CS 229r Information Theory in Computer Science Feb 26, 2019

Lecture 9

Instructor: Madhu Sudan *Scribes: Vinh-Kha Le and Amal Mattoo*

1 Overview

Here are some reminders for today.

1. Homework 2 is due at 8 PM.
2. Madhu has office hours after lecture.
3. Mitali has office hours at 4:30 PM.

This lecture wraps up our discussion of polar codes. We will cover the following topics.

1. Decoding
2. Towards the Analysis of Polarization
   (a) Martingales
   (b) Local and Global Polarization

2 Introduction

Recall from the previous lecture the polar encoding function E_n as defined on strings of length $n = 2^t$. We first recursively define a pre-encoding function $\tilde{E}_n$. Given strings U and V each of length $n/2$, we take

$$\tilde{E}_n(UV) = \tilde{E}_{n/2}(U \oplus V)\tilde{E}_{n/2}(V)$$

and $\tilde{E}_1 = id_{[2]}$. We visualize the pre-encoding step as a recursion tree of depth t, where at each step, we replace U with $U \oplus V$ and V with $V|U \oplus V$. Note that

$$H(U) + H(V) = H(U, V) = H(U \oplus V, V) = H(U \oplus V) + H(V|U \oplus V).$$

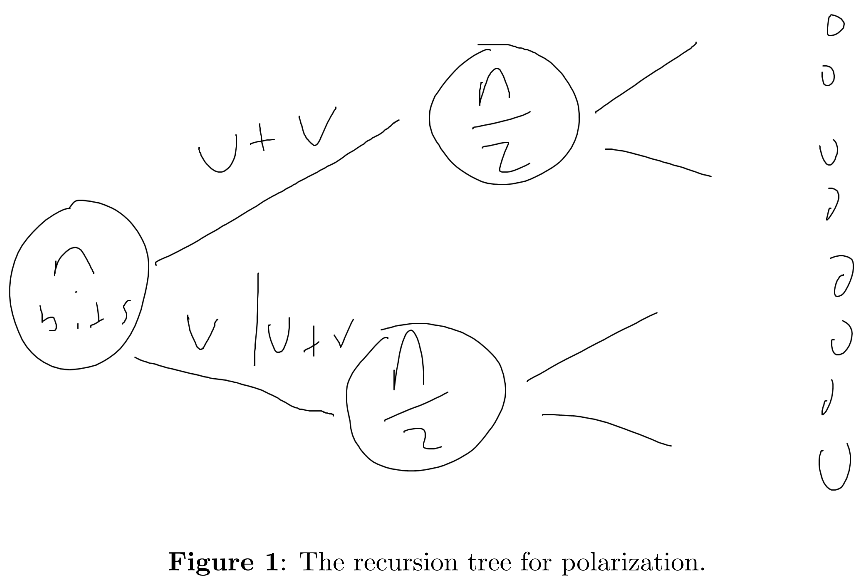

Figure 1: The recursion tree for polarization.

Given a string $Z \sim \text{Bernoulli}(p)^n$ of i.i.d. bits, let $W = \tilde{E}_n(Z)$. Because at each step of the recursion $U \oplus V$ has a lower index than $V|U \oplus V$, it is consistent to consider each bit W_j as conditioned upon all previous bits $W_{<j}$. Under this conditioning, each bit in $\tilde{E}_n(Z)$ has either small or large entropy with high probability. More precisely, for every $p \in [0,1]$ and $c > 0$, there exists a $\beta < 1$ such that for all $t \in \mathbb{N}$,

$$\mathbb{P}_{j \in [2^t]}[H(W_j|W_{<j}) \in (c^{-t}, 1 - c^{-t})] \leq O(\beta^t).$$

We call this condition (strong) polarization.

Let $S = \{j \in [2^t] \mid H(W_j|W_{<j}) \geq c^{-t}\}$ be the set of high-entropy bits. The encoding function removes the low-entropy bits from W and outputs $E_n(Z) = W_S$. We exhibit W_S as a member of $\{0,1,?\}^n$. Nonetheless, the encoder can transmit the bits in S without having to convey the set S because it only depends on the source Bernoulli(p), of which the encoder and decoder both have knowledge. Metaphorically, S resides in the hardware of the coding scheme.

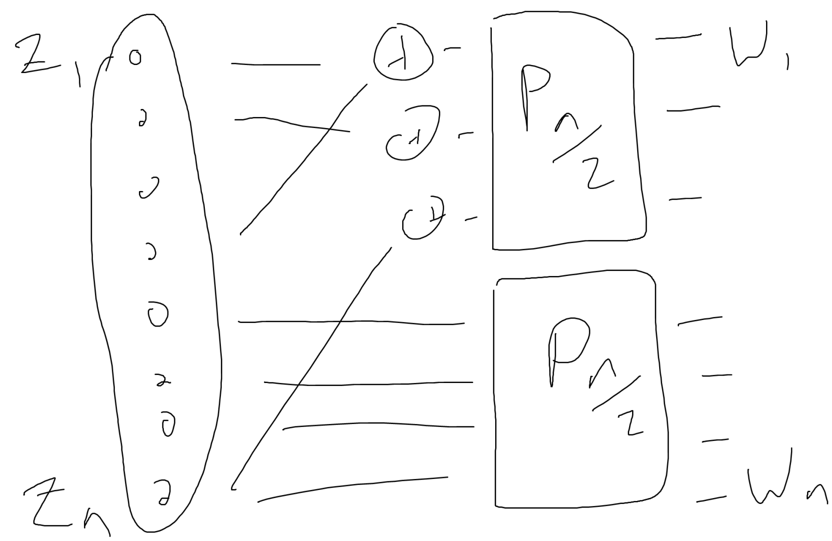

Figure 2: An alternative representation of the tree.

Exercise 1. *In terms of the length n of the input string, what is the complexity of encoding? Draw a circuit that computes E_n for the case where $n = 8$ and $p = 1/4$.*

Solution. There are n bit-wise operations at every recursion depth, and the maximum recursion depth is $t = \log n$. This means that the complexity of encoding is $O(n \log n)$. The circuit diagrams that compute E_n can be found in [Arikan '08] as Figures 1, 2, 3, 5, and 8. □

In Section 3, we introduce the polar decoding function and show that this function satisfies the relevant desiderata for a coding scheme. In Section 4, we begin a discussion on martingales and sketch a proof of polarization.

3 The Polar Decoding Function

Let D_n denote the polar decoding function as defined on strings of length $n = 2^t$. Given $W_S \in \{0,1,?\}^n$, we want $D_n(W_S)$ to be the $\hat{Z} \in [2]^n$ that maximizes the probability that $\tilde{E}_n(Z)_S = W_S$ given $Z = \hat{Z}$. We wish to define D_n recursively, but we face a minor obstruction.

Let $Z = UV$ be the concatenation of strings U and V at the first recursive step. The naïve recursion reconstructs $\hat{Z}$ based on the presumption that Z is a string of i.i.d. Bernoulli bits. Let $A = U \oplus V$ and

$B = V$. The goal is to take $\hat{Z} = (\hat{A} \oplus \hat{B}, \hat{B})$ upon reconstructing $\hat{A} = (U \oplus V)^\wedge$ and $\hat{B} = (V|U \oplus V)^\wedge$. Suppose we successfully decode $\hat{A}$ as to maximize the probability that $\tilde{E}_{n/2}(A)_S$ agrees with the first half of W_S. Indeed, A is a string of i.i.d. Bernoulli bits, and so far there should be no issue.

We now wish to decode $\hat{B}$ as to maximize the probability that $\tilde{E}_{n/2}(B)_S$ agrees with the last half of W_S given $U \oplus V = \hat{A}$. However, we cannot apply the recursive step here because B is not generally a string of i.i.d. bits. The distribution for B_i is distributed differently from B_j when $\hat{A}_i$ is not the same as $\hat{A}_j$.

Exercise 2. *Let U_i, U_j, V_i, and V_j be independent bits with probabilities p_i, p_j, $p_{n/2+i}$, and $p_{n/2+j}$. Let $\hat{A}_i = 0$ and $\hat{A}_j = 1$. Compute the distributions for B_i and B_j given $U \oplus V = \hat{A}$. Observe that B_i does not generally have the same distribution as B_j when $p_i = p_j = p_{n/2+i} = p_{n/2+j} = p$.*

Solution. Let $q_\bullet = 1 - p_\bullet$. Then B_i is a Bernoulli bit that is 1 with probability

$$\mathbb{P}[B_i = 1] = \mathbb{P}[V_i = 1 | U_i \oplus V_i = 0] = \frac{\mathbb{P}[U_i = 1, V_i = 1]}{\mathbb{P}[U_i = 1, V_i = 1] + \mathbb{P}[U_i = 0, V_i = 0]} = \frac{p_i p_{n/2+i}}{p_i p_{n/2+i} + q_i q_{n/2+i}}.$$

Similarly, B_j is a Bernoulli bit that is 1 with probability

$$\mathbb{P}[B_j = 1] = \mathbb{P}[V_j = 1 | U_j \oplus V_j = 1] = \frac{\mathbb{P}[U_j = 0, V_j = 1]}{\mathbb{P}[U_j = 1, V_j = 0] + \mathbb{P}[U_j = 0, V_j = 1]} = \frac{q_j p_{n/2+j}}{p_j q_{n/2+j} + q_j p_{n/2+j}}.$$

When $p_i = p_j = p_{n/2+i} = p_{n/2+j} = p$, these two expressions simplify to

$$\frac{p^2}{p^2 + q^2} \text{ and } \frac{1}{2}.$$

They are equal only when $p = 1/2$. In particular, for the case that $p = 0.01$, the probability that we get $B_i = 1$ is approximately 0.001. In this case, B_i has low entropy and B_j has high entropy. □

Because the inductive step from Z to A and B does not produce identical distributions, we must strengthen the inductive hypothesis to allow Z to be a string of independent (not necessarily identically distributed) Bernoulli bits. This is equivalent to providing D_n with additional data $(p_i)_{i=1}^n$ such that $Z_i \sim \text{Bernoulli}(p_i)$. The true algorithm proceeds as expected.

1. Compute $a_i = \mathbb{P}[U_i \oplus V_i = 1]$ and take $\hat{A} = D_{n/2}(W_1, W_2, \ldots, W_{n/2}; a_1, a_2, \ldots, a_{n/2})$.
2. Compute $b_i = \mathbb{P}[V_i = 1 | U_i \oplus V_i = \hat{A}_i]$ and take $\hat{B} = D_{n/2}(W_{n/2+1}, W_{n/2+2}, \ldots, W_n; b_1, b_2, \ldots, b_{n/2})$.
3. Output $D_n(W_1, W_1, \ldots, W_n; p_1, p_2, \ldots, p_n) = (\hat{A} \oplus \hat{B}, \hat{B})$.

Exercise 3. *The above description of D_n is incomplete. Express a_i and b_i in terms of p_i. Then use the desiderata for D_n to define the base case $D_1(W_1; p_1)$.*

Solution. We compute that

$$a_i = \mathbb{P}[U_i \oplus V_i = 1] = \mathbb{P}[U_i = 1, V_i = 0] + \mathbb{P}[U_i = 0, V_i = 1] = p_i q_{n/2+i} + q_i p_{n/2+i}.$$

We computed b_i in Exercise 2.

The base case algorithm $D_1(W_1; p_1)$ should decode the single bit $W_1 \in \{0, 1, ?\}$ given a prior p_1. If W_1 is known, output W_1. If W_1 is unknown, output 1 if $p_1 \geq 1/2$ and 0 otherwise. □

If we guess W correctly from W_S at the maximum recursion depth, the algorithm recovers Z exactly because $(U, V) \mapsto (U \oplus V, V)$ is idempotent and thus invertible. As a result, the decoding error is union bounded by the sum of the errors at each bit at the maximum recursion depth. With the correct definition of D_1, the error at bit j is bounded from above by the entropy $H(W_j | W_{<j})$ of that bit. By our choice of bits S, this entropy is bounded from above by c^{-t}. This gives us a total decoding error of nc^{-t}.

Exercise 4. *In terms of the length n of the input string, what is the complexity of successful decoding? Draw a ternary circuit that computes D_n for the case where $n = 8$ and $p = 1/4$.*

Solution. There are $O(n)$ bit-wise operations at every recursion depth, and the maximum recursion depth is $t = \log n$. This means that the complexity of encoding is $O(n \log n)$. The circuit diagrams that compute D_n can be found in [Arikan '08] as Figures 9 and 10. □

In this lecture, we specialize to the binary alphabet $\Sigma = [2]$. One would imagine that an analog of polar coding could be defined for arbitrary alphabets $\Sigma = [q]$ endowed with an additive group structure. However, for these larger alphabet sizes, it turns out that polar coding depends on field structure. Group structure does not suffice. In general, polar coding makes sense for finite fields $\Sigma = \mathbb{F}_q$. In other words, q must be a prime power. For this more general construction, we refer the reader to [Mori '14].

4 Polarization and Martingales

We take a random walk on the recursion tree. Let Z_j^i be node j at recursion depth i. We begin our walk at $Z_0^0 = Z$. If $Z_j^i = U_j^i V_j^i$ is the concatenation of strings as before, $Z_{2j}^{i+1} = U_j^i \oplus V_j^i$ and $Z_{2j+1}^{i+1} = V_j^i$. Just as we condition V on $U \oplus V$, we condition V_j^i on $U_j^i \oplus V_j^i$. Because Z_j^i appears after all the elements on which it is conditioned, we consider Z_j^i as conditioned upon $Z_{<j}^i$.

We define our walk as a sequence $(j_i)_{i=0}^t$ such that $j_i = 2j_{i-1} + b_i$ with uniformly random bits $(b_i)_{i=1}^t$. Then there is a correspondence between random walks on the recursion tree and random bits in W. Taking $Z_{j_t}^t = W_{j_t}$ gives us a uniform distribution on the bits in W, and taking $(b_i)_{i=1}^t$ to be the binary expansion of a uniformly distributed $j \in [n]$ gives us a uniform distribution on walks $(j_i)_{i=0}^t$.

Let $X_i = H(Z_{j_i}^i | Z_{<j_i}^i)$ so that $X_t = H(W_{j_t} | W_{<j_t})$. Polarization is equivalent to the condition that for every $p \in [0, 1]$ and $c > 0$, there exists a $\beta < 1$ such that for all $t \in \mathbb{N}$,

$$\mathbb{P}_{(X_0,\dots,X_t)}[X_t \in (c^{-t}, 1 - c^{-t})] \le O(\beta^t).$$

We first establish some important properties of $(X_i)_{i=0}^t$.

Exercise 5. *Show that $0 \le X_i \le 1$ for $0 \le i \le t$. A martingale is a sequence of random variables $X_i \in L^1(\Omega, \mathbb{P})$ such that*

$$\mathbb{E}[X_i | X_1, X_2, \dots, X_{i-1}] = X_{i-1}.$$

Show that $(X_i)_{i=0}^t$ is a Martingale.

Solution. Because X_i is the entropy of a single bit, it must lie inside the interval $[0, 1]$. The Martingale condition follows from the invertibility of $(U, V) \mapsto (U \oplus V, V)$ and the chain rule for entropy. □

These properties surprisingly are not sufficient to show convergence.

Exercise 6. *Find a bounded Martingale that fails to converge.*

Solution. Let $X_0 = 1/2$ and $X_i = X_{i-1} \pm 2^{-i-1}$ with uniform distribution on the sign. This Martingale is smeared over the whole interval $[0, 1]$ even though as time progresses, the step sizes get smaller. □

To show that the X_i converges, we want to show that it has variance in the middle and suction at the ends. Variance in the middle, i.e., away from 0 and 1, means that for all $\tau > 0$, there exists a $\sigma > 0$ such that for all $i \in [t]$,

$$X_{i-1} \in (\tau, 1 - \tau) \text{ implies that } \operatorname{Var}[X_i | X_{i-1}] \ge \sigma^2.$$

Suction at ends means that there exists a $\Theta > 0$ such that for all $c > 0$, there exists a $\tau > 0$ such that

$$X_{i-1} \le \tau \text{ implies that } \mathbb{P}\left[X_i < \frac{X_{i-1}}{c} \middle| X_{i-1}\right] \ge \Theta \text{ and}$$

$$X_{i-1} \ge 1 - \tau \text{ implies that } \mathbb{P}\left[1 - X_i < \frac{1 - X_{i-1}}{c} \middle| X_{i-1}\right] \ge \Theta.$$

Variance in the middle and suction at the ends are together called local polarization. We claim that X_i polarizes locally. We will prove this in the next lecture. It suffices to show that local polarization implies strong polarization. Before doing so, we first recall an important theorem about Martingales.

Doob's Martingale inequality. Let $(X_t)_0^T$ be a non-negative Martingale and $k > 0$. Given

$$X_{\sup} = \sup_{0 \le t \le T} X_t,$$

we have

$$\mathbb{P}(X_{\sup} \ge kX_0) \le \frac{1}{k}.$$

Exercise 7. *Prove Doob's Martingale inequality.*

Proof. Let τ be the earliest time for which $X_\tau \ge kX_0$. If no such time exists, let $\tau = T$. Conditioning on τ makes it clear that $\mathbb{E}X_\tau = \mathbb{E}X_T = X_0$. It follows that

$$X_0 = \mathbb{E}X_\tau = \mathbb{E}X_\tau \mathbf{1}_{X_{\sup} \ge kX_0} + \mathbb{E}X_\tau \mathbf{1}_{X_{\sup} < kX_0} \ge kX_0 \mathbb{P}(X_{\sup} \ge kX_0).$$

A bit of algebraic manipulation completes the proof. □

Exercise 8. *Here we sketch a proof that local polarization implies strong polarization. Suppose that X_t is locally polarized. Let $\phi_t = \min\{\sqrt{X_t}, \sqrt{1 - X_t}\}$ be a potential indicating the amount by which X_t is already polarized. Show that $\mathbb{E}[\phi_{t+1}|\phi_t] \le \alpha\phi_t$ for some $\alpha < 1$. Deduce via induction and Markov's inequality that $\mathbb{P}[\phi_t > \alpha^{t/2}] \le \alpha^{t/2}$ for fixed $t = t_0$. Conclude that X_t is strongly polarized by applying Doob's Martingale inequality to X_t for $t \in [t_0, 2t_0]$.*

Proof. It is clear from local polarization that α exists for fixed $t = t_0$. Induction and Markov's inequality give us

$$\mathbb{E}[\phi_t] \le \alpha^t \text{ and } \mathbb{P}[\phi_t > \alpha^{1/2}] \le \alpha^{1/2}.$$

We thus get

$$\mathbb{P}[X_t \in (\alpha^t, 1 - \alpha^t)] \le \alpha^{t/2}.$$

Recall the definition of X_t. This is precisely what we want. We wish to extend this result over a region of time steps $[t_0, 2t_0]$. Because of local polarization, we have $\mathbb{P}[X_{t+1} < X_t/100] \ge 1/2$ and $\mathbb{P}[X_t < X_{t+1} < 2X_t] \le 1/2$ for $t_0 \le t \le 2t_0$. Applying Doob's Martingale inequality completes the proof. □

Next time we will show that X_i polarizes locally, and that will be the end of polarization and channel coding.