General Strong Polarization

Madhu Sudan
Harvard University

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

This Talk: Martingales and Polarization

- [0,1]-Martingale: Sequence of r.v.s $X_0, X_1, X_2, \ldots$
  $X_t \in [0,1], \qquad \mathbb{E}[X_{t+1} | X_0, \ldots, X_t] = X_t$
- Well-studied in algorithms:
  - [MotwaniRaghavan, §4.4], [MitzenmacherUpfal, §12]
  - E.g., $X_t =$ Expected approx. factor of algorithm given first t random choices.
  - Usually study "concentration"
    - E.g. … whp $X_{\text{final}} \in [X_0 \pm \epsilon] \Leftrightarrow \lim_{t\to\infty} X_t = X_0$
- Today: Polarization:
  - … whp $X_{\text{final}} \notin (\epsilon, 1-\epsilon) \Leftrightarrow \lim_{t\to\infty} X_t = \text{Bern}(X_0)$
- Why? – Polar codes …

Main Result: Definition and Theorem

- Strong Polarization: (informally)
  $\Pr[X_t \in (\tau, 1-\tau)] \le \epsilon$ if $t \ge \max\{O(\log 1/\epsilon), o(\log 1/\tau)\}$
  formally $\forall \gamma > 0\ \exists \beta < 1, c\ s.t. \forall t\ \ \Pr[X_t \in (\gamma^t, 1-\gamma^t)] \le c \cdot \beta^t$
- A non-polarizing example: $X_0 = \frac{1}{2}$; $X_{t+1} = X_t \pm 2^{-t-2}$
- Local Polarization:
  - Variance in the middle: $X_t \in (\tau, 1-\tau)$
    $\forall \tau > 0\ \exists \sigma > 0\ s.t. \forall t, X_t \in (\tau, 1-\tau) \Rightarrow Var[X_{t+1} | X_t] \ge \sigma$
  - Suction at the ends: $X_t \notin (\tau, 1-\tau)$
    $\exists \theta > 0, \forall c < \infty, \exists \tau > 0\ s.t. X_t < \tau \Rightarrow \Pr\left[X_{t+1} < \frac{X_t}{c}\right] \ge \theta$
- Theorem: Local Polarizatio on.

Both definitions qualitative!

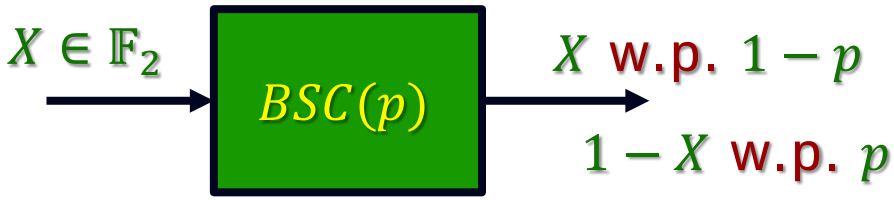
"low end" condition. Similar condition for high end

Rest of the talk

- Background/Motivation:
  - The Shannon Capacity Challenge
  - Resolution by [Arikan'08], [Guruswami-Xia'13], [Hassani,Alishahi,Urbanke'13]: Polar Codes!
  - Key ideas:
    - Polar Codes & Arikan-Martingale.
    - Implications of Weak/Strong Polarization.
- Our Contributions: Simplification & Generalization
  - Local Polarization $\Rightarrow$ Strong Polarization
  - Arikan-Martingale Locally Polarizes (generally)

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:



Rate?

Encode: $\mathbb{F}_2^{k_n} \to \mathbb{F}_2^n$

$\lim_{n\to\infty} \frac{k_n}{n} \geq R$

  - 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 $\mathrm{poly}\left(\frac{n}{\epsilon}\right)$?
    - Open till 2008
    - Arikan'08: Invented "Polar Codes" …
    - Final resolution: Guruswami+Xia'13, Hassani+Alishahi+Urbanke'13 – Strong analysis

Polar Codes and Polarization

- Arikan:
  - Defined Polar Codes, one for every t
  - Associated Martingale $X_0, \ldots, X_t, \ldots$
  - tth code: Length $n = 2^t$
  - If $\Pr[X_t \in (\gamma, 1-\delta)] \leq \epsilon$ then tth code is $(\epsilon + \delta)$-close to capacity, and
    - $\Pr[Decode(Encode(m) + error) \neq m] \leq n \cdot \gamma$
    - Need $\gamma = o\left(\frac{1}{n}\right)$ (strong polarization $\Rightarrow \gamma = neg(n)$)
    - Need $\epsilon = 1/n^{\Omega(1)}$ ($\Leftarrow$ strong polarization)

Context for today's talk

- [Arikan'08]: Martingale associated to general class of codes.
  - Martingale polarizes (weakly) implies code achieves capacity (weakly)
  - Martingale polarizes weakly for entire class.
- [GX'13, HAU'13]: Strong polarization for one specific code.
- Why so specific?
  - Bottleneck: Strong Polarization
- Rest of talk: Polar codes, Martingale and Strong Polarization

Lesson 0: Compression ⇒ Coding

- Linear Compression Scheme:
  - (H, D) for compression scheme for $\mathrm{bern}(p)^n$ if
    - Linear map $H: \mathbb{F}_2^n \to \mathbb{F}_2^m$
    - $\Pr_{Z \sim \mathrm{bern}(p)^n}[D(H \cdot Z) \neq Z] = o(1)$
    - Hopefully $\frac{m}{n} \leq H(p) + \epsilon$
- Compression ⇒ Coding
  - Let $H^\perp$ be such that $H \cdot H^\perp = 0$;
  - Encoder: $X \mapsto H^\perp \cdot X$
  - Error-Corrector: $Y = H^\perp \cdot X + \eta \mapsto Y - D(H \cdot Y)$

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

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 almost random, or almost 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:

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

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

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

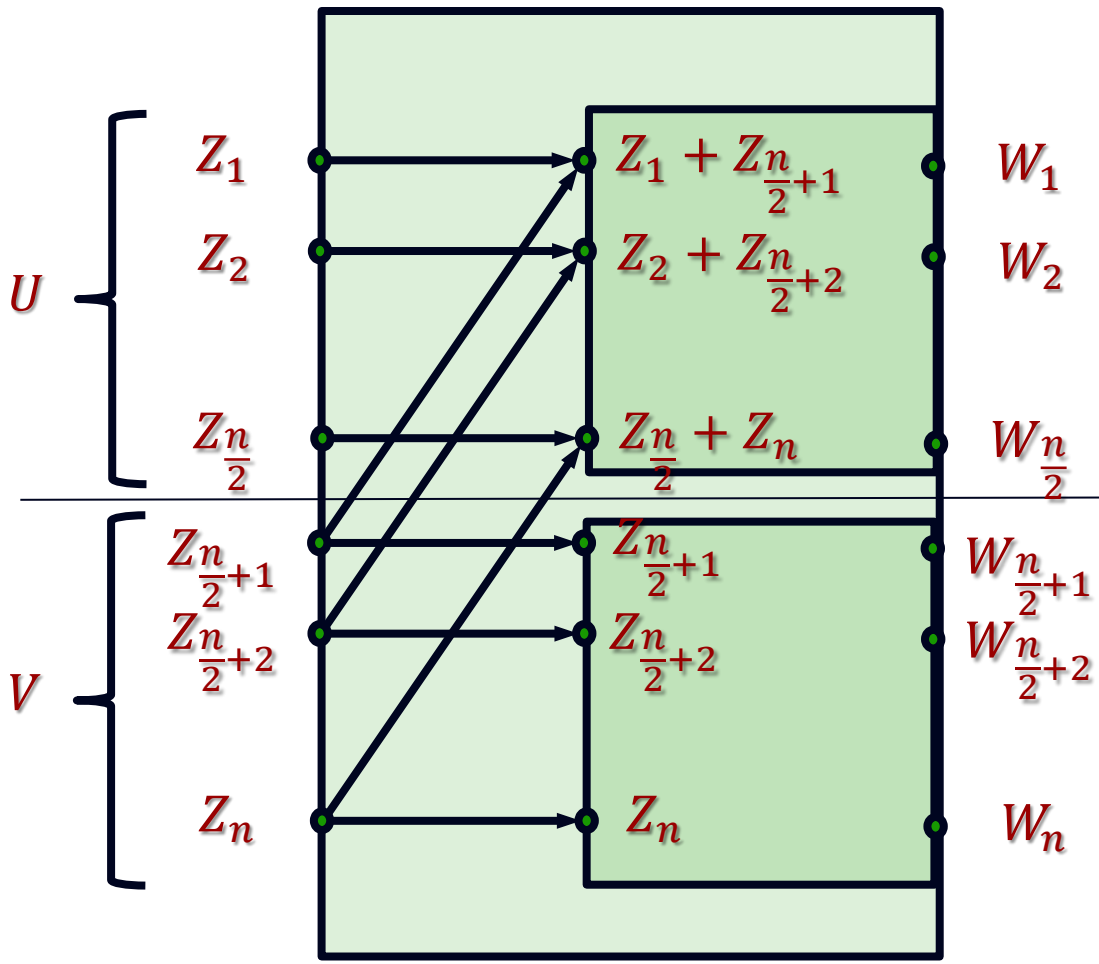
$$H\left(B + D + F + H \mid \begin{Bmatrix} A + B + C + D, \\ E + F + G + H \end{Bmatrix}\right)$$

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

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

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

The Polarization Butterfly



$$P_n(U,V) = \left(P_{\frac{n}{2}}(U+V), P_{\frac{n}{2}}(V)\right)$$

Polarization $\Rightarrow \exists S \subseteq [n]$

$$H\left(W_{[n]-S} \,\middle|\, W_S\right) \to 0$$

$$|S| \leq (H(p) + \epsilon) \cdot n$$

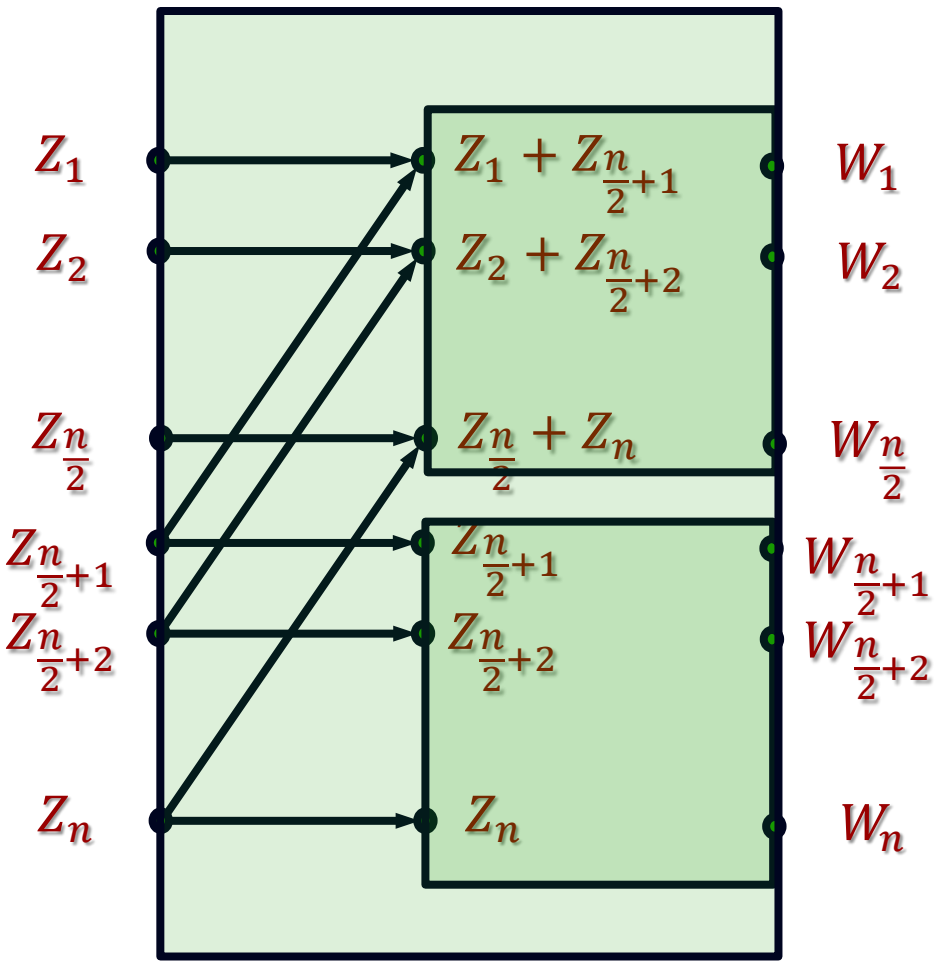
Compression $E(Z) = P_n(Z)_S$

Encoding time $= O(n \log n)$ (given S)

To be shown:

1. Decoding $= O(n \log n)$
2. Polarization!

The Polarization Butterfly: Decoding



Decoding idea:

Given $S, W_S = P_n(U,V)$

Compute $U + V \sim \mathrm{Bern}(2p - 2p^2)$

Determine $V \sim ?$

$$V|U+V \sim \prod_i \mathrm{Bern}(r_i)$$

Key idea: Stronger induction!

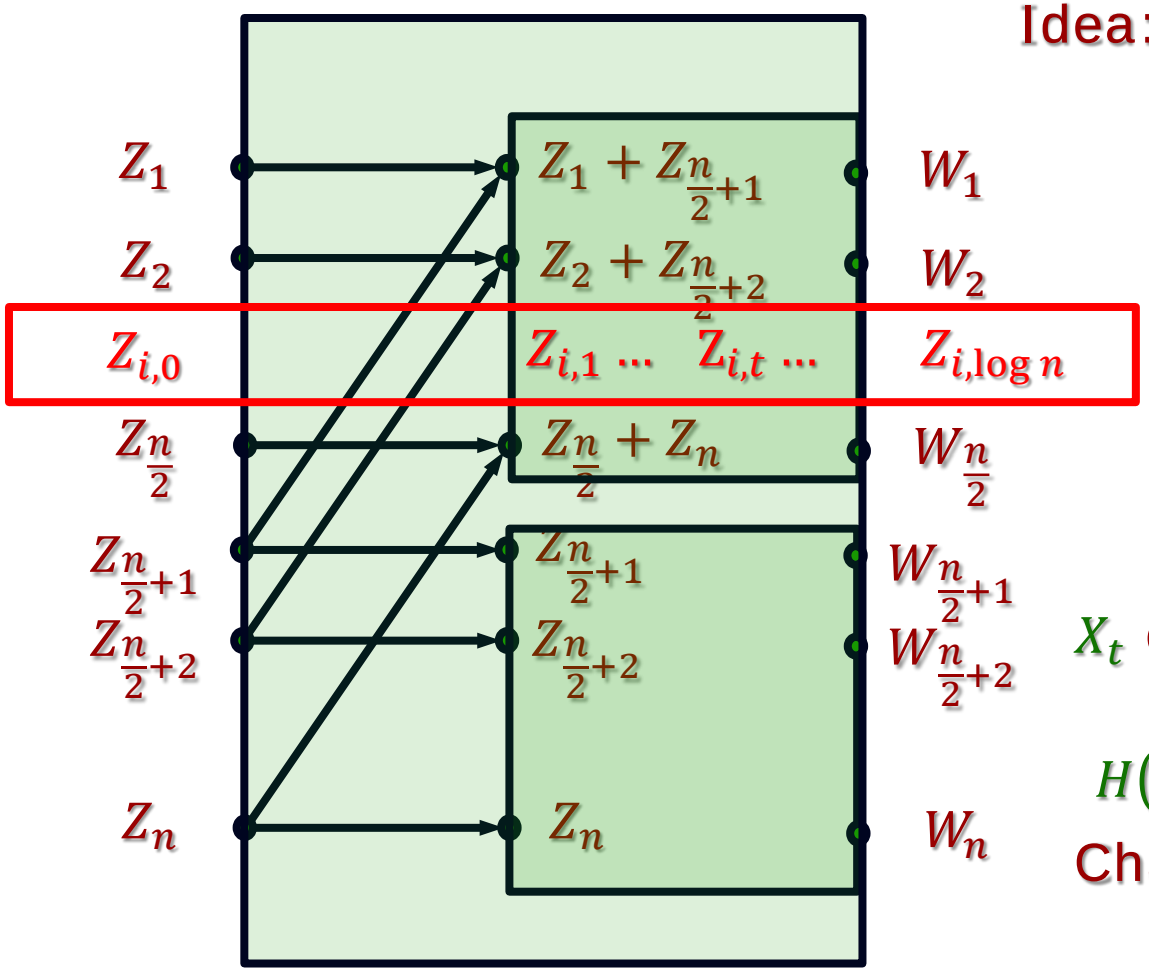
- non-identical product distribution

Decoding Algorithm:

Given $S, W_S = P_n(U,V), p_1, \dots, p_n$

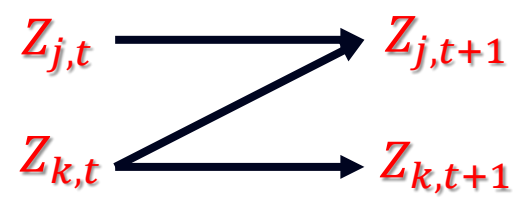
- Compute $q_1, .. q_{\frac{n}{2}} \leftarrow p_1 \dots p_n$
- Determine $U + V = D\left(W_S^+, q_1 \dots q_{\frac{n}{2}}\right)$
- Compute $r_1 \dots r_{\frac{n}{2}} \leftarrow p_1 \dots p_n; U + V$
- Determine $V = D\left(W_S^-, r_1 \dots r_{\frac{n}{2}}\right)$

Polarization? Martingale?



Idea: Follow $X_t = H(Z_{i,t} | Z_{<i,t})$ for random i

Martingale?



X_t can't distinguish $i = j$ from $i = k$

$$H(Z_{j,t}, Z_{k,t}) = H(Z_{j,t+1}, Z_{k,t+1})$$

Chain Rule …

$$\Rightarrow \mathbb{E}[X_{t+1} | X_t] = X_t$$

Remaining Tasks:

1. Prove that $X_t \stackrel{\text{def}}{=} H(Z_{i,t} | Z_{<i,t})$ polarizes locally
2. Prove that Local polarization ⇒ Strong Polarization.

- Recall
  - Strong Polarization: (informally)

$\Pr[X_t \in (\tau, 1-\tau)] \leq \epsilon$ if $t \geq \max\{O(\log 1/\epsilon), o(\log 1/\tau)\}$

formally $\forall \gamma > 0\ \exists \beta < 1, c\ s.t. \forall t\ \ \Pr[X_t \in (\gamma^t, 1-\gamma^t)] \leq c \cdot \beta^t$

  - Local Polarization:
    - Variance in the middle: $X_t \in (\tau, 1-\tau)$

$\forall \tau > 0\ \exists \sigma > 0\ s.t. \forall t, X_t \in (\tau, 1-\tau) \Rightarrow Var[X_{t+1} | X_t] \geq \sigma$

    - Suction at the ends: $X_t \notin (\tau, 1-\tau)$

$\exists \theta > 0, \forall c < \infty, \exists \tau > 0\ s.t. X_t < \tau \Rightarrow \Pr\left[X_{t+1} < \frac{X_t}{c}\right] \geq \theta$

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!

Polarization of Arikan Martingale

- Ignoring conditioning:
  - $X_t = H(p) \Rightarrow X_{t+1} = H(2p - 2p^2)$ w.p. ½
    $= 2H(p) - H(2p - 2p^2)$ w.p. ½
- Variance in the middle: Continuity of $H(\cdot)$ and separation of $2p - 2p^2$ from p
- Suction at high end:
  - $X_t = 1 - \epsilon \quad \Rightarrow X_{t+1} = 1 - O(\epsilon^2)$ w.p. ½
- Suction at low end:
  - $X_t = H(p) \quad \Rightarrow X_{t+1} = H\left(\frac{p}{\log p}\right)$ w.p. ½

General Strong Polarization?

- So far: $P_n = H_2^{\otimes \ell}$ for $H_2 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$
- What about other matrices?
  - E.g., $P_n = H_3^{\otimes \ell}$ for $H_3 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.
  - Open till our work!
  - Necessary conditions: H invertible, not-lower-triangular
  - Our work: Necessary conditions suffice (for local polarization)!

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!