CS294-226: Advances in Error-Correcting Codes **UC Berkeley, Fall 2022**
Lecture #14: Analysis of Polarization October 17, 2022
Lecturer: Venkatesan Guruswami Scribe: Hongxun Wu

1 Recap

Last Class We saw

- Efficient compression & decompression algorithms.
- The statement of polarization theorem.

We defined Arikan's polar code, which is the transformation

$$P_n = G_2^{\otimes m} = \begin{pmatrix} P_{n/2} & P_{n/2} \\ 0 & P_{n/2} \end{pmatrix}.$$

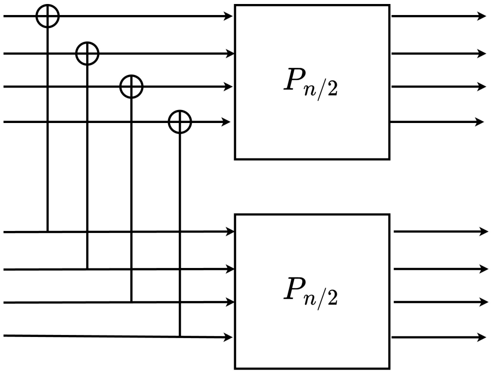

Figure 14.1: Construction of P_n.

For i.i.d. $z_1, z_2, \ldots, z_n \sim \text{Bernoulli}(p)$, we apply this transform and get $w = P_n z$. We gave the statement of the following theorem saying that the conditional entropy of $w_1, w_2, \ldots, w_n$ polarizes.

Theorem 14.1 (Polarization Theorem)**.** *$\forall \epsilon > 0$, $\exists n_0 \le \text{poly}(1/\epsilon)$ s.t. $\forall n \ge n_0$,*

$$\Pr_{i \in [n]} \left[H(w_i \mid w_{<i}) \ge \frac{1}{n^3} \right] \le h(p) + \epsilon.$$

Equivalently, if we define $S_{1/n^3} = \left\{ i \mid H(w_i \mid w_{<i}) \ge \frac{1}{n^3} \right\}$, we would have $|S_{1/n^3}| \le (h(p)+\epsilon)n$.

Note here the dependency of m and ϵ is $m = O(\log 1/\epsilon)$, i.e. exponentially fast convergence.

Instead of proving Theorem 14.1, we will prove a more symmetric version (Theorem 14.2).

Theorem 14.2 (Most entropies polarize to boundary)**.** *$\forall \epsilon > 0$, $\exists n_0 \le \text{poly}(1/\epsilon)$ s.t. $\forall n \ge n_0$,*

$$\left| \left\{ i \;\middle|\; H(w_i \mid w_{<i}) \in \left(1/n^3, 1 - 1/n^3\right) \right\} \right| \le \frac{\epsilon}{2} n$$

Proof of Theorem 14.1. We set $n_0 = 1/\epsilon$. By chain rule, we have $\mathbf{E}_{i \in [n]} \left[H(w_i \mid w_{<i}) \right] = h(p)$. From Markov inequality, we know that $\mathbf{Pr}_{i \in [n]}[H(w_i \mid w_{<i}) > 1 - \frac{1}{n^3}] \le \frac{h(p)}{1 - 1/n^3} \le h(p) + \frac{\epsilon}{2}$.

Together with Theorem 14.2, this implies Theorem 14.1. □

2 Reordering and Splitting Channels

Reordering Output To better present the proof of Theorem 14.2, we will be working with another ordering of $w_1, \ldots, w_n$ given by the following transformation R_n. This is also the ordering in Arikan's original paper [Ari09]. At the base of recursion, we have $R_1 = I$ and $R_2 = P_2$.

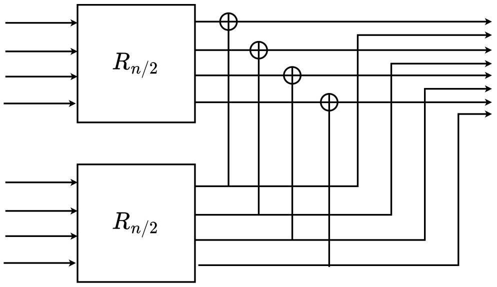

Figure 14.2: Construction of R_n.

- Previously: $P_n(a, b) = (P_{n/2}(a + b), P_{n/2}(b))$ (See Figure 14.1).
- Now: $R_n(a, b)_{2i-1} = R_{n/2}(a)_i + R_{n/2}(b)_i$ and $R_n(a, b)_{2i} = R_{n/2}(b)_i$ (See Figure 14.4).

We claim that up to a permutation of the outputs $R_n Z \equiv P_n Z$. Why?

- $R_n = B_n P_n$ where B_n is the bit-reversal permutation matrix, i.e. $B_n(b_1 b_2 \ldots b_n) = b_n b_{n-1} \ldots b_1$.
- $B_n P_n = P_n B_n$.
- $R_n = P_n B_n$, $R_n Z = P_n B_n Z = B_n P_n Z \equiv P_n Z$.

Recursive Analysis In our new ordering, let us denote output of $R_{n/2}$ by $R_{n/2}(a) = u_1 u_2 \ldots u_{n/2}$ and $R_{n/2}(b) = v_1 v_2 \ldots v_{n/2}$. Then $R_n(a, b) = w_1 w_2 \ldots w_n = (u_1 + v_1, v_1, u_2 + v_2, v_2, \ldots, u_{n/2} + v_{n/2}, v_{n/2})$.

Recall that we wish to analyze $H(w_{2i-1} \mid w_{<2i-1})$ and $H(w_{2i} \mid w_{<2i})$. The benefit of our new ordering is that conditioning on $(w_{2i-1}, w_{2i}) = (u_i + v_i, v_i)$ is equivalent to conditioning on (u_i, v_i). Hence,

- $H(w_{2i-1} \mid w_{<2i-1}) = H(u_i + v_i \mid u_{<i}, v_{<i})$
- $H(w_{2i} \mid w_{<2i}) = H(v_i \mid u_{<i}, v_{<i}, u_i + v_i)$.

Channel Splitting We can equivalently view this as channels. A channel is nothing but two correlated random variables. Fixing i, for i.i.d random bits $z = z_1 z_2 \ldots z_{n/2} \sim \text{Bernoulli}(p)^{n/2}$, we let $A = R_{n/2}(z)_i$ and $B = R_{n/2}(z)_{<i}$. These two correlated random variable give us a channel $\Lambda = (A; B)$. Then the quantity we care about, $H(R_{n/2}(z)_i \mid R_{n/2}(z)_{<i}) = H(A \mid B)$, is just the entropy of channel input conditioning on its output.

Note $(u_i; u_{<i})$ and $(v_i, v_{<i})$ is simply two independent samples from Λ. To emphasize this, let us denote them as $(A_1, B_1), (A_2, B_2)$ respectively. Applying the recursive analysis above, we see that $H(w_{2i-1} \mid w_{<2i-1})$ and $H(w_{2i} \mid w_{<2i})$ corresponds to the entropy of two different channels, Λ^+ and Λ^-. They are defined as Figure 14.4. Λ^+ is typically a worse channel with higher entropy, while Λ^- is typically a better channel. This is how entropy polarize locally.

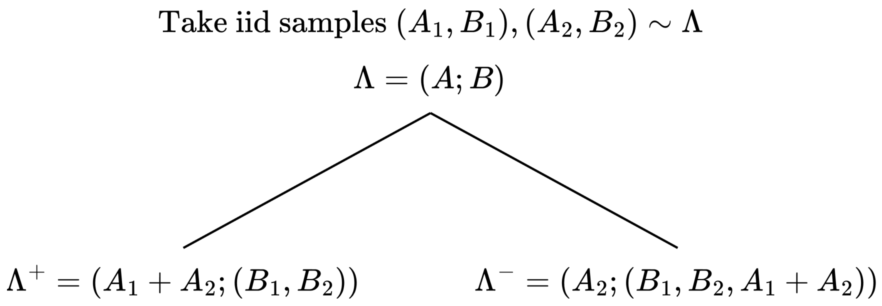
Take iid samples $(A_1, B_1), (A_2, B_2) \sim \Lambda$
$\Lambda = (A; B)$
$\Lambda^+ = (A_1 + A_2; (B_1, B_2))$
$\Lambda^- = (A_2; (B_1, B_2, A_1 + A_2))$

Figure 14.3: Channel Splitting.

Note here we always have $H(\Lambda^+) + H(\Lambda^-) = 2H(\Lambda)$ by chain rule. The factor of two comes from two independent samples.

Channel Splitting Tree At the very beginning, you only have one kind of random variable which is Bernoulli(p). This gives you the channel $\Lambda^{(0)}$ which is $A \sim$ Bernoulli$(p), B = \emptyset$. Then you repeatedly apply this procedure and get a binary tree of channels.

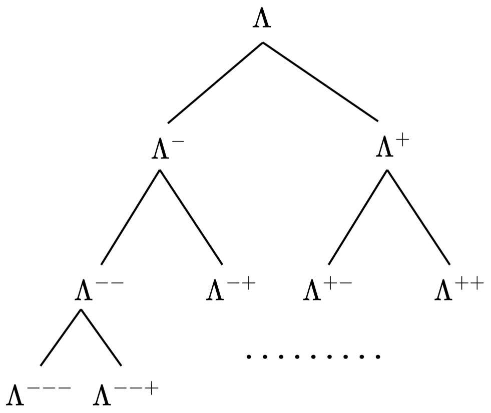

Figure 14.4: Iterative application of channel splitting.

On this tree, each bit w_i in our final output corresponds to the input of a leaf channel. The output of that channel contains exactly the same information as $w_{<i}$. This observation leads to the following claim.

Claim 14.3. *After apply m times, for $i \in \{1, \ldots, 2^m\}$,*

$$H(w_i \mid w_{<i}) = H(\Lambda^{(b_1, b_2, \ldots, b_m)})$$

where b_j is $+/-$ depending on the j-th bit of $i - 1$ being $0/1$.

3 Martingale convergence

Martingale We want to show that $\mathbf{Pr}[H(w_i \mid w_{<i})$ not close to $0] \to 0$. Equivalently speaking, deep in the tree, most channels have entropies $\approx 0/1$. To see this, we need a certain finite version of martingale convergence. We first define our martingale.

Definition 14.4. We define random variable X_t which is the entropy of a random channel at depth t.

$$X_t = H(\Lambda^{b_1, b_2, \ldots b_t}).$$

Here X_t is a random variable supported in $[0,1]$. determined by the random signs $b_1, b_2, \ldots, b_t \in \{+,-\}$.

Since we always have $H(\Lambda^+)+H(\Lambda^-) = 2H(\Lambda)$ by chain rule, conditioning on $b_1, b_2, \ldots, b_{t-1}$ (which determines X_{t-1}), we always have $\frac{1}{2}H(\Lambda^{b_1,b_2,\ldots,b_{t-1},b_t}) + \frac{1}{2}H(\Lambda^{b_1,b_2,\ldots,b_{t-1},\neg b_t}) = H(\Lambda^{b_1,b_2,\ldots b_{t-1}}) = X_{t-1}$.

This leads to the following observation.

Observation 14.5.

$$\mathbf{E}[X_t \mid X_{t-1} = a] = a$$

So X_t forms a martingale (bounded in $[0,1]$).

Intuitively, consider the first channel split from $\Lambda = (A; B)$, which is $(A_1 + A_2 \mid B_1, B_2)$. If we ignore the conditioning for now since XORing two copies always increase entropy unless (1) A already has maximum entropy and (2) A has zero entropy. So the only fix point is $h(\Lambda) = 0$ or $h(\Lambda) = 1$. This intuitively tells us the limiting distribution of X_t is Bernoulli. If $t \to \infty$, this proves the limiting analogue of Theorem 14.2.

But we need to argue about the finite case and care about the speed of convergence. We now rephrase our goal, Theorem 14.2 into martingale language.

Theorem 14.6 (Polynomially strong polarization). *$\forall \gamma$, $\exists \alpha < 1, \beta$ finite s.t.*

$$\forall t, \mathbf{Pr}[X_t \in (\gamma^t, 1-\gamma^t)] \le \beta \cdot \alpha^t.$$

Think of $n = 2^t$, $\gamma = 1/8$ (so that $\gamma^t \approx \frac{1}{n^3}$). Also think of $\alpha = 0.9$, so that we get $\epsilon = 0.9^t$ which is exponentially small in t.

Proof structure We first establish moderate polarization for first $t/2$ levels.

Claim 14.7 (Moderate Polarization).

$$\mathbf{Pr}[X_{t/2} \in (\alpha_1^t, 1-\alpha_1^t)] \le \alpha_1^t$$

for some $\alpha_1 < 1$ (say $\alpha_1 = 0.9$).

Note this is not enough for us, we at least need $\alpha_1 < 1/2$ to get $\alpha_1^t < \frac{1}{n}$ close to 0/1 boundary. But moderate polarization implies that at depth $t/2$, at least $1 - \alpha_1^{t/2}$ fraction of channels polarize relatively well.

Then we simply give up on the rest. This is fine because we are allowed to give up on an exponential fraction of channels. We continue to grow the subtree (split the channels) from the remaining channels.

Let us look at the case when $A \sim \text{Bernoulli}(p)$ and $B = \emptyset$. Since we already gave up on those channels that are not moderately polarized, we can think of p as 10^{-4}. It is not small enough as $\frac{1}{n}$, but it is already a pretty good start.

In this case, we split $\Lambda = (A, B)$ into $(A_1 + A_2; \emptyset) \sim (\text{Bernoulli}(2p(1-p)), \emptyset)$ with entropy $h(2p(1-p))$ and $(A_2; A_1 + A_2)$ which has entropy $2h(p) - h(2p(1-p))$ (because $H(\Lambda^+) + H(\Lambda^-) = 2H(\Lambda)$). The idea is that $2h(p) - h(2p(1-p)) \approx 2p \log 1/p - 2p(1-p) \log 1/p \approx h(p)^2$. In one branch, your entropy increases by a factor of 2, while in the other branch, your entropy goes down by $p = 10^{-4}$. Hence you will have extremely strong polarization.

Before we present the formal proof, we will first abstract a few key properties we used in the above argument. So that we can now forget about channels. Any martingale that satisfies these properties will have polynomially strong polarization (Theorem 14.6).

Definition 14.8. (Local polarization) A sequence of r.v. $X_0, X_1, \ldots, X_t \in [0,1]$ is locally polarizing if

- $\mathbf{E}[X_t \mid X_{t-1} = a] = a$ for all $a \in [0,1]$ and $t \geq 1$.
- (Variance in the middle) $\forall \tau$, $\exists \theta = \theta(t)$, if $X_t \in [\tau, 1-\tau]$, $|X_{t+1} - X_t| \geq \theta(\tau)$.
- (Suction at the ends) $\forall \Delta < \infty$, $\exists \tau = \tau(\Delta)$ s.t. $\mathbf{Pr}[X_{t+1} \leq \frac{X_t}{\Delta} \mid X_t \leq \tau] \geq \frac{1}{2}$. Likewise, $\mathbf{Pr}[1 - X_{t+1} \leq \frac{1-X_t}{\Delta} \mid X_t \geq 1 - \tau] \geq \frac{1}{2}$.

Here the second condition says that if X_t is in the middle, then in the next step, it gets a little variance. The third condition corresponds to the property that at one branch, your entropy X_{t+1} goes down by $1/\Delta = 10^{-4}$ given X_t is already moderately polarized.

We will need to prove the following two claims.

- The X_t in Definition 14.4 satisfy local polarization (Definition 14.8).
- Local polarization (Definition 14.8) implies polynomially strong polarization.

References

[Ari09] Erdal Arikan. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Trans. Inf. Theor.*, 55(7):30513073, jul 2009. 2