Lecture 8: Polar Codes

November 18, 2016
Polar Codes

1. Introduction
2. Coding
3. Decoding
1. Introduction

Polar codes, a class of codes which allows to
1. attain the capacity of all symmetric memoryless channels (≡those for which the capacity is attained for a uniform input distribution),
2. with an encoding algorithm of complexity $O(N \log N)$ ($N =$ code length),
3. with a decoding algorithm of complexity $O(N \log N)$.

This decoding algorithm borrows many ideas from the decoding algorithm used for LDPC codes.
Polar Codes

1. a coding architecture based on the Fast Fourier Transform by fixing some bits to 0,
2. a (suboptimal) decoding algorithm which computes the probability that the input bits are equal to 0 given the previous input bits and the probabilities of the output bits.
Encoding: example

positions in red = information
Polar Code : linear code

In the previous case, it is a code of generator matrix

\[ G = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \]
Decoding: example on the erasure channel
Decoding: using the properties of the "base code"
Decoding: an example where all erasures can be recovered
A configuration where erasures can be partially recovered
Another configuration where erasures can be partially recovered

\[ a \quad u \quad a \quad u \]

\[ ? \quad ? \quad a+u \quad ? \]
2. Encoding of a polar code

length : \( N = 2^n \), dimension : \( 0 \leq k \leq N \).

Choosing an ensemble \( \mathcal{F} \) of size \( N - k \) of positions \( \subset \{0, \ldots, N - 1\} \) fixed to 0.

\( B_t = \) subset of \( \{0, \ldots, N - 1\} \) whose \( t \)-th bit is equal to 0.

Encoding algorithm.

Input : \( u \in \{0, 1\}^N \), \( u_i = 0 \) if \( i \in \mathcal{F} \).
Output : \( x \) the codeword corresponding to \( u \).

\[
x \leftarrow u \\
\text{for } t = 0 \text{ to } n - 1 \text{ do} \\
\quad \text{for all } i \in B_t \text{ do} \\
\quad\quad x_i \leftarrow x_i \oplus x_{i+2t} \\
\quad \text{end for} \\
\text{end for} \\
\text{return } x
\]
3. Decoding

Decoding algorithm

Input : \( y \in A^N \) output of the channel corresponding to codeword \( x \)
Output : an estimate \( \hat{u} \) for \( u \).

\[
\text{for all } i \in \{0, 1, \ldots, N - 1\} \setminus \mathcal{F} \text{ do}
\]
\[
\text{Compute } p_i \overset{\text{def}}{=} \text{Prob}(u_i = 1|y, \hat{u}_0, \ldots, \hat{u}_{i-1})
\]
\[
\text{if } p_i > 0.5 \text{ then}
\]
\[
\hat{u}_i = 1
\]
\[
\text{else}
\]
\[
\hat{u}_i = 0
\]
\[
\text{end if}
\]
\[
\text{end for}
\]
Why does this work?

- How to choose $\mathcal{F}$?
- Can $p_i$ be computed efficiently?
- Why does this procedure work at all?
A simple computation

We know $p_1 = \text{Prob}(x_1 = 1|y_1)$ and $p_2 = \text{Prob}(x_2 = 1|y_2)$. We compute

$$q_1 = \text{Prob}(u_1 = 1|y_1, y_2).$$
The formula

**Lemma 1.** Let $X_1$ and $X_2$ be two independent binary random variables and denote by $r_i \overset{\text{def}}{=} \text{Prob}(X_1 = 1)$, then

$$\text{Prob}(Z_1 \oplus Z_2 = 1) = \frac{1 - (1 - 2r_1)(1 - 2r_2)}{2}$$

**Application:**

$$q_1 = \frac{1 - (1 - 2p_1)(1 - 2p_2)}{2}$$
Another simple computation

We know $p_1 = \text{Prob}(x_1 = 1|y_1)$, $p_2 = \text{Prob}(x_2 = 1|y_2)$ and $u_1$. We compute $q_2 = \text{Prob}(u_2 = 1|u_1, y_1, y_2)$. 
Lemma 2. A uniformly distributed random bit $B$ is sent through two memoryless channels, $y_1$ and $y_2$ are the corresponding outputs. If we denote by $r_i = \text{Prob}(B = 1|y_i)$, then

$$\text{Prob}(B = 1|y_1, y_2) = \frac{r_1r_2}{r_1r_2 + (1 - r_1)(1 - r_2)}.$$ 

Application:

$$q_2 = \frac{p_1p_2}{p_1p_2 + (1 - p_1)(1 - p_2)} \quad \text{if } u_1 = 0$$

$$q_2 = \frac{(1 - p_1)p_2}{(1 - p_1)p_2 + p_1(1 - p_2)} \quad \text{if } u_1 = 1$$
Notation

We denote by $u^t_i$ the input of the encoding circuit at level $t$ ($u^0_i = u_i$) and $p^t_i$ are the probabilities for these bits that are computed or given by the channel when $t = n$.
Decoding algorithm (full version)

for $i = 0$ to $N - 1$ do
    for $t = 1$ to $n - 1$ do
        Compute all the $u_j^t$'s that can be obtained directly from the already known $u_j^{t-1}$'s for $l < i$.
    end for
    for $t = n - 1$ to 0 do
        Compute all the $p_j^t$'s that can be obtained directly from the $p_j^{t+1}$'s and the known $u_j^t$'s (with $l < i$)
    end for
    if $i \notin \mathcal{F}$ then
        if $p_i^0 > 0.5$ then
            $u_i^0 = 1$
        else
            $u_i^0 = 0$
        end if
    end if
end for

Information Theory
Correction of the algorithm

Lemma 3. At step $i$ of the outer loop and step $t$ of the inner loop of the previous algorithm, the $u_{i}^{t}$’s that can be calculated correspond to the indices $j$ in the set
$$\{j: 0 \leq j \leq 2^{t}\left\lfloor \frac{i}{2^{t}} \right\rfloor - 1\}.$$  

Lemma 4. At step $i$ of the outer loop and step $t$ of the inner loop of the previous algorithm, the $p_{j}^{t}$’s that can be calculated correspond to the indices $j$ in the set
$$\{j: 0 \leq j \leq 2^{t}\left\lfloor \frac{i+2^{t}}{2^{t}} \right\rfloor - 1\}.$$  

Corollary 1. $p_{i}^{0}$ can always be computed at step $i$ of the outer loop.
Modeling the decoder

Decoding of the base code can be modeled by the following transmission over two different channels:

\[ u_1 \xrightarrow{\text{channel 1}} y_1, y_2 \]
\[ u_2 \xrightarrow{\text{channel 2}} u_1, y_1, y_2 \]

and we know the channels, they provide \( \text{Prob}(u_1 = 1 | y_1, y_2) \) and \( \text{Prob}(u_2 = 1 | u_1, y_1, y_2) \).
Case of an erasure channel

Assume that

\[
\begin{align*}
&x_1 \xrightarrow{\text{erasure channel of prob. } p_1} y_1 \\
&x_2 \xrightarrow{\text{erasure channel of prob. } p_2} y_2
\end{align*}
\]

\[\text{Prob}(u_1 \text{ is erased}) = \text{Prob}(x_1 \oplus x_2 \text{ erased }) \]
\[= \text{Prob}(x_1 \text{ or } x_2 \text{ erased }) \]
\[= 1 - (1 - p_1)(1 - p_2) \]
\[= p_1 + p_2 - p_1p_2 \]

\[\text{Prob}(u_2 \text{ is erased}) = \text{Prob}(x_1 \text{ and } x_2 \text{ erased }) \]
\[= p_1p_2 \]
Induced channel model in the case of the erasure channel

If we denote by $C(p)$ the capacity of the erasure channel of probability $p$ ($C(p) = 1 - p$) then

$$C(p_1) + C(p_2) = C(p_1 + p_2 - p_1p_2) + C(p_1p_2).$$
Equivalent models for $p = 0.25$ and $n = 3$

We choose the positions in red for $\mathcal{F}$. 
Equivalent models for $n \in \{5, 8, 16\}$

"$\text{Prob}(q > p)$" = $\frac{\#\{i: q > p\}}{2^n}$
Why the whole scheme works and attains the capacity of the erasure channel

Point 1: The equivalent channels “polarize”, the erasure probability is either close to 0 or 1.

Point 2: The “conservation law” (1) $C(p_1) + C(p_2) = C(p_1 + p_2 - p_1 p_2) + C(p_1 p_2)$ ensures that

$$\sum_{i=0}^{N-1} C(q_i) = \sum_{i=0}^{N-1} C(p_i) = NC(p)$$

with $q_i =$ capacity of the $i$–th equivalent channel at the input and $p_i =$ capacity of the $i$–th output channel.

Point 3: Since either $C(q_i) \approx 0$ or $C(q_i) \approx 1$,

$$k \stackrel{\text{def}}{=} N - |\mathcal{F}| \stackrel{\text{def}}{=} \# \{ i : C(q_i) \approx 1 \} \approx NC(p)$$
The general case: the basic scheme

Assumption: $U_1$ and $U_2$ independent and uniformly distributed in $\{0, 1\}$.

\[ I(U_1; Y_1, Y_2) + I(U_2; U_1, Y_1, Y_2) = I(X_1; Y_1) + I(X_2; Y_2). \]
A lemma on the independence of random variables

**Lemma 5.** $U_1$ and $U_2$ independent and uniformly distributed, 
$\Rightarrow X_1$ and $X_2$ independent and uniformly distributed 
$\Rightarrow Y_1$ and $Y_2$ independent.

**proof:** $X_1$ and $X_2$ independent and uniformly distributed (obvious).
Proof (cont’d)

\[
P(Y_1 = a, Y_2 = b) = \sum_{c,d} P(Y_1 = a, Y_2 = b | X_1 = c, X_2 = d) P(X_1 = c, X_2 = d)
\]

\[
= \sum_{c,d} P(Y_1 = a | X_2 = c) P(Y_2 = b | X_2 = d) P(X_1 = c) P(X_2 = d)
\]

\[
= S_1 S_2 \quad \text{with}
\]

\[
S_1 = \sum_c P(Y_1 = a | X_1 = c) P(X_1 = c) = P(Y_1 = a)
\]

\[
S_2 = \sum_d P(Y_2 = b | X_2 = d) P(X_2 = d) = P(Y_2 = b)
\]

Hence

\[
P(Y_1 = a, Y_2 = b) = P(Y_1 = a) P(Y_2 = b)
\]
An important lemma in information theory

**Lemma 6.** If $Y_i$ is the output corresponding to $X_i$ after transmission through a memoryless channel

\[ I(X_1, X_2; Y_1, Y_2) \leq I(X_1; Y_1) + I(X_2; Y_2). \]

If $Y_1$ and $Y_2$ are independent

\[ I(X_1, X_2; Y_1, Y_2) = I(X_1; Y_1) + I(X_2; Y_2). \]
Proof

\[ I(X_1, X_2; Y_1, Y_2) = H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) \]
\[ = H(Y_1) + H(Y_2) - H(Y_1|X_1, X_2) - H(Y_2|X_1, X_2, Y_1) \]
\[ (\text{independence of the } Y_i \text{'s}) \]
\[ = H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \]
\[ (\text{memoryless channel}) \]
\[ = I(X_1; Y_1) + I(X_2; Y_2) \]
\[ (\text{definition of mutual information}) \]
Proof of Theorem 1

\[ I(X_1; Y_1) + I(X_2; Y_2) = I(X_1, X_2; Y_1, Y_2) \]
\[ = I(U_1, U_2; Y_1, Y_2) \]
\[ = H(U_1, U_2) - H(U_1, U_2|Y_1, Y_2) \]
\[ = H(U_1) + H(U_2) - H(U_1|Y_1, Y_2) - H(U_2|U_1, Y_1, Y_2) \]
\[ = I(U_1; Y_1, Y_2) + I(U_2; U_1, Y_1, Y_2) \]