Lecture 8: Analysis of Polar Codes

Debsoumya Chakraborti* and Mihir Hasabnis†

November 8, 2018

1 Introduction

Let us recall what we studied in last class.

- Goal was to achieve the capacity for BSC_p using (explicit) codes with block length and encoding/decoding complexity that are $poly(\frac{1}{\epsilon})$ where ϵ is the gap to capacity. $\epsilon = 1 - h(p) -$ rate.
- We also reduced the above task to linear compression of $Ber(p)^{\otimes n}$ which is n i.i.d. Bernoulli R.V.s.

Remember the compression and decompression technique we learnt in the last lecture. If z is from $Ber(p)^{\otimes n}$ distribution, then one can find a matrix H such that $W = Hz \in \{0,1\}^{(h(p)+\epsilon)n}$ with high probability.

We have seen that if P is a polarizing transform, then $z \to Pz \in \{0,1\}^n$, where $H(W_i|W_{<i})$ is either very close to 0 or very close to 1. Hence it implies that P is a linear compression which keeps entries W_i such that $H(W_i|W_{<i}) \geq \frac{1}{n^c}$.

We will now show that $P_n = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]^{\otimes m}$ is a polarizing transform where $n = 2^m$. This work is originally due to Arikan [1]. We have already seen that the encoding and decoding schemes are efficient and run in $O(n \log n)$ and have proved everything except the P_n is the (ϵ, τ)-polarizing.

In order to show P_n polarizing, we introduced a *slightly* different matrix R_n to analyze the polarization. Our idea is to show that if R_n is (ϵ, τ)-polarizing then so is P_n. Refer to [3] section 8 for a detailed discussion about R_n.

2 Polarizing transform: R_n is polarizing

Theorem 2.1. *For all C and for all $\epsilon > 0$, $\exists n_0$ polynomial such that for all $n \geq n_0(\frac{1}{\epsilon})$ we have*

$$\mathbb{P}_{i \in [n]} \left[H(W_i|W_{<i}) \geq \frac{1}{n^c} \right] \leq h(p) + \epsilon.$$

*Department of Mathematical Sciences, Carnegie Mellon University. Email : dchakrab@cmu.edu

†Department of Mathematical Sciences, Carnegie Mellon University. Email : mhasabni@andrew.cmu.edu

where $W = R_n Z$ *with* $Z \sim Ber(p)^n$.

Proof. Recall $S = S_\tau = \{i : H(W_i|W_{<i}) \geq \tau\}$ and $|S| \leq (h(p) + \epsilon)n$. The statement we'll prove is that

$$\# \left\{ i : H(W_i|W_{<i}) \in \left(\frac{1}{n^c}, 1 - \frac{1}{n^c} \right) \right\} \leq \frac{\epsilon}{2} n.$$

The above statement is a nicer statement than the theorem. This symmetric looking statement also justifies the name polarizing.

Note that $H(W) = \sum_{i=1}^{n} H(W_i|W_{<i}) = h(p)n = H(Z)$. Hence an elementary averaging argument proves the theorem. □

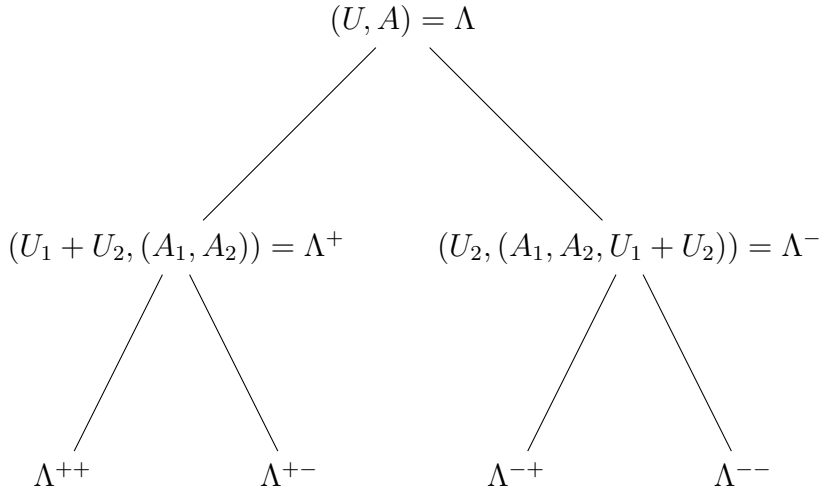
Now as we have mentioned in the last lecture, break the input into 2 parts $(B_0, B_1, ..., B_{n/2-1})$ and $(C_0, C_1, ..., C_{n/2-1})$, where each B_i, C_i are independent and identically distributed. For $0 \leq i \leq n/2 - 1$, define $D_{2i} = B_i + C_i$ and $D_{2i+1} = C_i$. For convenience let's introduce the notation $A_{<i}$ denoting the sequence $A_0, A_1, ..., A_{i-1}$.

Fact 1.

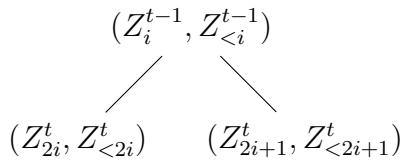
$$H(D_{2i}|D_{<2i}) = H(D_{2i}|B_{<i}, C_{<i})$$
$$H(D_{2i+1}|D_{<2i+1}) = H(D_{2i+1}|B_{<i}, C_{<i}, D_{2i})$$

3 Evolution of pair of random variables

Let (U, A) be such that U is a distribution on $\mathbb{F}_2$ and A is a set of finitely many random variables. Let (U_i, A_i) be i.i.d.s according to the distribution of (U, A). In the diagram below Λ denotes a pair of random variables.



Let's define a tree process starting from (U, A) as shown above, the left child of (U, A) is $(U_1 + U_2, (A_1, A_2))$ and the right child is $(U_2, (A_1, A_2, U_1 + U_2))$. We will now analyze the process starting with the pair of random variable $(Z, \phi) = (Ber(p), \phi)$. It's clear that the vertices at height t will look like $(Z_i^t, Z_{<i}^t)$ for $0 \leq i \leq 2^t - 1$.



Observe that by using induction on t, the two children of the pair $(Z_i^{t-1}, Z_{<i}^{t-1})$ in height $t-1$ will be of the form $(Z_{2i}^t, Z_{<2i}^t)$ and $(Z_{2i+1}^t, Z_{<2i+1}^t)$ as depicted above. Note that the number of conditional entropies in level t is exactly 2^t.

Definition 1. *For all $t \geq 0$, define $X_t = H(Z_i^{(t)}|Z_{<i}^{(t)})$ for i chosen uniformly at random in $\{0, 1, 2 \dots 2^t - 1\}$. Note that $X_t \in [0, 1]$.*

The real motivation behind the above definition is to be able to show that when t is large enough, all the conditional entropies in height t are either very close to 0 or very close to 1. Rigorously speaking, we want to show that as $t \to \infty, X_t \to$ Ber $r.v$ in some sense which is made formal below, i.e. $X_t \in (\zeta, 1-\zeta) \to 0$ w.h.p
Formally we would like to show the following:

Theorem 3.1. *Polynomially Strong Polarization: $\forall t\ \forall \gamma > 0\ \exists \alpha < 1\ \beta < \infty$ such that*

$$\mathbb{P}[X_t \in (\gamma^t, 1-\gamma^t)] \leq \beta\alpha^t$$

Note that the probability here is over the evolution of X_t, i.e. as you walk down the tree randomly, see where you end up at the t-th level, and that's the value of X_t. Note that by the Conservation of Entropy, $H(U_1 + U_2|A_1, A_2) + H(U_2|A_1, A_2, U_1 + U_2) = 2H(U|A)$. Hence the expected value of X_t stays the same, moreover $E(X_t|X_{t-1}) = E(X_{t-1})$ which is a property of a martingale.

It turns out that γ is related to the decoding error probability, and we will need some γ to be less than $\frac{1}{2}$ to make sure that the small values in n-th level which are close to 0 don't add up to exceed 1. α is related to the polynomial block length, which is a function of gap ϵ to capacity. Now we'll take the following three steps one by one.

- Showing abstract local polarization to capture the essence of local evolution of $\{X_t\}$.
- Implying strong (global) polarization from the local polarization.
- Verifying that X_t defined above satisfies local polarization.

Definition 2. *Local Polarization: A sequence of r.v.s $X_0, X_1, \dots X_t \dots$ is locally polarizing if*

- *(Unbiased) $\mathbb{E}[X_{j+1}|X_j = a] = a$ for all j and for all a*
- *(Variance in the middle) $\forall \tau\ \exists \theta = \theta(\tau)$ such that if $X_j \in (\tau, 1-\tau)$ then almost surely*

$$|X_{j+1} - X_j| \geq \theta \tag{3.1}$$

- *(Suction at the ends)* $\forall c < \infty \; \exists \tau = \tau(c) > 0$ *such that*

$$X_j \leq \tau \;\; \textit{and} \;\; \mathbb{P}\left[X_{j+1} \leq \frac{X_j}{c}\right] \geq \frac{1}{2} \tag{3.2}$$

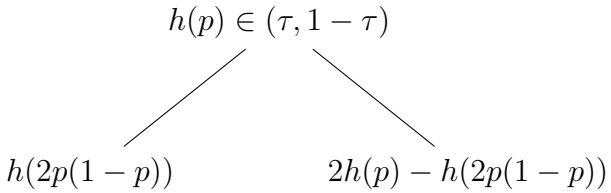
$$X_j \geq 1 - \tau \;\; \textit{and} \;\; \mathbb{P}\left[1 - X_{j+1} \leq 1 - \frac{X_j}{c}\right] \geq \frac{1}{2} \tag{3.3}$$

Eventually we want to know the subsets with $X_t \geq \tau$. The polarization pattern is not so simple analytically. As a remark let's mention that τ- polarizing set S_τ can be estimated by a randomized algorithm. For details, look at the problem 3 of the homework set 3.

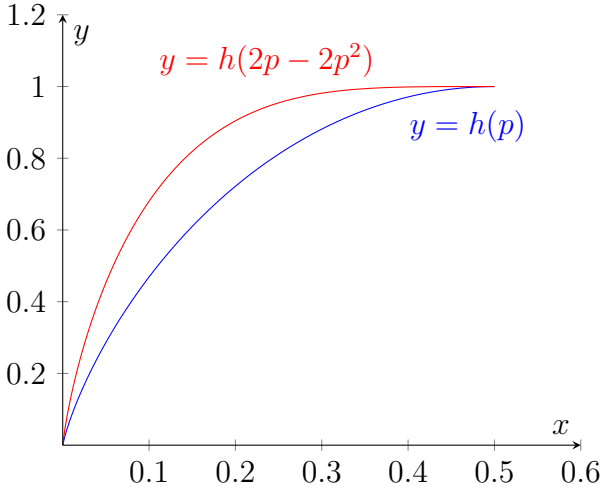
We want to show that X_t in Definition 1 satisfies local polarization and then show that local polarization $\implies$ strong polarization.

The "unbiased" property follows by using the chain rule, i.e. $H(U_1 + U_2|A_1, A_2) + H(U_2|A_1, A_2, U_1 + U_2) = H(U_1 + U_2, U_2|A_1, A_2) = H(U_1, U_2|A_1, A_2) = 2H(U|A)$.

For variance in the middle, we'll just present a rough sketch as the original detailed proof is quite messy and can be found in []. Let's first try to prove it for the special case $A = \phi$, i.e. unconditioned case and $U = Ber(p)$. Let $H(U) = h(p) \in (\tau, 1 - \tau)$, then the first child of (U, A) is $Ber(2p(1 - p))$, hence the entropies in the two children nodes of (U, A) will be $h(2p(1 - p))$ and $2h(p) - h(2p(1 - p))$.



Now the "variance in middle" statement will be clear form the plot of entropy function as below. The argument can be made formal by using some basic calculus.



Now let's try to handle the general situation, i.e. general A. We have the following lemma which essentially shows that minimum change in the value of entropy happens in the unconditioned case.

Lemma 3.2 (Mrs. Gerber's Lemma)**.** *Let (U_i, A_i) are i.i.d. and $H(U_i|A_i) = \alpha$. Then $H(U_1 + U_2|A_1, A_2) - H(U|A)$ is minimized when $\forall a \in supp(A), H(U|A = a) = \alpha$.*

We define $a * b = a(1-b) + (1-a)b$. The proof of this lemma uses some clever convexity arguments, in particular it uses the strict convexity of the function $f_a(x) = h(a * h^{-1}(x))$ for $x \in [0, 1]$.

We'll try to make this lemma a little bit intuitive by mentioning an equivalent form of it. Suppose X^n is a sequence of bits from some distribution and it has some reasonable entropy, i.e. $H(X^n) \geq n\alpha$. Now send these n bits over a BSC channel to get $Y^n = X^n + Z^n$ where $Z^n \in Ber(p)^{\otimes n}$. Then $H(Y^n) \geq n \cdot h(p * h^{-1}(\alpha))$ where equality holds if and only if X^n is i.i.d. with $X_i \sim Ber(h^{-1}(\alpha))$. This is some sort of entropic version of isoperimetric inequality.

Now let's talk about "suction at the ends". By Mrs. Gerber's Lemma, it's enough to prove for the unconditioned case. Also, remember that the unconditional case basically gives us the minimal separation.

- Let $h(\zeta) \approx \tau$ i.e. $\zeta \approx \frac{\tau}{\log 1/\tau}$. The lower entropy would be $2h(\zeta) - h(2\zeta(1-\zeta))$. Now,
$$\begin{aligned} 2h(\zeta) - h(2\zeta(1-\zeta)) &\approx 2\zeta \log \frac{1}{\zeta} - 2\zeta(1-\zeta) \log \frac{1}{2\zeta(1-\zeta)} \\ &= 2\zeta(1-\zeta) - 2\zeta(1-\zeta) \log \frac{1}{1-\zeta} \\ &\leq 2\zeta \\ &\leq \frac{2\tau}{\log \frac{1}{\tau}} \end{aligned}$$
- Let $h(\frac{1}{2} - \eta) \approx 1 - \tau$ i.e. $\eta \approx \sqrt{\tau}$. The higher entropy would be $h\left(2\left(\frac{1}{2} - \eta\right)\left(\frac{1}{2} + \eta\right)\right)$.
$$\begin{aligned} h\left(2\left(\frac{1}{2} - \eta\right)\left(\frac{1}{2} + \eta\right)\right) &= h\left(\frac{1}{2} - 2\eta^2\right) \\ &= 1 - \theta(\eta^4) \\ &= 1 - \theta(\tau^2) \end{aligned}$$

  Hence we have verified that X_t satisfies the properties of local polarization.

4 Local Polarization implies Poly Strong Polarization

Given γ, we'll pick c suitably large enough, then pick $\tau(c) > 0$ depending on c, and lastly $\theta(\tau) > 0$. The dependence will look like $\gamma \to c \to \tau(c) \to \theta(\tau)$. Note that if γ is a given constant, all other parameters are just constants.

We will now do a 2-phase analysis.

- Phase 1: Moderate Polarization
  $\mathbb{P}[X_{t/2} \in (\alpha_1{}^t, 1 - \alpha_1{}^t)] \leq \alpha_1{}^t$ for some $\alpha_1 < 1$. Sadly α_1 can be closer to 1, remember that we need this constant to be below $\frac{1}{2}$ which will be achieved in the next phase.
- Phase 2: Conditioned on moderate polarization, the sequence gets highly polarized, i.e. the probability of $X_t \in (\gamma^t, 1 - \gamma^t)$ is bounded by $\exp\{-t\}$, for any $\gamma > 0$, as desired.

4.1 Analysis of First Phase

We will use a simple potential based argument.
Define $\phi_j = \max\{\sqrt{X_j}, \sqrt{1-X_j}\}$. Note that ϕ_j is a concave function. Our goal is to show that ϕ_j is expected to drop by a constant factor at each step.

Lemma 4.1. $\mathbb{E}[\phi_{j+1}|\phi_j = a] \leq \left(1 - \frac{\theta^2}{16}\right) a$

Proof. WLOG $X_j \leq \frac{1}{2}$. $a = \phi_j = \sqrt{X_j}$. Now for some $\delta > 0$,

$$X_{j+1} = \begin{cases} a^2 + \delta, & \text{with probability } = 1/2 \\ a^2 - \delta, & \text{with probability } = 1/2 \end{cases}$$

Using local evolution, $\delta \geq \theta a^2$ because we have $\delta \geq \theta$ from 3.1 and $\delta \geq (1 - \frac{1}{c})a^2$ from 3.2.
$\mathbb{E}[\phi_{j+1}] = \frac{1}{2}\left(\sqrt{a^2+\delta} + \sqrt{a^2-\delta}\right) \leq \frac{1}{2}\left(\sqrt{a^2+\theta a^2} + \sqrt{a^2-\theta a^2}\right) \leq a\left(1 - \frac{\theta^2}{16}\right)$ (Using Concavity)
Hence, $\mathbb{E}[\phi_{t/2}] \leq \left(\sqrt{1-\frac{\theta^2}{16}}\right)^t = {\beta_1}^t$
Using Markov's inequality we get that $\mathbb{P}[\phi_{t/2} \geq {\alpha_1}^t] \leq \left(\frac{\beta_1}{\alpha_1}\right)^t = {\alpha_1}^t$ □

We take $\alpha_1 = \sqrt{\beta_1}$.

4.2 Analysis of Second Phase

Now after $\frac{t}{2}$ steps if we reach a value in $(\alpha_1^t, 1-\alpha_1^t)$ i.e. $X_{t/2} \in (\alpha_1^t, 1-\alpha_1^t)$, we give up. Otherwise our strategy will be to show the following:

- If $X_{t/2} \leq \lambda << $ (say $\lambda = \tau^2$), then X_j is unlikely to become $> \tau$ for any $j \in [\frac{t}{2}, t]$. Same thing for $X_{t/2} \geq 1 - \tau$ case as well.
- If X_j never becomes $> \tau$, then it falls rapidly ($\leq \gamma^t$) w.h.p.

Let's do the second step first. We define random variables.

$$\Lambda_i = \begin{cases} 1 & \text{if } X_{t/2+i+1} < X_{t/2+i} \\ 0 & \text{else} \end{cases} \tag{4.1}$$

Define $\Lambda = \sum_{i=0}^{t/2} \Lambda_i$. Intuitively if $\Lambda_i = 1$, then it took a good branch in $t/2 + i$-th level. So Λ is keeping a count of good branches. As Λ_i s are independent, by using Chernoff bound we get $\mathbb{P}[\Lambda \leq t/8] \leq \exp(-3t)$. If $X_j \leq \tau$ for all $j \in [\frac{t}{2}, t]$, then using the fact that $X_{i+1} \leq 2X_i$ and 3.2, $X_t \leq 2^{\frac{t}{2}} \cdot c^{-\Lambda} \cdot X_{t/2}$. Hence with probability $\geq 1 - \exp(-3t)$, $X_t \leq 2^{\frac{t}{2}} \cdot c^{-\frac{t}{8}} \leq \gamma^t$ for a suitable choice of c.

Lemma 4.2. *Special case of Doob's inequality for martingales : If $X_0 \leq \lambda$, then the probability that there exists j such that $X_j \geq \tau$ is at most $\frac{\lambda}{\tau}$.*

Proof. We want to show that $\forall T > 0,\ \mathbb{P}[\max_{t \le T} x_t \ge \tau] < \frac{\lambda}{\tau}$
Define Y_i as follows

$$Y_0 = X_0$$

$$Y_i = \begin{cases} X_i & \text{if } Y_{i-1} < \tau \\ Y_{i-1} & \text{else} \end{cases}$$

Note $\mathbb{E}[Y_{i+1}|Y_i = a] = a$
$\mathbb{P}[\max_{t \le T} x_t \ge \tau] = \mathbb{P}[Y_T \ge \tau] \le \frac{\mathbb{E}[Y_T]}{\tau}$
$\mathbb{E}[Y_T] = \mathbb{E}[Y_{T-1}] = \cdots = \mathbb{E}[X_0] = \lambda$. So we are done. □

Now we can finish the whole argument as below. $\mathbb{P}[X_t \in (\gamma_1^t, 1-\gamma_1^t)] \le \alpha_1^t + \frac{\lambda}{\tau} + \eta^t$ where we choose $\tau = (\sqrt{\alpha_1})^t$.

References

[1] Erdal Arikan. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on Information Theory,* 55(7):30513073, 2009.

[2] Venkatesan Guruswami and Patrick Xia. Polar Codes: Speed of polarization and polynomial gap to capacity. *IEEE Transactions on Information Theory,* 61(1):316, 2015.

[3] Riazanov, Andrii and Sandeep, Sai. Lecture 7: Coding theory (*CMU: Fall 2018*) `http://www.cs.cmu.edu/~venkatg/teaching/au18-coding-theory/lec-scribes/polar-part-1.pdf`