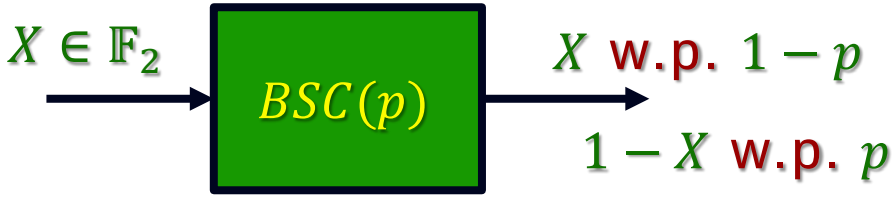
General Strong Polarization

Madhu Sudan
Harvard University

Joint work with Jaroslaw Blasiok (Harvard), Venkatesan Gurswami (CMU), Preetum Nakkiran (Harvard) and Atri Rudra (Buffalo)

Shannon and Channel Capacity

- Shannon ['48]: For every (noisy) channel of communication, ∃ capacity C s.t. communication at rate $R < C$ is possible and $R > C$ is not.
  - Needs lot of formalization – omitted.
- Example:



  - Acts independently on bits
  - Capacity = 1 – $H(p)$; $H(p)$ = binary entropy!
  - $H(p) = p \cdot \log\frac{1}{p} + (1-p) \cdot \log\frac{1}{1-p}$

"Achieving" Channel Capacity

- Commonly used phrase. Many mathematical interpretations.
- Some possibilities:
  - How small can n be?

    Shannon '48 : $n = O(C - R)^{-2}$
  - Get $R > C - \epsilon$ with polytime algorithms?

    Forney '66 :time $= poly\left(n, 2^{\frac{1}{\epsilon^2}}\right)$
  - [Luby et al.'95] Make running time $\text{poly}\left(\frac{1}{\epsilon}\right)$?
    - Open till 2008
    - Arikan'08: Invented "Polar Codes" …
    - Final resolution: Guruswami+Xia'13, Hassani+Alishahi+Urbanke'13 – Strong analysis

Context for today's talk

- [Arikan'08]: General class of codes.
- [GX'13, HAU'13]: One specific code within.
- Why?
  - Bottleneck: Strong Polarization
    - Arikan et al.: General class polarizes
    - [GX,HAU]: Specific code polarizes strongly!
- This work/talk:
  - Introduce Local Polarization
  - Local $\Rightarrow$ Strong Polarization
  - Thm: All known codes that polarize also polarize strongly!

This talk:

- What are Polar Codes
- “Local vs. Global” (Strong) Polarization
- Analysis of Polarization

Lesson 0: Compression ⇒ Coding

- Linear Compression Scheme:
  - (H, D) for compression scheme for $\mathrm{bern}(p)^n$ if
    - Linear map $H: \mathbb{F}_2^n \to \mathbb{F}_2^m$
    - $\Pr_{Z \sim \mathrm{bern}(p)^n}[D(H \cdot Z) \neq Z] \leq \epsilon$
    - Hopefully $\frac{m}{n} \to H(p)$
- Compression ⇒ Coding
  - Let $H^\perp$ be such that $H \cdot H^\perp = 0$;
  - Encoder: $X \mapsto H^\perp \cdot X$
  - Error-Corrector: $Y = H^\perp \cdot X + \eta \mapsto Y - D(H \cdot Y)$

$$=_{w.p.(1-\epsilon)} H^\perp \cdot X$$

Question: How to compress?

- Arikan's key idea:
  - Start with 2×2 "Polarization Transform":
    $$(U,V) \to (U+V,V)$$
    - Invertible – so does nothing?
    - If U,V independent,
      - then $U+V$ "more random" than either
      - $V \mid U+V$ "less random" than either
  - Iterate (ignoring conditioning)
    - End with bits that are completely random, or completed determined (by others).
    - Output "totally random part" to get compression!

Information Theory Basics

- Entropy:
  - Defn: for random variable $X \in \Omega$,

$$H(X) = \sum_{x \in \Omega} p_x \log \frac{1}{p_x} \text{ where } p_x = \Pr[X = x]$$

  - Bounds: $0 \leq H(X) \leq \log|\Omega|$
- Conditional Entropy
  - Defn: for jointly distributed (X, Y)

$$H(X|Y) = \sum_{y} p_y \, H(X|Y = y)$$

  - Chain Rule: $H(X, Y) = H(X) + H(Y|X)$
  - Conditioning does not increase entropy: $H(X|Y) \leq H(X)$
    - Equality $\Leftrightarrow$ Independence: $H(X|Y) = H(X) \Leftrightarrow \quad X \perp Y$

Iteration by Example:

- $(A, B, C, D) \rightarrow (A + B + C + D, B + D, C + D, D)$
- $(E, F, G, H) \rightarrow (E + F + G + H, F + H, G + H, H)$

$$B + D \longrightarrow B + D + F + H$$

$$H(B + D \mid A + B + C + D)$$

$$H\left(B + D + F + H \mid \begin{Bmatrix} A + B + C + D, \\ E + F + G + H \end{Bmatrix}\right)$$

$$F + H \longrightarrow F + H$$

$$H(F + H \mid E + F + G + H)$$

$$H\left(F + H \mid \begin{Bmatrix} A + B + C + D, \\ E + F + G + H. \\ B + D + F + H \end{Bmatrix}\right)$$

Algorithmics

- Summary: Iterating t times yields $n \times n$-polarizing transform for $n = 2^t$:
  - $(Z_1, \ldots, Z_n) \to (W_1, \ldots, W_n) \to W_S \quad (|S| \approx H(p) \cdot n)$
- Encoding/Compression: Clearly poly time.
- Decoding/Decompression:
  - "Successive cancellation decoder": implementable in poly time.
    - Can compute $\Pr_{Z \sim \mathrm{bern}(p)}[W_i = 1 \mid W_{<i}]$ recursively
- Can reduce both run times to $O(n \log n)$
- Only barrier: Determining S
  - Resolved by [Tal-Vardy'13]

Analysis: The Polarization martingale

- Martingale: Sequence of r.v., $X_0, X_1, \ldots$
  - $\forall i, x_0, \ldots, x_{i-1},\ \mathbb{E}[X_i | x_0, \ldots, x_{i-1}] = x_{i-1}$
- Polarization martingale:
  - X_i = Conditional entropy of random output at ith stage.
- Why martingale?
  - At ith stage two inputs take (U, A); (V, B)
  - Output = $(U + V, A \cup B)$ and $(V, A \cup B \cup U + V)$
  - Input entropies = x_{i-1} $\quad (H(U|A) = H(V|B) = x_{i-1})$

  $\Rightarrow$ Output entropies average to x_{i-1}

  $$H(U + V | A, B) + H(V | A, B, U + V) = 2x_{i-1}$$

Quantitative issues

- How "quickly" does this martingale converge? (λ)
- How "well" does it converge? (ϵ)
- Formally, want:

$$\Pr\left[X_t \in (\lambda, 1-\lambda)\right] \leq \epsilon$$

  - ϵ: Gives gap to capacity
  - λ: Governs decoding failure $\approx n \cdot \sqrt{\lambda}$
  - Reminder: $n = 2^t$
  - For useable code, need: $\lambda \ll 4^{-t}$
  - For poly$\left(\frac{1}{\epsilon}\right)$ block length, need: $\epsilon \leq \alpha^t, \quad \alpha < 1$

Regular and Strong Polarization

- (Regular) Polarization:
  $$\forall \gamma > 0, \qquad \mathrm{Lim}_{t\to\infty} \{\Pr[X_t \in (\gamma^t, 1-\gamma^t)]\} \to 0.$$
- Strong Polarization:
  $$\forall \gamma > 0\ \exists \alpha > 0\ \forall t, \qquad \Pr[X_t \in (\gamma^t, 1-\gamma^t)] \le \alpha^t$$
- Both "global" properties (large t behavior)
- Construction yields: Local Property
  - How does X_i relate to X_{i-1} ?
- Local vs. Global Relationship?
  - Regular polarization well understood.
  - Strong understood only $(U,V) \mapsto (U+V,V)$
  - What about $(U,V,W) \mapsto (U+V,V,W)$?, more generally?

General Polarization

- Basic underlying ingredient $M \in \mathbb{F}_2^{k \times k}$
  - Local polarization step: Pick $(U_j, A_j)_{j \in [k]}$ i.i.d.
  - One step: generate $(L_1, \ldots, L_k) = M \cdot (U_1, \ldots, U_k)$
  - $X_{i-1} = H(U_j | \cup_\ell A_\ell, U_{<j}) = H(U_j | A_j)$
  - $X_i = H(L_j | \cup_\ell A_\ell, L_{<j})$
- Examples
  - $H_2 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$; $H_3' = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
  - Which matrices polarize? Strongly?

Mixing matrices

- Observation 1: Identity matrix does not polarize.
- Observation 1': Upper triangular matrix does not polarize.
- Observation 2: (Row) permutation of non-polarizing matrix does not polarize!
  - Corresponds to permuting $(U_1, \ldots, U_k)$
- Definition: M <u>mixing</u> if no row permutation is upper-triangular.
- [KSU]: M <u>mixing</u> $\Rightarrow$ martingale polarizes.
- [Our main theorem]: M <u>mixing</u> $\Rightarrow$ martingale polarizes <u>strongly</u>.

Key Concept: Local (Strong) Polarization

- Martingale $X_0, X_1, \ldots$ <u>locally polarizes</u> if it exhibits
  - Variance in the middle:

    $\forall \tau > 0\ \exists \sigma > 0, \qquad \mathrm{Var}[X_i \mid X_{i-1} \in (\tau, 1-\tau)] > \sigma$
  - Suction at the end:

    $\exists \theta > 0\ \forall c < \infty, \exists \tau > 0 \quad \Pr\left[X_i < \frac{X_{i-1}}{c} \mid X_{i-1} \leq \tau\right] \geq \theta$

    (similarly with $1 - X_i$)
- Definition local: compares X_i with X_{i-1}
- Implications global:
- Thm 1: Local Polarization ⇒ Strong Polarization.
- Thm 2: Mixing matrices ⇒ Local Polarization.

Mixing Matrices ⇒ Local Polarization

- 2×2 case:
  - Ignoring conditioning, we have

$$H(p) \to \left\{ \begin{array}{ll} H\big(2p(1-p)\big) & \text{w.p. } \frac{1}{2} \\ 2H(p) - H\big(2p(1-p)\big) & \text{w.p. } \frac{1}{2} \end{array} \right\}$$

  - Variance, Suction ⇐ Taylor series for $\log(.)$ ...

- $k \times k$ case:
  - Reduction to 2×2 case: $\frac{1}{2}$ probs. $\to \sim \frac{1}{k}$

Local ⇒ (Global) Strong Polarization

- Step 1: Medium Polarization:
  - $\exists \alpha, \beta < 1 \; \Pr[\, X_t \in (\beta^t, 1-\beta^t)\,] \le \alpha^t$
  - Proof: Potential $\Phi_t = \mathbb{E}\left[\min\left\{\sqrt{X_t}, \sqrt{1-X_t}\right\}\right]$
    - Variance+Suction ⇒ $\exists \eta < 1$ s.t. $\Phi_t \le \eta \cdot \phi_{t-1}$
    - (Aside: Why $\sqrt{X_t}$? Clearly X_t won't work. And X_t^2 increases!)
    - $\Rightarrow \Phi_t \le \eta^t \quad \Rightarrow \Pr[\, X_t \in (\beta^t, 1-\beta^t)\,] \le \left(\frac{\eta}{\beta}\right)^t$

Local ⇒ (Global) Strong Polarization

- Step 1: Medium Polarization:
  - $\exists \alpha, \beta < 1 \Pr[\, X_t \in (\beta^t, 1-\beta^t)\,] \le \alpha^t$
- Step 2: Medium Polarization + t-more steps ⇒ Strong Polarization:
- Proof: Assume $X_0 \le \alpha^t$; Want $X_t \le \gamma^t$ w.p. $1 - O(\alpha_1^t)$
  - Pick c large & τ s.t. $\Pr\left[X_i < \frac{X_{i-1}}{c} \mid X_{i-1} \le \tau\right] \ge \theta$
  - 2.1: $\Pr[\, \exists i \; s.t. X_i \ge \tau\,] \le \tau^{-1} \cdot \alpha^t$ (Doob's Inequality)
  - 2.2: Let $Y_i = \log X_i$; Assuming $X_i \le \tau$ for all i
    - $Y_i - Y_{i-1} \le -\log c$ w.p. $\ge \theta$; & $\ge k$ w.p. $\le 2^{-k}$.
    - $\mathbb{E}\mathrm{xp}[Y_t] \le -t\left(\frac{\log c}{\theta} + 1\right) \le 2t \cdot \log \gamma$
    - Concentration! $\Pr[Y_t \ge t \cdot \log \gamma] \le \alpha_2^t$ QED!

Conclusion

- Simple General Analysis of Strong Polarization
- But main reason to study general polarization:
  - Maybe some matrices polarize even faster!
  - This analysis does not get us there.
  - But hopefully following the (simpler) steps in specific cases can lead to improvements.

Thank You!