ary modulation, treat it as an ordinary q-ary input memoryless channel and apply a suitable polarization transform.
- Theory of q-ary polarization exists:
  - Şaşoğlu, E., E. Telatar, and E. Arıkan. "Polarization for arbitrary discrete memoryless channels." IEEE ITW 2009.
  - Sahebi, A. G. and S. S. Pradhan, "Multilevel polarization of polar codes over arbitrary discrete memoryless channels." IEEE Allerton, 2011.
  - Park, W.-C. and A. Barg. "Polar codes for q-ary channels," IEEE Trans. Inform. Theory, 2013.
  - ...
- Deserves to be studied further.

Direct Method

- Idea: Given a system with q-ary modulation, treat it as an ordinary q-ary input memoryless channel and apply a suitable polarization transform.
- Theory of q-ary polarization exists:
  - Şasoğlu, E., E. Telatar, and E. Arıkan. "Polarization for arbitrary discrete memoryless channels." IEEE ITW 2009.
  - Sahebi, A. G. and S. S. Pradhan, "Multilevel polarization of polar codes over arbitrary discrete memoryless channels." IEEE Allerton, 2011.
  - Park, W.-C. and A. Barg. "Polar codes for q-ary channels," IEEE Trans. Inform. Theory, 2013.
  - ...
- Deserves to be studied further.

Direct Method

- Idea: Given a system with q-ary modulation, treat it as an ordinary q-ary input memoryless channel and apply a suitable polarization transform.
- Theory of q-ary polarization exists:
  - Şasoğlu, E., E. Telatar, and E. Arıkan. "Polarization for arbitrary discrete memoryless channels." IEEE ITW 2009.
  - Sahebi, A. G. and S. S. Pradhan, "Multilevel polarization of polar codes over arbitrary discrete memoryless channels." IEEE Allerton, 2011.
  - Park, W.-C. and A. Barg. "Polar codes for q-ary channels," IEEE Trans. Inform. Theory, 2013.
  - ...
- Deserves to be studied further.

Multi-Level Modulation (Imai and Hirakawa, 1977)

- Represent (if possible) each channel input symbol as a vector $X = (X_1, X_2, \ldots, X_r)$; then the capacity can be written as a sum of capacities of smaller channels by the chain rule:

$$
\begin{aligned}
I(X; Y) &= I(X_1, X_2, \ldots, X_r; Y) \\
&= \sum_{i=1}^{r} I(X_i; Y | X_1, \ldots, X_{i-1}).
\end{aligned}
$$

- This splits the original channel into r parallel channels, which are encoded independently and decoded using successive cancellation decoding.
- Polarization is a natural complement to MLM.

Multi-Level Modulation (Imai and Hirakawa, 1977)

- Represent (if possible) each channel input symbol as a vector $X = (X_1, X_2, \ldots, X_r)$; then the capacity can be written as a sum of capacities of smaller channels by the chain rule:

$$\begin{aligned} I(X;Y) &= I(X_1, X_2, \ldots, X_r; Y) \\ &= \sum_{i=1}^{r} I(X_i; Y | X_1, \ldots, X_{i-1}). \end{aligned}$$

- This splits the original channel into r parallel channels, which are encoded independently and decoded using successive cancellation decoding.
- Polarization is a natural complement to MLM.

Multi-Level Modulation (Imai and Hirakawa, 1977)

- Represent (if possible) each channel input symbol as a vector $X = (X_1, X_2, \ldots, X_r)$; then the capacity can be written as a sum of capacities of smaller channels by the chain rule:

$$I(X;Y) = I(X_1, X_2, \ldots, X_r; Y) = \sum_{i=1}^{r} I(X_i; Y | X_1, \ldots, X_{i-1}).$$

- This splits the original channel into r parallel channels, which are encoded independently and decoded using successive cancellation decoding.
- Polarization is a natural complement to MLM.

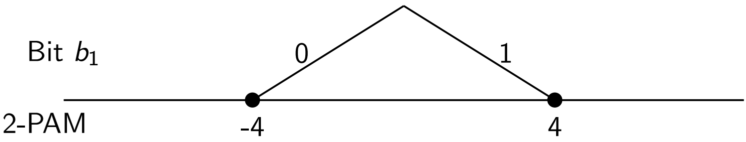
Polar coding with multi-level modulation

Already a well-studied subject:

- Arıkan, E., “Polar Coding,” Plenary Talk, ISIT 2011.
- Seidl, M., Schenk, A., Stierstorfer, C., and Huber, J. B. “Polar-coded modulation,” IEEE Trans. Comm. 2013.
- Seidl, M., Schenk, A., Stierstorfer, C., and Huber, J. B. “Multilevel polar-coded modulation‘,” IEEE ISIT 2013
- Ionita, Corina, et al. "On the design of binary polar codes for high-order modulation." IEEE GLOBECOM, 2014.
- Beygi, L., Agrell, E., Kahn, J. M., and Karlsson, M., “Coded modulation for fiber-optic networks,” IEEE Sig. Proc. Mag., 2014.
- ...

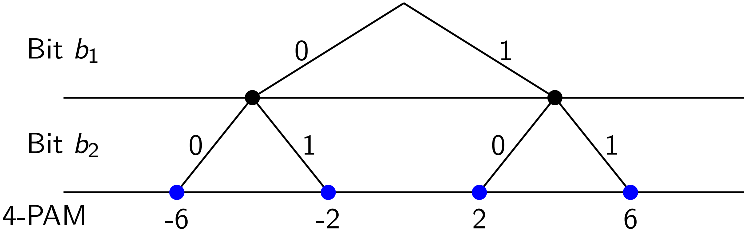
Example: 8-PAM as 3 bit channels

- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1



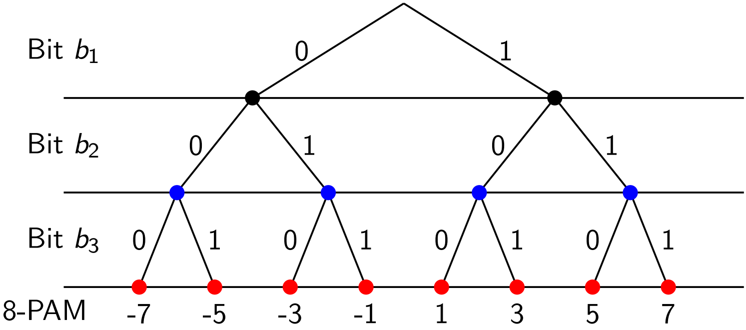
Example: 8-PAM as 3 bit channels

- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1



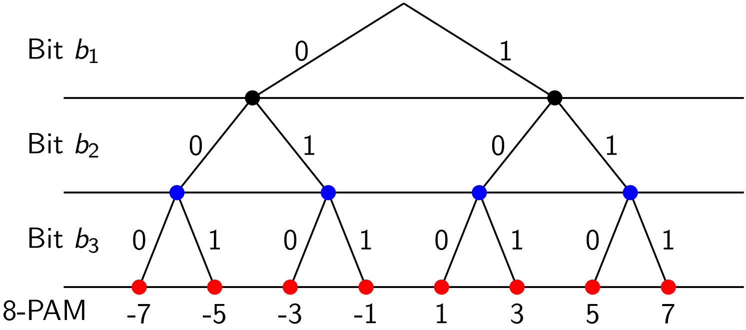
Example: 8-PAM as 3 bit channels

- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1



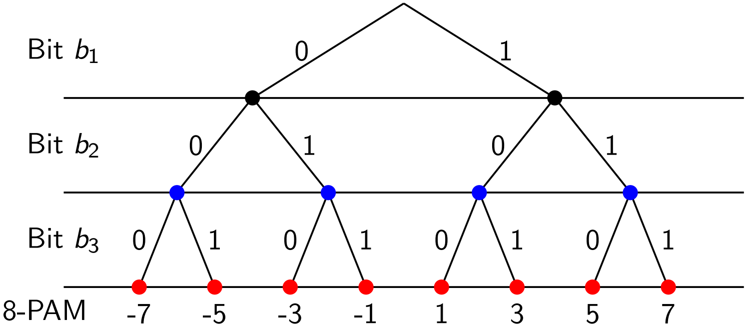
Example: 8-PAM as 3 bit channels

- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1



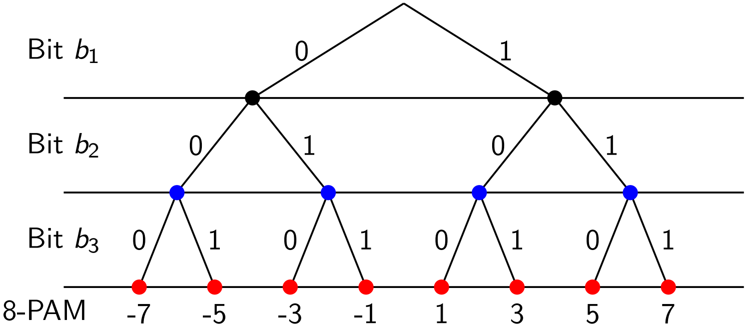
Example: 8-PAM as 3 bit channels

- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1

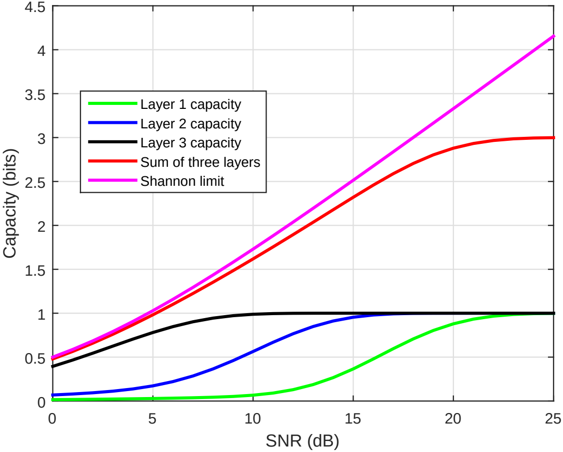


Example: 8-PAM as 3 bit channels

- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1



Polarization across layers by natural labeling



Most coding work needs to be done at the least significant bits.

Performance comparison: Polar vs. Turbo

Turbo code

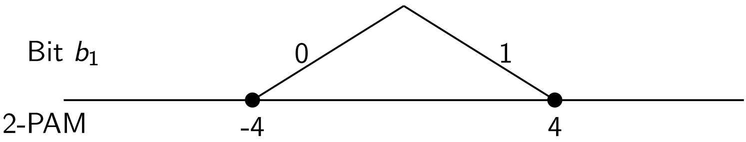
- WiMAX CTC
- Duobinary, memory 3
- QAM over AWGN channel
- Gray mapping
- BICM
- Simulator: "Coded Modulation Library"

Polar code

- Standard construction
- Successive cancellation decoding
- QAM over AWGN channel
- Natural mapping
- Multi-level PAM
- PAM over AWGN channel

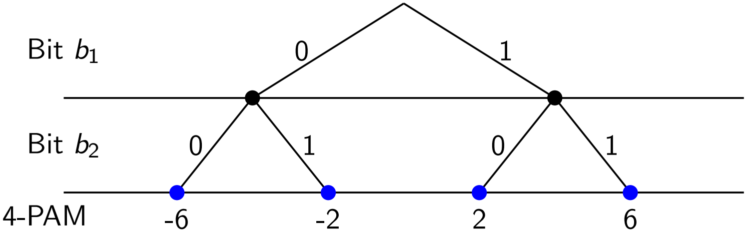
Example: 8-PAM as 3 bit channels

- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1



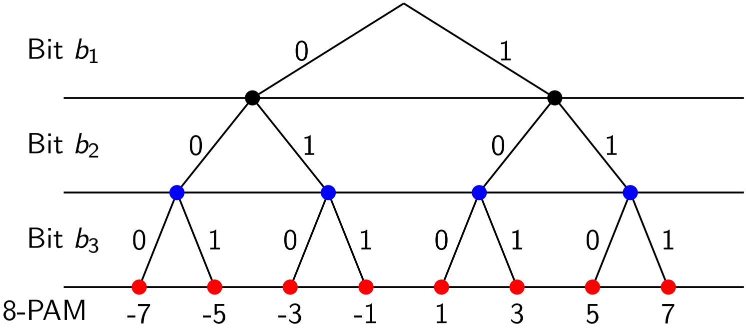
Example: 8-PAM as 3 bit channels

- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1



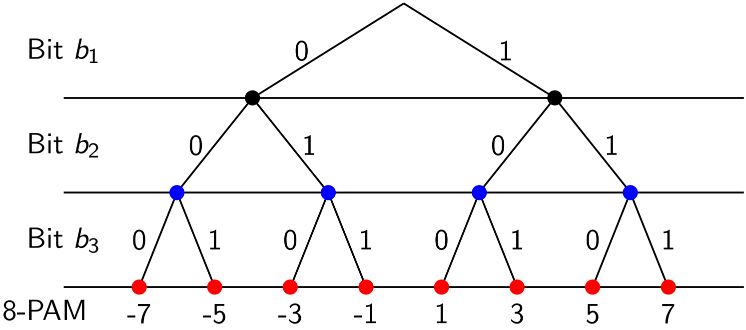
Example: 8-PAM as 3 bit channels

- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1



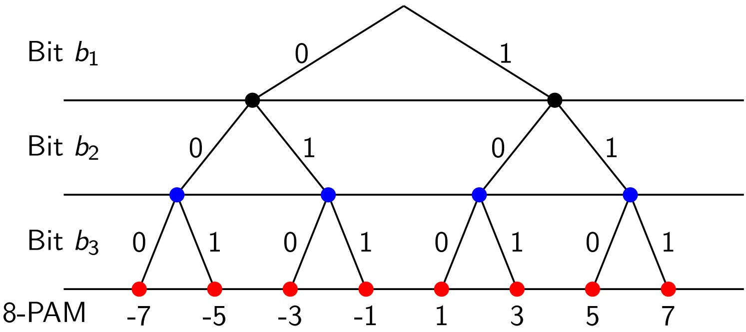
Example: 8-PAM as 3 bit channels

- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1



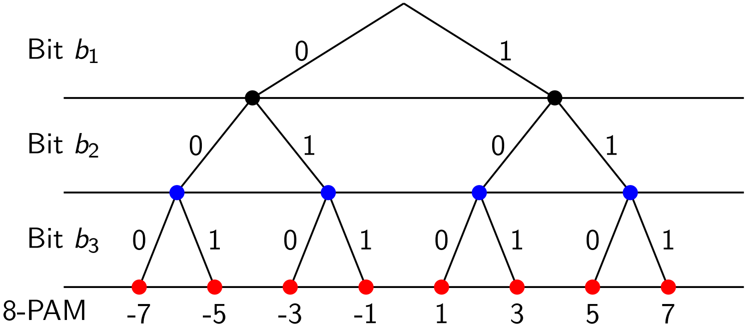
Example: 8-PAM as 3 bit channels

- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1

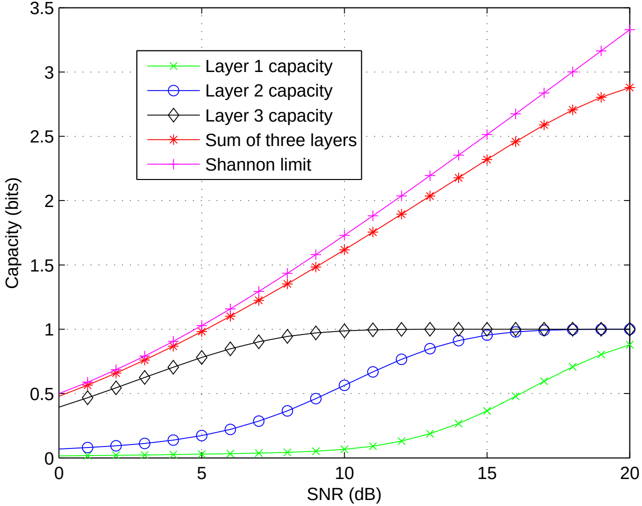


Example: 8-PAM as 3 bit channels

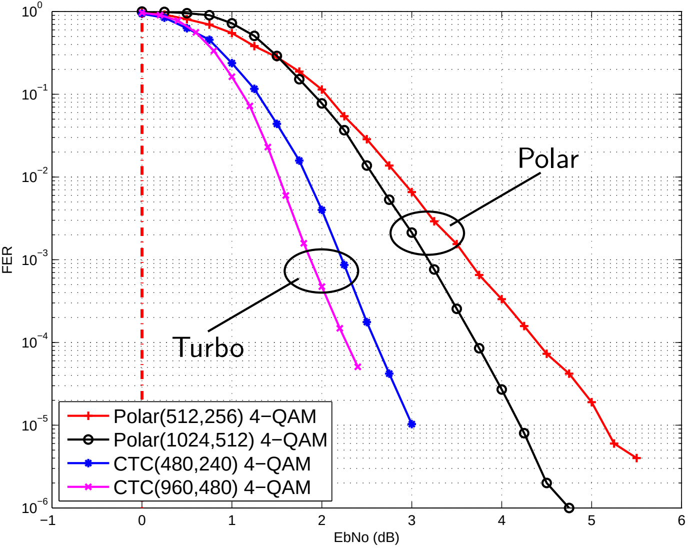
- PAM signals selected by three bits (b_1, b_2, b_3)
- Three layers of binary channels created
- Each layer encoded independently
- Layers decoded in the order b_3, b_2, b_1



Multi-layering jump-starts polarization

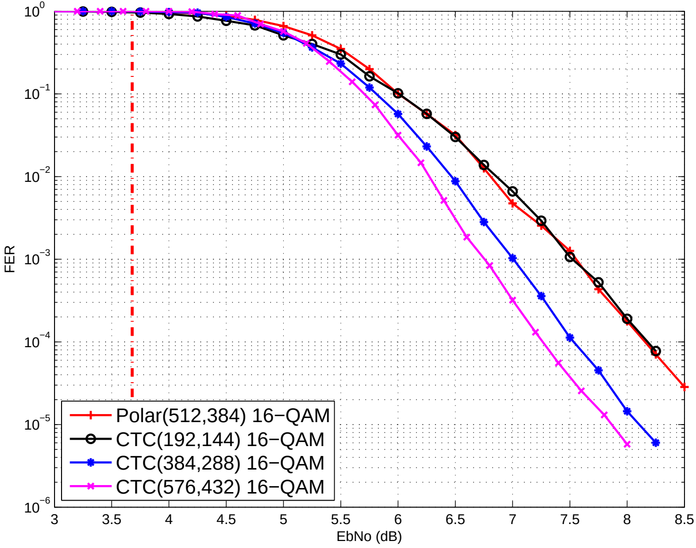


4-QAM, Rate 1/2

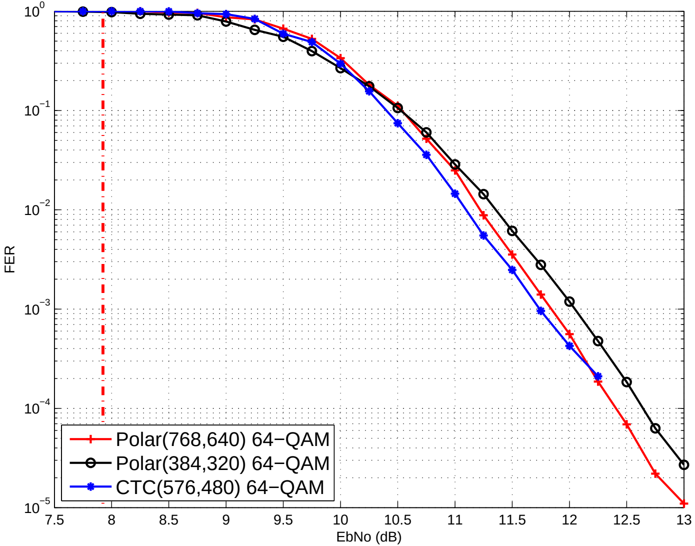


FER
EbNo (dB)
Polar
Turbo
Polar(512,256) 4−QAM
Polar(1024,512) 4−QAM
CTC(480,240) 4−QAM
CTC(960,480) 4−QAM

16-QAM, Rate 3/4



64-QAM, Rate 5/6



Complexity comparison: 64-QAM, Rate 5/6

Average decoding time in milliseconds per codeword (ms/cw)

E_b/N_0	CTC(576,432)	Polar(768,640)	Polar(384,320)
10 dB	6.23	0.92	0.48
11 dB	1.83	1.01	0.53

Both decoders implemented as MATLAB mex functions. Polar decoder is a successive cancellation decoder. CTC decoder is a public domain decoder (CML). Profiling done by MATLAB Profiler. Iteration limit for CTC decoder was 10; average no of iterations was 10 at 10 dB and 3.3 at 11 dB. CTC decoder used a linear approximation to log-MAP while polar decoder used exact log-MAP.

Complexity comparison: 64-QAM, Rate 5/6

Average decoding time in milliseconds per codeword (ms/cw)

E_b/N_0	CTC(576,432)	Polar(768,640)	Polar(384,320)
10 dB	6.23	0.92	0.48
11 dB	1.83	1.01	0.53

Polar codes show a complexity advantage against CTC codes.

Both decoders implemented as MATLAB mex functions. Polar decoder is a successive cancellation decoder. CTC decoder is a public domain decoder (CML). Profiling done by MATLAB Profiler. Iteration limit for CTC decoder was 10; average no of iterations was 10 at 10 dB and 3.3 at 11 dB. CTC decoder used a linear approximation to log-MAP while polar decoder used exact log-MAP.

Lattices and polar coding

Yan, Cong, and Liu explored the connection between lattices and polar coding.

- Yan, Yanfei, and L. Cong, "A construction of lattices from polar codes." IEEE 2012 ITW.
- Yan, Yanfei, Ling Liu, Cong Ling, and Xiaofu Wu. "Construction of capacity-achieving lattice codes: Polar lattices." arXiv preprint arXiv:1411.0187 (2014)

Lattices and polar coding

Yan *et al* used the Barnes-Wall lattice contructions such as

$$\mathsf{BW}_{16} = \mathsf{RM}(1,4) + 2\mathsf{RM}(3,4) + 4(\mathbb{Z}^{16})$$

as a template for constructing polar lattices of the type

$$\mathsf{P}_{16} = \mathsf{P}(1,4) + 2\mathsf{P}(3,4) + 4(\mathbb{Z}^{16})$$

and demonstrated by simulations that polar lattices perform better.

BICM

BICM [Zehavi, 1991], [Caire, Taricco, Biglieri, 1998] is the dominant technique in modern wireless standards such as LTE.

BICM

BICM [Zehavi, 1991], [Caire, Taricco, Biglieri, 1998] is the dominant technique in modern wireless standards such as LTE.

As in MLM, BICM splits the channel input symbols into a vector $X = (X_1, X_2, \ldots, X_r)$ but strives to do so such that

$$\begin{aligned} I(X;Y) &= I(X_1, X_2, \ldots, X_r; Y) \\ &= \sum_{i=1}^{r} I(X_i; Y | X_1, \ldots, X_{i-1}) \\ &\approx \sum_{i=1}^{r} I(X_i; Y). \end{aligned}$$

BICM vs Multi Level Modulation

Why has BICM won over MLM and other techniques in practice?

- MLM is provably capacity-achieving; BICM is suboptimal but the rate penalty is tolerable.
- MLM has to do delicate rate-matching at individual layers, which is difficult with turbo and LDPC codes.
- BICM is well-matched to iterative decoding methods used with turbo and LDPC codes.
- MLM suffers extra latency due to multi-stage decoding (mitigated in part by the lack of need for protecting the upper layers by long codes)
- With MLM, the overall code is split into shorter codes which weakens performance (one may mix and match the block lengths of each layer to alleviate this problem).

BICM vs Multi Level Modulation

Why has BICM won over MLM and other techniques in practice?

- MLM is provably capacity-achieving; BICM is suboptimal but the rate penalty is tolerable.
- MLM has to do delicate rate-matching at individual layers, which is difficult with turbo and LDPC codes.
- BICM is well-matched to iterative decoding methods used with turbo and LDPC codes.
- MLM suffers extra latency due to multi-stage decoding (mitigated in part by the lack of need for protecting the upper layers by long codes)
- With MLM, the overall code is split into shorter codes which weakens performance (one may mix and match the block lengths of each layer to alleviate this problem).

BICM vs Multi Level Modulation

Why has BICM won over MLM and other techniques in practice?

- MLM is provably capacity-achieving; BICM is suboptimal but the rate penalty is tolerable.
- MLM has to do delicate rate-matching at individual layers, which is difficult with turbo and LDPC codes.
- BICM is well-matched to iterative decoding methods used with turbo and LDPC codes.
- MLM suffers extra latency due to multi-stage decoding (mitigated in part by the lack of need for protecting the upper layers by long codes)
- With MLM, the overall code is split into shorter codes which weakens performance (one may mix and match the block lengths of each layer to alleviate this problem).

BICM vs Multi Level Modulation

Why has BICM won over MLM and other techniques in practice?

- MLM is provably capacity-achieving; BICM is suboptimal but the rate penalty is tolerable.
- MLM has to do delicate rate-matching at individual layers, which is difficult with turbo and LDPC codes.
- BICM is well-matched to iterative decoding methods used with turbo and LDPC codes.
- MLM suffers extra latency due to multi-stage decoding (mitigated in part by the lack of need for protecting the upper layers by long codes)
- With MLM, the overall code is split into shorter codes which weakens performance (one may mix and match the block lengths of each layer to alleviate this problem).

BICM vs Multi Level Modulation

Why has BICM won over MLM and other techniques in practice?

- MLM is provably capacity-achieving; BICM is suboptimal but the rate penalty is tolerable.
- MLM has to do delicate rate-matching at individual layers, which is difficult with turbo and LDPC codes.
- BICM is well-matched to iterative decoding methods used with turbo and LDPC codes.
- MLM suffers extra latency due to multi-stage decoding (mitigated in part by the lack of need for protecting the upper layers by long codes)
- With MLM, the overall code is split into shorter codes which weakens performance (one may mix and match the block lengths of each layer to alleviate this problem).

BICM and Polar Coding

This subject, too, has been studied in connection with polar codes.

- Mahdavifar, H. and El-Khamy, M. and Lee, J. and Kang, I., "Polar Coding for Bit-Interleaved Coded Modulation," IEEE Trans. Veh. Tech., 2015.
- Afser, H., N. Tirpan, H. Delic, and M. Koca, "Bit-interleaved polar-coded modulation," Proc. IEEE WCNC, 2014.
- Chen, Kai, Kai Niu, and Jia-Ru Lin. "An efficient design of bit-interleaved polar coded modulation." IEEE PIMRC 2013.
- ...

2.2 Polar codes for future applications

- Objective: Review the literature on polar coding for selected applications
- Topics
  - 60 GHz wireless
  - Optical access networks
  - 5G
    - Ultra reliable low latency communications (URLLC)
    - Machine type communications (MTC)
    - 5G channel coding at Gb/s throughput

2.2 Polar codes for future applications

- Objective: Review the literature on polar coding for selected applications
- Topics
  - 60 GHz wireless
  - Optical access networks
  - 5G
    - Ultra reliable low latency communications (URLLC)
    - Machine type communications (MTC)
    - 5G channel coding at Gb/s throughput

2.2 Polar codes for future applications

- Objective: Review the literature on polar coding for selected applications
- Topics
  - 60 GHz wireless
  - Optical access networks
  - 5G
    - Ultra reliable low latency communications (URLLC)
    - Machine type communications (MTC)
    - 5G channel coding at Gb/s throughput

2.2 Polar codes for future applications

- Objective: Review the literature on polar coding for selected applications
- Topics
  - 60 GHz wireless
  - Optical access networks
  - 5G
    - Ultra reliable low latency communications (URLLC)
    - Machine type communications (MTC)
    - 5G channel coding at Gb/s throughput

2.2 Polar codes for future applications

- Objective: Review the literature on polar coding for selected applications
- Topics
  - 60 GHz wireless
  - Optical access networks
  - 5G
    - Ultra reliable low latency communications (URLLC)
    - Machine type communications (MTC)
    - 5G channel coding at Gb/s throughput

2.2 Polar codes for future applications

- Objective: Review the literature on polar coding for selected applications
- Topics
  - 60 GHz wireless
  - Optical access networks
  - 5G
    - Ultra reliable low latency communications (URLLC)
    - Machine type communications (MTC)
    - 5G channel coding at Gb/s throughput

2.2 Polar codes for future applications

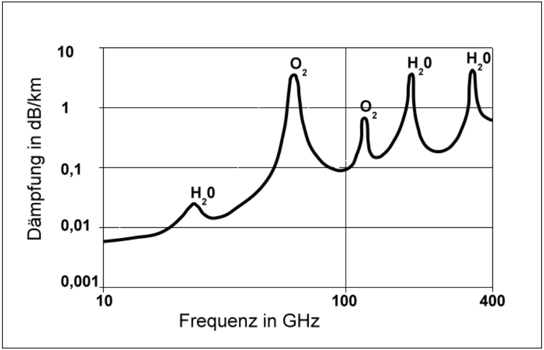
- Objective: Review the literature on polar coding for selected applications
- Topics
  - 60 GHz wireless
  - Optical access networks
  - 5G
    - Ultra reliable low latency communications (URLLC)
    - Machine type communications (MTC)
    - 5G channel coding at Gb/s throughput

2.2 Polar codes for future applications

- Objective: Review the literature on polar coding for selected applications
- Topics
  - 60 GHz wireless
  - Optical access networks
  - 5G
    - Ultra reliable low latency communications (URLLC)
    - Machine type communications (MTC)
    - 5G channel coding at Gb/s throughput

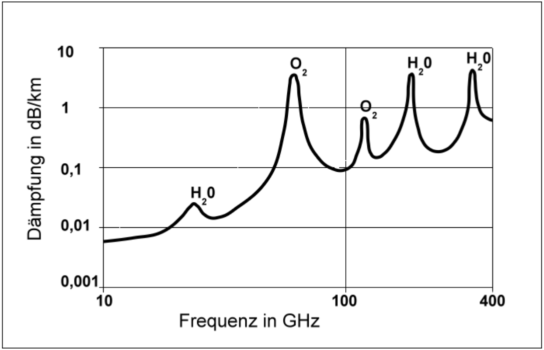
Millimeter Wave 60 GHz Communications

- 7 GHz of bandwidth available (57-64 GHz allocated in the US)
- Free-space path loss $(4\pi d/\lambda)^2$ is high at $\lambda = 5$ mm but compensated by large antenna arrays.
- Propagation range limited severely by O_2 absorption. Cells confined to rooms.



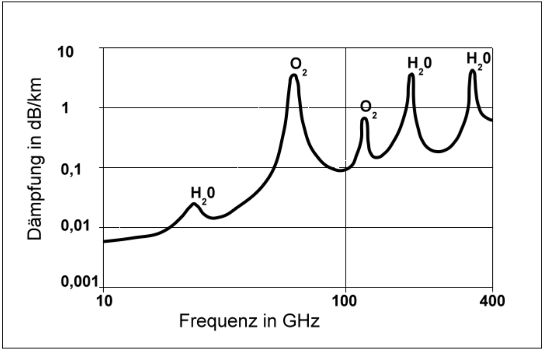
Millimeter Wave 60 GHz Communications

- 7 GHz of bandwidth available (57-64 GHz allocated in the US)
- Free-space path loss $(4\pi d/\lambda)^2$ is high at $\lambda = 5$ mm but compensated by large antenna arrays.
- Propagation range limited severely by O_2 absorption. Cells confined to rooms.



Millimeter Wave 60 GHz Communications

- 7 GHz of bandwidth available (57-64 GHz allocated in the US)
- Free-space path loss $(4\pi d/\lambda)^2$ is high at $\lambda = 5$ mm but compensated by large antenna arrays.
- Propagation range limited severely by O_2 absorption. Cells confined to rooms.



Millimeter Wave 60 GHz Communications

- Recent IEEE 802.11.ad Wi-Fi standard operates at 60 GHz ISM band and uses an LDPC code with block length 672 bits, rates 1/2, 5/8, 3/4, 13/16.
- Two papers compare polar codes that study polar coding for 60 GHz applications:
  - Z. Wei, B. Li, and C. Zhao, "On the polar code for the 60 GHz millimeter-wave systems," EURASIP, JWCN, 2015.
  - Youn Sung Park, "Energy-Effcient Decoders of Near-Capacity Channel Codes," PhD Dissertation, The University of Michigan, 2014.

Millimeter Wave 60 GHz Communications

- Recent IEEE 802.11.ad Wi-Fi standard operates at 60 GHz ISM band and uses an LDPC code with block length 672 bits, rates 1/2, 5/8, 3/4, 13/16.
- Two papers compare polar codes that study polar coding for 60 GHz applications:
  - Z. Wei, B. Li, and C. Zhao, “On the polar code for the 60 GHz millimeter-wave systems,” EURASIP, JWCN, 2015.
  - Youn Sung Park, “Energy-Effcient Decoders of Near-Capacity Channel Codes,” PhD Dissertation, The University of Michigan, 2014.

Millimeter Wave 60 GHz Communications

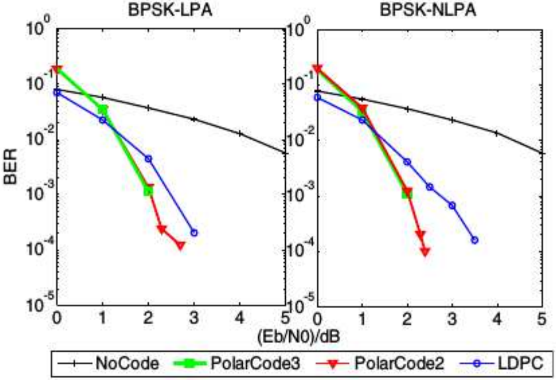
- Recent IEEE 802.11.ad Wi-Fi standard operates at 60 GHz ISM band and uses an LDPC code with block length 672 bits, rates 1/2, 5/8, 3/4, 13/16.
- Two papers compare polar codes that study polar coding for 60 GHz applications:
  - Z. Wei, B. Li, and C. Zhao, "On the polar code for the 60 GHz millimeter-wave systems," EURASIP, JWCN, 2015.
  - Youn Sung Park, "Energy-Effcient Decoders of Near-Capacity Channel Codes," PhD Dissertation, The University of Michigan, 2014.

Millimeter Wave 60 GHz Communications

- Recent IEEE 802.11.ad Wi-Fi standard operates at 60 GHz ISM band and uses an LDPC code with block length 672 bits, rates 1/2, 5/8, 3/4, 13/16.
- Two papers compare polar codes that study polar coding for 60 GHz applications:
  - Z. Wei, B. Li, and C. Zhao, "On the polar code for the 60 GHz millimeter-wave systems," EURASIP, JWCN, 2015.
  - Youn Sung Park, "Energy-Effcient Decoders of Near-Capacity Channel Codes," PhD Dissertation, The University of Michigan, 2014.

Millimeter Wave 60 GHz Communications

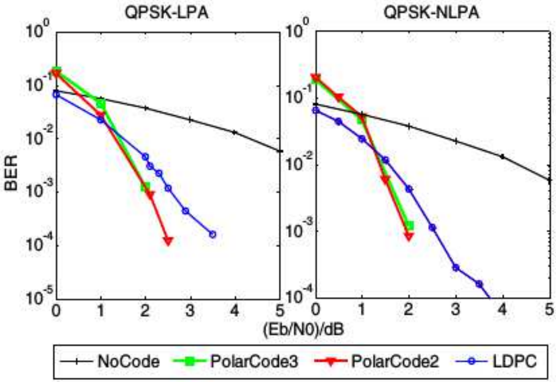
Wei et al compare polar codes with the LDPC codes used in the standard using a nonlinear channel model



Wei, B. Li, and C. Zhao, "On the polar code for the 60 GHz millimeter-wave systems," EURASIP, JWCN, 2015.

Millimeter Wave 60 GHz Communications

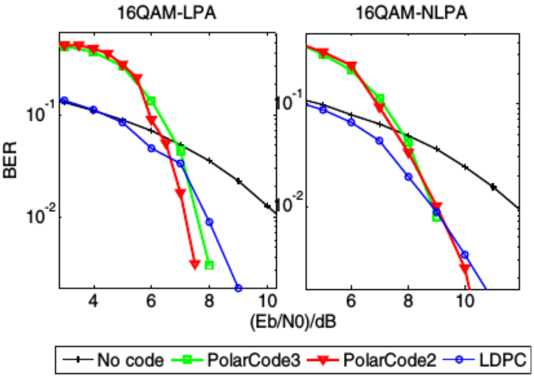
Wei et al compare polar codes with the LDPC codes used in the standard using a nonlinear channel model



Wei, B. Li, and C. Zhao, "On the polar code for the 60 GHz millimeter-wave systems," EURASIP, JWCN, 2015.

Millimeter Wave 60 GHz Communications

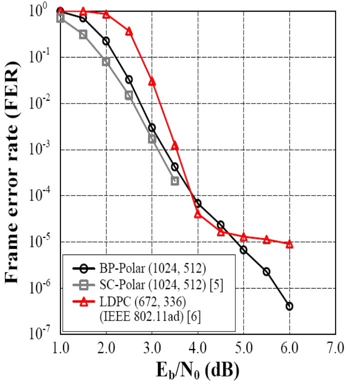
Wei et al compare polar codes with the LDPC codes used in the standard using a nonlinear channel model



Wei, B. Li, and C. Zhao, "On the polar code for the 60 GHz millimeter-wave systems," EURASIP, JWCN, 2015.

Polar codes vs IEEE 802.11ad LDPC codes

Park (2014) gives the following performance comparison.



(Park's result on LDPC conflicts with reference IEEE 802.11-10/0432r2. Whether there exists an error floor as shown needs to be confirmed independently.)

Source: Youn Sung Park, "Energy-Effcient Decoders of Near-Capacity Channel Codes," PhD Dissertation, The University of Michigan, 2014.

Polar codes vs IEEE 802.11ad LDPC codes

In terms of implementation complexity and throughput, Park (2014) gives the following figures.

	LPDC			Polar	
Throughput Gb/s	0.5	6	9	0.779	4.676
Energy efficiency (pJ/b)	21	61.7	89.5	23.8	102.1
Area efficiency (Gb/s/mm2)	0.31	3.75	5.63	0.528	3.168

Source: Youn Sung Park, "Energy-Efficient Decoders of Near-Capacity Channel Codes," PhD Dissertation, The University of Michigan, 2014.

Optical access/transport network

- 10-100 Gb/s at 1E-12 BER
- OTU4 (100 Gb/s Ethernet) and ITU G.975.1 standards use Reed-Solomon (RS) codes
- The challenge is to provide high reliability at low hardware complexity.

Optical access/transport network

- 10-100 Gb/s at 1E-12 BER
- OTU4 (100 Gb/s Ethernet) and ITU G.975.1 standards use Reed-Solomon (RS) codes
- The challenge is to provide high reliability at low hardware complexity.

Optical access/transport network

- 10-100 Gb/s at 1E-12 BER
- OTU4 (100 Gb/s Ethernet) and ITU G.975.1 standards use Reed-Solomon (RS) codes
- The challenge is to provide high reliability at low hardware complexity.

Polar codes for optical access/transport

There have been some studies of polar codes fore optical transmission.

- A. Eslami and H. Pishro-Nik, "A practical approach to polar codes," ISIT 2011. (Considers a polar-LDPC concatenated code and compares it with OTU4 RS codes.)
- Z. Wu and B. Lankl, "Polar codes for low-complexity forward error correction in optical access networks," ITG-Fachbericht 248: Photonische Netze - 05, 06.05.2014, Leipzig. (Compares polar codes with G.975.1 RS codes.)
- T. Ahmad, "Polar codes for optical communications", MS Thesis, Bilkent University, May 2016.
- L. Beygi, E. Agrell, J. M. Kahn, and M. Karlsson, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014. (Coded modulation for optical transport.)

Polar codes for optical access/transport

There have been some studies of polar codes fore optical transmission.

- A. Eslami and H. Pishro-Nik, "A practical approach to polar codes," ISIT 2011. (Considers a polar-LDPC concatenated code and compares it with OTU4 RS codes.)
- Z. Wu and B. Lankl, "Polar codes for low-complexity forward error correction in optical access networks," ITG-Fachbericht 248: Photonische Netze - 05, 06.05.2014, Leipzig. (Compares polar codes with G.975.1 RS codes.)
- T. Ahmad, "Polar codes for optical communications", MS Thesis, Bilkent University, May 2016.
- L. Beygi, E. Agrell, J. M. Kahn, and M. Karlsson, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014. (Coded modulation for optical transport.)

Polar codes for optical access/transport

There have been some studies of polar codes fore optical transmission.

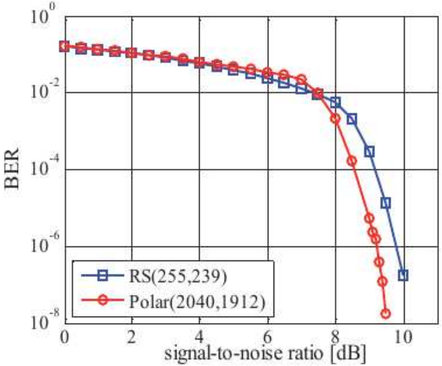
- A. Eslami and H. Pishro-Nik, "A practical approach to polar codes," ISIT 2011. (Considers a polar-LDPC concatenated code and compares it with OTU4 RS codes.)
- Z. Wu and B. Lankl, "Polar codes for low-complexity forward error correction in optical access networks," ITG-Fachbericht 248: Photonische Netze - 05, 06.05.2014, Leipzig. (Compares polar codes with G.975.1 RS codes.)
- T. Ahmad, "Polar codes for optical communications", MS Thesis, Bilkent University, May 2016.
- L. Beygi, E. Agrell, J. M. Kahn, and M. Karlsson, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014. (Coded modulation for optical transport.)

Polar codes for optical access/transport

There have been some studies of polar codes fore optical transmission.

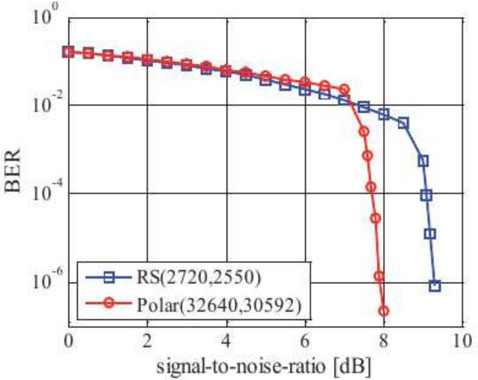
- A. Eslami and H. Pishro-Nik, "A practical approach to polar codes," ISIT 2011. (Considers a polar-LDPC concatenated code and compares it with OTU4 RS codes.)
- Z. Wu and B. Lankl, "Polar codes for low-complexity forward error correction in optical access networks," ITG-Fachbericht 248: Photonische Netze - 05, 06.05.2014, Leipzig. (Compares polar codes with G.975.1 RS codes.)
- T. Ahmad, "Polar codes for optical communications", MS Thesis, Bilkent University, May 2016.
- L. Beygi, E. Agrell, J. M. Kahn, and M. Karlsson, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014. (Coded modulation for optical transport.)

Comparison of polar codes with G.975.1 RS codes



Source: Z. Wu and B. Lankl, above reference.

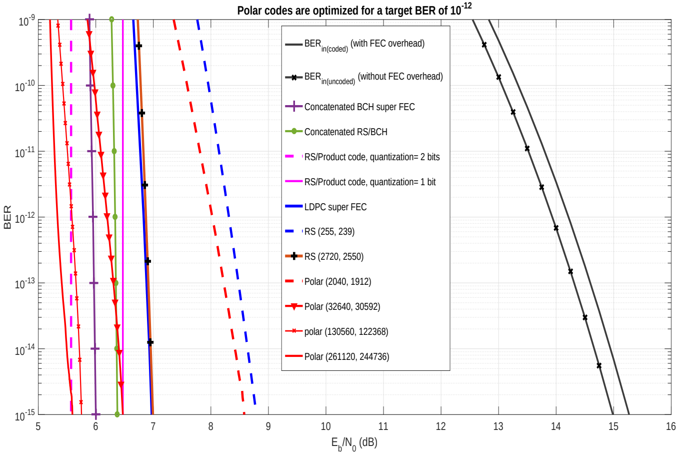
Comparison of polar codes with G.975.1 RS codes



Source: Z. Wu and B. Lankl, above reference.

Comparison of polar codes with all codes in G.975.1

In a recent MS thesis, T. Ahmad compared polar codes with G.975.1 codes.



Comparison of polar codes with all codes in G.975.1

The conclusion of Ahmad (2016) is that polar codes perform better than all G.975.1 FEC schemes.

FEC Code	BER_{in}	**NCG (dB)**	**CG (dB)**	**Q (dB)**	$\frac{Eb}{No}$ **(dB)**
RS (255, 239)	1.82E-04	5.62	5.90	11.04	8.31
LDPC super FEC code	1.33E-03	7.10	7.39	9.56	6.83
RS (2720, 2550)	1.26E-03	7.06	7.34	9.60	6.87
Conc. RS/CSOC code(24.5%OH)	5.80E-03	7.95	8.90	8.04	5.31
Concatenated BCH code	3.30E-03	7.98	8.26	8.68	5.95
Conc. RS/BCH code	2.26E-03	7.63	7.91	9.06	6.34
Conc. RS/Product code	4.60E-03	8.40	8.68	8.30	5.57
Polar (2040, 1912)	2.81E-04	5.91	6.19	10.75	8.02
Polar (32640, 30592)	2.60E-03	7.74	8.02	8.92	6.20
Polar (130560, 122368)	4.61E-03	8.35	8.63	8.31	5.58
Polar (261120, 244736)	5.72E-03	8.60	8.89	8.06	5.33

Comparison of polar codes with 3rd Generation FEC for optical transport

Ahmad's study finds that polar codes fall short of beating 3G FEC proposed for optical transport.

FEC code	NCG (dB)	Comments
Polar (32640, 27200)	10.07	Ahmad (2016)
Polar (130560, 108800)	10.79	Ahmad (2016)
Polar (261120, 217600)	11.07	Ahmad (2016)
Polar (522240, 435200)	11.30	Ahmad (2016)
CC-LDPC (10032, 4, 24)	11.50	3G FEC, 12 iterations
QC-LDPC (18360, 15300)	11.30	3G FEC, 12 iterations

Coded modulation for fiber-optic communication

Main reference for this part is the paper:

L. Beygi, E. Agrell, J. M. Kahn, and M. Karlsson, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014.

- Data rates 100 Gb/s and beyond
- BER 1E-15
- Channel model: Self-interfering nonlinear distortion, additive Gaussian noise

Coded modulation for fiber-optic communication

Main reference for this part is the paper:

L. Beygi, E. Agrell, J. M. Kahn, and M. Karlsson, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014.

- Data rates 100 Gb/s and beyond
- BER 1E-15
- Channel model: Self-interfering nonlinear distortion, additive Gaussian noise

Coded modulation for fiber-optic communication

Main reference for this part is the paper:

L. Beygi, E. Agrell, J. M. Kahn, and M. Karlsson, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014.

- Data rates 100 Gb/s and beyond
- BER 1E-15
- Channel model: Self-interfering nonlinear distortion, additive Gaussian noise

Coded modulation: BICM approach

Split the 2^q'ary channel into q bit channels and decode them independently.

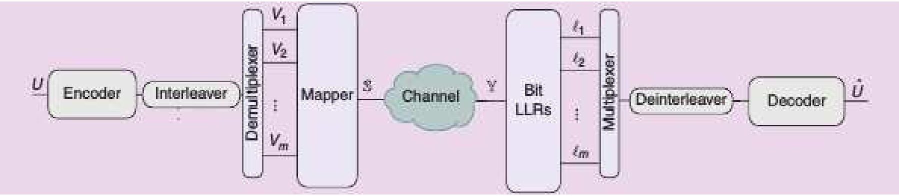


Figure source: Beygi, L., et al, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014.

Coded modulation: Multi-level approach

Split the 2^q'ary channel into q bit channels and decode them successively.

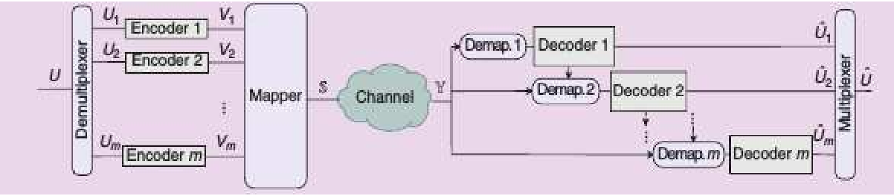


Figure source: Beygi, L., et al, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014.

Coded modulation: BICM approach

Split the 2^q'ary channel into q bit channels and decode them independently.

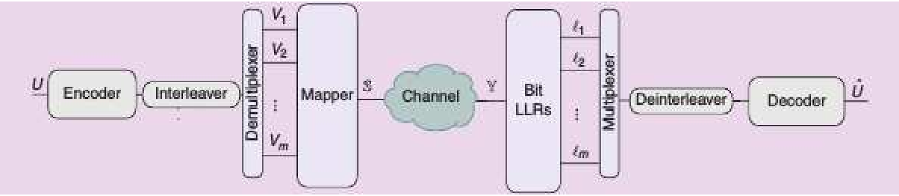


Figure source: Beygi, L., et al, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014.

Coded modulation: TCM approach

Split the 2^q'ary channels into two classes and encode the low-order channels using a trellis hand-crafted for large Euclidean distance and ML-decoded

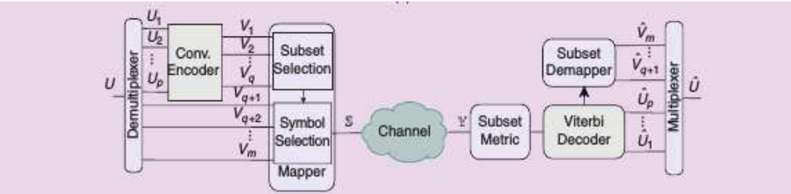


Figure source: Beygi, L., et al, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014.

Coded modulation: q'ary coding

No splitting; 2^q'ary processing applied; too complex

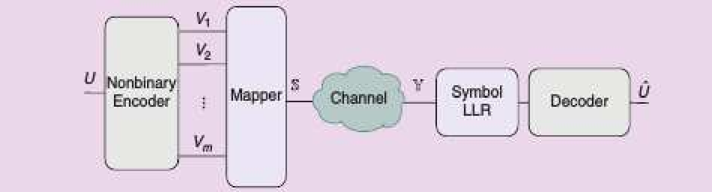


Figure source: Beygi, L., et al, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014.

Coded modulation: Polar approach

Split the 2^q'ary channel into "good", "mediocre", and "bad" bit channels; apply coding only to mediocre channels

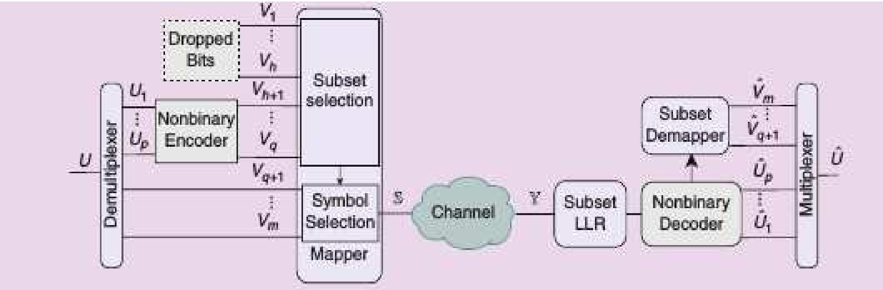
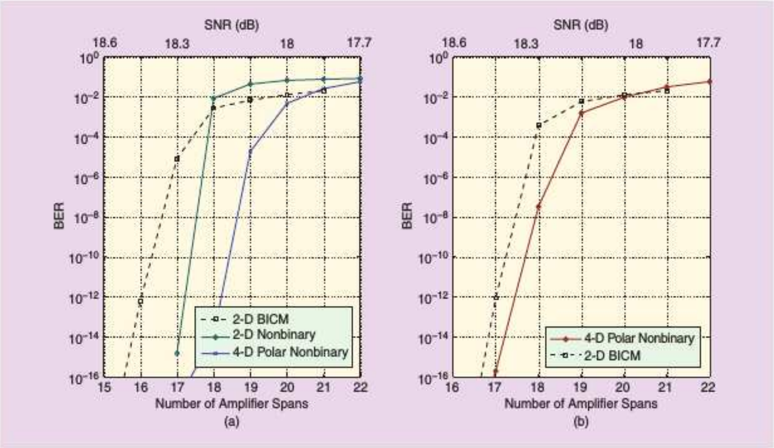


Figure source: Beygi, L., et al, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014.

Coded modulation: performance comparison



[FIG6] (a) The BER of three CM schemes with information-block-length-constraint. (b) The BER of 2-D and 4-D CM schemes with binary and nonbinary LDPC codes, respectively, and similar complexity. All the CM schemes use PM 64-QAM with 21% coding overhead and have therefore the same spectral efficiency.

Figure source: Beygi, L., et al, "Coded modulation for fiber-optic networks," IEEE Sig. Proc. Mag., Mar. 2014.

What is 5G?

Andrews *et al.*[1] answer this question as follows.

- It willl *not* be an incremental advance over 4G.
- Will be characterized by
  - Very high frequencies and massive bandwidths with very large no of antennas
  - Extreme base station and device connectivity
  - Universal connectivity between 5G new air interfaces, LTE, WiFi, etc.

[1] Andrews et al., "What will 5G be?" JSAC 2014

What is 5G?

Andrews *et al.*[1] answer this question as follows.

- It willl *not* be an incremental advance over 4G.
- Will be characterized by
  - Very high frequencies and massive bandwidths with very large no of antennas
  - Extreme base station and device connectivity
  - Universal connectivity between 5G new air interfaces, LTE, WiFi, etc.

[1] Andrews et al., "What will 5G be?" JSAC 2014

What is 5G?

Andrews *et al.*[1] answer this question as follows.

- It willl *not* be an incremental advance over 4G.
- Will be characterized by
  - Very high frequencies and massive bandwidths with very large no of antennas
  - Extreme base station and device connectivity
  - Universal connectivity between 5G new air interfaces, LTE, WiFi, etc.

[1] Andrews et al., "What will 5G be?" JSAC 2014

What is 5G?

Andrews *et al.*[1] answer this question as follows.

- It willl *not* be an incremental advance over 4G.
- Will be characterized by
  - Very high frequencies and massive bandwidths with very large no of antennas
  - Extreme base station and device connectivity
  - Universal connectivity between 5G new air interfaces, LTE, WiFi, etc.

[1]Andrews et al., "What will 5G be?" JSAC 2014

What is 5G?

Andrews *et al.*[1] answer this question as follows.

- It willl *not* be an incremental advance over 4G.
- Will be characterized by
  - Very high frequencies and massive bandwidths with very large no of antennas
  - Extreme base station and device connectivity
  - Universal connectivity between 5G new air interfaces, LTE, WiFi, etc.

[1] Andrews et al., "What will 5G be?" JSAC 2014

Technical requirements for 5G

Again, according to Andrews *et al.*, 5G will have to meet the following requirements (not all at once):

- Data rates compared to 4G
  - Aggregate: 1000 times more capacity/km2 compared to 4G
  - Cell-edge: 100 - 1000 Mb/s/user with 95% guarantee
  - Peak: 10s of Gb/s/user
- Round-trip latency: Some applications (tactile Internet, two-way gaming, virtual reality) will require 1 ms latency compared to 10-15 ms that 4G can provide
- Energy and cost: Link energy consumption should remain the same as data rates increase, meaning that a 100-times more energy-efficient link is required
- No of devices: 10,000 more low-rate devices for M2M communications, along with traditional high-rate users

Technical requirements for 5G

Again, according to Andrews *et al.*, 5G will have to meet the following requirements (not all at once):

- Data rates compared to 4G
  - Aggregate: 1000 times more capacity/km2 compared to 4G
  - Cell-edge: 100 - 1000 Mb/s/user with 95% guarantee
  - Peak: 10s of Gb/s/user
- Round-trip latency: Some applications (tactile Internet, two-way gaming, virtual reality) will require 1 ms latency compared to 10-15 ms that 4G can provide
- Energy and cost: Link energy consumption should remain the same as data rates increase, meaning that a 100-times more energy-efficient link is required
- No of devices: 10,000 more low-rate devices for M2M communications, along with traditional high-rate users

Technical requirements for 5G

Again, according to Andrews *et al.*, 5G will have to meet the following requirements (not all at once):

- Data rates compared to 4G
  - Aggregate: 1000 times more capacity/km2 compared to 4G
  - Cell-edge: 100 - 1000 Mb/s/user with 95% guarantee
  - Peak: 10s of Gb/s/user
- Round-trip latency: Some applications (tactile Internet, two-way gaming, virtual reality) will require 1 ms latency compared to 10-15 ms that 4G can provide
- Energy and cost: Link energy consumption should remain the same as data rates increase, meaning that a 100-times more energy-efficient link is required
- No of devices: 10,000 more low-rate devices for M2M communications, along with traditional high-rate users

Technical requirements for 5G

Again, according to Andrews *et al.*, 5G will have to meet the following requirements (not all at once):

- Data rates compared to 4G
  - Aggregate: 1000 times more capacity/km2 compared to 4G
  - Cell-edge: 100 - 1000 Mb/s/user with 95% guarantee
  - Peak: 10s of Gb/s/user
- Round-trip latency: Some applications (tactile Internet, two-way gaming, virtual reality) will require 1 ms latency compared to 10-15 ms that 4G can provide
- Energy and cost: Link energy consumption should remain the same as data rates increase, meaning that a 100-times more energy-efficient link is required
- No of devices: 10,000 more low-rate devices for M2M communications, along with traditional high-rate users

Technical requirements for 5G

Again, according to Andrews *et al.*, 5G will have to meet the following requirements (not all at once):

- Data rates compared to 4G
  - Aggregate: 1000 times more capacity/km2 compared to 4G
  - Cell-edge: 100 - 1000 Mb/s/user with 95% guarantee
  - Peak: 10s of Gb/s/user
- Round-trip latency: Some applications (tactile Internet, two-way gaming, virtual reality) will require 1 ms latency compared to 10-15 ms that 4G can provide
- Energy and cost: Link energy consumption should remain the same as data rates increase, meaning that a 100-times more energy-efficient link is required
- No of devices: 10,000 more low-rate devices for M2M communications, along with traditional high-rate users

Technical requirements for 5G

Again, according to Andrews *et al.*, 5G will have to meet the following requirements (not all at once):

- Data rates compared to 4G
  - Aggregate: 1000 times more capacity/km2 compared to 4G
  - Cell-edge: 100 - 1000 Mb/s/user with 95% guarantee
  - Peak: 10s of Gb/s/user
- Round-trip latency: Some applications (tactile Internet, two-way gaming, virtual reality) will require 1 ms latency compared to 10-15 ms that 4G can provide
- Energy and cost: Link energy consumption should remain the same as data rates increase, meaning that a 100-times more energy-efficient link is required
- No of devices: 10,000 more low-rate devices for M2M communications, along with traditional high-rate users

Technical requirements for 5G

Again, according to Andrews *et al.*, 5G will have to meet the following requirements (not all at once):

- Data rates compared to 4G
  - Aggregate: 1000 times more capacity/km2 compared to 4G
  - Cell-edge: 100 - 1000 Mb/s/user with 95% guarantee
  - Peak: 10s of Gb/s/user
- Round-trip latency: Some applications (tactile Internet, two-way gaming, virtual reality) will require 1 ms latency compared to 10-15 ms that 4G can provide
- Energy and cost: Link energy consumption should remain the same as data rates increase, meaning that a 100-times more energy-efficient link is required
- No of devices: 10,000 more low-rate devices for M2M communications, along with traditional high-rate users

Key technology ingredients for 5G

It is generally agreed that the 1000x aggregate data rate increase will be possible through a combination of three types of gains.

- Densification of network access nodes
- Increased bandwidth (move to mm waves)
- Increased spectral efficiency through new communication techniques:
  - advanced MIMO
  - improved multi-access
  - better interference management
  - improved coding and modulation schemes

Key technology ingredients for 5G

It is generally agreed that the 1000x aggregate data rate increase will be possible through a combination of three types of gains.

- Densification of network access nodes
- Increased bandwidth (move to mm waves)
- Increased spectral efficiency through new communication techniques:
  - advanced MIMO
  - improved multi-access
  - better interference management
  - improved coding and modulation schemes

Key technology ingredients for 5G

It is generally agreed that the 1000x aggregate data rate increase will be possible through a combination of three types of gains.

- Densification of network access nodes
- Increased bandwidth (move to mm waves)
- Increased spectral efficiency through new communication techniques:
  - advanced MIMO
  - improved multi-access
  - better interference management
  - improved coding and modulation schemes

Key technology ingredients for 5G

It is generally agreed that the 1000x aggregate data rate increase will be possible through a combination of three types of gains.

- Densification of network access nodes
- Increased bandwidth (move to mm waves)
- Increased spectral efficiency through new communication techniques:
  - advanced MIMO
  - improved multi-access
  - better interference management
  - improved coding and modulation schemes

Key technology ingredients for 5G

It is generally agreed that the 1000x aggregate data rate increase will be possible through a combination of three types of gains.

- Densification of network access nodes
- Increased bandwidth (move to mm waves)
- Increased spectral efficiency through new communication techniques:
  - advanced MIMO
  - improved multi-access
  - better interference management
  - improved coding and modulation schemes

Key technology ingredients for 5G

It is generally agreed that the 1000x aggregate data rate increase will be possible through a combination of three types of gains.

- Densification of network access nodes
- Increased bandwidth (move to mm waves)
- Increased spectral efficiency through new communication techniques:
  - advanced MIMO
  - improved multi-access
  - better interference management
  - improved coding and modulation schemes

Key technology ingredients for 5G

It is generally agreed that the 1000x aggregate data rate increase will be possible through a combination of three types of gains.

- Densification of network access nodes
- Increased bandwidth (move to mm waves)
- Increased spectral efficiency through new communication techniques:
  - advanced MIMO
  - improved multi-access
  - better interference management
  - improved coding and modulation schemes

Summary

- With list-decoding and CRC polar codes deliver comparable performance to LDPC and Turbo codes used in present wireless standards
- SoA in coding is already close to theoretical limits for low-order modulation, leaving little margin for improvement
- The biggest asset of polar coding compared to SoA is its universal, flexible, and versatile nature
  - Universal: the same hardware can be used with different code lengths, rates, channels
  - Flexible: the code rate can be adjusted readily to any number between 0 and 1
  - Versatile: can be used in multi-terminal coding scenarios

Summary

- With list-decoding and CRC polar codes deliver comparable performance to LDPC and Turbo codes used in present wireless standards
- SoA in coding is already close to theoretical limits for low-order modulation, leaving little margin for improvement
- The biggest asset of polar coding compared to SoA is its universal, flexible, and versatile nature
  - Universal: the same hardware can be used with different code lengths, rates, channels
  - Flexible: the code rate can be adjusted readily to any number between 0 and 1
  - Versatile: can be used in multi-terminal coding scenarios

Summary

- With list-decoding and CRC polar codes deliver comparable performance to LDPC and Turbo codes used in present wireless standards
- SoA in coding is already close to theoretical limits for low-order modulation, leaving little margin for improvement
- The biggest asset of polar coding compared to SoA is its universal, flexible, and versatile nature
  - Universal: the same hardware can be used with different code lengths, rates, channels
  - Flexible: the code rate can be adjusted readily to any number between 0 and 1
  - Versatile: can be used in multi-terminal coding scenarios

Summary

- With list-decoding and CRC polar codes deliver comparable performance to LDPC and Turbo codes used in present wireless standards
- SoA in coding is already close to theoretical limits for low-order modulation, leaving little margin for improvement
- The biggest asset of polar coding compared to SoA is its universal, flexible, and versatile nature
  - Universal: the same hardware can be used with different code lengths, rates, channels
  - Flexible: the code rate can be adjusted readily to any number between 0 and 1
  - Versatile: can be used in multi-terminal coding scenarios

Summary

- With list-decoding and CRC polar codes deliver comparable performance to LDPC and Turbo codes used in present wireless standards
- SoA in coding is already close to theoretical limits for low-order modulation, leaving little margin for improvement
- The biggest asset of polar coding compared to SoA is its universal, flexible, and versatile nature
  - Universal: the same hardware can be used with different code lengths, rates, channels
  - Flexible: the code rate can be adjusted readily to any number between 0 and 1
  - Versatile: can be used in multi-terminal coding scenarios

Summary

- With list-decoding and CRC polar codes deliver comparable performance to LDPC and Turbo codes used in present wireless standards
- SoA in coding is already close to theoretical limits for low-order modulation, leaving little margin for improvement
- The biggest asset of polar coding compared to SoA is its universal, flexible, and versatile nature
  - Universal: the same hardware can be used with different code lengths, rates, channels
  - Flexible: the code rate can be adjusted readily to any number between 0 and 1
  - Versatile: can be used in multi-terminal coding scenarios

Outlook

- There is need for new FEC techniques as we move to 5G scenarios that call for very high spectral efficiencies and advanced multi-user and multi-antenna techniques
- Extensive research is needed before any FEC method can be declared a winner for 5G scenarios; the field is wide open for introducing new techniques
- It is likely that the winner will emerge based on a trade-off between the overall communication performance under a diverse set of application scenarios and a number of implementation metrics such as complexity and energy efficiency

Outlook

- There is need for new FEC techniques as we move to 5G scenarios that call for very high spectral efficiencies and advanced multi-user and multi-antenna techniques
- Extensive research is needed before any FEC method can be declared a winner for 5G scenarios; the field is wide open for introducing new techniques
- It is likely that the winner will emerge based on a trade-off between the overall communication performance under a diverse set of application scenarios and a number of implementation metrics such as complexity and energy efficiency

Outlook

- There is need for new FEC techniques as we move to 5G scenarios that call for very high spectral efficiencies and advanced multi-user and multi-antenna techniques
- Extensive research is needed before any FEC method can be declared a winner for 5G scenarios; the field is wide open for introducing new techniques
- It is likely that the winner will emerge based on a trade-off between the overall communication performance under a diverse set of application scenarios and a number of implementation metrics such as complexity and energy efficiency

Polar Coding

Part 3: Origin of Polar Coding

Prof. Erdal Arıkan

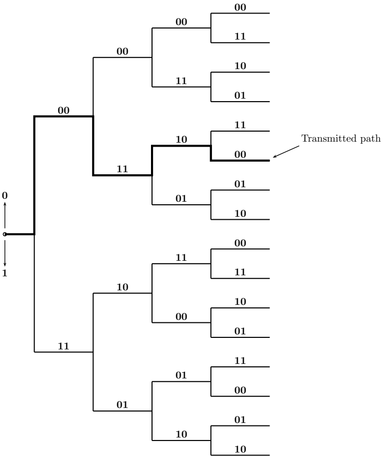
Electrical-Electronics Engineering Department,
Bilkent University, Ankara, Turkey

Indian Institute of Science and Technology,
Bangalore, 27 June - 1 July 2016

Table of Contents

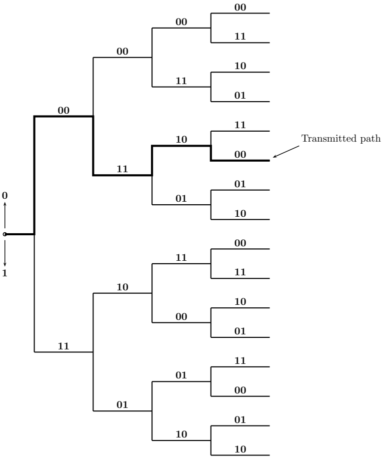
Tree coding and sequential decoding (SD)

- Consider a tree code (of rate $1/2$)
- A path is chosen and transmitted
- Given the channel output, search the tree for the correct (transmitted) path
- The tree structure turns the ML decoding problem into a tree search problem
- A depth-first search algorithm exists called **sequential decoding** (SD)



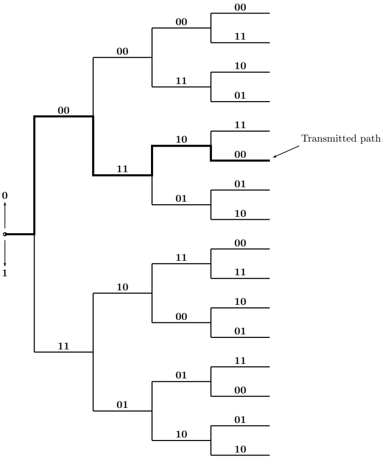
Tree coding and sequential decoding (SD)

- Consider a tree code (of rate 1/2)
- A path is chosen and transmitted
- Given the channel output, search the tree for the correct (transmitted) path
- The tree structure turns the ML decoding problem into a tree search problem
- A depth-first search algorithm exists called **sequential decoding** (SD)



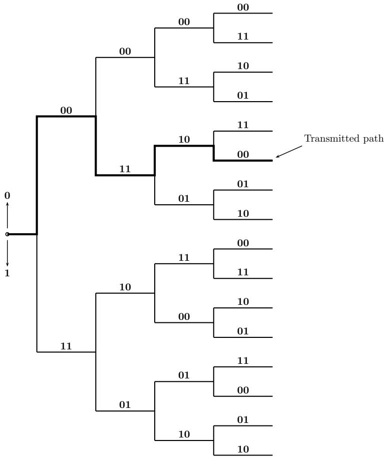
Tree coding and sequential decoding (SD)

- Consider a tree code (of rate 1/2)
- A path is chosen and transmitted
- Given the channel output, search the tree for the correct (transmitted) path
- The tree structure turns the ML decoding problem into a tree search problem
- A depth-first search algorithm exists called **sequential decoding** (SD)



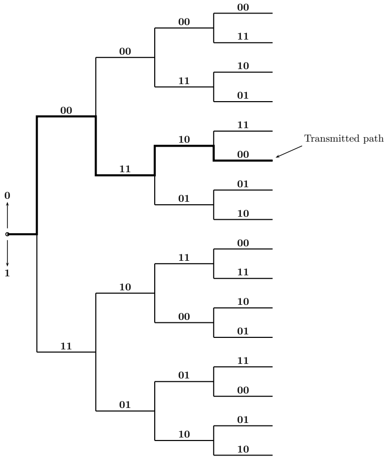
Tree coding and sequential decoding (SD)

- Consider a tree code (of rate 1/2)
- A path is chosen and transmitted
- Given the channel output, search the tree for the correct (transmitted) path
- The tree structure turns the ML decoding problem into a tree search problem
- A depth-first search algorithm exists called **sequential decoding** (SD)



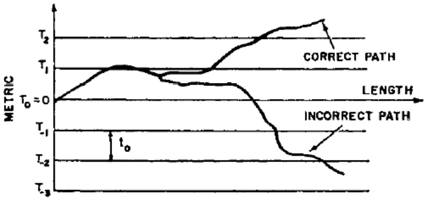
Tree coding and sequential decoding (SD)

- Consider a tree code (of rate 1/2)
- A path is chosen and transmitted
- Given the channel output, search the tree for the correct (transmitted) path
- The tree structure turns the ML decoding problem into a tree search problem
- A depth-first search algorithm exists called **sequential decoding** (SD)



Search metric

SD uses a "metric" to distinguish the correct path from the incorrect ones



Fano's metric:

$$\Gamma(y^n, x^n) = \log \frac{P(y^n|x^n)}{P(y^n)} - nR$$

path length	n
candidate path	x^n
received sequence	y^n
code rate	R

History

- Tree codes were introduced by Elias (1955) with the aim of reducing the complexity of ML decoding (the tree structure makes it possible to use search heuristics for ML decoding)
- Sequential decoding was introduced by Wozencraft (1957) as part of his doctoral thesis
- Fano (1963) simplified the search algorithm and introduced the above metric



History

- Tree codes were introduced by Elias (1955) with the aim of reducing the complexity of ML decoding (the tree structure makes it possible to use search heuristics for ML decoding)
- Sequential decoding was introduced by Wozencraft (1957) as part of his doctoral thesis
- Fano (1963) simplified the search algorithm and introduced the above metric

History

- Tree codes were introduced by Elias (1955) with the aim of reducing the complexity of ML decoding (the tree structure makes it possible to use search heuristics for ML decoding)
- Sequential decoding was introduced by Wozencraft (1957) as part of his doctoral thesis
- Fano (1963) simplified the search algorithm and introduced the above metric



Drift properties of the metric

- On the correct path, the expectation of the metric per channel symbol is

$$\sum_{y,x} p(x,y)\left[\log \frac{p(y|x)}{P(y)} - R\right] = I(X;Y) - R.$$

- On any incorrect path, the expectation is

$$\sum_{x,y} p(x)p(y)\left[\log \frac{p(y|x)}{p(y)} - R\right] \leq -R$$

- A properly designed SD scheme – given enough time – identifies the correct path with probability one at any rate $R < I(X;Y)$.

Drift properties of the metric

- On the correct path, the expectation of the metric per channel symbol is

$$\sum_{y,x} p(x,y)\left[\log \frac{p(y|x)}{P(y)} - R\right] = I(X;Y) - R.$$

- On any incorrect path, the expectation is

$$\sum_{x,y} p(x)p(y)\left[\log \frac{p(y|x)}{p(y)} - R\right] \leq -R$$

- A properly designed SD scheme – given enough time – identifies the correct path with probability one at any rate $R < I(X;Y)$.

Drift properties of the metric

- On the correct path, the expectation of the metric per channel symbol is

$$\sum_{y,x} p(x,y)\left[\log \frac{p(y|x)}{P(y)} - R\right] = I(X;Y) - R.$$

- On any incorrect path, the expectation is

$$\sum_{x,y} p(x)p(y)\left[\log \frac{p(y|x)}{p(y)} - R\right] \leq -R$$

- A properly designed SD scheme – given enough time – identifies the correct path with probability one at any rate $R < I(X;Y)$.

Computation problem in sequential decoding

- Computation in sequential decoding is a random quantity, depending on the code rate R and the noise realization
- Bursts of noise create barriers for the depth-first search algorithm, necessitating excessive backtracking in the search
- Still, the average computation per decoded digit in sequential decoding can be kept bounded provided the code rate R is below the *cutoff rate*

$$R_0 \triangleq -\log \sum_y \left(\sum_x Q(x) \sqrt{W(y|x)} \right)^2$$

- So, SD solves the coding problem for rates below R_0
- Indeed, SD was the method of choice in space communications, albeit briefly

Computation problem in sequential decoding

- Computation in sequential decoding is a random quantity, depending on the code rate R and the noise realization
- Bursts of noise create barriers for the depth-first search algorithm, necessitating excessive backtracking in the search
- Still, the average computation per decoded digit in sequential decoding can be kept bounded provided the code rate R is below the *cutoff rate*

$$R_0 \triangleq -\log \sum_y \left(\sum_x Q(x) \sqrt{W(y|x)} \right)^2$$

- So, SD solves the coding problem for rates below R_0
- Indeed, SD was the method of choice in space communications, albeit briefly

Computation problem in sequential decoding

- Computation in sequential decoding is a random quantity, depending on the code rate R and the noise realization
- Bursts of noise create barriers for the depth-first search algorithm, necessitating excessive backtracking in the search
- Still, the average computation per decoded digit in sequential decoding can be kept bounded provided the code rate R is below the *cutoff rate*

$$R_0 \stackrel{\Delta}{=} -\log \sum_y \left(\sum_x Q(x) \sqrt{W(y|x)} \right)^2$$

- So, SD solves the coding problem for rates below R_0
- Indeed, SD was the method of choice in space communications, albeit briefly

Computation problem in sequential decoding

- Computation in sequential decoding is a random quantity, depending on the code rate R and the noise realization
- Bursts of noise create barriers for the depth-first search algorithm, necessitating excessive backtracking in the search
- Still, the average computation per decoded digit in sequential decoding can be kept bounded provided the code rate R is below the *cutoff rate*

$$R_0 \stackrel{\Delta}{=} -\log \sum_{y} \left(\sum_{x} Q(x) \sqrt{W(y|x)} \right)^2$$

- So, SD solves the coding problem for rates below R_0
- Indeed, SD was the method of choice in space communications, albeit briefly

Computation problem in sequential decoding

- Computation in sequential decoding is a random quantity, depending on the code rate R and the noise realization
- Bursts of noise create barriers for the depth-first search algorithm, necessitating excessive backtracking in the search
- Still, the average computation per decoded digit in sequential decoding can be kept bounded provided the code rate R is below the *cutoff rate*

$$R_0 \stackrel{\Delta}{=} -\log \sum_y \left(\sum_x Q(x) \sqrt{W(y|x)} \right)^2$$

- So, SD solves the coding problem for rates below R_0
- Indeed, SD was the method of choice in space communications, albeit briefly

References on complexity of sequential decoding

- Achievability: Wozencraft (1957), Reiffen (1962), Fano (1963), Stiglitz and Yudkin (1964)
- Converse: Jacobs and Berlekamp (1967)
- Refinements: Wozencraft and Jacobs (1965), Savage (1966), Gallager (1968), Jelinek (1968), Forney (1974), Arıkan (1986), Arıkan (1994)

Sequential decoding and the cutoff rate

Guessing and cutoff rate

Boosting the cutoff rate

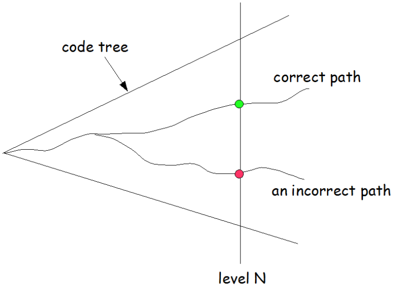
Pinsker's scheme

Massey's scheme

Polar coding

A computational model for sequential decoding

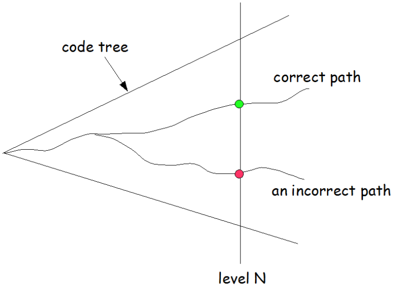
- SD visits nodes at level N in a certain order



- No "look-ahead" assumption: SD forgets what it saw beyond level N upon backtracking
- Complexity measure G_N: The number of nodes searched (visited) at level N until the correct node is visited for the first time

A computational model for sequential decoding

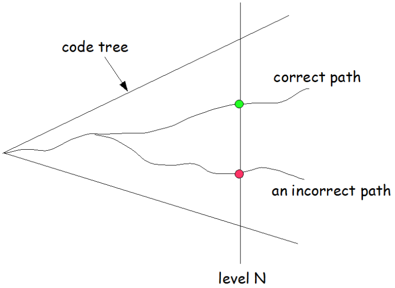
- SD visits nodes at level N in a certain order



- No "look-ahead" assumption: SD forgets what it saw beyond level N upon backtracking
- Complexity measure G_N: The number of nodes searched (visited) at level N until the correct node is visited for the first time

A computational model for sequential decoding

- SD visits nodes at level N in a certain order



- No "look-ahead" assumption: SD forgets what it saw beyond level N upon backtracking
- Complexity measure G_N: The number of nodes searched (visited) at level N until the correct node is visited for the first time

A bound of computational complexity

- Let R be a fixed code rate.
- There exist tree codes of rate R such that

$$E[G_N] \leq 1 + 2^{-N(R_0 - R)}.$$

- Conversely, for any tree code of rate R,

$$E[G_N] \gtrsim 1 + 2^{-N(R_0 - R)}$$

A bound of computational complexity

- Let R be a fixed code rate.
- There exist tree codes of rate R such that

$$E[G_N] \leq 1 + 2^{-N(R_0 - R)}.$$

- Conversely, for any tree code of rate R,

$$E[G_N] \gtrsim 1 + 2^{-N(R_0 - R)}$$

A bound of computational complexity

- Let R be a fixed code rate.
- There exist tree codes of rate R such that

$$E[G_N] \leq 1 + 2^{-N(R_0 - R)}.$$

- Conversely, for any tree code of rate R,

$$E[G_N] \gtrsim 1 + 2^{-N(R_0 - R)}$$

The Guessing Problem

- Alice draws a sample of a random variable $X \sim P$.
- Bob wishes to determine X by asking questions of the form

  "Is X equal to x ?"

  which are answered truthfully by Alice.
- Bob's goal is to minimize the expected number of questions until he gets a YES answer.

The Guessing Problem

- Alice draws a sample of a random variable $X \sim P$.
- Bob wishes to determine X by asking questions of the form

  "Is X equal to x ?"

  which are answered truthfully by Alice.
- Bob's goal is to minimize the expected number of questions until he gets a YES answer.

The Guessing Problem

- Alice draws a sample of a random variable $X \sim P$.
- Bob wishes to determine X by asking questions of the form

  "Is X equal to x ?"

  which are answered truthfully by Alice.
- Bob's goal is to minimize the expected number of questions until he gets a YES answer.

Guessing with Side Information

- Alice samples $(X, Y) \sim P(x, y)$.
- Bob observes Y and is to determine X by asking the same type of questions

  "Is X equal to x ?"

- The goal is to minimize the expected number of quesses.

Guessing with Side Information

- Alice samples $(X, Y) \sim P(x, y)$.
- Bob observes Y and is to determine X by asking the same type of questions

  "Is X equal to x ?"

- The goal is to minimize the expected number of quesses.

Guessing with Side Information

- Alice samples $(X, Y) \sim P(x, y)$.
- Bob observes Y and is to determine X by asking the same type of questions

  "Is X equal to x ?"

- The goal is to minimize the expected number of quesses.

Optimal guessing strategies

- Let G be the number of guesses to determine X.
- The expected no of guesses is given by

$$\mathbb{E}[G] = \sum_{x \in \mathcal{X}} P(x)G(x)$$

- A guessing strategy minimizes $\mathbb{E}[G]$ if

$$P(x) > P(x') \Longrightarrow G(x) < G(x').$$

Optimal guessing strategies

- Let G be the number of guesses to determine X.
- The expected no of guesses is given by

$$\mathbb{E}[G] = \sum_{x \in \mathcal{X}} P(x)G(x)$$

- A guessing strategy minimizes $\mathbb{E}[G]$ if

$$P(x) > P(x') \Longrightarrow G(x) < G(x').$$

Optimal guessing strategies

- Let G be the number of guesses to determine X.
- The expected no of guesses is given by

$$\mathbb{E}[G] = \sum_{x \in \mathcal{X}} P(x) G(x)$$

- A guessing strategy minimizes $\mathbb{E}[G]$ if

$$P(x) > P(x') \Longrightarrow G(x) < G(x').$$

Upper bound on guessing effort

For any *optimal* guessing function

$$\mathbb{E}[G^*(X)] \leq \left[\sum_x \sqrt{P(x)}\right]^2$$

Proof.

$$G^*(x) \leq \sum_{\text{all } x'} \sqrt{P(x')/P(x)} = \sum_{i=1}^{M} i p_G(i)$$

$$\mathbb{E}[G^*(X)] \leq \sum_x P(x) \sum_{x'} \sqrt{P(x')/P(x)} = \left[\sum_x \sqrt{P(x)}\right]^2.$$

Lower bound on guessing effort

For any guessing function for a target r.v. X with M possible values,

$$\mathbb{E}[G(X)] \geq (1 + \ln M)^{-1} \left[\sum_x \sqrt{P(x)} \right]^2$$

For the proof we use the following variant of Hölder's inequality.

Lemma

Let a_i, p_i be positive numbers.

$$\sum_i a_i p_i \geq \left[\sum_i a_i^{-1}\right]^{-1} \left[\sum_i \sqrt{p_i}\right]^2.$$

Proof. Let $\lambda = 1/2$ and put $A_i = a_i^{-1}$, $B_i = a_i^{\lambda} p_i^{\lambda}$, in Hölder's inequality

$$\sum_i A_i B_i \leq \left[\sum_i A_i^{1/(1-\lambda)}\right]^{1-\lambda} \left[\sum_i B_i^{1/\lambda}\right]^{\lambda}.$$

Proof of Lower Bound

$$
\begin{aligned}
\mathbb{E}[G(X) &= \sum_{i=1}^{M} i p_G(i) \\
&\geq \left(\sum_{i=1}^{M} 1/i \right)^{-1} \left(\sum_{i=1}^{M} \sqrt{p_G(i)} \right)^2 \\
&= \left(\sum_{i=1}^{M} 1/i \right)^{-1} \left(\sum_{x} \sqrt{P(x)} \right)^2 \\
&\geq (1 + \ln M)^{-1} \left(\sum_{x} \sqrt{P(x)} \right)^2
\end{aligned}
$$

Essense of the inequalities

For any set of real numbers $p_1 \geq p_2 \geq \cdots \geq p_M > 0$,

$$1 \geq \frac{\sum_{i=1}^{M} i\, p_i}{\left[\sum_{i=1}^{M} \sqrt{p_i}\right]^2} \geq (1 + \ln M)^{-1}$$

Guessing Random Vectors

- Let $\mathbf{X} = (X_1, \ldots, X_n) \sim P(x_1, \ldots, x_n)$.
- Guessing $\mathbf{X}$ means asking questions of the form

  $$\text{“Is } \mathbf{X} = \mathbf{x} \text{ ?”}$$

  for possible values $\mathbf{x} = (x_1, \ldots, x_n)$ of $\mathbf{X}$.
- Notice that coordinate-wise probes of the type

  $$\text{“Is } X_i = x_i \text{ ?”}$$

  are not allowed.

Complexity of Vector Guessing

Suppose X_i has M_i possible values, $i = 1, \ldots, n$. Then,

$$1 \;\geq\; \frac{\mathbb{E}[G^*(X_1, \ldots, X_n)]}{\left[\sum_{x_1,\ldots,x_n} \sqrt{P(x_1, \ldots, x_n)}\right]^2} \;\geq\; [1 + \ln(M_1 \cdots M_n)]^{-1}$$

In particular, if $X_1, \ldots, X_n$ are i.i.d. $\sim P$ with a common alphabet $\mathcal{X}$,

$$1 \;\geq\; \frac{\mathbb{E}[G^*(X_1, \ldots, X_n)]}{\left[\sum_{x\in\mathcal{X}} \sqrt{P(x)}\right]^{2n}} \;\geq\; [1 + n \ln |\mathcal{X}|]^{-1}$$

Guessing with Side Information

- (X, Y) a pair of random variables with a joint distribution $P(x, y)$.
- Y known. X to be guessed as before.
- $G(x|y)$ the number of guesses when $X = x$, $Y = y$.

Lower Bound

For any guessing strategy and any $\rho > 0$,

$$\mathbb{E}[G(X|Y)] \geq (1 + \ln M)^{-1} \sum_y \left[\sum_x \sqrt{P(x,y)} \right]^2$$

where M is the number of possible values of X.

Proof.

$$\mathbb{E}[G(X|Y)] = \sum_y P(y) \mathbb{E}[G(X|Y = y)]$$

$$\geq \sum_y P(y)(1 + \ln M)^{-1} \left[\sum_x \sqrt{P(x|y)} \right]^2$$

$$= (1 + \ln M)^{-1} \sum_y \left[\sum_x \sqrt{P(x,y)} \right]^2$$

Upper bound

Optimal guessing functions satisfy

$$\mathbb{E}[G^*(X|Y)] \leq \sum_y \left[\sum_x \sqrt{P(x,y)}\right]^2.$$

Proof.

$$\begin{aligned}
\mathbb{E}[G^*(X|Y)] &= \sum_y P(y) \sum_x P(x|y) G^*(x|y) \\
&\leq \sum_y P(y) \left[\sum_x \sqrt{P(x|y)}\right]^2 \\
&= \sum_y \left[\sum_x \sqrt{P(x,y)}\right]^2.
\end{aligned}$$

Generalization to Random Vectors

For optimal guessing functions, for $\rho > 0$,

$$1 \geq \frac{\mathbb{E}[G^*(X_1, \ldots, X_k | Y_1, \ldots, Y_n)]}{\sum_{y_1,\ldots,y_n} \left[\sum_{x_1,\ldots,x_k} \sqrt{P(x_1, \ldots, x_k, y_1, \ldots, y_n)}\right]^2} \geq [1 + \ln(M_1 \cdots M_k)]^{-1}$$

where M_i denotes the number of possible values of X_i.

A "guessing" decoder

- Consider a block code with M codewords $\mathbf{x}_1, \ldots, \mathbf{x}_M$ of block length N.
- Suppose a codeword is chosen at random and sent over a channel W
- Given the channel output $\mathbf{y}$, a "guessing decoder" decodes by asking questions of the form

  "Is the correct codeword the mth one?"

  to which it receives a truthful YES or NO answer.
- On a NO answer it repeats the question with a new m.
- The complexity C for this decoder is the number of questions until a YES answer.

A "guessing" decoder

- Consider a block code with M codewords $\mathbf{x}_1, \ldots, \mathbf{x}_M$ of block length N.
- Suppose a codeword is chosen at random and sent over a channel W
- Given the channel output $\mathbf{y}$, a "guessing decoder" decodes by asking questions of the form

  "Is the correct codeword the mth one?"

  to which it receives a truthful YES or NO answer.
- On a NO answer it repeats the question with a new m.
- The complexity C for this decoder is the number of questions until a YES answer.

A "guessing" decoder

- Consider a block code with M codewords $\mathbf{x}_1, \ldots, \mathbf{x}_M$ of block length N.
- Suppose a codeword is chosen at random and sent over a channel W
- Given the channel output $\mathbf{y}$, a "guessing decoder" decodes by asking questions of the form

  "Is the correct codeword the mth one?"

  to which it receives a truthful YES or NO answer.
- On a NO answer it repeats the question with a new m.
- The complexity C for this decoder is the number of questions until a YES answer.

A "guessing" decoder

- Consider a block code with M codewords $\mathbf{x}_1, \ldots, \mathbf{x}_M$ of block length N.
- Suppose a codeword is chosen at random and sent over a channel W
- Given the channel output $\mathbf{y}$, a "guessing decoder" decodes by asking questions of the form

  "Is the correct codeword the mth one?"

  to which it receives a truthful YES or NO answer.
- On a NO answer it repeats the question with a new m.
- The complexity C for this decoder is the number of questions until a YES answer.

A "guessing" decoder

- Consider a block code with M codewords $\mathbf{x}_1, \ldots, \mathbf{x}_M$ of block length N.
- Suppose a codeword is chosen at random and sent over a channel W
- Given the channel output $\mathbf{y}$, a "guessing decoder" decodes by asking questions of the form

  "Is the correct codeword the mth one?"

  to which it receives a truthful YES or NO answer.
- On a NO answer it repeats the question with a new m.
- The complexity C for this decoder is the number of questions until a YES answer.

Optimal guessing decoder

An optimal guessing decoder is one that minimizes the expected complexity $E[C]$.
Clearly, $E[C]$ is minimized by generating the guesses in decreasing order of likelihoods $W(\mathbf{y}|\mathbf{x}_m)$.

$$\begin{aligned}
&\mathbf{x}_{i_1} \leftarrow \text{1st guess (the most likely codeword given } \mathbf{y}\text{)}\\
&\mathbf{x}_{i_2} \leftarrow \text{2nd guess (2nd most likely codeword given } \mathbf{y}\text{)}\\
&\vdots\\
&\mathbf{x}_L \leftarrow \text{correct codeword obtained; guessing stops}
\end{aligned}$$

Complexity C equals the number of guesses L

Application to the guessing decoder

- A block code $\mathcal{C} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ with $M = e^{NR}$ codewords of block length N.
- A codeword $\mathbf{X}$ chosen at random and sent over a DMC W.
- Given the channel output vector $\mathbf{Y}$, the decoder guesses $\mathbf{X}$. A special case of guessing with side information where

$$P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = e^{-NR} \prod_{i=1}^{N} W(y_i|x_i), \quad \mathbf{x} \in \mathcal{C}$$

Cutoff rate bound

$$
\begin{aligned}
\mathbb{E}[G^*(\mathbf{X}|\mathbf{Y})] &\geq [1+NR]^{-1} \sum_{\mathbf{y}} \left[\sum_{\mathbf{x}} \sqrt{P(\mathbf{x},\mathbf{y})} \right]^2 \\
&= [1+NR]^{-1}\, e^{NR} \sum_{\mathbf{y}} \left[\sum_{\mathbf{x}} Q_N(\mathbf{x}) \sqrt{W_N(\mathbf{x},\mathbf{y})} \right]^{2N} \\
&\geq [1+NR]^{-1}\, e^{N(R-R_0(W))}
\end{aligned}
$$

where

$$
R_0(W) = \max_Q \left\{ -\ln \sum_y \left[\sum_x Q(x) \sqrt{W(y|x)} \right]^2 \right\}
$$

is the channel *cutoff rate*.

Sequential decoding and the cutoff rate

Guessing and cutoff rate

Boosting the cutoff rate

Pinsker's scheme

Massey's scheme

Polar coding

Boosting the cutoff rate

- It was clear almost from the beginning that R_0 was at best shaky in its role as a limit to practical communications
- There were many attempts to boost the cutoff rate by devising clever schemes for searching a tree
- One striking example is Pinsker's scheme that displayed the strange nature of R_0

Boosting the cutoff rate

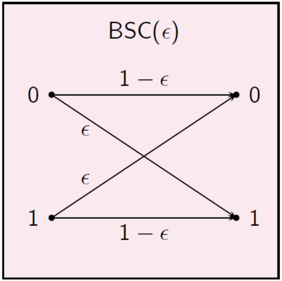
- It was clear almost from the beginning that R_0 was at best shaky in its role as a limit to practical communications
- There were many attempts to boost the cutoff rate by devising clever schemes for searching a tree
- One striking example is Pinsker's scheme that displayed the strange nature of R_0

Boosting the cutoff rate

- It was clear almost from the beginning that R_0 was at best shaky in its role as a limit to practical communications
- There were many attempts to boost the cutoff rate by devising clever schemes for searching a tree
- One striking example is Pinsker's scheme that displayed the strange nature of R_0

Binary Symmetric Channel

We will describe Pinsker's scheme using the BSC example:



- Capacity

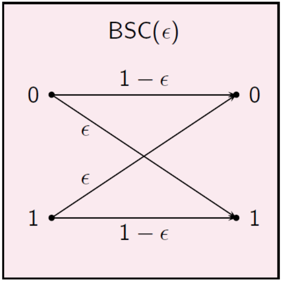
$$C = 1 + \epsilon \log_2(\epsilon) + (1-\epsilon)\log_2(1-\epsilon)$$

- Cutoff rate

$$R_0 = \log_2 \frac{2}{1 + 2\sqrt{\epsilon(1-\epsilon)}}$$

Binary Symmetric Channel

We will describe Pinsker's scheme using the BSC example:



- Capacity

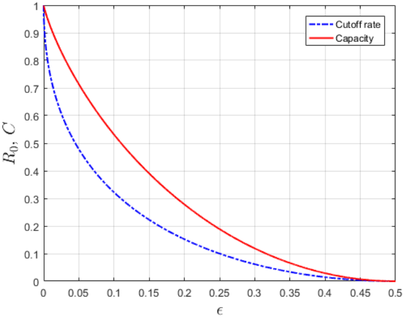
$$C = 1 + \epsilon \log_2(\epsilon) + (1-\epsilon)\log_2(1-\epsilon)$$

- Cutoff rate

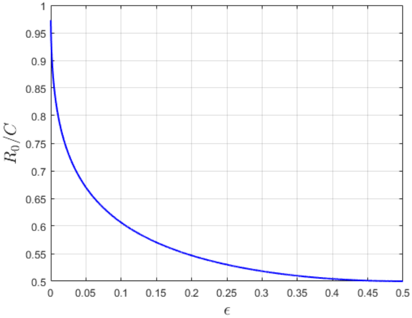
$$R_0 = \log_2 \frac{2}{1 + 2\sqrt{\epsilon(1-\epsilon)}}$$

Capacity and cutoff rate for the BSC

R_0 and C



R_0/C



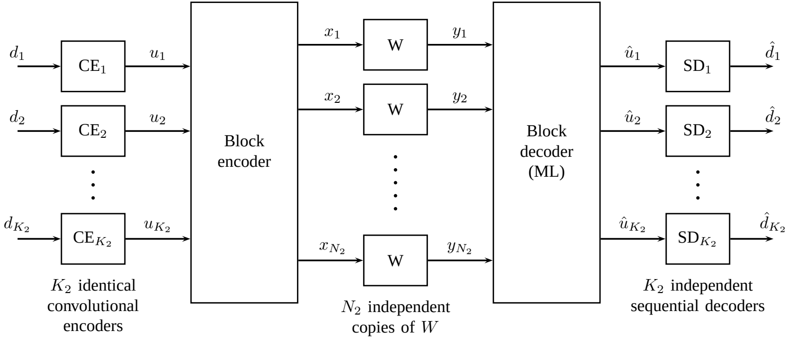
Pinsker's scheme

Based on the observations that as $\epsilon \to 0$

$$\frac{R_0(\epsilon)}{C(\epsilon)} \to 1 \quad \text{and} \quad R_0(\epsilon) \to 1,$$

Pinsker (1965) proposed concatenation scheme that achieved capacity within constant average cost per decoded bit irrespective of the level of reliability

Pinsker's scheme



The inner block code does the initial clean-up at huge but finite complexity; the outer convolutional encoding (CE) and sequential decoding (SD) boosts the reliability at little extra cost.

Discussion

- Although Pinsker's scheme made a very strong theoretical point, it was not practical.
- There were many more attempts to go around the R_0 barrier in 1960s:
  - D. Falconer, "A Hybrid Sequential and Algebraic Decoding Scheme," Sc.D. thesis, Dept. of Elec. Eng., M.I.T., 1966.
  - I. Stiglitz, Iterative sequential decoding, IEEE Transactions on Information Theory, vol. 15, no. 6, pp. 715721, Nov. 1969.
  - F. Jelinek and J. Cocke, "Bootstrap hybrid decoding for symmetrical binary input channels," Inform. Contr., vol. 18, no. 3, pp. 261-298, Apr. 1971.
- It is fair to say that none of these schemes had any practical impact

Discussion

- Although Pinsker's scheme made a very strong theoretical point, it was not practical.
- There were many more attempts to go around the R_0 barrier in 1960s:
  - D. Falconer, "A Hybrid Sequential and Algebraic Decoding Scheme," Sc.D. thesis, Dept. of Elec. Eng., M.I.T., 1966.
  - I. Stiglitz, Iterative sequential decoding, IEEE Transactions on Information Theory, vol. 15, no. 6, pp. 715721, Nov. 1969.
  - F. Jelinek and J. Cocke, "Bootstrap hybrid decoding for symmetrical binary input channels," Inform. Contr., vol. 18, no. 3, pp. 261-298, Apr. 1971.
- It is fair to say that none of these schemes had any practical impact

Discussion

- Although Pinsker's scheme made a very strong theoretical point, it was not practical.
- There were many more attempts to go around the R_0 barrier in 1960s:
  - D. Falconer, "A Hybrid Sequential and Algebraic Decoding Scheme," Sc.D. thesis, Dept. of Elec. Eng., M.I.T., 1966.
  - I. Stiglitz, Iterative sequential decoding, IEEE Transactions on Information Theory, vol. 15, no. 6, pp. 715721, Nov. 1969.
  - F. Jelinek and J. Cocke, "Bootstrap hybrid decoding for symmetrical binary input channels," Inform. Contr., vol. 18, no. 3, pp. 261-298, Apr. 1971.
- It is fair to say that none of these schemes had any practical impact

Discussion

- Although Pinsker's scheme made a very strong theoretical point, it was not practical.
- There were many more attempts to go around the R_0 barrier in 1960s:
  - D. Falconer, "A Hybrid Sequential and Algebraic Decoding Scheme," Sc.D. thesis, Dept. of Elec. Eng., M.I.T., 1966.
  - I. Stiglitz, Iterative sequential decoding, IEEE Transactions on Information Theory, vol. 15, no. 6, pp. 715721, Nov. 1969.
  - F. Jelinek and J. Cocke, "Bootstrap hybrid decoding for symmetrical binary input channels," Inform. Contr., vol. 18, no. 3, pp. 261-298, Apr. 1971.
- It is fair to say that none of these schemes had any practical impact

Discussion

- Although Pinsker's scheme made a very strong theoretical point, it was not practical.
- There were many more attempts to go around the R_0 barrier in 1960s:
  - D. Falconer, "A Hybrid Sequential and Algebraic Decoding Scheme," Sc.D. thesis, Dept. of Elec. Eng., M.I.T., 1966.
  - I. Stiglitz, Iterative sequential decoding, IEEE Transactions on Information Theory, vol. 15, no. 6, pp. 715721, Nov. 1969.
  - F. Jelinek and J. Cocke, "Bootstrap hybrid decoding for symmetrical binary input channels," Inform. Contr., vol. 18, no. 3, pp. 261-298, Apr. 1971.
- It is fair to say that none of these schemes had any practical impact

Discussion

- Although Pinsker's scheme made a very strong theoretical point, it was not practical.
- There were many more attempts to go around the R_0 barrier in 1960s:
  - D. Falconer, "A Hybrid Sequential and Algebraic Decoding Scheme," Sc.D. thesis, Dept. of Elec. Eng., M.I.T., 1966.
  - I. Stiglitz, Iterative sequential decoding, IEEE Transactions on Information Theory, vol. 15, no. 6, pp. 715721, Nov. 1969.
  - F. Jelinek and J. Cocke, "Bootstrap hybrid decoding for symmetrical binary input channels," Inform. Contr., vol. 18, no. 3, pp. 261-298, Apr. 1971.
- It is fair to say that none of these schemes had any practical impact

R_0 as practical capacity

- The failure to beat the cutoff rate bound in a meaningful manner despite intense efforts elevated R_0 to the status of a "realistic" limit to reliable communications
- R_0 appears as the key figure-of-merit for communication system design in the influential works of the period:
  - Wozencraft and Jacobs, *Principles of Communication Engineering*, 1965
  - Wozencraft and Kennedy, "Modulation and demodulation for probabilistic coding," IT Trans.,1966
  - Massey, "Coding and modulation in digital communications," Zürich, 1974
- Forney (1995) gives a first-hand account of this situation in his Shannon Lecture "Performance and Complexity"

R_0 as practical capacity

- The failure to beat the cutoff rate bound in a meaningful manner despite intense efforts elevated R_0 to the status of a "realistic" limit to reliable communications
- R_0 appears as the key figure-of-merit for communication system design in the influential works of the period:
  - Wozencraft and Jacobs, *Principles of Communication Engineering*, 1965
  - Wozencraft and Kennedy, "Modulation and demodulation for probabilistic coding," IT Trans.,1966
  - Massey, "Coding and modulation in digital communications," Zürich, 1974
- Forney (1995) gives a first-hand account of this situation in his Shannon Lecture "Performance and Complexity"

R_0 as practical capacity

- The failure to beat the cutoff rate bound in a meaningful manner despite intense efforts elevated R_0 to the status of a "realistic" limit to reliable communications
- R_0 appears as the key figure-of-merit for communication system design in the influential works of the period:
  - Wozencraft and Jacobs, *Principles of Communication Engineering*, 1965
  - Wozencraft and Kennedy, "Modulation and demodulation for probabilistic coding," IT Trans.,1966
  - Massey, "Coding and modulation in digital communications," Zürich, 1974
- Forney (1995) gives a first-hand account of this situation in his Shannon Lecture "Performance and Complexity"

R_0 as practical capacity

- The failure to beat the cutoff rate bound in a meaningful manner despite intense efforts elevated R_0 to the status of a "realistic" limit to reliable communications
- R_0 appears as the key figure-of-merit for communication system design in the influential works of the period:
  - Wozencraft and Jacobs, *Principles of Communication Engineering*, 1965
  - Wozencraft and Kennedy, "Modulation and demodulation for probabilistic coding," IT Trans.,1966
  - Massey, "Coding and modulation in digital communications," Zürich, 1974
- Forney (1995) gives a first-hand account of this situation in his Shannon Lecture "Performance and Complexity"

R_0 as practical capacity

- The failure to beat the cutoff rate bound in a meaningful manner despite intense efforts elevated R_0 to the status of a "realistic" limit to reliable communications
- R_0 appears as the key figure-of-merit for communication system design in the influential works of the period:
  - Wozencraft and Jacobs, *Principles of Communication Engineering*, 1965
  - Wozencraft and Kennedy, "Modulation and demodulation for probabilistic coding," IT Trans.,1966
  - Massey, "Coding and modulation in digital communications," Zürich, 1974
- Forney (1995) gives a first-hand account of this situation in his Shannon Lecture "Performance and Complexity"

R_0 as practical capacity

- The failure to beat the cutoff rate bound in a meaningful manner despite intense efforts elevated R_0 to the status of a "realistic" limit to reliable communications
- R_0 appears as the key figure-of-merit for communication system design in the influential works of the period:
  - Wozencraft and Jacobs, *Principles of Communication Engineering*, 1965
  - Wozencraft and Kennedy, "Modulation and demodulation for probabilistic coding," IT Trans.,1966
  - Massey, "Coding and modulation in digital communications," Zürich, 1974
- Forney (1995) gives a first-hand account of this situation in his Shannon Lecture "Performance and Complexity"

Other attempts to boost the cutoff rate

Efforts to beat the cutoff rate continues to this day

- D. J. Costello and F. Jelinek, 1972.
- P. R. Chevillat and D. J. Costello Jr., 1977.
- F. Hemmati, 1990.
- B. Radosavljevic, E. Arıkan, B. Hajek, 1992.
- J. Belzile and D. Haccoun, 1993.
- S. Kallel and K. Li, 1997.
- E. Arıkan, 2006
- ...

Other attempts to boost the cutoff rate

Efforts to beat the cutoff rate continues to this day

- D. J. Costello and F. Jelinek, 1972.
- P. R. Chevillat and D. J. Costello Jr., 1977.
- F. Hemmati, 1990.
- B. Radosavljevic, E. Arıkan, B. Hajek, 1992.
- J. Belzile and D. Haccoun, 1993.
- S. Kallel and K. Li, 1997.
- E. Arıkan, 2006
- ...

In fact, polar coding originates from such attempts.

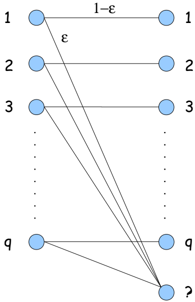
The R_0 debate

A case study by McEliece (1980) cast a big doubt on the significance of R_0 as a practical limit

- McEliece's study was concerned with a Pulse Position Modulation (PPM) scheme, modeled as a q-ary erasure channel
- Capacity: $C(q) = (1 - \epsilon) \log q$
- Cutoff rate: $R_0(q) = \log \frac{q}{1+(q-1)\epsilon}$
- As the bandwidth (q) grew,

$$\frac{R_0(q)}{C(q)} \to 0$$

- Algebraic coding (Reed-Solomon) scored a big win over probabilistic coding!



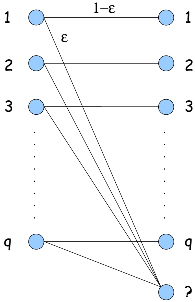
The R_0 debate

A case study by McEliece (1980) cast a big doubt on the significance of R_0 as a practical limit

- McEliece's study was concerned with a Pulse Position Modulation (PPM) scheme, modeled as a q-ary erasure channel
- Capacity: $C(q) = (1-\epsilon)\log q$
- Cutoff rate: $R_0(q) = \log \frac{q}{1+(q-1)\epsilon}$
- As the bandwidth (q) grew,

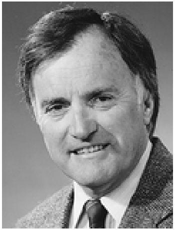
$$\frac{R_0(q)}{C(q)} \to 0$$

- Algebraic coding (Reed-Solomon) scored a big win over probabilistic coding!



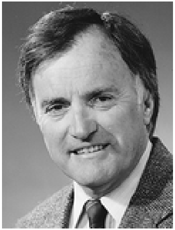
Massey meets the challenge

- Massey (1981) showed that there was a different way of doing coding and modulation on a q-ary erasure channel that boosted R_0 effortlessly
- Paradoxically, as Massey restored the status of R_0, he exhibited the "flaky" nature of this parameter

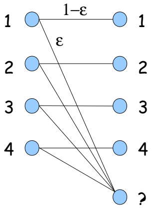


Massey meets the challenge

- Massey (1981) showed that there was a different way of doing coding and modulation on a q-ary erasure channel that boosted R_0 effortlessly
- Paradoxically, as Massey restored the status of R_0, he exhibited the "flaky" nature of this parameter

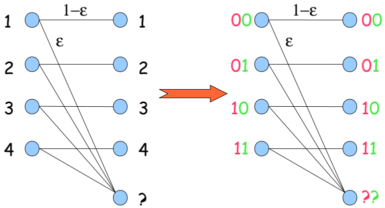


Channel splitting to boost cutoff rate (Massey, 1981)



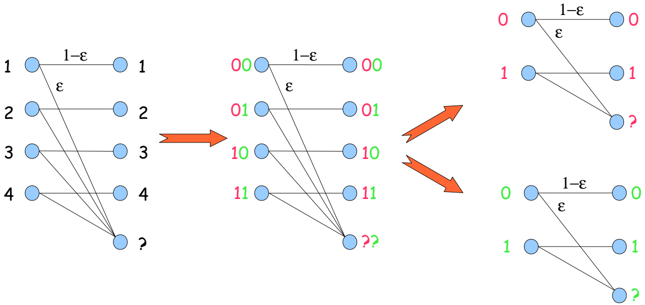
- Begin with a quaternary erasure channel (QEC)

Channel splitting to boost cutoff rate (Massey, 1981)



- Relabel the inputs

Channel splitting to boost cutoff rate (Massey, 1981)

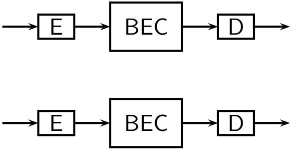


- Split the QEC into two binary erasure channels (BEC)
- BECs fully correlated: erasures occur jointly

Capacity, cutoff rate for one QEC vs two BECs

Ordinary coding of QEC

Independent coding of BECs



$C(\text{QEC}) = 2(1-\epsilon)$

$R_0(\text{QEC}) = \log \frac{4}{1+3\epsilon}$

$C(\text{BEC}) = (1-\epsilon)$

$R_0(\text{BEC}) = \log \frac{2}{1+\epsilon}$

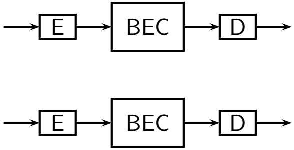
Capacity, cutoff rate for one QEC vs two BECs

Ordinary coding of QEC

$C(\text{QEC}) = 2(1 - \epsilon)$

$R_0(\text{QEC}) = \log \frac{4}{1+3\epsilon}$

Independent coding of BECs



$C(\text{BEC}) = (1 - \epsilon)$

$R_0(\text{BEC}) = \log \frac{2}{1+\epsilon}$

- $C(\text{QEC}) = 2 \times C(\text{BEC})$

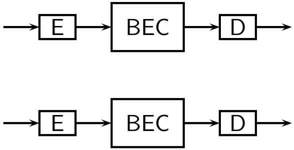
Capacity, cutoff rate for one QEC vs two BECs

Ordinary coding of QEC

$$C(\text{QEC}) = 2(1-\epsilon)$$

$$R_0(\text{QEC}) = \log \frac{4}{1+3\epsilon}$$

Independent coding of BECs

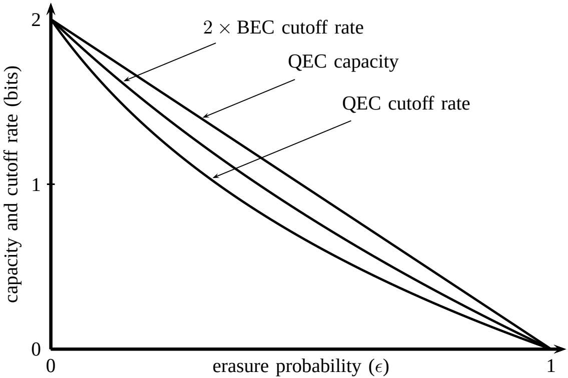


$$C(\text{BEC}) = (1-\epsilon)$$

$$R_0(\text{BEC}) = \log \frac{2}{1+\epsilon}$$

- $C(\text{QEC}) = 2 \times C(\text{BEC})$
- $R_0(\text{QEC}) \leq 2 \times R_0(\text{BEC})$ with equality iff $\epsilon = 0$ or 1.

Cutoff rate improvement by splitting



Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

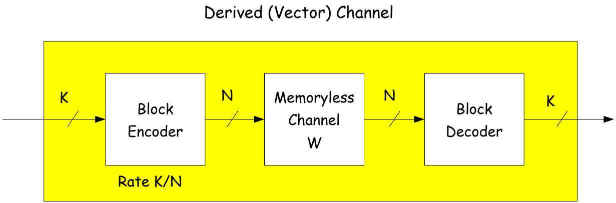
Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

Comparison of Pinsker's and Massey's schemes

- Pinsker
  - Construct a superchannel by combining independent copies of a given DMC W
  - Split the superchannel into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Can be used universally
  - Can achieve capacity
  - Not practical
- Massey
  - Split the given DMC W into correlated subchannels
  - Ignore correlations between the subchannels, encode and decode them independently
  - Applicable only to specific channels
  - Cannot achieve capacity
  - Practical

A conservation law for the cutoff rate

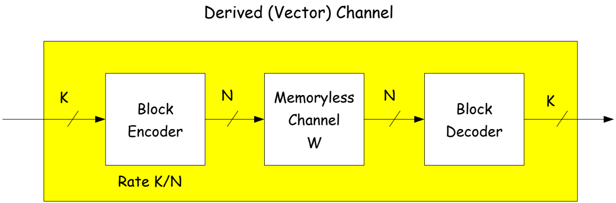


- "Parallel channels" theorem (Gallager, 1965)

$$R_0(\text{Derived vector channel}) \leq N\, R_0(W)$$

- "Cleaning up" the channel by pre-/post-processing can only hurt R_0
- Shows that boosting cutoff rate requires more than one sequential decoder

A conservation law for the cutoff rate

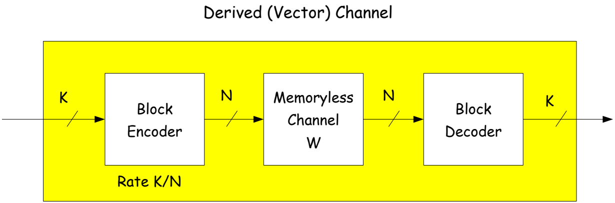


- "Parallel channels" theorem (Gallager, 1965)

$$R_0(\text{Derived vector channel}) \leq N\, R_0(W)$$

- "Cleaning up" the channel by pre-/post-processing can only hurt R_0
- Shows that boosting cutoff rate requires more than one sequential decoder

A conservation law for the cutoff rate

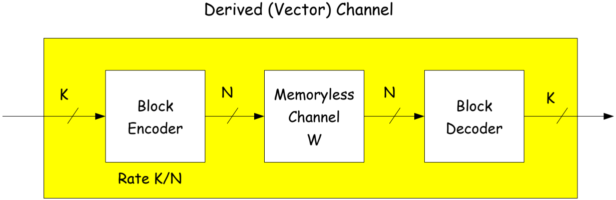


- "Parallel channels" theorem (Gallager, 1965)

$$R_0(\text{Derived vector channel}) \leq N\, R_0(W)$$

- "Cleaning up" the channel by pre-/post-processing can only hurt R_0
- Shows that boosting cutoff rate requires more than one sequential decoder

A conservation law for the cutoff rate



- "Parallel channels" theorem (Gallager, 1965)

$$R_0(\text{Derived vector channel}) \leq N\, R_0(W)$$

- "Cleaning up" the channel by pre-/post-processing can only hurt R_0
- Shows that boosting cutoff rate requires more than one sequential decoder

Sequential decoding and the cutoff rate

Guessing and cutoff rate

Boosting the cutoff rate

Pinsker's scheme

Massey's scheme

Polar coding

Recap of Part 1

- There is a decoding algorithm for tree codes called *sequential decoding* that more or less solves the coding problem for rates below a certain cutoff rate R_0
- As Pinsker's example and Massey's example showed, cutoff rate is not a conserved quantity
- It can be boosted by employing independent sequential decoding across artificially created correlated bit channels

Recap of Part 1

- There is a decoding algorithm for tree codes called *sequential decoding* that more or less solves the coding problem for rates below a certain cutoff rate R_0
- As Pinsker's example and Massey's example showed, cutoff rate is not a conserved quantity
- It can be boosted by employing independent sequential decoding across artificially created correlated bit channels

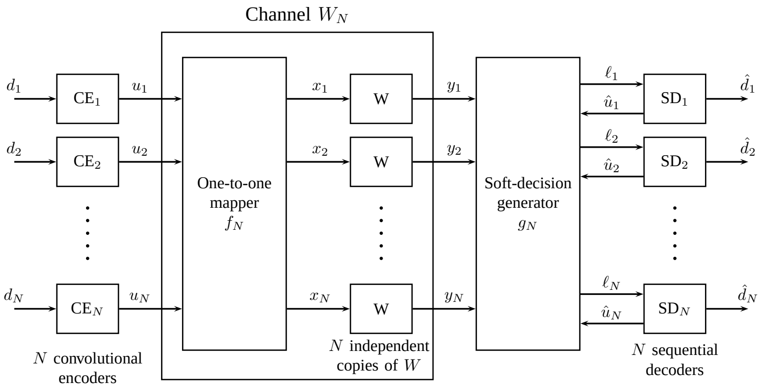
Recap of Part 1

- There is a decoding algorithm for tree codes called *sequential decoding* that more or less solves the coding problem for rates below a certain cutoff rate R_0
- As Pinsker's example and Massey's example showed, cutoff rate is not a conserved quantity
- It can be boosted by employing independent sequential decoding across artificially created correlated bit channels

Prescription for a new scheme

- Consider small constructions
- Retain independent encoding for the subchannels
- Do not ignore correlations between subchannels at the expense of capacity
- This points to multi-level coding and successive cancellation decoding

Multi-stage decoding architecture



Prescription for a new scheme

- Consider small constructions
- Retain independent encoding for the subchannels
- Do not ignore correlations between subchannels at the expense of capacity
- This points to multi-level coding and successive cancellation decoding

Notation

- Let $V : \mathbb{F}_2 \triangleq \{0,1\} \to \mathcal{Y}$ be an arbitrary binary-input memoryless channel
- Let (X, Y) be an input-output ensemble for channel V with X uniform on $\mathbb{F}_2$
- The (symmetric) capacity is defined as

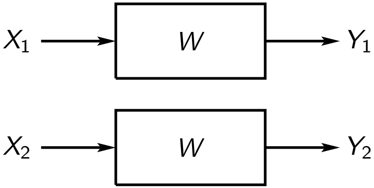
$$I(V) \triangleq I(X;Y) \triangleq \sum_{y \in \mathcal{Y}} \sum_{x \in \mathbb{F}_2} \tfrac{1}{2} V(y|x) \log \frac{V(y|x)}{\tfrac{1}{2} V(y|0) + \tfrac{1}{2} V(y|1)}$$

- The (symmetric) cutoff rate is defined as

$$R_0(V) \triangleq R_0(X;Y) \triangleq -\log \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathbb{F}_2} \tfrac{1}{2} \sqrt{V(y|x)} \right]^2$$

The basic construction

Given two copies of a binary input channel $W : \mathbb{F}_2 \triangleq \{0, 1\} \to \mathcal{Y}$



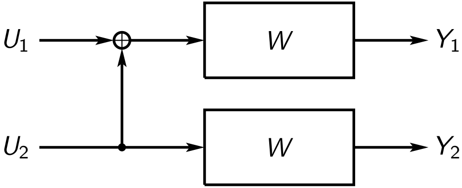
consider the transformation above to generate two channels $W^- : F_2 \to \mathcal{Y}^2$ and $W^+ : F_2 \to \mathcal{Y}^2 \times F_2$ with

$$W^-(y_1 y_2 | u_1) = \sum_{u_2} \tfrac{1}{2} W(y_1 | u_1 + u_2) W(y_2 | u_2)$$

$$W^+(y_1 y_2 u_1 | u_2) = \tfrac{1}{2} W(y_1 | u_1 + u_2) W(y_2 | u_2)$$

The basic construction

Given two copies of a binary input channel $W : \mathbb{F}_2 \triangleq \{0,1\} \to \mathcal{Y}$



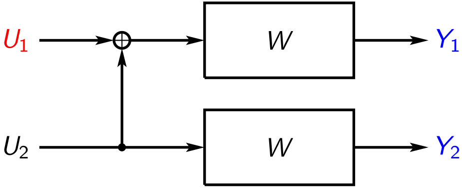
consider the transformation above to generate two channels $W^- : F_2 \to \mathcal{Y}^2$ and $W^+ : F_2 \to \mathcal{Y}^2 \times F_2$ with

$$W^-(y_1 y_2 | u_1) = \sum_{u_2} \frac{1}{2} W(y_1 | u_1 + u_2) W(y_2 | u_2)$$

$$W^+(y_1 y_2 u_1 | u_2) = \frac{1}{2} W(y_1 | u_1 + u_2) W(y_2 | u_2)$$

The basic construction

Given two copies of a binary input channel $W : \mathbb{F}_2 \triangleq \{0,1\} \to \mathcal{Y}$



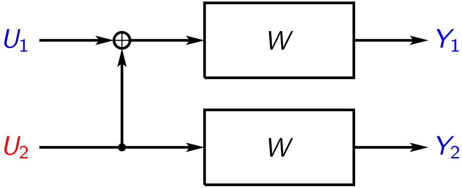
consider the transformation above to generate two channels $W^- : F_2 \to \mathcal{Y}^2$ and $W^+ : F_2 \to \mathcal{Y}^2 \times F_2$ with

$$W^-(y_1 y_2 | u_1) = \sum_{u_2} \frac{1}{2} W(y_1 | u_1 + u_2) W(y_2 | u_2)$$

$$W^+(y_1 y_2 u_1 | u_2) = \frac{1}{2} W(y_1 | u_1 + u_2) W(y_2 | u_2)$$

The basic construction

Given two copies of a binary input channel $W : \mathbb{F}_2 \triangleq \{0,1\} \to \mathcal{Y}$



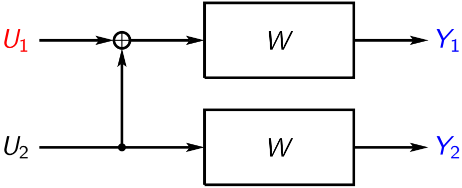
consider the transformation above to generate two channels $W^- : F_2 \to \mathcal{Y}^2$ and $W^+ : F_2 \to \mathcal{Y}^2 \times F_2$ with

$$W^-(y_1 y_2 | u_1) = \sum_{u_2} \frac{1}{2} W(y_1 | u_1 + u_2) W(y_2 | u_2)$$

$$W^+(y_1 y_2 u_1 | u_2) = \frac{1}{2} W(y_1 | u_1 + u_2) W(y_2 | u_2)$$

The basic construction

Given two copies of a binary input channel $W : \mathbb{F}_2 \stackrel{\Delta}{=} \{0,1\} \to \mathcal{Y}$



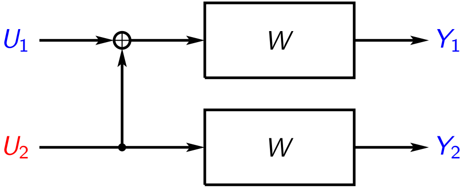
consider the transformation above to generate two channels $W^- : F_2 \to \mathcal{Y}^2$ and $W^+ : F_2 \to \mathcal{Y}^2 \times F_2$ with

$$W^-(y_1 y_2 | u_1) = \sum_{u_2} \tfrac{1}{2} W(y_1 | u_1 + u_2) W(y_2 | u_2)$$

$$W^+(y_1 y_2 u_1 | u_2) = \tfrac{1}{2} W(y_1 | u_1 + u_2) W(y_2 | u_2)$$

The basic construction

Given two copies of a binary input channel $W : \mathbb{F}_2 \stackrel{\Delta}{=} \{0,1\} \to \mathcal{Y}$



consider the transformation above to generate two channels $W^- : F_2 \to \mathcal{Y}^2$ and $W^+ : F_2 \to \mathcal{Y}^2 \times F_2$ with

$$W^-(y_1 y_2|u_1) = \sum_{u_2} \tfrac{1}{2} W(y_1|u_1+u_2)W(y_2|u_2)$$

$$W^+(y_1 y_2 u_1|u_2) = \tfrac{1}{2} W(y_1|u_1+u_2)W(y_2|u_2)$$

The 2x2 transformation is information lossless

- With independent, uniform U_1, U_2,

$$I(W^-) = I(U_1; Y_1 Y_2),$$
$$I(W^+) = I(U_2; Y_1 Y_2 U_1).$$

- Thus,

$$\begin{aligned} I(W^-) + I(W^+) &= I(U_1 U_2; Y_1 Y_2) \\ &= 2I(W), \end{aligned}$$

- and $I(W^-) \leq I(W) \leq I(W^+)$.

The 2x2 transformation "creates" cutoff rate

With independent, uniform U_1, U_2,

$$R_0(W^-) = R_0(U_1; Y_1 Y_2),$$
$$R_0(W^+) = R_0(U_2; Y_1 Y_2 U_1).$$

Theorem (2005)

Correlation helps create cutoff rate:

$$R_0(W^-) + R_0(W^+) \geq 2R_0(W)$$

with equality iff W is a perfect channel, $I(W) = 1$, or a pure noise channel, $I(W) = 0$. Cutoff rates start polarizing:

$$R_0(W^-) \leq R_0(W) \leq R_0(W^+)$$

The 2x2 transformation "creates" cutoff rate

With independent, uniform U_1, U_2,

$$R_0(W^-) = R_0(U_1; Y_1 Y_2),$$
$$R_0(W^+) = R_0(U_2; Y_1 Y_2 U_1).$$

Theorem (2005)

Correlation helps create cutoff rate:

$$R_0(W^-) + R_0(W^+) \geq 2R_0(W)$$

with equality iff W is a perfect channel, $I(W) = 1$, or a pure noise channel, $I(W) = 0$. Cutoff rates start polarizing:

$$R_0(W^-) \leq R_0(W) \leq R_0(W^+)$$

The 2x2 transformation "creates" cutoff rate

With independent, uniform U_1, U_2,

$$R_0(W^-) = R_0(U_1; Y_1 Y_2),$$
$$R_0(W^+) = R_0(U_2; Y_1 Y_2 U_1).$$

Theorem (2005)

Correlation helps create cutoff rate:

$$R_0(W^-) + R_0(W^+) \geq 2R_0(W)$$

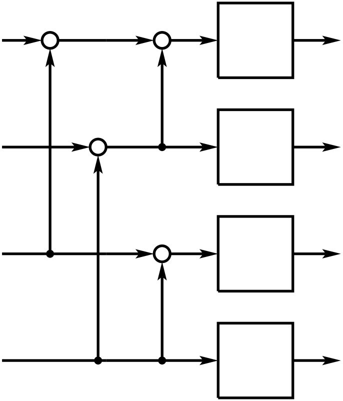
with equality iff W is a perfect channel, $I(W) = 1$, or a pure noise channel, $I(W) = 0$. Cutoff rates start polarizing:

$$R_0(W^-) \leq R_0(W) \leq R_0(W^+)$$

Recursive continuation

Do the same recursively: Given W,

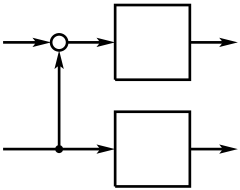
- Duplicate W and obtain W^- and W^+.
- Duplicate W^- (W^+),
- and obtain W^{--} and W^{-+} $(W^{+-}$ and $W^{++})$.
- Duplicate W^{--} $(W^{-+}$, W^{+-}, $W^{++})$ and obtain W^{---} and W^{--+} $(W^{-+-}$, W^{-++}, W^{+--}, W^{+-+}, W^{++-}, $W^{+++})$.
- ...



Recursive continuation

Do the same recursively: Given W,

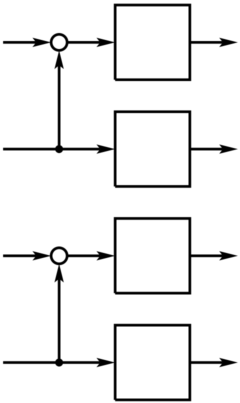
- Duplicate W and obtain W^- and W^+.
- Duplicate W^- (W^+),
- and obtain W^{--} and W^{-+} (W^{+-} and W^{++}).
- Duplicate W^{--} (W^{-+}, W^{+-}, W^{++}) and obtain W^{---} and W^{--+} (W^{-+-}, W^{-++}, W^{+--}, W^{+-+}, W^{++-}, W^{+++}).
- ...



Recursive continuation

Do the same recursively: Given W,

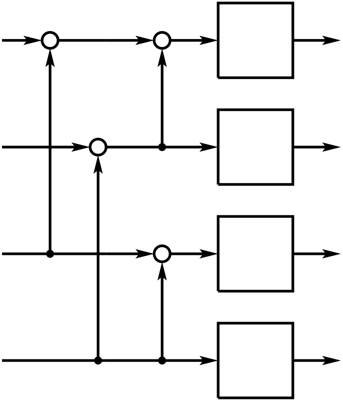
- Duplicate W and obtain W^- and W^+.
- Duplicate W^- (W^+),
- and obtain W^{--} and W^{-+} (W^{+-} and W^{++}).
- Duplicate W^{--} (W^{-+}, W^{+-}, W^{++}) and obtain W^{---} and W^{--+} (W^{-+-}, W^{-++}, W^{+--}, W^{+-+}, W^{++-}, W^{+++}).
- ...



Recursive continuation

Do the same recursively: Given W,

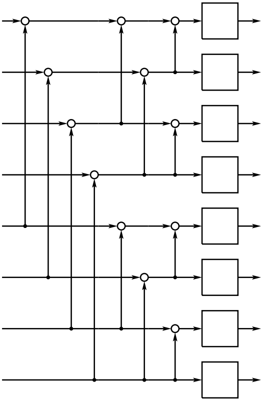
- Duplicate W and obtain W^- and W^+.
- Duplicate W^- (W^+),
- and obtain W^{--} and W^{-+} (W^{+-} and W^{++}).
- Duplicate W^{--} (W^{-+}, W^{+-}, W^{++}) and obtain W^{---} and W^{--+} (W^{-+-}, W^{-++}, W^{+--}, W^{+-+}, W^{++-}, W^{+++}).
- ...



Recursive continuation

Do the same recursively: Given W,

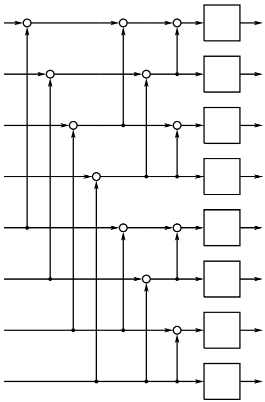
- Duplicate W and obtain W^- and W^+.
- Duplicate W^- (W^+),
- and obtain W^{--} and W^{-+} (W^{+-} and W^{++}).
- Duplicate W^{--} (W^{-+}, W^{+-}, W^{++}) and obtain W^{---} and W^{--+} (W^{-+-}, W^{-++}, W^{+--}, W^{+-+}, W^{++-}, W^{+++}).
- ...



Recursive continuation

Do the same recursively: Given W,

- Duplicate W and obtain W^- and W^+.
- Duplicate W^- (W^+),
- and obtain W^{--} and W^{-+} (W^{+-} and W^{++}).
- Duplicate W^{--} (W^{-+}, W^{+-}, W^{++}) and obtain W^{---} and W^{--+} (W^{-+-}, W^{-++}, W^{+--}, W^{+-+}, W^{++-}, W^{+++}).
- ...



Polarization Process

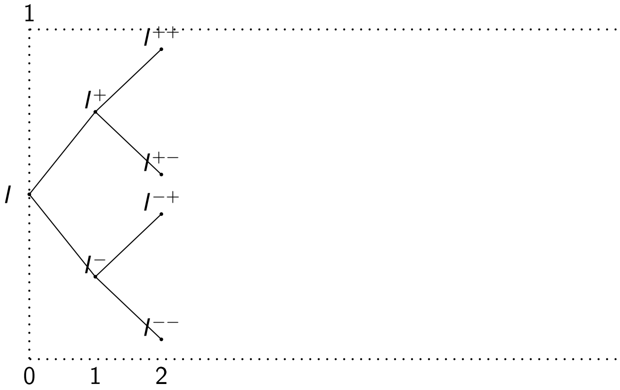
Evolution of $I = I(W)$, $I^+ = I(W^+)$, $I^- = I(W^-)$, etc.

Polarization Process

Evolution of $I = I(W)$, $I^+ = I(W^+)$, $I^- = I(W^-)$, etc.

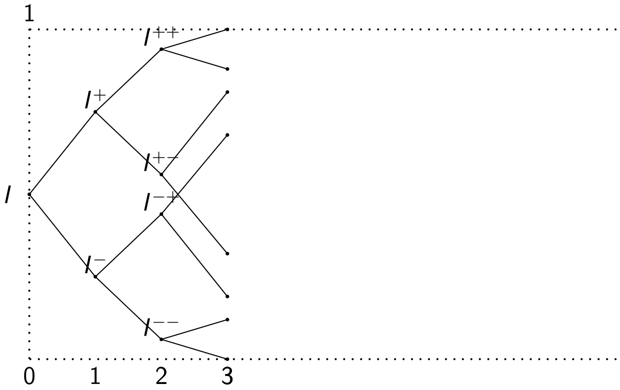
Polarization Process

Evolution of $I = I(W)$, $I^+ = I(W^+)$, $I^- = I(W^-)$, etc.



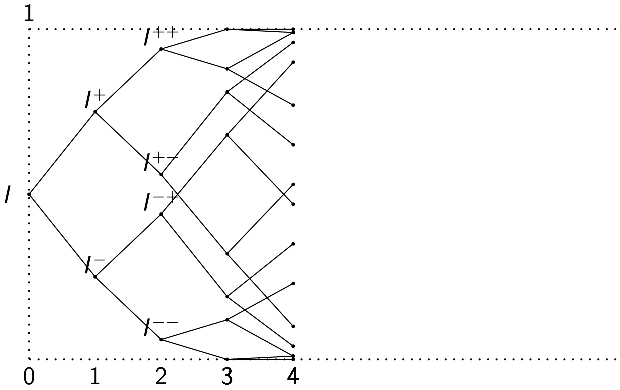
Polarization Process

Evolution of $I = I(W)$, $I^+ = I(W^+)$, $I^- = I(W^-)$, etc.



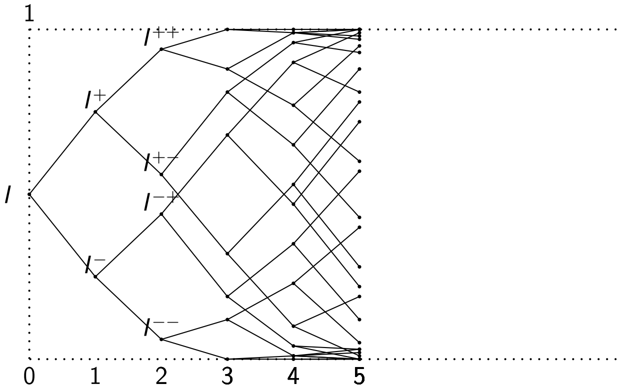
Polarization Process

Evolution of $I = I(W)$, $I^+ = I(W^+)$, $I^- = I(W^-)$, etc.



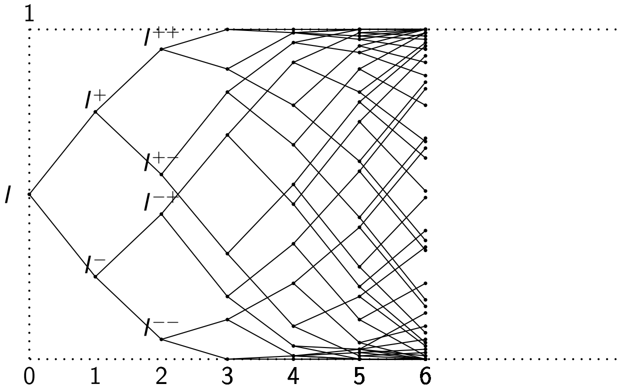
Polarization Process

Evolution of $I = I(W)$, $I^+ = I(W^+)$, $I^- = I(W^-)$, etc.



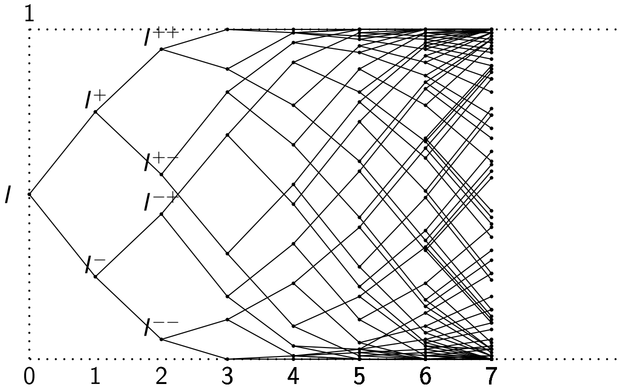
Polarization Process

Evolution of $I = I(W)$, $I^+ = I(W^+)$, $I^- = I(W^-)$, etc.



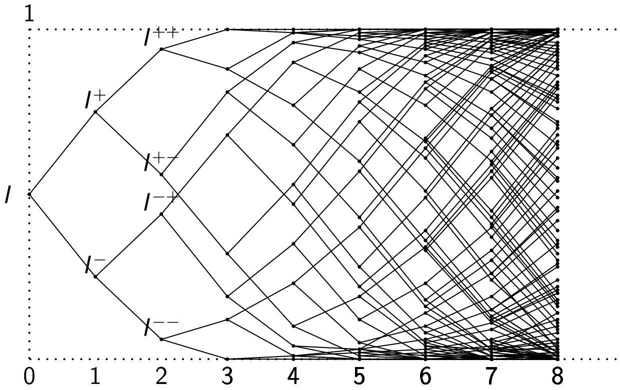
Polarization Process

Evolution of $I = I(W)$, $I^+ = I(W^+)$, $I^- = I(W^-)$, etc.



Polarization Process

Evolution of $I = I(W)$, $I^+ = I(W^+)$, $I^- = I(W^-)$, etc.



Cutoff Rate Polarization

Theorem (2006)

The cutoff rates $\{R_0(U_i; Y^N U^{i-1})\}$ *of the channels created by the recursive transformation converge to their extremal values,* *i.e.,*

$$\frac{1}{N}\#\{i : R_0(U_i; Y^N U^{i-1}) \approx 1\} \to I(W)$$

and

$$\frac{1}{N}\#\{i : R_0(U_i; Y^N U^{i-1}) \approx 0\} \to 1 - I(W).$$

Remark: $\{I(U_i; Y^N U^{i-1})\}$ also polarize.

Cutoff Rate Polarization

Theorem (2006)

The cutoff rates $\{R_0(U_i; Y^N U^{i-1})\}$ *of the channels created by the recursive transformation converge to their extremal values, i.e.,*

$$\frac{1}{N}\#\{i : R_0(U_i; Y^N U^{i-1}) \approx 1\} \to I(W)$$

and

$$\frac{1}{N}\#\{i : R_0(U_i; Y^N U^{i-1}) \approx 0\} \to 1 - I(W).$$

Remark: $\{I(U_i; Y^N U^{i-1})\}$ also polarize.

Cutoff Rate Polarization

Theorem (2006)

The cutoff rates $\{R_0(U_i; Y^N U^{i-1})\}$ *of the channels created by the recursive transformation converge to their extremal values, i.e.,*

$$\frac{1}{N}\#\{i : R_0(U_i; Y^N U^{i-1}) \approx 1\} \to I(W)$$

and

$$\frac{1}{N}\#\{i : R_0(U_i; Y^N U^{i-1}) \approx 0\} \to 1 - I(W).$$

Remark: $\{I(U_i; Y^N U^{i-1})\}$ also polarize.

Cutoff Rate Polarization

Theorem (2006)

The cutoff rates $\{R_0(U_i; Y^N U^{i-1})\}$ *of the channels created by the recursive transformation converge to their extremal values, i.e.,*

$$\frac{1}{N}\#\{i : R_0(U_i; Y^N U^{i-1}) \approx 1\} \to I(W)$$

and

$$\frac{1}{N}\#\{i : R_0(U_i; Y^N U^{i-1}) \approx 0\} \to 1 - I(W).$$

Remark: $\{I(U_i; Y^N U^{i-1})\}$ also polarize.

Sequential decoding with successive cancellation

- Use the recursive construction to generate N bit-channels with cutoff rates $R_0(U_i; Y^N U^{i-1})$, $1 \leq i \leq N$.
- Encode the bit-channels independently using convolutional coding
- Decode the bit-channels one by one using sequential decoding and successive cancellation
- Achievable sum cutoff rate is

$$\sum_{i=1}^{N} R_0(U_i; Y^N U^{i-1})$$

which approaches $N\,I(W)$ as N increases.

Final step: Doing away with sequential decoding

- Due to polarization, rate loss is negligible if one does not use the “bad” bit-channels
- Rate of polarization is strong enough that a vanishing frame error rate can be achieved even if the “good” bit-channels are used uncoded
- The resulting system has no convolutional encoding and sequential decoding, only successive cancellation decoding

Polar coding

To communicate at rate $R < I(W)$:

- Pick N, and $K = NR$ good indices i such that $I(U_i; Y^N U^{i-1})$ is high,
- let the transmitter set U_i to be uncoded binary data for good indices, and set U_i to random but publicly known values for the rest,
- let the receiver decode the U_i successively: U_1 from Y^N; U_i from $Y^N \hat{U}^{i-1}$.

Polar coding

To communicate at rate $R < I(W)$:

- Pick N, and $K = NR$ good indices i such that $I(U_i; Y^N U^{i-1})$ is high,
- let the transmitter set U_i to be uncoded binary data for good indices, and set U_i to random but publicly known values for the rest,
- let the receiver decode the U_i successively: U_1 from Y^N; U_i from $Y^N \hat{U}^{i-1}$.

Polar coding

To communicate at rate $R < I(W)$:

- Pick N, and $K = NR$ good indices i such that $I(U_i; Y^N U^{i-1})$ is high,
- let the transmitter set U_i to be uncoded binary data for good indices, and set U_i to random but publicly known values for the rest,
- let the receiver decode the U_i successively: U_1 from Y^N; U_i from $Y^N \hat{U}^{i-1}$.

Polar coding

To communicate at rate $R < I(W)$:

- Pick N, and $K = NR$ good indices i such that $I(U_i; Y^N U^{i-1})$ is high,
- let the transmitter set U_i to be uncoded binary data for good indices, and set U_i to random but publicly known values for the rest,
- let the receiver decode the U_i successively: U_1 from Y^N; U_i from $Y^N \hat{U}^{i-1}$.

Polar coding complexity and performance

Theorem (2007)

With the particular one-to-one mapping described here and with the successive cancellation decoding, polar codes achieve the capacity $I(W)$ with

- *encoding complexity $N \log N$,*
- *decoding complexity $N \log N$,*
- *and probability of frame error better than $2^{-N^{0.49}}$*

Polar coding complexity and performance

Theorem (2007)

With the particular one-to-one mapping described here and with the successive cancellation decoding, polar codes achieve the capacity $I(W)$ with

- *encoding complexity* $N \log N$,
- *decoding complexity* $N \log N$,
- *and probability of frame error better than* $2^{-N^{0.49}}$

Polar coding complexity and performance

Theorem (2007)

With the particular one-to-one mapping described here and with the successive cancellation decoding, polar codes achieve the capacity $I(W)$ with

- *encoding complexity* $N \log N$,
- *decoding complexity* $N \log N$,
- *and probability of frame error better than* $2^{-N^{0.49}}$

Polar coding complexity and performance

Theorem (2007)

With the particular one-to-one mapping described here and with the successive cancellation decoding, polar codes achieve the capacity $I(W)$ with

- *encoding complexity* $N \log N$,
- *decoding complexity* $N \log N$,
- *and probability of frame error better than* $2^{-N^{0.49}}$

Polar coding complexity and performance

Theorem (2007)

With the particular one-to-one mapping described here and with the successive cancellation decoding, polar codes achieve the capacity $I(W)$ with

- *encoding complexity $N \log N$,*
- *decoding complexity $N \log N$,*
- *and probability of frame error better than $2^{-N^{0.49}}$*

Polar codes: nits and grits

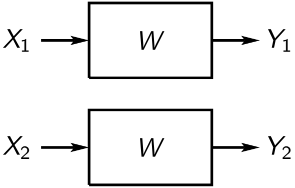
Erdal Arıkan, Emre Telatar

Bilkent U., EPFL

Cambridge — July 1, 2012

Building block

Given two copies of a binary input channel $W\colon \mathbb{F}_2 \to \mathcal{Y}$



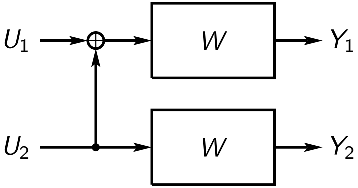
Building block

Given two copies of a binary input channel $W \colon \mathbb{F}_2 \to \mathcal{Y}$

- Set

$$X_1 = U_1 + U_2$$
$$X_2 = U_2$$

with U_1, U_2 i.i.d., uniform on $\mathbb{F}_2$.



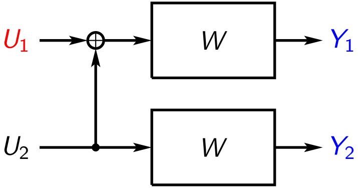
Building block

Given two copies of a binary input channel $W : \mathbb{F}_2 \to \mathcal{Y}$

- Set

$$X_1 = U_1 + U_2$$
$$X_2 = U_2$$

with U_1, U_2 i.i.d., uniform on $\mathbb{F}_2$.

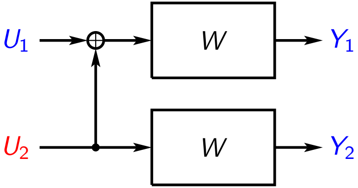


- This induces two synthetic channels $W^- : \mathbb{F}_2 \to \mathcal{Y}^2$

Building block

Given two copies of a binary input channel $W\colon \mathbb{F}_2 \to \mathcal{Y}$

- Set
$$X_1 = U_1 + U_2$$
$$X_2 = U_2$$
with U_1, U_2 i.i.d., uniform on $\mathbb{F}_2$.

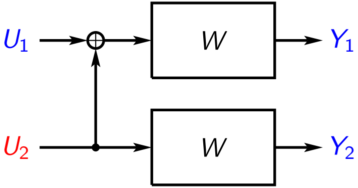


- This induces two synthetic channels $W^-\colon \mathbb{F}_2 \to \mathcal{Y}^2$ and $W^+\colon \mathbb{F}_2 \to \mathcal{Y}^2 \times \mathbb{F}_2$.

Building block

Given two copies of a binary input channel $W\colon \mathbb{F}_2 \to \mathcal{Y}$

- Set
$$X_1 = U_1 + U_2$$
$$X_2 = U_2$$
with U_1, U_2 i.i.d., uniform on $\mathbb{F}_2$.



- This induces two synthetic channels $W^-\colon \mathbb{F}_2 \to \mathcal{Y}^2$ and $W^+\colon \mathbb{F}_2 \to \mathcal{Y}^2 \times \mathbb{F}_2$.
- How come U_1 appears at the output of W^+? Are we being cheated?

Building block: successive decoding

Consider successively decoding $U_1, U_2, \ldots, U_N$ from Y

(a) with a genie-aided decoder:

$$\hat{U}_1 = \phi_1(Y)$$

$$\hat{U}_2 = \phi_2(Y, U_1)$$

$$\hat{U}_3 = \phi_3(Y, U^2)$$

$$\ldots$$

$$\hat{U}_N = \phi_N(Y, U^{N-1})$$

Building block: successive decoding

Consider successively decoding $U_1, U_2, \ldots, U_N$ from Y

(a) with a genie-aided decoder:

$$\hat{U}_1 = \phi_1(Y)$$
$$\hat{U}_2 = \phi_2(Y, U_1)$$
$$\hat{U}_3 = \phi_3(Y, U^2)$$
$$\ldots$$
$$\hat{U}_N = \phi_N(Y, U^{N-1})$$

vs

(b) a Standalone decoder:

$$\hat{U}_1 = \phi_1(Y)$$
$$\hat{U}_2 = \phi_2(Y, \hat{U}_1)$$
$$\hat{U}_3 = \phi_3(Y, \hat{U}^2)$$
$$\ldots$$
$$\hat{U}_N = \phi_N(Y, \hat{U}^{N-1}).$$

Building block: successive decoding

Consider successively decoding $U_1, U_2, \ldots, U_N$ from Y

(a) with a genie-aided decoder:

$$\hat{U}_1 = \phi_1(Y)$$
$$\hat{U}_2 = \phi_2(Y, U_1)$$
$$\hat{U}_3 = \phi_3(Y, U^2)$$
$$\ldots$$
$$\hat{U}_N = \phi_N(Y, U^{N-1})$$

vs

(b) a Standalone decoder:

$$\hat{U}_1 = \phi_1(Y)$$
$$\hat{U}_2 = \phi_2(Y, \hat{U}_1)$$
$$\hat{U}_3 = \phi_3(Y, \hat{U}^2)$$
$$\ldots$$
$$\hat{U}_N = \phi_N(Y, \hat{U}^{N-1}).$$

If the genie-aided decoder makes no errors, then, the standalone decoder makes no errors.

Building block: successive decoding

Consider successively decoding $U_1, U_2, \ldots, U_N$ from Y

(a) with a genie-aided decoder:

$$\hat{U}_1 = \phi_1(Y)$$
$$\hat{U}_2 = \phi_2(Y, U_1)$$
$$\hat{U}_3 = \phi_3(Y, U^2)$$
$$\ldots$$
$$\hat{U}_N = \phi_N(Y, U^{N-1})$$

vs

(b) a Standalone decoder:

$$\hat{U}_1 = \phi_1(Y)$$
$$\hat{U}_2 = \phi_2(Y, \hat{U}_1)$$
$$\hat{U}_3 = \phi_3(Y, \hat{U}^2)$$
$$\ldots$$
$$\hat{U}_N = \phi_N(Y, \hat{U}^{N-1}).$$

If the genie-aided decoder makes no errors, then, the standalone decoder makes no errors. The block error events of the two decoders are the same.

Building block: successive decoding

Consider successively decoding $U_1, U_2, \ldots, U_N$ from Y

(a) with a genie-aided decoder:

$$\hat{U}_1 = \phi_1(Y)$$
$$\hat{U}_2 = \phi_2(Y, U_1)$$
$$\hat{U}_3 = \phi_3(Y, U^2)$$
$$\ldots$$
$$\hat{U}_N = \phi_N(Y, U^{N-1})$$

vs

(b) a Standalone decoder:

$$\hat{U}_1 = \phi_1(Y)$$
$$\hat{U}_2 = \phi_2(Y, \hat{U}_1)$$
$$\hat{U}_3 = \phi_3(Y, \hat{U}^2)$$
$$\ldots$$
$$\hat{U}_N = \phi_N(Y, \hat{U}^{N-1}).$$

If the genie-aided decoder makes no errors, then, the standalone decoder makes no errors. The block error events of the two decoders are the same. As long as the block error probability of the genie-aided decoder is shown to be small, we are not cheated.

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y = ?$ otherwise.

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y =?$ otherwise.

- W^- has input U_1, output $(Y_1, Y_2) =$

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y =?$ otherwise.

- W^- has input U_1, output $(Y_1, Y_2) = (U_1 + U_2, U_2)$

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y =?$ otherwise.

- W^- has input U_1, output $(Y_1, Y_2) = (U_1 + U_2, \; ? \;)$

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y =?$ otherwise.

- W^- has input U_1, output $(Y_1, Y_2) = ($? $, U_2)$

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y =?$ otherwise.

- W^- has input U_1, output $(Y_1, Y_2) = ($? , ? $)$

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y =?$ otherwise.

- W^- is a BEC($2p - p^2$).

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y = ?$ otherwise.

- W^- is a BEC($2p - p^2$).
- W^+ has input U_2, output $(Y_1, Y_2, U_1) =$

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y =?$ otherwise.

- W^- is a BEC($2p - p^2$).
- W^+ has input U_2, output $(Y_1, Y_2, U_1) = (U_1 + U_2, U_2, U_1)$

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y = ?$ otherwise.

- W^- is a BEC($2p - p^2$).
- W^+ has input U_2, output $(Y_1, Y_2, U_1) = ($? $, U_2, U_1)$

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y =?$ otherwise.

- W^- is a BEC($2p - p^2$).
- W^+ has input U_2, output $(Y_1, Y_2, U_1) = (U_1 + U_2, \; ? \;, U_1)$

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y =?$ otherwise.

- W^- is a BEC($2p - p^2$).
- W^+ has input U_2, output $(Y_1, Y_2, U_1) = (\quad ? \quad , ? , U_1)$

Polarization Example: Erasure channel

Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y =?$ otherwise.

- W^- is a BEC($2p - p^2$).
- W^+ is a BEC(p^2).

Polarization Example: Erasure channel

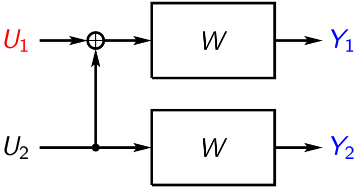
Suppose W is a BEC(p), i.e., $Y = X$ with probabilty $1 - p$, $Y =?$ otherwise.

- W^- is a BEC($2p - p^2$).
- W^+ is a BEC(p^2).
- We already begin to see some extremalization: W^+ is better than W, while W^- is worse.

Building block: properties

Properties of $W \mapsto (W^-, W^+)$:

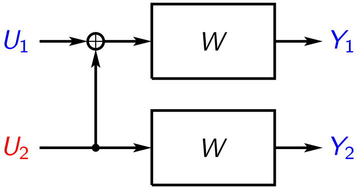
$$I(W^-) = I(U_1; Y_1 Y_2)$$



Building block: properties

Properties of $W \mapsto (W^-, W^+)$:

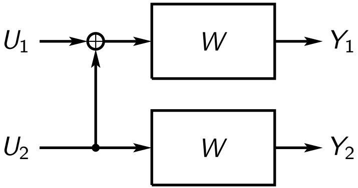
$$I(W^-) = I(U_1; Y_1 Y_2)$$
$$I(W^+) = I(U_2; Y_1 Y_2 U_1)$$



Building block: properties

Properties of $W \mapsto (W^-, W^+)$:

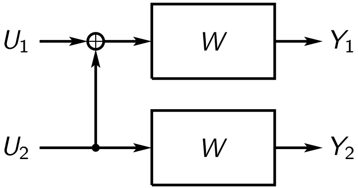
$$
\begin{aligned}
I(W^-) &= I(U_1; Y_1 Y_2) \\
I(W^+) &= I(U_2; Y_1 Y_2 U_1) \\
I(W^-) + I(W^+) &= I(U_1 U_2; Y_1 Y_2)
\end{aligned}
$$



Building block: properties

Properties of $W \mapsto (W^-, W^+)$:

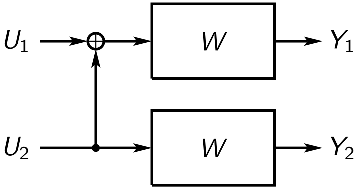
$$\begin{aligned} I(W^-) &= I(U_1; Y_1 Y_2) \\ I(W^+) &= I(U_2; Y_1 Y_2 U_1) \\ I(W^-) + I(W^+) &= I(U_1 U_2; Y_1 Y_2) \\ &= I(X_1 X_2; Y_1 Y_2) \end{aligned}$$



Building block: properties

Properties of $W \mapsto (W^-, W^+)$:

$$\begin{aligned} I(W^-) &= I(U_1; Y_1 Y_2) \\ I(W^+) &= I(U_2; Y_1 Y_2 U_1) \\ I(W^-) + I(W^+) &= I(U_1 U_2; Y_1 Y_2) \\ &= I(X_1 X_2; Y_1 Y_2) \end{aligned}$$

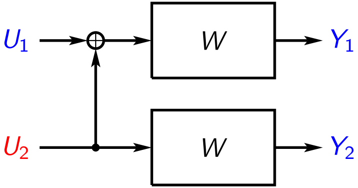


- $\frac{1}{2} I(W^-) + \frac{1}{2} I(W^+) = I(W)$.

Building block: properties

Properties of $W \mapsto (W^-, W^+)$:

$$\begin{aligned} I(W^-) &= I(U_1; Y_1 Y_2) \\ I(W^+) &= I(U_2; Y_1 Y_2 U_1) \\ I(W^-) + I(W^+) &= I(U_1 U_2; Y_1 Y_2) \\ &= I(X_1 X_2; Y_1 Y_2) \end{aligned}$$

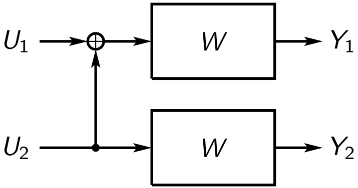


- $\frac{1}{2} I(W^-) + \frac{1}{2} I(W^+) = I(W)$.
- $I(W^+) \geq I(W)$

Building block: properties

Properties of $W \mapsto (W^-, W^+)$:

$$\begin{aligned} I(W^-) &= I(U_1; Y_1 Y_2) \\ I(W^+) &= I(U_2; Y_1 Y_2 U_1) \\ I(W^-) + I(W^+) &= I(U_1 U_2; Y_1 Y_2) \\ &= I(X_1 X_2; Y_1 Y_2) \end{aligned}$$



- $\frac{1}{2} I(W^-) + \frac{1}{2} I(W^+) = I(W)$.
- $I(W^+) \geq I(W) \geq I(W^-)$.

Building block: properties

Properties of $W \mapsto (W^-, W^+)$:

- $\frac{1}{2}I(W^-) + \frac{1}{2}I(W^+) = I(W)$.
- $I(W^+) \geq I(W) \geq I(W^-)$.

Building block: properties

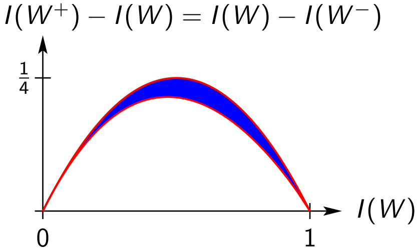
Properties of $W \mapsto (W^-, W^+)$:

- $\frac{1}{2}I(W^-) + \frac{1}{2}I(W^+) = I(W)$.
- $I(W^+) \geq I(W) \geq I(W^-)$.

Building block: properties

Properties of $W \mapsto (W^-, W^+)$:

- $\frac{1}{2}I(W^-) + \frac{1}{2}I(W^+) = I(W)$.
- $I(W^+) \geq I(W) \geq I(W^-)$.

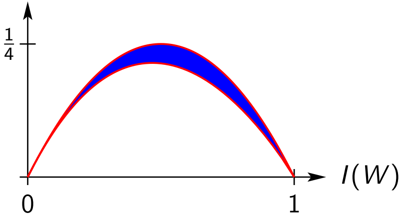


Building block: properties

Properties of $W \mapsto (W^-, W^+)$:

- $\frac{1}{2}I(W^-) + \frac{1}{2}I(W^+) = I(W)$.
- $I(W^+) \geq I(W) \geq I(W^-)$.
- 'Guaranteed progress' unless already extremal.

$I(W^+) - I(W) = I(W) - I(W^-)$



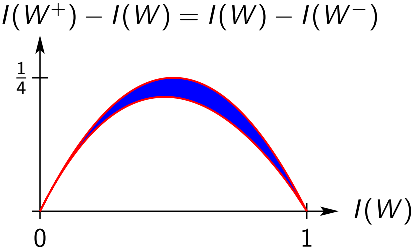
Building block: properties

Properties of $W \mapsto (W^-, W^+)$:

- $\frac{1}{2}I(W^-) + \frac{1}{2}I(W^+) = I(W)$.
- $I(W^+) \geq I(W) \geq I(W^-)$.
- 'Guaranteed progress' unless already extremal.
- $|I(W^{\pm}) - I(W)| < \delta$ implies

$$I(W) \notin (\epsilon, 1-\epsilon),$$

with $\epsilon(\delta) \to 0$ as $\delta \to 0$.



Guaranteed progress

Notation: $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$, denotes the binary entropy function.
Define $p * q := p(1-q) + (1-p)q$; handy when expressing the distribution of the mod-2 sum of independent binary RVs.

Guaranteed progress

Notation: $h(p) = -p \log_2 p - (1-p)\log_2(1-p)$, denotes the binary entropy function.
Define $p * q := p(1-q) + (1-p)q$; handy when expressing the distribution of the mod-2 sum of independent binary RVs.

Lemma

If (X_1, Y_1) and (X_2, Y_2) are independent, X_1 and X_2 are binary, $H(X_1|Y_1) = h(p_1)$, and $H(X_2|Y_2) = h(p_2)$, then,

$$H(X_1 + X_2 | Y_1 Y_2) \geq h(p_1 * p_2).$$

Guaranteed progress

Notation: $h(p) = -p \log_2 p - (1-p)\log_2(1-p)$, denotes the binary entropy function.
Define $p * q := p(1-q) + (1-p)q$; handy when expressing the distribution of the mod-2 sum of independent binary RVs.

Lemma

If (X_1, Y_1) and (X_2, Y_2) are independent, X_1 and X_2 are binary, $H(X_1|Y_1) = h(p_1)$, and $H(X_2|Y_2) = h(p_2)$, then,

$$H(X_1 + X_2 | Y_1 Y_2) \geq h(p_1 * p_2).$$

Proof (Lazy).

This is just Mrs Gerber's Lemma. □

Guaranteed progress

Corollary

If $I(W) = 1 - h(p)$*, then* $I(W^-) \leq 1 - h(p * p)$*, and thus* $I(W) - I(W^-) \geq h(p * p) - h(p)$*.*

Guaranteed progress

Corollary

If $I(W) = 1 - h(p)$*, then* $I(W^-) \leq 1 - h(p * p)$*, and thus* $I(W) - I(W^-) \geq h(p * p) - h(p)$.

Proof.

From $I(W) = 1 - h(p)$ we find $H(X_i|Y_i) = h(p)$. Consequently,

$$\begin{aligned} I(W^-) &= I(U_1; Y_1 Y_2) \\ &= 1 - H(U_1|Y_1 Y_2) \\ &= 1 - H(X_1 + X_2|Y_1 Y_2) \\ &\leq 1 - h(p * p) \end{aligned}$$

□

Guaranteed progress

Corollary

For every $\epsilon > 0$*, there exists* $\delta > 0$ *such that*

$$\left| I(W) - I(W^{\pm}) \right| < \delta$$

implies

$$I(W) \notin (\epsilon, 1 - \epsilon).$$

Proof.

See figure. □

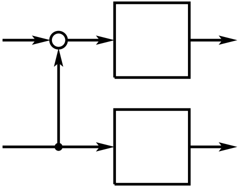
Polarization: why?

Recall the polar construction:

Polarization: why?

Recall the polar construction:

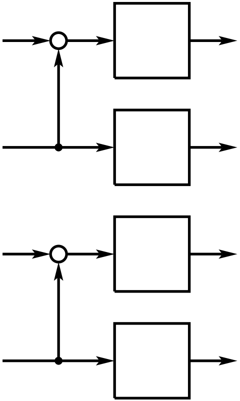
- Duplicate W and obtain W^- and W^+.



Polarization: why?

Recall the polar construction:

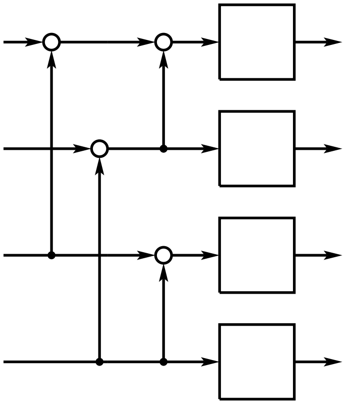
- Duplicate W and obtain W^- and W^+.
- Duplicate W^- (and W^+),



Polarization: why?

Recall the polar construction:

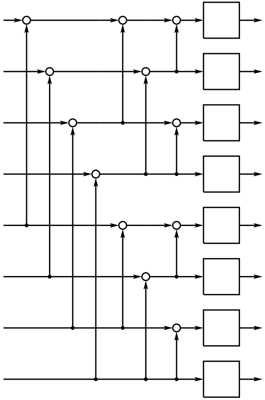
- Duplicate W and obtain W^{-} and W^{+}.
- Duplicate W^{-} (and W^{+}),
- and obtain W^{--} and W^{-+} (and W^{+-} and W^{++}).



Polarization: why?

Recall the polar construction:

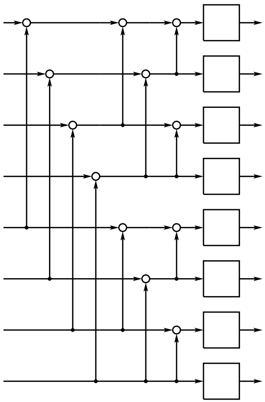
- Duplicate W and obtain W^- and W^+.
- Duplicate W^- (and W^+),
- and obtain W^{--} and W^{-+} (and W^{+-} and W^{++}).
- Duplicate W^{--} (and W^{-+}, W^{+-}, W^{++}) and obtain W^{---} and W^{--+} (and W^{-+-}, W^{-++}, W^{+--}, W^{+-+}, W^{++-}, W^{+++}).



Polarization: why?

Recall the polar construction:

- Duplicate W and obtain W^- and W^+.
- Duplicate W^- (and W^+),
- and obtain W^{--} and W^{-+} (and W^{+-} and W^{++}).
- Duplicate W^{--} (and W^{-+}, W^{+-}, W^{++}) and obtain W^{---} and W^{--+} (and W^{-+-}, W^{-++}, W^{+--}, W^{+-+}, W^{++-}, W^{+++}).
- ...



Polarization: why?

At the nth level into this process we have transformed $N = 2^n$ uses of the channel W to one use each of the 2^n channels

$$W^{b_1 \dots b_n}, \quad b_j \in \{+, -\}.$$

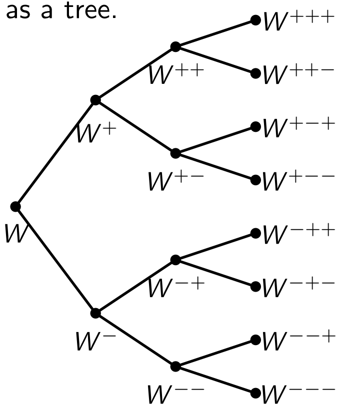
The meaning of polarizatoin is that the 2^n quantities

$$I(W^{-\cdots-}), \dots, I(W^{+\cdots+})$$

are all close to 0 or 1 except for a vanishing fraction (as n grows).

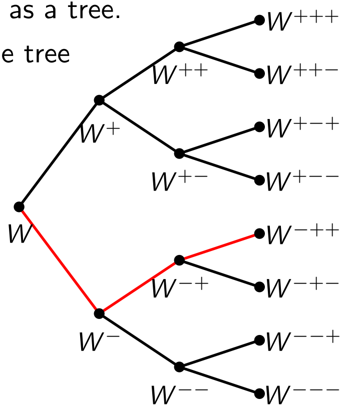
Polarization: why?

- Organize the synthetic channels as a tree.



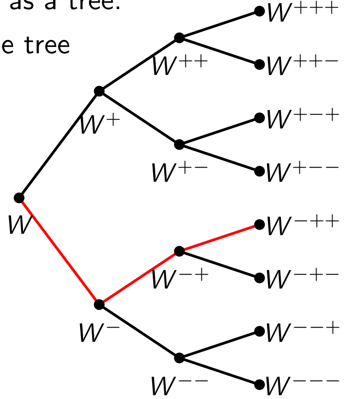
Polarization: why?

- Organize the synthetic channels as a tree.
- Pick a random path climbing the tree according to fair coin flips. This path uniformly samples the nodes at any level n.



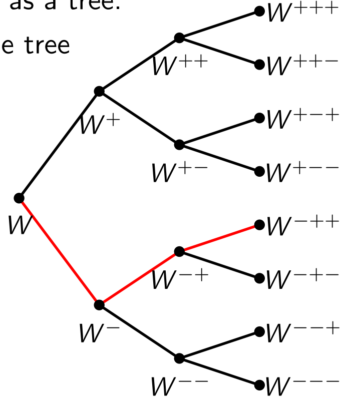
Polarization: why?

- Organize the synthetic channels as a tree.
- Pick a random path climbing the tree according to fair coin flips. This path uniformly samples the nodes at any level n.
- The $I(\cdot)$ sequence we encounter satisfies $E\left[I_{n+1} \mid I_0, \ldots, I_n\right] = I_n$.



Polarization: why?

- Organize the synthetic channels as a tree.
- Pick a random path climbing the tree according to fair coin flips. This path uniformly samples the nodes at any level n.
- The $I(\cdot)$ sequence we encounter satisfies $E\left[I_{n+1} \mid I_0, \ldots, I_n\right] = I_n$.
- Thus, the differences $J_n = I_{n+1} - I_n$ are zero mean, uncorrelated random variables.



Polarization: why?

- $1 \geq (I_n - I_0)^2 = \left(\sum_{k=0}^{n-1} J_k\right)^2 = \sum_{i,k=0}^{n-1} J_i J_k$

Polarization: why?

- $1 \geq (I_n - I_0)^2 = \left(\sum_{k=0}^{n-1} J_k\right)^2 = \sum_{i,k=0}^{n-1} J_i J_k$
- Thus $1 \geq \sum_{k=0}^{n-1} E[J_k^2]$.

Polarization: why?

- $1 \geq (I_n - I_0)^2 = \left(\sum_{k=0}^{n-1} J_k\right)^2 = \sum_{i,k=0}^{n-1} J_i J_k$
- Thus $1 \geq \sum_{k=0}^{n-1} E[J_k^2]$.
- So, $E[J_n^2] \to 0$, thus, for any $\delta > 0$, $\Pr(|J_n| > \delta) \to 0$.

Polarization: why?

- $1 \geq (I_n - I_0)^2 = \left(\sum_{k=0}^{n-1} J_k\right)^2 = \sum_{i,k=0}^{n-1} J_i J_k$
- Thus $1 \geq \sum_{k=0}^{n-1} E[J_k^2]$.
- So, $E[J_n^2] \to 0$, thus, for any $\delta > 0$, $\Pr(|J_n| > \delta) \to 0$.
- By 'guaranteed progress property' the event $\{|J_n| > \delta\}$ includes the event $\{I_n \in (\epsilon, 1-\epsilon)\}$.

Polarization: why?

- $1 \geq (I_n - I_0)^2 = \left(\sum_{k=0}^{n-1} J_k\right)^2 = \sum_{i,k=0}^{n-1} J_i J_k$
- Thus $1 \geq \sum_{k=0}^{n-1} E[J_k^2]$.
- So, $E[J_n^2] \to 0$, thus, for any $\delta > 0$, $\Pr(|J_n| > \delta) \to 0$.
- By 'guaranteed progress property' the event $\{|J_n| > \delta\}$ includes the event $\{I_n \in (\epsilon, 1-\epsilon)\}$.
- Thus the fraction paths for which $I_n \in (\epsilon, 1-\epsilon)$ approaches zero as n gets large. Done! Thanks: H.A. Loeliger

Polarization

- We have shown that $\lim_n \Pr\{I_n \in (\epsilon, 1-\epsilon)\} = 0$.

Polarization

- We have shown that $\lim_n \Pr\{I_n \in (\epsilon, 1-\epsilon)\} = 0$.
- Together with $E[I_n] = I)(W)$ this implies

$$\Pr(I_n \geq 1-\epsilon) \to I(W) \quad \text{and} \quad \Pr(I_n \leq \epsilon) \to 1 - I(W).$$

Polarization

- We have shown that $\lim_n \Pr\{I_n \in (\epsilon, 1-\epsilon)\} = 0$.
- Together with $E[I_n] = I)(W)$ this implies

$$\Pr(I_n \geq 1-\epsilon) \to I(W) \quad \text{and} \quad \Pr(I_n \leq \epsilon) \to 1 - I(W).$$

- Even stronger statements can be made by appealing to the martingale convergence theorem:

$$\Pr\{\lim_n I_n = 1\} = I(W) \quad \text{and} \quad \Pr\{\lim_n I_n = 0\} = 1 - I(W).$$

Polarization speed

- We have seen that polarization takes place.

Polarization speed

- We have seen that polarization takes place.
- But how fast? Fast enough to arrest error propagation?

Polarization speed

- We have seen that polarization takes place.
- But how fast? Fast enough to arrest error propagation?
- Introduce the Bhattacharyya parameter

$$Z(W) = \sum_{y} \sqrt{W(y|0)W(y|1)}$$

as a companion to $I(W)$. Note that this is an upper bound on probability of error for uncoded transmission over W.

A useful representation

$$
\begin{aligned}
I(W) &= 1 - H(X|Y) \\
&= \sum_y W(y)\left[1 - H(X|Y=y)\right] \\
&= \sum_y W(y)\left[1 - h\big(W(0|y)\big)\right]
\end{aligned}
$$

A useful representation

$$\begin{aligned} I(W) &= 1 - H(X|Y) \\ &= \sum_y W(y)\big[1 - H(X|Y=y)\big] \\ &= \sum_y W(y)\big[1 - h\big(W(0|y)\big)\big] \end{aligned}$$

Similarly

$$\begin{aligned} Z(W) &= \sum_y \sqrt{W(y|0)W(y|1)} \\ &= \sum_y W(y)\sqrt{4W(0|y)W(1|y)} \\ &= \sum_y W(y)\sqrt{4W(0|y)(1 - W(0|y))} \end{aligned}$$

A useful representation

$$I(W) = 1 - H(X|Y)$$
$$= \sum_y W(y)\big[1 - H(X|Y = y)\big]$$
$$= \sum_y W(y)\big[1 - h\big(W(0|y)\big)\big]$$

So

$$I(W) = E[1 - h(\Delta)]$$
$$Z(W) = E[\sqrt{4\Delta(1 - \Delta)}]$$

Similarly

$$Z(W) = \sum_y \sqrt{W(y|0)W(y|1)}$$
$$= \sum_y W(y)\sqrt{4W(0|y)W(1|y)}$$
$$= \sum_y W(y)\sqrt{4W(0|y)(1 - W(0|y))}$$

A useful representation

$$
\begin{aligned}
I(W) &= 1 - H(X|Y) \\
&= \sum_y W(y)\big[1 - H(X|Y = y)\big] \\
&= \sum_y W(y)\big[1 - h\big(W(0|y)\big)\big]
\end{aligned}
$$

Similarly

$$
\begin{aligned}
Z(W) &= \sum_y \sqrt{W(y|0)W(y|1)} \\
&= \sum_y W(y)\sqrt{4W(0|y)W(1|y)} \\
&= \sum_y W(y)\sqrt{4W(0|y)(1 - W(0|y))}
\end{aligned}
$$

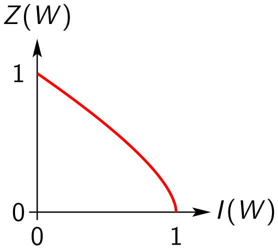
So

$$
\begin{aligned}
I(W) &= E[1 - h(\Delta)] \\
Z(W) &= E[\sqrt{4\Delta(1 - \Delta)}]
\end{aligned}
$$

Consequently $\big(I(W), Z(W)\big)$ belongs to the Convex hull of the curve

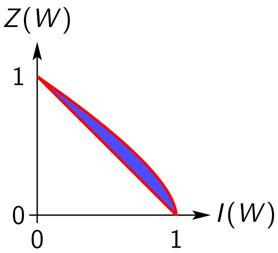
$$
\Big\{\big(1 - h(\delta), \sqrt{4\delta(1 - \delta)}\big) : \delta \in [0, 1]\Big\}
$$

Polarization speed



Polarization speed

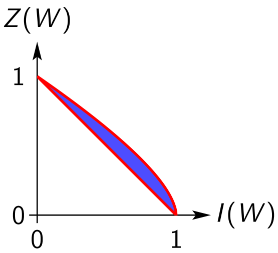
Properties of $Z(W)$:



Polarization speed

Properties of $Z(W)$:

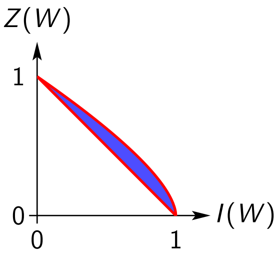
- $Z(W) \approx 0$ iff $I(W) \approx 1$.



Polarization speed

Properties of $Z(W)$:

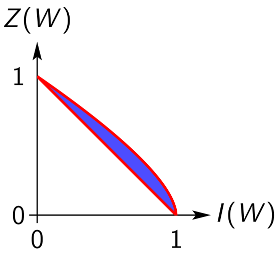
- $Z(W) \approx 0$ iff $I(W) \approx 1$.
- $Z(W) \approx 1$ iff $I(W) \approx 0$.



Polarization speed

Properties of $Z(W)$:

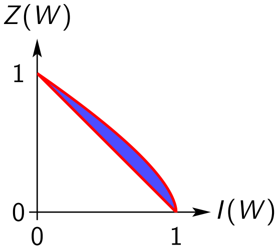
- $Z(W) \approx 0$ iff $I(W) \approx 1$.
- $Z(W) \approx 1$ iff $I(W) \approx 0$.
- $Z(W^+) = Z(W)^2$.



Polarization speed

Properties of $Z(W)$:

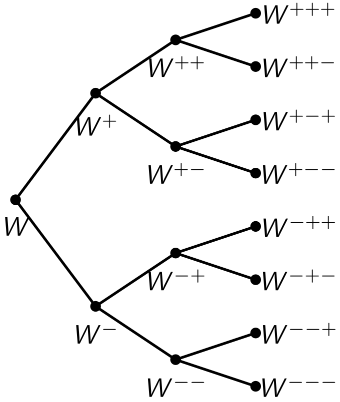
- $Z(W) \approx 0$ iff $I(W) \approx 1$.
- $Z(W) \approx 1$ iff $I(W) \approx 0$.
- $Z(W^+) = Z(W)^2$.
- $Z(W^-) \leq 2Z(W)$.



Since $Z(W)$ upper bounds on probability of error for uncoded transmission over W, we can choose the good indices on the basis of $Z(W)$. The sum of the Z's of the chosen channels will upper bound the block error probability. Good reason to study the polarization speed of Z.

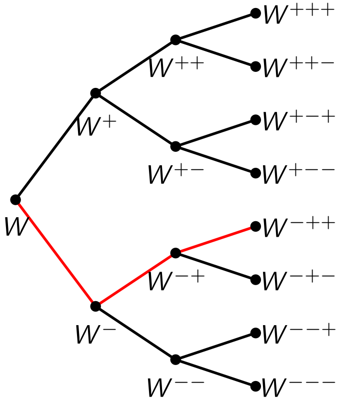
Polarization speed

- Recall the channels organized in a tree.



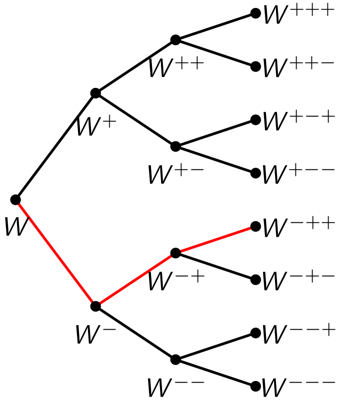
Polarization speed

- Recall the channels organized in a tree.
- Let $Z_0, Z_1, \ldots$ be the $Z(\cdot)$ values we encouter we climb the tree.



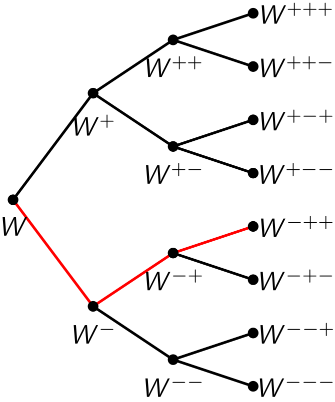
Polarization speed

- Recall the channels organized in a tree.
- Let $Z_0, Z_1, \ldots$ be the $Z(\cdot)$ values we encouter we climb the tree.
- We know that $P(Z_n \to 0) = I(W)$.



Polarization speed

- Recall the channels organized in a tree.
- Let $Z_0, Z_1, \ldots$ be the $Z(\cdot)$ values we encouter we climb the tree.
- We know that $P(Z_n \to 0) = I(W)$.
- We want to show that when $Z_n \to 0$ it does so fast.



Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$
\begin{aligned}
V_{n+1} &= 2V_n && \text{for a `plus' move} \\
V_{n+1} &\leq V_n + 1 && \text{for a `minus' move}
\end{aligned}
$$

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$\begin{aligned} V_{n+1} &= 2V_n && \text{for a `plus' move} \\ V_{n+1} &\leq V_n + 1 && \text{for a `minus' move} \end{aligned}$$

- E.g., starting with $V_m = -20$, and sequence moves: $-$, $-$, $+$, $-$, $-$, $+$, $+$, $-$, we will see a sequence dominated by -20,

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$\begin{aligned} V_{n+1} &= 2V_n && \text{for a 'plus' move} \\ V_{n+1} &\leq V_n + 1 && \text{for a 'minus' move} \end{aligned}$$

- E.g., starting with $V_m = -20$, and sequence moves: $-$, $-$, $+$, $-$, $-$, $+$, $+$, $-$, we will see a sequence dominated by $-20, -19,$

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$\begin{aligned} V_{n+1} &= 2V_n && \text{for a 'plus' move} \\ V_{n+1} &\leq V_n + 1 && \text{for a 'minus' move} \end{aligned}$$

- E.g., starting with $V_m = -20$, and sequence moves: $-$, $-$, $+$, $-$, $-$, $+$, $+$, $-$, we will see a sequence dominated by $-20, -19, -18,$

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$\begin{aligned} V_{n+1} &= 2V_n && \text{for a `plus' move} \\ V_{n+1} &\leq V_n + 1 && \text{for a `minus' move} \end{aligned}$$

- E.g., starting with $V_m = -20$, and sequence moves: $-$, $-$, $+$, $-$, $-$, $+$, $+$, $-$, we will see a sequence dominated by $-20, -19, -18, -36,$

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$\begin{aligned} V_{n+1} &= 2V_n && \text{for a `plus' move} \\ V_{n+1} &\leq V_n + 1 && \text{for a `minus' move} \end{aligned}$$

- E.g., starting with $V_m = -20$, and sequence moves: $-$, $-$, $+$, $-$, $-$, $+$, $+$, $-$, we will see a sequence dominated by $-20, -19, -18, -36, -35,$

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$V_{n+1} = 2V_n \qquad \text{for a `plus' move}$$
$$V_{n+1} \leq V_n + 1 \qquad \text{for a `minus' move}$$

- E.g., starting with $V_m = -20$, and sequence moves: $-$, $-$, $+$, $-$, $-$, $+$, $+$, $-$, we will see a sequence dominated by -20, -19,-18,-36,-35,-34,

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$V_{n+1} = 2V_n \qquad \text{for a 'plus' move}$$
$$V_{n+1} \leq V_n + 1 \qquad \text{for a 'minus' move}$$

- E.g., starting with $V_m = -20$, and sequence moves: $-$, $-$, $+$, $-$, $-$, $+$, $+$, $-$, we will see a sequence dominated by $-20, -19, -18, -36, -35, -34, -68,$

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$\begin{aligned} V_{n+1} &= 2V_n && \text{for a `plus' move} \\ V_{n+1} &\leq V_n + 1 && \text{for a `minus' move} \end{aligned}$$

- E.g., starting with $V_m = -20$, and sequence moves: $-$, $-$, $+$, $-$, $-$, $+$, $+$, $-$, we will see a sequence dominated by $-20, -19, -18, -36, -35, -34, -68, -136,$

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$\begin{aligned} V_{n+1} &= 2V_n && \text{for a `plus' move} \\ V_{n+1} &\leq V_n + 1 && \text{for a `minus' move} \end{aligned}$$

- E.g., starting with $V_m = -20$, and sequence moves: $-$, $-$, $+$, $-$, $-$, $+$, $+$, $-$, we will see a sequence dominated by $-20, -19, -18, -36, -35, -34, -68, -136, -135,$

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$\begin{aligned} V_{n+1} &= 2V_n && \text{for a 'plus' move} \\ V_{n+1} &\leq V_n + 1 && \text{for a 'minus' move} \end{aligned}$$

- E.g., starting with $V_m = -20$, and sequence moves: $-$, $-$, $+$, $-$, $-$, $+$, $+$, $-$, we will see a sequence dominated by $-20, -19, -18, -36, -35, -34, -68, -136, -135, \ldots$

Polarization speed

- It is more convenient to work with $V_n = \log_2 Z_n$. This takes values in $(-\infty, 0]$, We already know that $V_n \to -\infty$ with probability $I(W)$, and want to show that it goes to $-\infty$ fast when it does.
- V_n's obey

$$\begin{aligned} V_{n+1} &= 2V_n && \text{for a ‘plus’ move} \\ V_{n+1} &\leq V_n + 1 && \text{for a ‘minus’ move} \end{aligned}$$

- E.g., starting with $V_m = -20$, and sequence moves: $-$, $-$, $+$, $-$, $-$, $+$, $+$, $-$, we will see a sequence dominated by $-20, -19, -18, -36, -35, -34, -68, -136, -135, \ldots$
- The amounts the ‘minus’ moves change the V values are negligible compared to the changes made by the ‘plus’ moves.

Polarizaton speed: heuristics

- To the first approximation, V_n process behaves like

$$\tilde{V}_{n+1} = 2\tilde{V}_n \qquad \text{for a 'plus' move}$$
$$\tilde{V}_{n+1} = \tilde{V}_n \qquad \text{for a 'minus' move}$$

Polarizaton speed: heuristics

- To the first approximation, V_n process behaves like

$$\tilde{V}_{n+1} = 2\tilde{V}_n \qquad \text{for a 'plus' move}$$
$$\tilde{V}_{n+1} = \tilde{V}_n \qquad \text{for a 'minus' move}$$

- In a long sequence of moves we will typlically see an almost equal number of $+$ and $-$'s, thus

$$\tilde{V}_n = O\left(-2^{n/2}\right) = O(-\sqrt{N}).$$

Polarizaton speed: heuristics

- To the first approximation, V_n process behaves like

$$\tilde{V}_{n+1} = 2\tilde{V}_n \qquad \text{for a `plus' move}$$
$$\tilde{V}_{n+1} = \tilde{V}_n \qquad \text{for a `minus' move}$$

- In a long sequence of moves we will typlically see an almost equal number of $+$ and $-$'s, thus

$$\tilde{V}_n = O\left(-2^{n/2}\right) = O(-\sqrt{N}).$$

- So we expect Z_n to behave roughly like $2^{-\sqrt{N}}$.

Polarization speed: more formally

- In going from V_m to V_n we make $n - m$ moves. If $S_{m,n}$ of these are 'plus' moves, then

  $$V_n \leq [V_m + (n - m - S_{m,n})]2^{S_{m,n}}$$

Polarization speed: more formally

- In going from V_m to V_n we make $n - m$ moves. If $S_{m,n}$ of these are 'plus' moves, then

$$V_n \leq [V_m + (n - m - S_{m,n})]2^{S_{m,n}} \leq [V_m + n - m]2^{S_{m,n}}$$

Polarization speed: more formally

- In going from V_m to V_n we make $n - m$ moves. If $S_{m,n}$ of these are 'plus' moves, then
$$V_n \leq [V_m + (n - m - S_{m,n})]2^{S_{m,n}} \leq [V_m + n - m]2^{S_{m,n}}$$
- Note that the bound is useful only when $n \leq m - V_m$. So one cannot show too strong a convergence speed based on this alone.

Polarization speed: more formally

- In going from V_m to V_n we make $n - m$ moves. If $S_{m,n}$ of these are 'plus' moves, then
$$V_n \leq [V_m + (n - m - S_{m,n})]2^{S_{m,n}} \leq [V_m + n - m]2^{S_{m,n}}$$
- Note that the bound is useful only when $n \leq m - V_m$. So one cannot show too strong a convergence speed based on this alone.
- But using the bound twice by introducing an intermediate destination k:

$$V_n \leq [V_k + n - k]2^{S_{k,n}}$$

Polarization speed: more formally

- In going from V_m to V_n we make $n - m$ moves. If $S_{m,n}$ of these are 'plus' moves, then
$$V_n \leq [V_m + (n - m - S_{m,n})]2^{S_{m,n}} \leq [V_m + n - m]2^{S_{m,n}}$$
- Note that the bound is useful only when $n \leq m - V_m$. So one cannot show too strong a convergence speed based on this alone.
- But using the bound twice by introducing an intermediate destination k:

$$\begin{aligned} V_n &\leq [V_k + n - k]2^{S_{k,n}} \\ &\leq \Big[[V_m + k - m]2^{S_{m,k}} + n - k\Big]2^{S_{k,n}} \end{aligned}$$

Polarization speed: more formally

If V_m were less than $-2m$, we could take $k = 2m$, and $n = m^2$ to obtain

$$\begin{aligned} V_{m^2} &\le [-m2^{S_{m,2m}} + m^2 - 2m]2^{S_{2m,m^2}} \\ &= [-m2^{m(1-\epsilon)} + m^2 - 2m]2^{(m^2-m)(1-\epsilon)/2} \qquad \text{(typically)} \\ &= O\big(-2^{m^2(0.5-\epsilon)}\big) \end{aligned}$$

Equivalently,

$$V_n \le O(-N^{0.5-\epsilon})$$

Polarization speed: more formally

- Only thing left to show is that $V_m \leq -2m$ is a typical event for the paths where $V_n \to -\infty$.

Polarization speed: more formally

- Only thing left to show is that $V_m \leq -2m$ is a typical event for the paths where $V_n \to -\infty$.
- On such paths, there will come a time n_0 so that $V_n \leq -11$ for all $n \geq n_0$. The evolution of V_n then satisfies

$$V_{n+1} \leq 2V_n \leq V_n - 11 \qquad \text{'plus' moves}$$
$$V_{n+1} \leq V_n + 1 \qquad \text{'minus' moves}$$

Polarization speed: more formally

- Only thing left to show is that $V_m \leq -2m$ is a typical event for the paths where $V_n \to -\infty$.
- On such paths, there will come a time n_0 so that $V_n \leq -11$ for all $n \geq n_0$. The evolution of V_n then satisfies

$$V_{n+1} \leq 2V_n \leq V_n - 11 \qquad \text{'plus' moves}$$
$$V_{n+1} \leq V_n + 1 \qquad \text{'minus' moves}$$

- Thus from n_0 onwards, V_n is dominated by a random walk with average drift -5.

Polarization speed: more formally

- Only thing left to show is that $V_m \leq -2m$ is a typical event for the paths where $V_n \to -\infty$.
- On such paths, there will come a time n_0 so that $V_n \leq -11$ for all $n \geq n_0$. The evolution of V_n then satisfies

$$V_{n+1} \leq 2V_n \leq V_n - 11 \qquad \text{'plus' moves}$$
$$V_{n+1} \leq V_n + 1 \qquad \text{'minus' moves}$$

- Thus from n_0 onwards, V_n is dominated by a random walk with average drift -5.
- Thus at time $m = 2n_0$ the typical value of V_m is dominated by $-5n_0 = -2.5m \leq -2m$, which is what we want (with room to spare).

Construction complexity

Let $V \preceq W$ denote that V is stochastically degraded with respect to W.

Lemma

If $V \preceq W$ then $V^{\pm} \preceq W^{\pm}$.

Proof.

Obvious. □

Construction complexity

Lemma

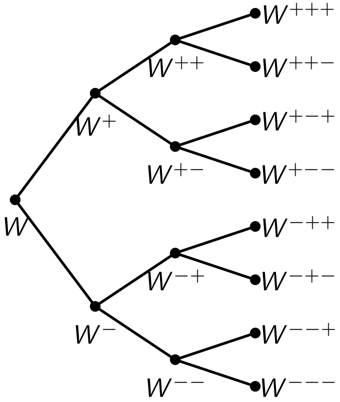
Given any symmetric channel W, and $\delta > 0$ there is a symmetric channel V such that

- $V \preceq W$
- $I(W) - I(V) \leq \delta$
- *V has an output alphabet of cardinality $\leq 2/\delta$.*

Moreover, one can efficiently find such a V.

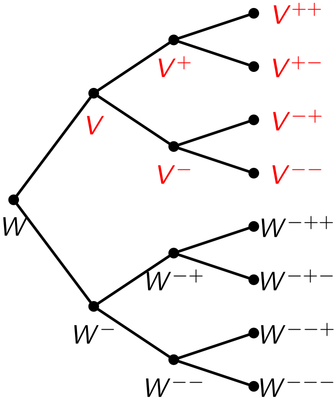
Construction complexity

- If we take the tree of channels,



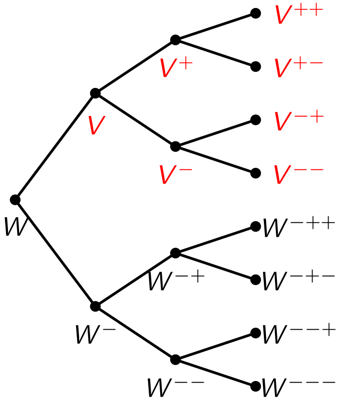
Construction complexity

- If we take the tree of channels,
- Replace a channel on a node by a stochastically degraded version (E.g., replace W^+ by a $V \preceq W^+$) whose mutual information is differs from the original by δ, (E.g, $I(W^+) - I(V) = \delta$)

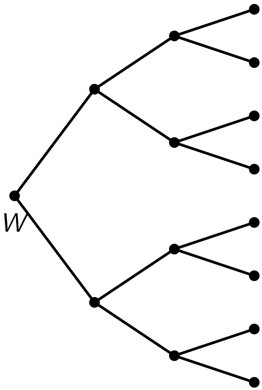


Construction complexity

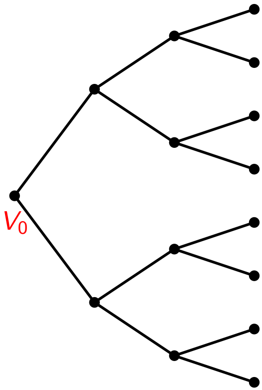
- If we take the tree of channels,
- Replace a channel on a node by a stochastically degraded version (E.g., replace W^+ by a $V \preceq W^+$) whose mutual information is differs from the original by δ, (E.g, $I(W^+) - I(V) = \delta$)
- Then the average loss of mutual information the descendants of this node at any level equals δ.



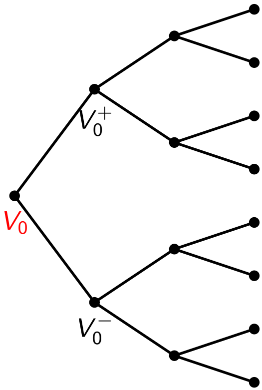
Construction complexity



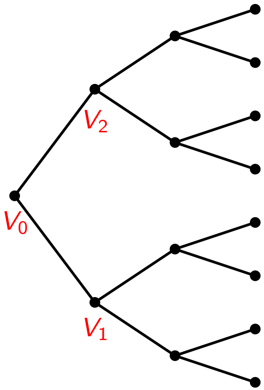
Construction complexity



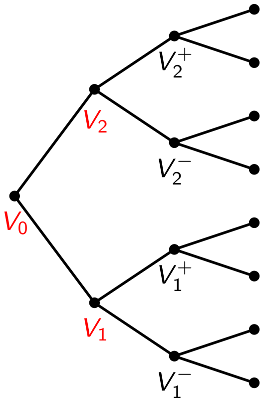
Construction complexity



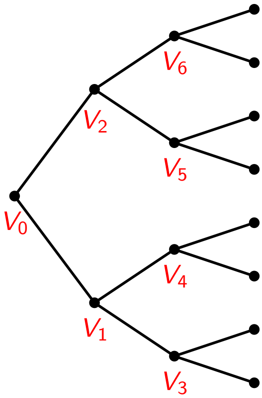
Construction complexity



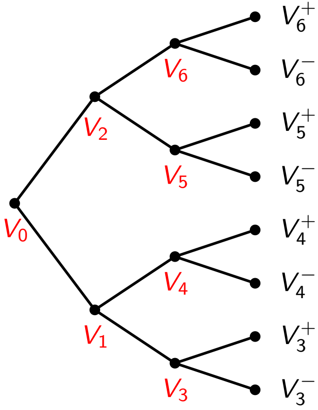
Construction complexity



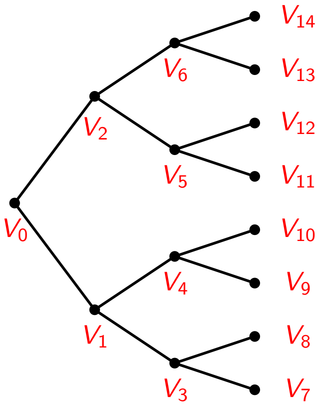
Construction complexity



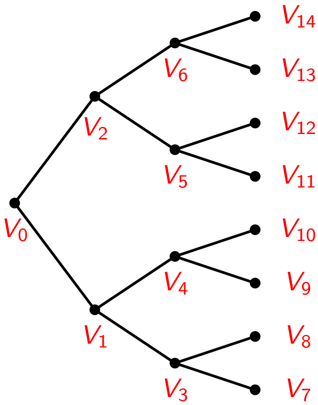
Construction complexity



Construction complexity

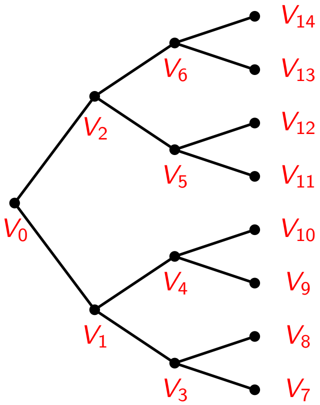


Construction complexity



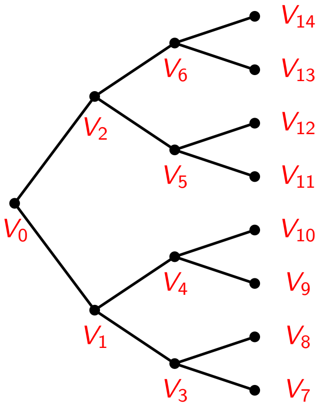
- If each of the replacements are as in the lemma, their total effect on average loss of mutual information on the *n*th level of the tree is $(n+1)\delta$

Construction complexity



- If each of the replacements are as in the lemma, their total effect on average loss of mutual information on the nth level of the tree is $(n+1)\delta$
- Choosing $\delta = 1/(n+1)n$ ensures that the average loss is at most $1/n$.

Construction complexity



- If each of the replacements are as in the lemma, their total effect on average loss of mutual information on the nth level of the tree is $(n+1)\delta$
- Choosing $\delta = 1/(n+1)n$ ensures that the average loss is at most $1/n$.
- In particular the fraction of channels that suffer a loss more than $1/\sqrt{n}$ is less than $1/\sqrt{n}$.