Polar codes : information theoretic analysis and performances

Meryem Benammar†

† ISAE-Supaéro

Joint work with V.Bioglio, F.Gabry and I.Land (Huawei Paris)

† ISAE Institut Supérieur de l'Aéronautique et de l'Espace SUPAERO

Arikan's Polar Codes

New Proof of Polarization

Error Exponents

Tools : Information Theory

- Let X be a random variable on $\mathcal{X}$ with associated pmf P_X. The entropy of X is defined by

$$H(X) = -\sum_{x} P_X(x) \log_2(P_X(x))$$

  - Minimized when X is deterministic, i.e. constant
  - Maximized when P_X is the uniform law, i.e, $P_X(x) = \frac{1}{||\mathcal{X}||}$

Tools : Information Theory

- Let X be a random variable on $\mathcal{X}$ with associated pmf P_X.
  The entropy of X is defined by

$$H(X) = -\sum_{x} P_X(x) \log_2(P_X(x))$$

- Let (X, Y) be correlated random variables on $\mathcal{X} \times \mathcal{Y}$ with pmf $P_{X,Y}$.
  The conditional entropy of X to Y is defined by

$$H(X|Y) = -\sum_{x,y} P_{X,Y}(x,y) \log_2(P_{X|Y}(x|y))$$

  - Minimized when X is a function of Y
  - Maximized when X and Y are independent

Tools : Information Theory

- Let X be a random variable on $\mathcal{X}$ with associated pmf P_X.
  The entropy of X is defined by

$$H(X) = -\sum_{x} P_X(x) \log_2(P_X(x))$$

- Let (X, Y) be correlated random variables on $\mathcal{X} \times \mathcal{Y}$ with pmf $P_{X,Y}$.
  The conditional entropy of X to Y is defined by

$$H(X|Y) = -\sum_{x,y} P_{X,Y}(x,y) \log_2(P_{X|Y}(x|y))$$

- The mutual information between X and Y is defined by

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

  - Minimized when X and Y are independent
  - Maximized when X and Y are equal

Tools : Information Theory

- Let X be a random variable on $\mathcal{X}$ with associated pmf P_X.
  The entropy of X is defined by

$$H(X) = -\sum_{x} P_X(x) \log_2(P_X(x))$$

- Let (X, Y) be correlated random variables on $\mathcal{X} \times \mathcal{Y}$ with pmf $P_{X,Y}$.
  The conditional entropy of X to Y is defined by

$$H(X|Y) = -\sum_{x,y} P_{X,Y}(x,y) \log_2(P_{X|Y}(x|y))$$

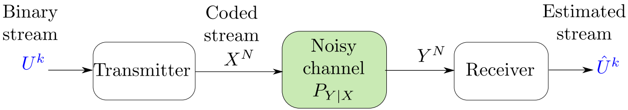
- The mutual information between X and Y is defined by

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

Useful to predict the performance of communication schemes

Context

- Consider a point-to-point communication channel



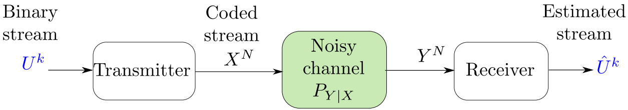
- A bit stream $U^k = (U_1, \ldots, U_k)$
- A coded stream $X^N = (X_1, \ldots, X_N)$
- N memoryless channel uses of $W \sim P_{Y|X}$
- Target : error free communication, i.e.,

$$\mathbb{P}(U^k \neq \hat{U}^k) \to 0 \text{ as } N \to \infty$$

- What is the optimal ratio $\dfrac{k}{N}$: number of bits per channel use ?

Context

- Consider a point-to-point communication channel



- A bit stream $U^k = (U_1, \ldots, U_k)$
- A coded stream $X^N = (X_1, \ldots, X_N)$
- N memoryless channel uses of $W \sim P_{Y|X}$
- Target : error free communication, i.e.,

$$\mathbb{P}(U^k \neq \hat{U}^k) \to 0 \text{ as } N \to \infty$$

- What is the optimal ratio $\frac{k}{N}$: number of bits per channel use ?

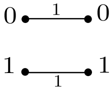
Channel capacity

The capacity of the channel $\mathcal{W} : \mathcal{X} \to \mathcal{Y}$ is given by

$$C(W) = \lim_{N \to \infty} \frac{k}{N} = \max_{P_X} I(X;Y) \leq 1 = H(U)$$

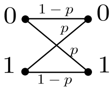
Standard channels

- For a noiseless channel $W \sim \mathbb{1}(y = x)$



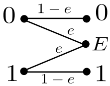
$C(W) = 1$

- For a Binary Symmetric Channel (BSC) $W \sim Bern(p)$



$C(W) = 1 - h_2(p)$

- For a Binary Erasure Channel (BEC) $W \sim Err(e)$



$C(W) = 1 - e$

- For a binary input Gaussian channel $W \sim \mathcal{N}(\mu, \sigma^2)$

$$C(W) > C_{\text{cstr}}(W)$$

A bit of History

For infinite blocklengths N

- Shannon enunciated the capacity formula in 1948
- Since then, subsequent work on error correction coding
- Multi-level codes (Ungerbock* 1976, Imai & Hirakawa 1977)
- LDPC codes (Gallager* 1960s, McKay 2000)
- Turbo-codes (Glavieux & Bérou 1990)
- Polar codes (Arikan* 2008)

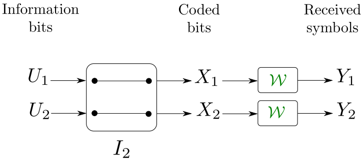
However, at finite blocklength N

- Bounded probablity of error
- Capacity formula enunicated (Polianski & Verdu* 2011)
- Finite blocklength capacity achieving codes are still unknown

* are recipients of the Shannon award

Before polarization

- Original transformation : identity



$$(X_1, X_2) = (U_1, U_2) \cdot I_2 = (U_1, U_2)$$

- Memoryless channel : Markov chains

$$U_1 -\!\!\!\circ\!\!\!- Y_1 -\!\!\!\circ\!\!\!- (Y_2, U_2) \text{ and } U_2 -\!\!\!\circ\!\!\!- Y_2 -\!\!\!\circ\!\!\!- (Y_1, U_1)$$

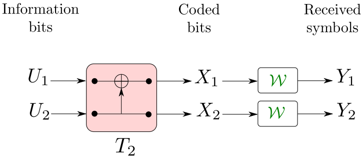
- Information conservation property

$$\begin{aligned} I(U_1U_2; Y_1Y_2) &= I(U_1; Y_1Y_2) + I(U_2; Y_1Y_2|U_1) \\ &= I(U_1; Y_1) + I(U_2; Y_2) \\ &= 2.I(X; Y) \end{aligned}$$

- Each bit U_i experiences the same channel $W_0 : X_i \to Y_i$

Arikan's kernel T_2

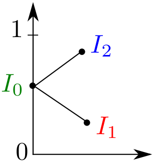
- Transformation matrix T_2 [Arikan'08]



$$(X_1, X_2) = (U_1, U_2) \cdot T_2 = (U_1 + U_2, U_2)$$

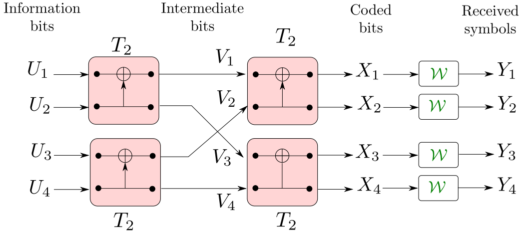
- Coding introduces memory $T_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$
- Information conservation property (chain rule)

$$\begin{aligned} I(U_1U_2; Y_1Y_2) &= I(U_1; Y_1Y_2) + I(U_2; Y_1Y_2|U_1) \\ &= I(X_1X_2; Y_1Y_2) \\ &= 2.I(X;Y) \\ &= 2.I_0 \end{aligned}$$



- Two new channels $W_1 : U_1 \to (Y_1, Y_2)$ and $W_2 : U_2 \to (Y_1, Y_2, U_1)$

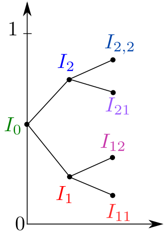
Arikan's construction : two iterations



$$(X_1, X_2, X_3, X_4) = (U_1, U_2, U_3, U_4) \cdot T_2^{\otimes 2}$$

2-th fold Kronecker product of T_2

$$T_2^{\otimes 2} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

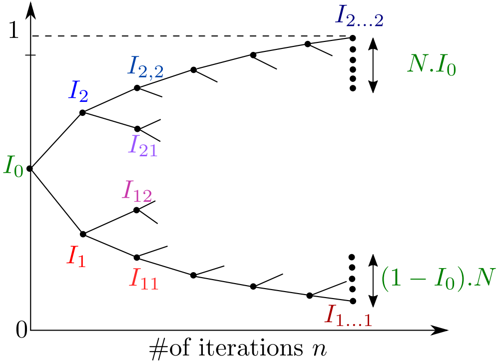


Polarization

- Recursive construction : n-th fold Kronecker product of T_2
- Encoding rule

$$(X_1, \ldots, X_N) = (U_1, \ldots, U_N).T_2^{\otimes n} \text{ where } N = 2^n$$

- Polarization tree



Properties of polar codes

Properties of the polar code

- Block linear code
- Low complexity encoding
- Successive cancellation decoding
- Capacity achieving for a binary input channel ($\mathcal{X} = \{0, 1\}$)

Polar codes : capacity achieving

As $N \to \infty$, among the N input bits, $k = I_0.N$ bits have noiseless channels

$$\lim_{N \to \infty} \frac{k}{N} = I_0 = \max_{P_X} I(X;Y)$$

- Define these k bits as reliable bits
- Define the $N - k$ remaining ones as frozen bits
- Input sequence to the code $U_1^N = (u_1, 0, 0, u_2, 0, u_3, ..., u_k, 0, 0)$

Polar codes design

Challenge : Locate the frozen bits at a given length $N = 2^n$ and rate k/N ?

- For a Binary Erasure Channel (BEC), with erasure prob e : explicit solution

$$I_0 = 1 - e \qquad \Rightarrow \qquad I_1 = 1 - e^2 \;\text{ and }\; I_2 = (1-e)^2$$

- For Gaussian channels : density evolution with Gaussian approximation
- For arbitrary channels : Monte-Carlo simulations

Frozen set known by both encoder and decoder

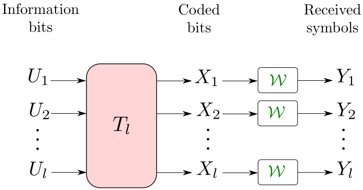
- Set the a priori information of the frozen bits to 0
- Successively decode the bits u_1, then u_2, then .... u_k
- At each decoding step i, decoder knows $v_{1,\ldots,i-1}$
- Perform a local MAP for the bit U_i

$$\arg\max_{u=0,1} \mathbb{P}\left(U_i = u | y_1, ..., y_N, u_1, ..., u_{i-1}\right)$$

- Decoding equations : hard to obtain

Polar codes : arbitrary kernels

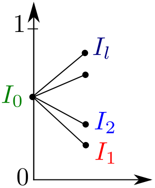
- Transformation matrix T_l [KoradaSasogluUrbanke'10]



$$(X_1, X_2, \dots, X_l) = (U_1, U_2, \dots, U_l) \cdot T_l$$

- Information conservation property (chain rule)

$$\begin{aligned} I(U_1^n; Y_1^n) &= \sum_{i=1}^{l} I(U_i; Y_1^n | U_1^{i-1}) \\ &= l.I(X;Y) \\ &= l.I_0 \end{aligned}$$

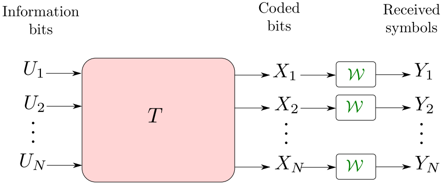


- Recursive construction

$$T = T_l^{\otimes n} \text{where } N = l^n$$

Multi-kernel construction

Consider N channel uses of a discrete memoryless channel $\mathcal{W} : \mathcal{X} \to \mathcal{Y}$



$$(X_1, X_2, \dots, X_N) = (U_1, U_2, \dots, U_N) \cdot T$$

where

$$T = T_{l_1} \otimes \cdots \otimes T_{l_m} \text{ where } N = l_1 \times \cdots \times l_m$$

[BioglioGabryLandBelfiore'16]

Polarization conditions and error exponent for multi-kernel polar codes ?

Arikan's Polar Codes

New Proof of Polarization

Error Exponents

Polarization principle

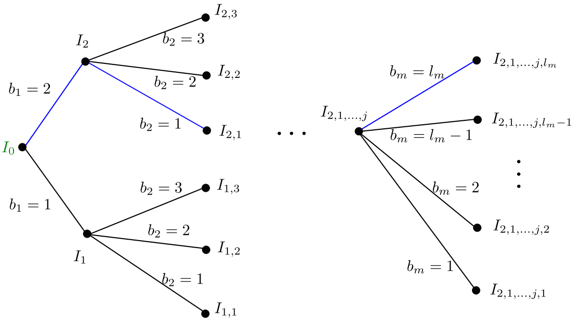
- Assume that the transformation matrix is given by

$$T = T_{l_1} \otimes \cdots \otimes T_{l_m}$$

- Let $(B_1, \ldots, B_m)$ be m random variables such that

$$B_j \sim \text{Unif}([1 : l_j])$$

- Each channel $W_i : U_i \to (Y_1^N, U_1^{i-1})$ by $W_{b_1,,b_m}$
- Corresponding mutual information denoted by $I_{b_1,\ldots,b_m} = I_m$



Polarization proof

The proof of polarization is two fold :

1) Convergence : The sequence $(I_m)_m$ is a bounded martingale and thus converges to I_∞
2) Limit distribution : The random variable I_∞ follows a Bernoulli(I_0) distribution

Polarization proof

The proof of polarization is two fold :

1) Convergence : The sequence $(I_m)_m$ is a bounded martingale and thus converges to I_∞
2) Limit distribution : The random variable I_∞ follows a Bernoulli(I_0) distribution

Proof of convergence :

- The sequence $(I_m)_m$ is bounded, $\forall m \in \mathbb{N}\ ,\ 0 \leq I_m \leq 1$
- The sequence $(I_m)_m$ is a martingale w.r.t. $(B_1, \ldots, B_m)$

$$\forall m \in \mathbb{N}\ ,\ \mathbb{E}_{B_{m+1}}(I_{m+1}|B_1, \ldots, B_m) = I_m$$

  using the information conservation property
- The expected value of I_m is constant

$$\forall m \in \mathbb{N}\ ,\ \mathbb{E}(I_m) = \mathbb{E}(I_{m-1}) = \cdots = \mathbb{E}(I_0) = I_0$$

$(I_m)_m$ converges to a random variable I_∞ and $\mathbb{E}(I_\infty) = I_0$

Limit distribution : new inequality

Limit distribution : sufficient condition on the kernels $(T_{l_1}, \ldots, T_{l_{m+1}})$

If, for all kernels $(T_{l_1}, \ldots, T_{l_{m+1}})$, we have that

$$\forall m > 0 \ , \ \forall (b_1, \ldots, b_m, b_{m+1}) \in \bigotimes_{j=1}^{m+1} [1 : l_j]$$

$$|I_{b_1,\ldots,b_m,b_{m+1}} - I_{b_1,\ldots,b_m}| \geq I^{\alpha}_{b_1,\ldots,b_m} (1 - I_{b_1,\ldots,b_m})^{\beta}$$

where $\alpha, \beta > 1$, then I_∞ is a binary random variable.

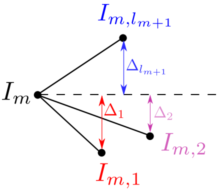
This implies, that with probability 1

$$|I_{m+1} - I_m| \geq I_m^{\alpha} (1 - I_m)^{\beta}$$

which would imply, since $(I_m)_m$ is convergent, that

$$I_\infty^{\alpha} (1 - I_\infty)^{\beta} = 0$$

which yields that $I_\infty = 0$ or $I_\infty = 1$.



Example of kernels : T_2

Consider the kernel $T_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$

- Let $I_0 = I(X;Y) = 1 - H(X|Y) = 1 - H_0$
- We need to prove that for some $\alpha, \beta > 1$

$$\begin{aligned} |I(U_1;Y_1Y_2) - I(X;Y)| &\geq I(X;Y)^{\alpha}(1 - I(X;Y))^{\beta} \\ |I(U_2;Y_1Y_2|U_1) - I(X;Y)| &\geq I(X;Y)^{\alpha}(1 - I(X;Y))^{\beta}, \end{aligned}$$

- Amounts to proving that

$$H(X_1 \oplus X_2|Y_1Y_2) - H_0 \geq H_0^{\beta}.(1 - H_0)^{\alpha}$$

Example of kernels : T_2

Consider the kernel $T_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$

- Let $I_0 = I(X;Y) = 1 - H(X|Y) = 1 - H_0$
- We need to prove that for some $\alpha, \beta > 1$

$$\begin{aligned} |I(U_1;Y_1Y_2) - I(X;Y)| &\geq I(X;Y)^{\alpha}(1 - I(X;Y))^{\beta} \\ |I(U_2;Y_1Y_2|U_1) - I(X;Y)| &\geq I(X;Y)^{\alpha}(1 - I(X;Y))^{\beta}, \end{aligned}$$

- Amounts to proving that

$$H(X_1 \oplus X_2|Y_1Y_2) - H_0 \geq H_0^{\beta}.(1 - H_0)^{\alpha}$$

Mrs gerber's Lemma : $H(X_1 \oplus X_2|Y_1^2) \geq h_2\left(h_2^{-1}(H_0) \star h_2^{-1}(H_0)\right)$

An entropy inequality : $h_2(a \star a) - h_2(a) \geq h_2^2(a) \cdot (1 - h_2(a)) \geq 0$

Yields the sufficient inequality with $\beta = 2$ and $\alpha = 1$

Arikan's Polar Codes

New Proof of Polarization

Error Exponents

Error exponent and Bhattacharyya parameter

- For a given channel $W : X \to Y$, we define the Bhattacharyya parameter

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

- Link to mutual information

$$Z(W) = 0 \Leftrightarrow I(W) = 1, \text{ and } Z(W) = 1 \Leftrightarrow I(W) = 0$$

- Define a sequence of random Bhattacharyya parameters

$$Z_m \triangleq Z(W_{B_1,\dots,B_m}) = \sum_{z \in \mathcal{Z}} \sqrt{W_{B_1,\dots,B_m}(z|1)W_{B_1,\dots,B_m}(z|0)}.$$

A polar code has error exponent E iif ([KoradaSasogluUrbanke'10])

1. For all $\gamma \geq E$,
$$\lim_{m \to \infty} \mathbb{P}(Z_m \geq 2^{-N^{\gamma}}) = 1;$$
2. for all $0 < \gamma \leq E$
$$\lim_{m \to \infty} \mathbb{P}(Z_m \leq 2^{-N^{\gamma}}) = I_0.$$

Error exponents and partial distance

- Assume that a transformation matrix T_l writes as

$$T_l = (\mathbf{t}_1^\dagger, \dots, \mathbf{t}_i^\dagger, \dots, \mathbf{t}_l^\dagger)^\dagger$$

- Define the partial distances $(D_1, \dots, D_l)$ of a matrix T_l as

$$D_i \triangleq dist(\mathbf{t}_i, < \mathbf{t}_{i+1}, \dots, \mathbf{t}_l >)$$

where $< \mathbf{t}_{i+1}, \dots, \mathbf{t}_l >$ is the linear code spanned by the remaining rows of T_l

The error exponent of a one kernel, T_l, polar code is given

$$E_l \triangleq \frac{1}{l} \sum_{i=1}^{l} \log_l(D_i)$$

Example : For $T_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, $D_1 = 1$ and $D_2 = 2$, thus

$$E_2 = \frac{1}{2}$$

Error exponent for multi-kernel polar codes

- Assume in a mutli-kernel construction $T = T_{l_1} \otimes \cdots \otimes T_{l_m}$
- Kernels are chosen from a pool of s distinct kernels, $l_j \in [1 : s]$
- Each kernel T_{l_j} appears with a frequency p_j
- Each kernel T_{l_j} has an associated error exponent E_{l_j}

The error exponent of a multi-kernel polar code is given by by

$$E = \sum_{j=1}^{s} \alpha_j \cdot E_{l_j}$$

where

$$\alpha_j \triangleq \frac{p_j \log_2(l_j)}{\sum_{j'} p_{j'} \log_2(l_{j'})} \quad \text{and} \quad \sum_{j=1}^{s} \alpha_j = 1$$

Weighted sum of the error exponents E_{l_j}

Idea of proof : indirect part (1)

- Key inequality : for all m, [KoradaSasogluUrbanke'10]

$$\forall (b_1, \dots, b_m) \qquad , \qquad Z_{m-1}^{D_{b_m}} \leq Z_m \leq 2^{l_m - b_m} Z_{m-1}^{D_{b_m}}$$

- To prove the indrect part : for all $\gamma \geq E$,

$$\lim_{m \to \infty} \mathbb{P}(Z_m \geq 2^{-N^\gamma}) = 1;$$

we use the LHS of this inequality

$$Z_m \geq Z_{m-1}^{D_{b_m}} \geq Z_{m-2}^{D_{b_m} . D_{b_{m-1}}} \geq \cdots \geq Z_0^{\prod_{k=1}^m D_{b_m}}$$

- This yields

$$Z_m \geq 2^{-N^{E+o(N)}}$$

Thus,

$$\lim_{m \to \infty} \mathbb{P}(Z_m \geq 2^{-N^E}) = 1;$$

Idea of proof : direct part (2)

- Key inequality : for all m, [KoradaSasogluUrbanke'10]

$$\forall(b_1,\ldots,b_m) \qquad , \qquad Z_{m-1}^{D_{b_m}} \leq Z_m \leq 2^{l_m - b_m} Z_{m-1}^{D_{b_m}}$$

- Presence of $2^{l_m - b_m} > 1$ renders it challenging
- The sequence $(Z_m)_m$ converges to 0 exponentially in $-N$
- The probability of convergence to 0 is equal to I_0 (Bernoulli distribution)
- Annihilate the role of the constant $2^{l_m - b_m}$ as N grows infinite
- Generalization, and little corrections, of the proof of [Sasoglu'12]

Conclusions

Thank you !

Questions ?