Polar Codes: **Speed of Polarization & Polynomial Gap to Capacity**

Venkatesan Guruswami

Carnegie Mellon University

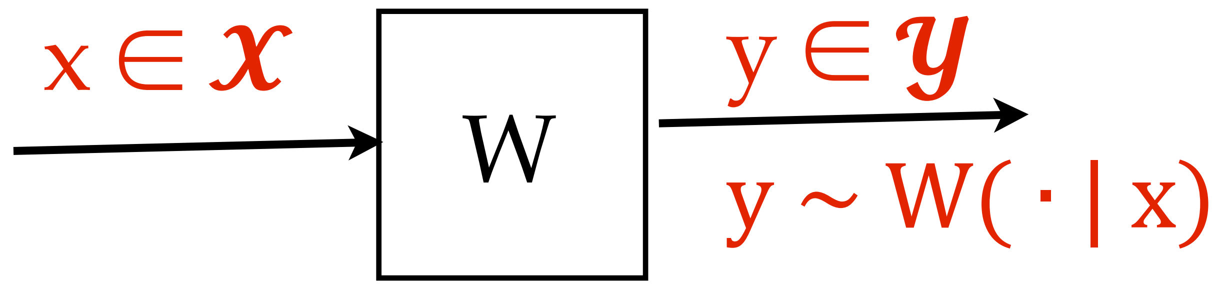
(currently visiting Microsoft Research New England)

Based on joint work with **Patrick Xia**

Charles River Science of Information Day
MIT, April 28, 2014

Discrete Memoryless Channel

Discrete Channel
Input alphabet $\mathcal{X}$
Finite output alphabet $\mathcal{Y}$



	a	b	c	d
0	0.1	0.4	0.2	0.3
1	0.4	0.1	0.3	0.2

Memoryless channels:
Channel's behavior on i'th bit independent of rest

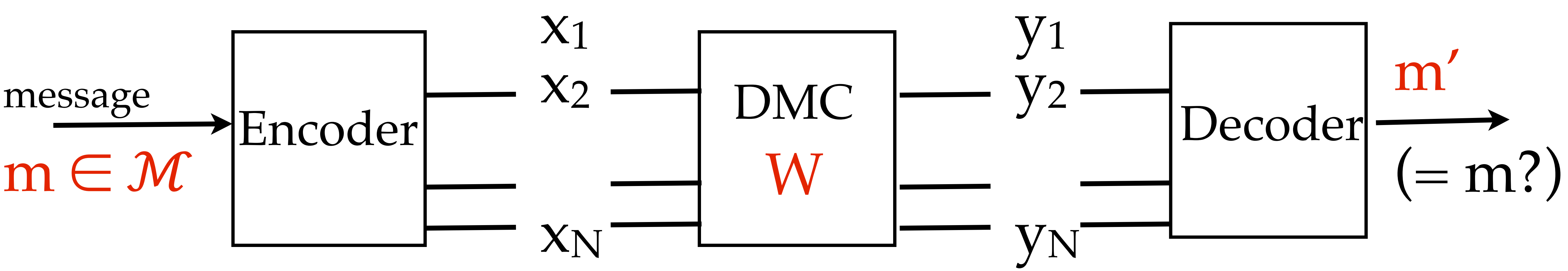
$$W^n(y_1 y_2 \ldots y_n | x_1 x_2 \ldots x_n) = \prod_{i=1}^{n} W(y_i | x_i)$$

Noisy Coding theorem

[Shannon'48] Every discrete memoryless channel W has a **capacity** I(W) such that one can communicate at *asymptotic **rate*** I(W) - ε with vanishing probability of miscommunication (for any desired *gap to capacity* ε > 0)

Conversely, reliable communication is ***not*** possible at rate I(W)+ε.

Asymptotic rate: Communicate (I(W)-ε)N bits in N uses of the channel in limit of large *block length* N



Rate = $(\log |\mathcal{M}|)/N$

Shannon's Theorem

Shows that (if channel isn't completely noisy) ***constant factor*** overhead suffices for ***negligible*** decoding error probability, provided we tolerate some ***delay***

- Delay / block length $N \approx 1/\varepsilon^2$ suffices for rate within ε of capacity
- Miscommunication prob. $\approx \exp(-\varepsilon^2 N)$

Binary Memoryless Symmetric (BMS) channel

- $\mathcal{X} = \{0,1\}$ (binary inputs)
- *Symmetric*
  - Output symbols can be paired up $\{y,y'\}$ such that $W(y \mid b) = W(y' \mid 1-b)$

Most important example:

BSC_p (binary symmetric channel with crossover probability p)

	0	1
0	1-p	p
1	p	1-p

Capacity of BMS channels

Denote $H(W) := H(X \mid Y)$

where $X \sim U_{\{0,1\}}$; $Y \sim W(\,\cdot \mid X)$

$\boxed{\textit{Shannon capacity } I(W) = 1 - H(W)}$

Two well-known examples

BSC_p

	0	1
0	1-p	p
1	p	1-p

Capacity = 1 - h(p)

BEC_α

	0	1	?
0	$1-\alpha$	0	α
1	0	$1-\alpha$	α

Capacity = $1 - \alpha$

Realizing Shannon

- Shannon's theorem *non-constructive*
  - random codes, exponential time decoding

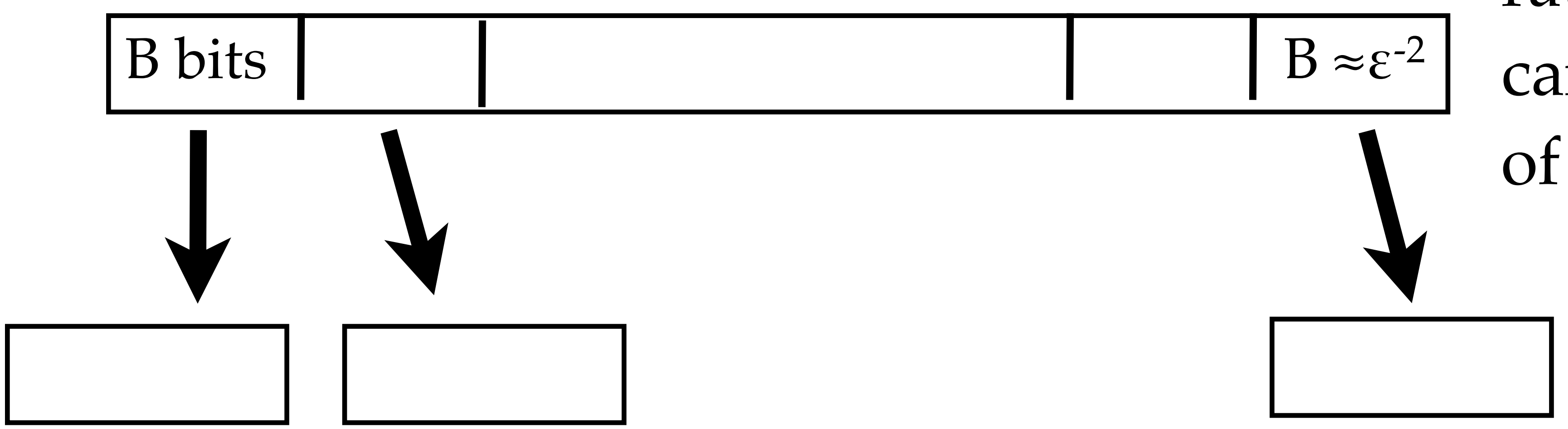
★ Challenge: *Explicit* coding schemes with *efficient* encoding/decoding algorithms to communicate at information rates ≈ capacity

- Has occupied coding & information theorists for 60+ years

“Achieving” capacity

In the asymptotic limit of large block lengths N, not hard to approach capacity within any fixed $\varepsilon > 0$

- ✦ Code concatenation (Forney’66)



rate $\approx 1 - \varepsilon$ outer code, can correct a positive frac. of *worst-case* errors

Ensemble of inner codes of rate $\approx$ capacity - ε

Decoding time $\approx$ N $\exp(1/\varepsilon^2)$

(brute force max. likelihood decoding of inner blocks)

Complexity scales poorly with gap ε to capacity

Achieving capacity: A precise theoretical formalism

Given channel W and desired gap to capacity ε,

Construct $\mathrm{Enc} : \{0,1\}^{RN} \to \{0,1\}^{N}$ & $\mathrm{Dec} : \{0,1\}^{N} \to \{0,1\}^{RN}$

for rate $R = I(W) - \varepsilon$ such that

- $\forall$ msg. m, $\Pr[\,\mathrm{Dec}(W(\mathrm{Enc}(m))) \neq m\,] \ll \varepsilon$ (say ε^{100})
- Block length $N \leq \mathrm{poly}(1/\varepsilon)$
- Runtime of Enc and Dec bounded by $\mathrm{poly}(1/\varepsilon)$

That is, seek complexity ***polynomially bounded*** in single parameter, ***gap ε to capacity***

Our Main Result

Polar codes [Arikan, 2008] give a solution to this challenge

Deterministic polytime constructible binary linear codes for approaching capacity of BMS channels W within ε with complexity $O(N \log N)$ for $N \geq (1/\varepsilon)^c$

- c = absolute constant independent of W
- Decoding error probability $\exp(-N^{0.49})$

✦ The ***first*** *(and so far only)* construction to achieve capacity with such a theoretically proven guarantee.

✦ Provides a complexity-theoretic basis for the statement ``*polar codes are the first constructive capacity achieving codes*"

Other "capacity achievers"

- Forney's concatenated codes (1966)
  - Decoding complexity $\exp(1/\varepsilon)$ due to brute-force inner decoder
- LDPC codes + variants (Gallager 1963, revived ~ 1995 onwards)
  - Proven to approach capacity *arbitrarily closely* **only** for erasures
  - *Ensemble* to draw from, rather than explicit codes
- Turbo codes (1993)
  - Excellent empirical performance. *Not known* to approach capacity arbitrarily closely as block length $N \to \infty$
- Spatially coupled LDPC codes (Kudekar-Richardson-Urbanke, 2012)
  - Asymptotically achieves capacity of all BMS channels!
  - Polynomial convergence to limit not yet known

Weren't polar codes already shown to achieve capacity?

- Yes, in the limit of large block length
  - Can approach rate I(W) as $N \to \infty$ [Arikan]
- We need to bound the *speed* of convergence to capacity
  - Block length $N = N(\varepsilon)$ needed for rate $I(W) - \varepsilon$?
- We show **$N(\varepsilon) \leq \mathrm{poly}(1/\varepsilon)$**
  - Mentioned as an open problem, eg. in [Korada'09; Kudekar-Richardson-Urbanke'12; Shpilka'12; Tal-Vardy'13]
  - Independently shown in [Hassani-Alishahi-Urbanke'13]

Finite length analysis

- Asymptotic nature of previous analyses due to use of convergence theorem for supermartingales
- We give an elementary analysis, leading to effective bounds on the speed of convergence

Roadmap

- Polarizing matrices & capacity-achieving codes
- Arikan's recursive polarizing matrix construction
- Analysis: Rough polarization
- Remaining issues, fine polarization

Source coding setting & Polarization

Focus on BSC_p.

Suppose $C \subset \{0,1\}^N$ is a *linear* code of rate $R \approx 1- h(p)$

C is the kernel of a $(1-R)N \times N$ parity check matrix H_N:

$$C = \{ c \in \{0,1\}^N : H_N\, c = 0 \}$$

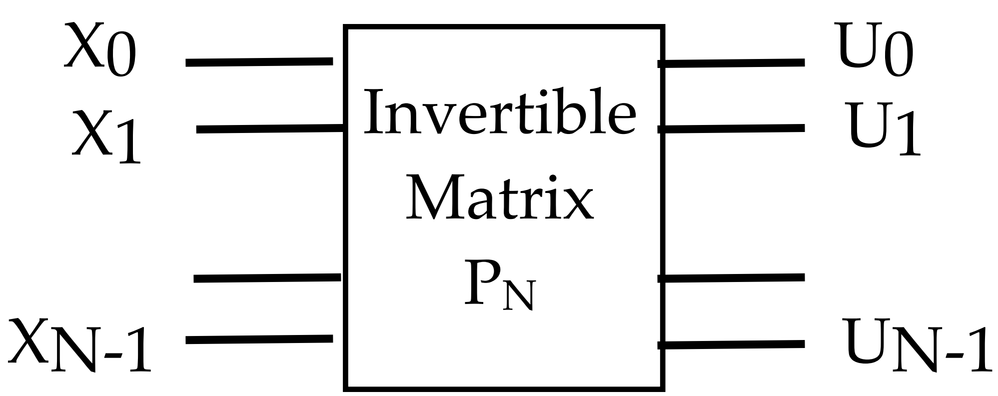
C is a good *channel code* for BSC_p $\Longleftrightarrow$ H_N gives a optimal lossless *source code* for *compressing* Bernoulli(p) source:

- $x_0\, x_1 \ldots\, x_{N-1}$ i.i.d samples from source X = Bernoulli(p)
- They can be recovered w.h.p from $\approx h(p)N$ bits $H_N(x_0\, x_1 \ldots\, x_{N-1})^T$

If we complete the rows of H_N to a basis, resulting $N \times N$ invertible matrix P_N is ``***polarizing***''

$$P_N = \left[\frac{H_N}{A}\right]$$

Coding needs Polarization



Source coding setting

- $X_0\ X_1 \ldots X_{N-1}$ **i.i.d copies of X**
  - (For general channel coding, work with conditional r.v's $X_i \mid Y_i$ + handle some subtleties)

P_N has the following ***polarizing property***:

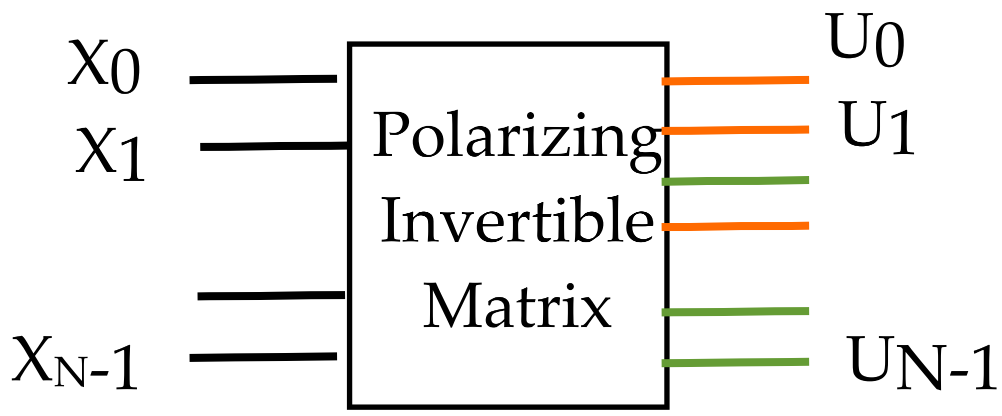
$$\forall \epsilon > 0 \quad \frac{1}{N} \cdot |\{i : H(U_i|U_0^{i-1}) \in (\epsilon, 1-\epsilon)\}| \to 0 \text{ as } N \to \infty.$$

$$\text{Good} = \{i : H(U_i|U_0^{i-1}) \approx 0\} \text{ has size} \approx (1 - H(X))N$$

$$\text{Bad} = \{i : H(U_i|U_0^{i-1}) \approx 1\} \text{ has size} \approx H(X)N$$

Polarizing matrices are implied by linear capacity-achieving codes

Insights in Polar Coding



1. *Sufficiency* of such matrices
   - No need to output U_i for *good* indices i (when $H(U_i \mid U_0 \ldots U_{i-1}) \approx 0$)

2. *Recursive* construction of polarizing matrices, along with *low-complexity decoder*

2 x 2 polarization

$$U_0 = X_0 + X_1$$
$$U_1 = X_1$$

$$\begin{pmatrix} U_0 \\ U_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_0 \\ X_1 \end{pmatrix}$$

Suppose X ~ Bernoulli(p)

$$H(U_0) = h(2p(1-p)) \quad > h(p)$$
$$H(U_1 \mid U_0) = 2h(p) - H(U_0) \quad < h(p)$$

(unless h(p)=0 or 1)

If X is not fully deterministic or random, the output entropies are separated from each other

An explicit polarizing matrix [Arikan]

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^{\otimes n}$$

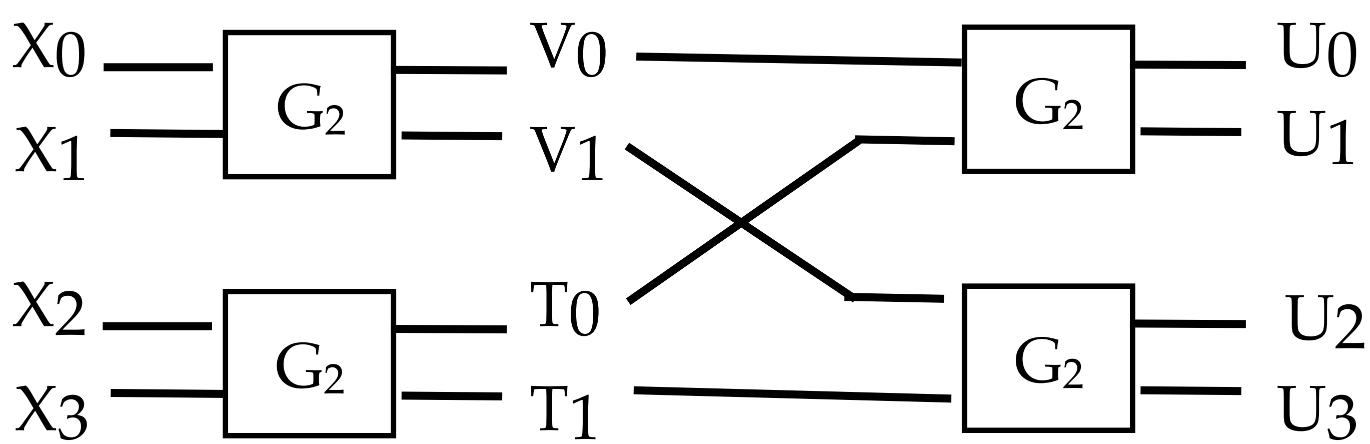
(for N = 2^n)

$$\begin{array}{c c} & \begin{array}{cccc} \mathbf{x_1 x_2} & \mathbf{x_2} & \mathbf{x_1} & \mathbf{1} \end{array} \\ \begin{array}{c} (1,1) \\ (0,1) \\ (1,0) \\ (0,0) \end{array} & \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

n=2

Recursive Polarization

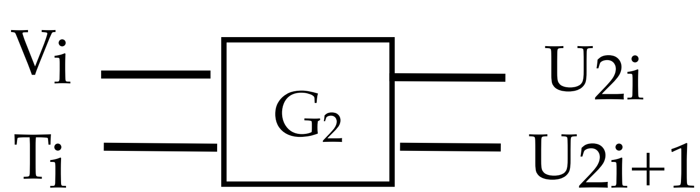
$$G_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$



(V_0, V_1) & (T_0, T_1) i.i.d

<u>*General recursion*</u>

$(V_0, V_1, \ldots, V_{M-1})$ and $(T_0, T_1, \ldots, T_{M-1})$ i.i.d copies of $(U_0, U_1, \ldots, U_{M-1})$



$$H(U_{2i}|U_0^{2i-1}) + H(U_{2i+1}|U_0^{2i})$$
$$= H(V_i|V_0^{i-1}) + H(T_i|T_0^{i-1})$$

$$U_0^{N-1} = B_n G_2^{\otimes n} X_0^{N-1}$$

for $N = 2^n$

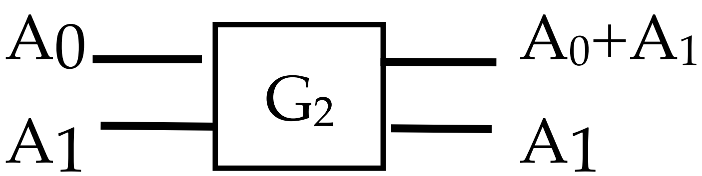
B_n = bit reversal permutation

Proof idea

Channel = pair $\mathcal{W}$ of *(correlated) random variables* (A;B) (with $A \in \{0,1\}$)

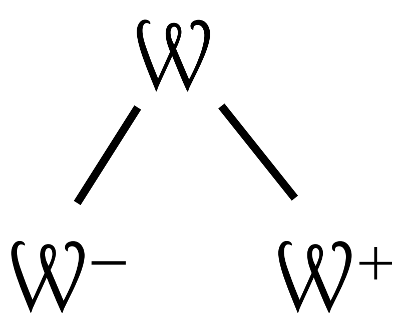
- Channel entropy $H(\mathcal{W}) = H(A \mid B)$

Abstracting each step in recursion:



- Given a channel $\mathcal{W} = (A; B)$
- Take ***two*** i.i.d copies $(A_0; B_0)$ and $(A_1; B_1)$ of $\mathcal{W}$
- Output two pairs $\mathcal{W}^- = (A_0+A_1; B_0,B_1)$ and $\mathcal{W}^+ =(A_1; A_0+A_1,B_0,B_1)$

Channel splitting



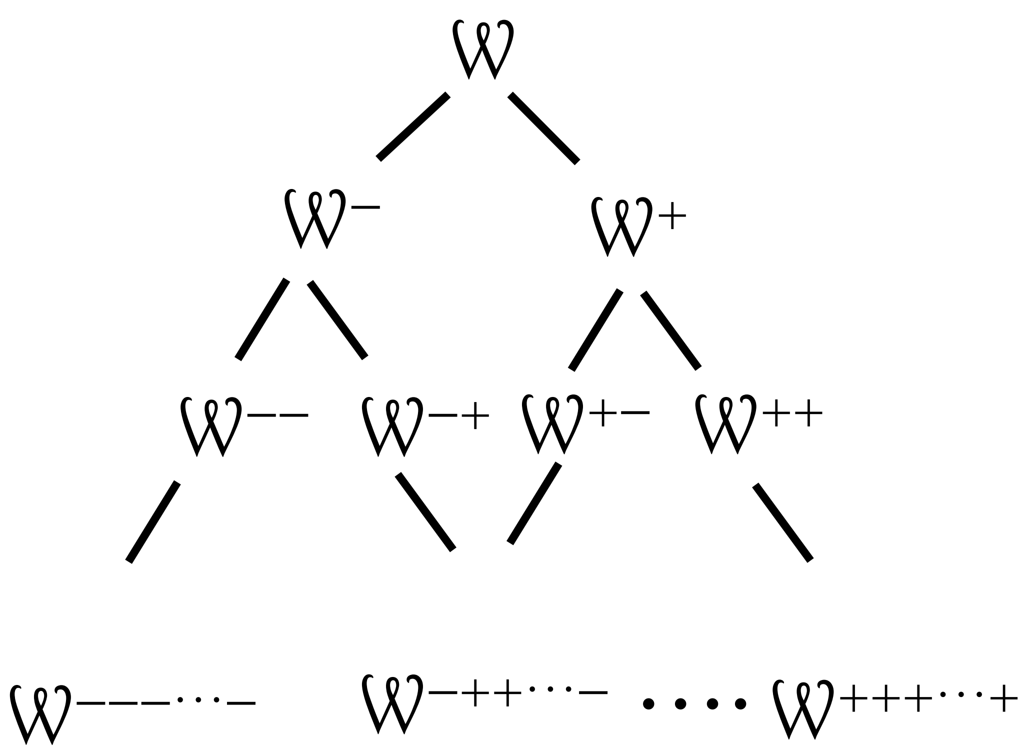
$$H(\mathcal{W}^-) + H(\mathcal{W}^+) = 2\, H(\mathcal{W})$$

$$H(\mathcal{W}^+) \leq H(\mathcal{W}) \leq H(\mathcal{W}^-)$$

Channels produced by recursion

<u>Input</u> = 2^n i.i.d copies of $\mathcal{W}$ (= (X; 0) where X is the source, H($\mathcal{W}$) = H(X))

The channels at various levels of recursion evolve as follows:



Therefore,

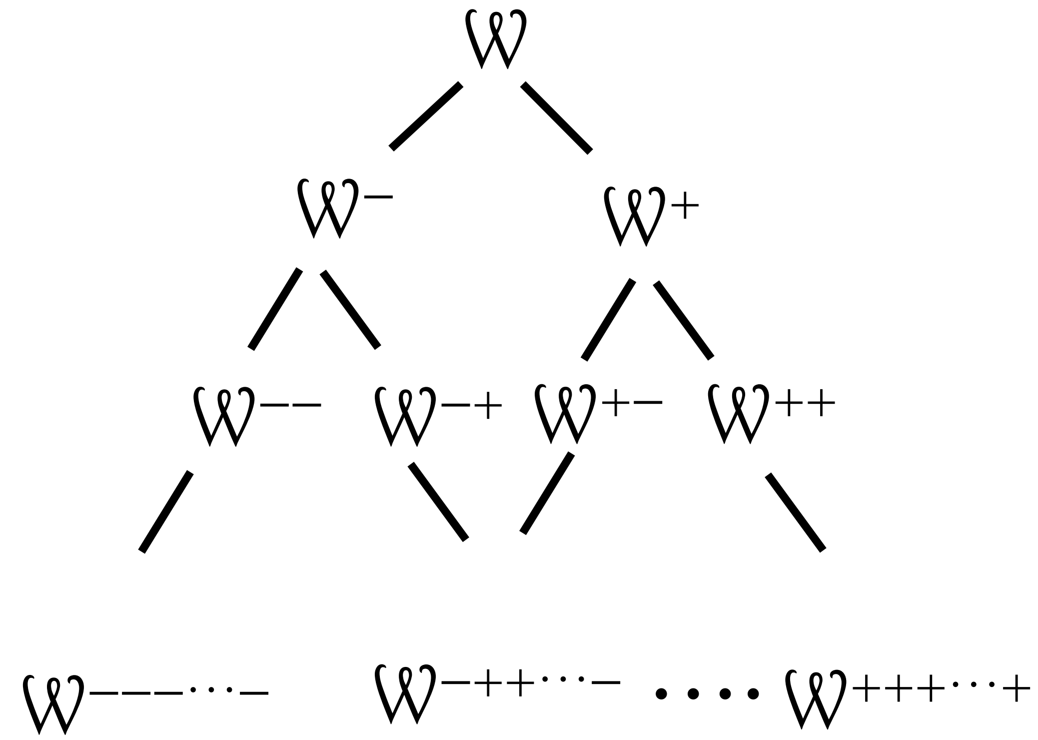
$H(U_i \mid U_0, U_1, \ldots, U_{i-1}) = H(\mathcal{W}^{s_1 s_2 \ldots s_n})$

where $s_j = \begin{cases} - & \text{if } b_j = 0 \\ + & \text{if } b_j = 1 \end{cases}$

and $i = (b_1, b_2, \ldots, b_n)$ in binary

Polarization: Asymptotic Analysis

Consider random walk down the tree, moving left/right randomly at each step



Let H_n be the r.v. equal to entropy of the channel at depth n.

- $H_0, H_1, H_2, \ldots.$ is a *bounded martingale*

$\Longrightarrow$ Converges almost surely to a r.v. H_∞ (martingale convergence theorem)

H_∞ is {0,1}-valued

Only fixed points for entropy evolution $H(W) \rightarrow H(W^{-})$ are 0,1 (deterministic/fully noisy channels)

Entropy increase lemma

[Sasoglu] If $H(\mathbb{W}) \in (\delta, 1-\delta)$ for some $\delta > 0$, then

$$H(\mathbb{W}^{-}) \geq H(\mathbb{W}) + \gamma(\delta) \quad \text{for some } \gamma(\delta) > 0$$

That is,

If (X_1, Y_1), (X_2, Y_2) are i.i.d with $X_i \in \{0,1\}$ & $H(X_i \mid Y_i) \in (\delta, 1-\delta)$, then

$$H(X_1 + X_2 \mid Y_1, Y_2) \geq H(X_1 \mid Y_1) + \gamma(\delta)$$

<u>*Note:*</u> We saw this for $X_i \sim$ Bernoulli(p) (without any Y_i) earlier.

- $h(2p(1-p)) > h(p)$ unless $h(p) \in \{0,1\}$

Polarization: A direct analysis

<u>*Lemma*</u>: There is a $\Lambda < 1$ such that for all "channels" W

$$(\circledast) \quad \mathbb{E}_{s\in\{-,+\}}\left[\sqrt{H(W^s)(1-H(W^s))}\right] \leq \Lambda\sqrt{H(W)(1-H(W))}$$

<u>Corollary</u>: $n = O(\log(1/\varepsilon))$ recursive steps (and thus $N=\mathrm{poly}(1/\varepsilon)$) suffice for $\Pr[H_n(1-H_n) \geq \varepsilon] \leq \varepsilon$ (and $\therefore$ $\Pr[H_n \leq \varepsilon] \geq 1-H(X)-\varepsilon$)

- *rough polarization*

Proof of *Lemma* has two steps:

1. $H(W^-) - H(W) \geq \theta\, H(W)(1-H(W))$ for some $\theta > 0$
   - *quantitative* version of "entropy increase lemma"
2. Use 1. + calculations to deduce $(\circledast)$

Polarization to (source) codes

Invertible polarizing map
$X_0^{N-1} \to U_0^{N-1}$

To compress (encode) X_0^{N-1}

- output U_i, $i \notin$ Good where Good = $\{ i \mid H(U_i \mid U_0^{i-1}) < \delta \}$

To decompress (decode): For i=0,1,…, N-1,

- If $i \notin$ Good, we know U_i from the encoder
- If $i \in$ Good, set U_i to more likely bit (based on estimated prefix U_0^{i-1})
  - (this can be efficiently computed based on the recursive construction)

Prob[decoder is incorrect on U_i, given correct U_0^{i-1}] $\leq H(U_i \mid U_0^{i-1}) < \delta$

$\therefore$ Prob. that decoder doesn't recover U_0^{N-1} (and thus X_0^{N-1}) correctly $\leq \sum_{i \in \text{Good}} H(U_i | U_0^{i-1}) \leq \delta N$

✳ Would like $\delta \ll 1/N$

Getting a code: Issues

Have polarization, but still need:

$H(U_i \mid U_0^{i-1}) \ll 1/N$ for a subset Good of $\approx (1\text{-}H(X)\text{-}\varepsilon)N$ indices i (ε = gap to capacity)

1. for $N \le \mathrm{poly}(1/\varepsilon)$

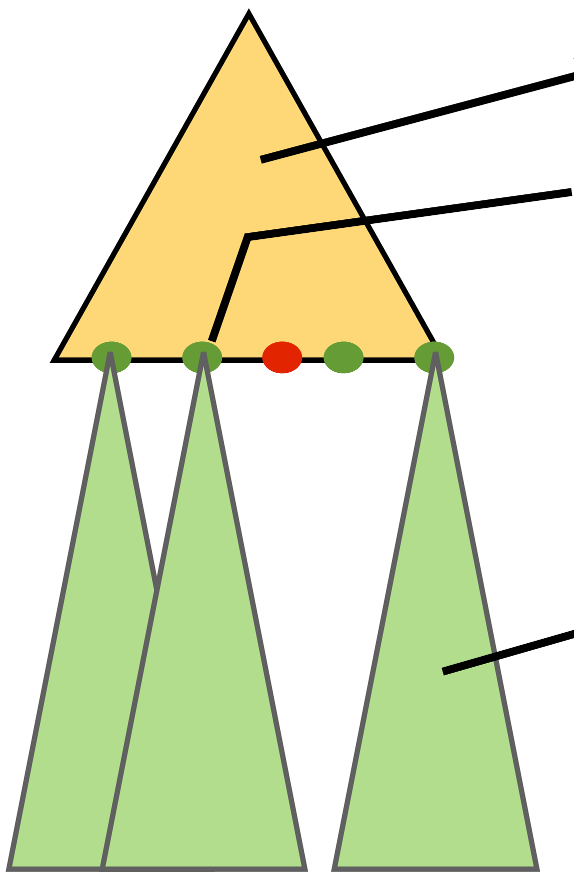
2. with *efficient computation* of the set Good

Amplifying to fine polarization

<u>Recall</u>: we'd like $1\text{-}H(X)\text{-}\varepsilon$ frac. of H_n's to be $\ll 1/N = 2^{-n}$

(to survive union bound)

<u>High level structure of analysis</u>



Rough polarization

$\approx 1\text{-}H(X)\text{-}\varepsilon$ entropies $< \varepsilon$

Give up on rest (red nodes)

Fine polarization

Most paths drive down conditional entropies at leaves to very small values

Rapid polarization of near-zero entropies

To get adequate decrease in H_n, track the *Bhattacharyya parameter* $Z(W)$ of various channels

Quadratically tied to entropy: $Z(W)^2 \leq H(W) \leq Z(W)$

Lemma [Arikan]: $Z(W^+) = Z(W)^2$
$Z(W^-) \leq 2\, Z(W)$

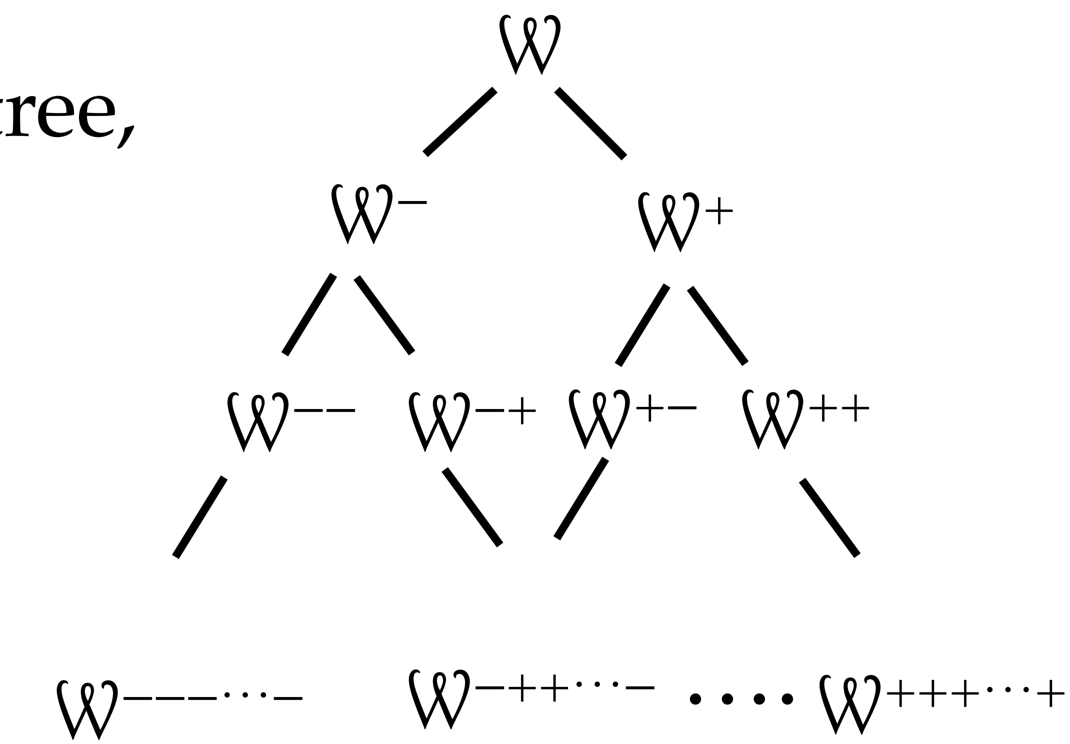
Rapid improvement in the better (+) channel !

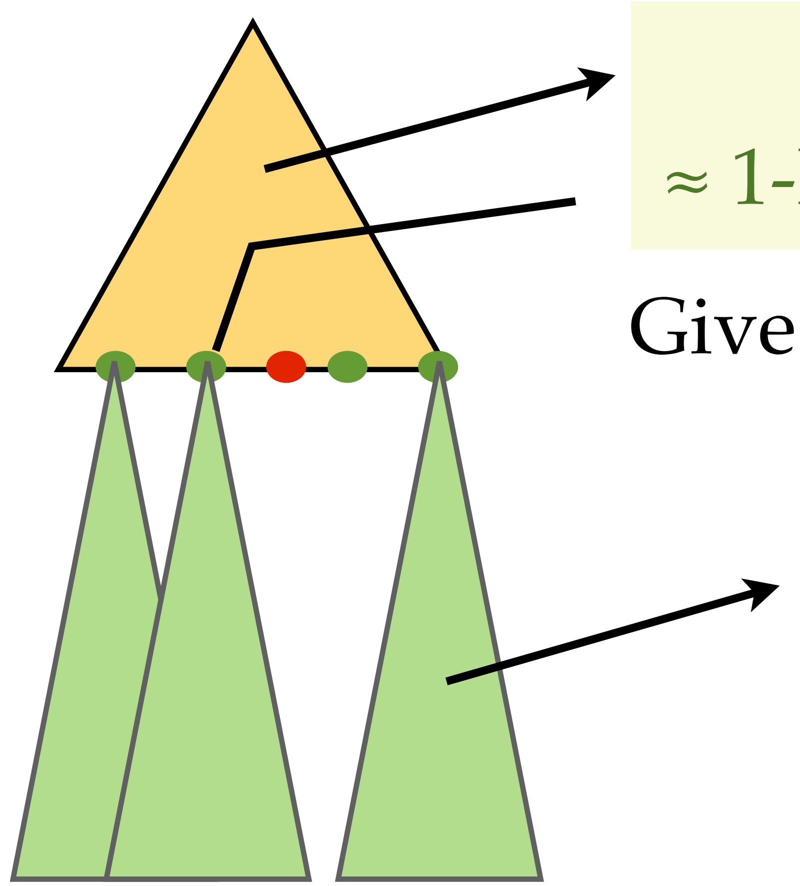
Rapid polarization of near-zero entropies

Fix $\beta < ½$. In an n-long path down the tree, w.h.p Z-parameter squares $> \beta n$ times.
Thus, w.h.p

$$Z_n \lessapprox \exp(-2^{\beta n}) = \exp(-N^{\beta})$$

(after some care to handle the doublings)



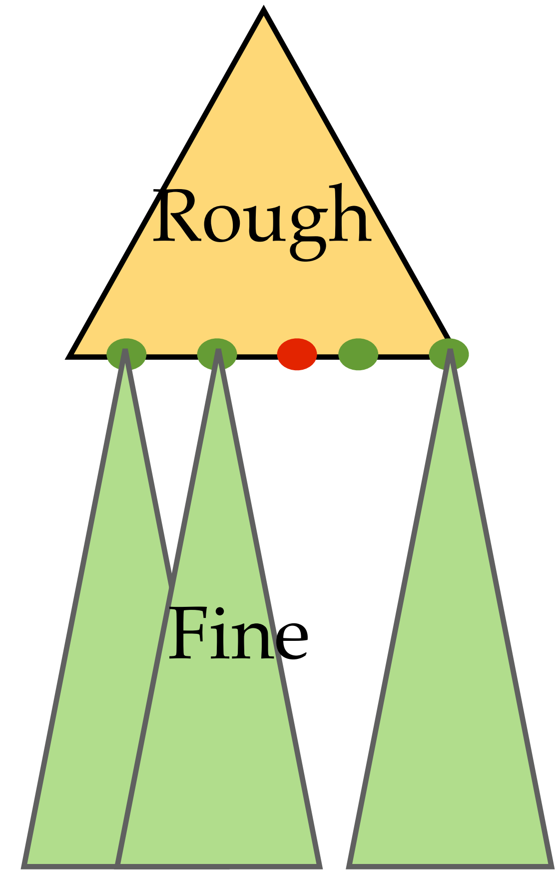


Rough polarization
$\approx 1\text{-}H(X)\text{-}\varepsilon$ entropies $< \varepsilon$

Give up on rest (red nodes)

Fine polarization
Most paths in green subtrees are good,
ending at leaves with
very small conditional entropies

Computing the good channels



Fine polarization (driving down entropy of good channels):

✓ can explicitly pick paths with roughly balanced + and – branches

Rough polarization (first $O(\log(1/\varepsilon))$ steps):

- (Approximately) compute the entropies?

Challenge: Combat increase in output alphabet size

✦ (A;B) ➡ $(A_0 + A_1 ; B_0, B_1)$ squares size of B-space

Idea: Slightly degrade channel by merging output symbols (to reduce output alphabet size after each recursive step)

Channel approximation lemma: $\forall$ channels $W = (A; B)$ and integers $k \geq 1$, $\exists$ channel $\widetilde{W} = (A; \widetilde{B})$ with $\mathsf{supp}(\widetilde{B}) \leq k$ and computable in time $\mathsf{poly}(\mathsf{supp}(B), k)$ such that

$$H(W) \leq H(\widetilde{W}) \leq H(W) + \frac{2 \log k}{k} .$$

- Originally proposed by [Tal-Vardy], variants analyzed in [Pedersani-Hassani-Tal-Teletar]

Concluding remarks

- Exponent μ in $N(\varepsilon) = O(1/\varepsilon^{\mu})$ in our analysis likely much larger than the empirical suggestion $\mu \approx 4$ [Korada-Montanari-Teletar-Urbanke]
  - "Lower bound" of ≈ 3.55 on μ [Goli-Hassani-Urbanke]
- Extend to larger alphabets?
- Connections to binary Reed-Muller codes?