Reduction to bare minimum depth

- By the efficient $O(\lg d)$-depth reduction we know that: to prove hardness results for a degree d polynomial f it suffices to study $O(\lg d)$-depth.

- Now we will reduce this, further, to depth-4.

Theorem [Agrawal-Vinay '08, Koiran '12, Tavenas '15]: Let f be a degree d polynomial computed by a size s circuit. Then, for all $t \in [d]$, f has a homogeneous $\Sigma\Pi^{O(d/t)}\Sigma\Pi^{t}$ circuit of top fanin $s^{O(d/t)}$ & size $s^{O(t+d/t)}$.

[$\Sigma^k\Pi^{d'}\Sigma\Pi^t$ circuit looks like $\sum_{i=1}^{k}\prod_{j=1}^{d'} f_{ij}$ where k is the top fanin & the bottom fanin t bounds the degree of f_{ij}'s.]

[To optimize the size one could take $t = \sqrt{d}$, giving $k \approx \text{size} \approx s^{O(\sqrt{d})}$ which is nontrivial!]

Proof: • We will use Saptharishi (2016)'s version.

• Let C be the $O(\lg d)$-depth circuit, of size s, computing f. Wlog, for each internal gate g of C we have a homogeneous expr.:

$$g = \sum_{i \in [s]} g_{i1} \cdots g_{i5} \text{,} \qquad \text{-----(1)}$$

where for each lower gate $\deg g_{ij} \leq \frac{\deg g}{2}$.

[Recall that C can be computed in randomized poly(snd)-time.]

• In particular, the above expression (1) gives a $\Sigma^{s}\Pi\Sigma\Pi^{d/2}$ circuit computing f.

To reach to $\Sigma\Pi\Sigma\Pi^{t}$, we will incrementally "open" it up:

i) For each summand $g_{i1} \cdots g_{ir}$, with some $\deg g_{ij} > t$, expand g_{ij} one step further (& g) using the expression (1).

ii) Repeat this process till all g_{ij}'s on the RHS have degree $\leq t$.

• Each expansion, like (i), grows the top fanin by a multiple of s.

We intend to show that this can happen only $O(d/t)$ times.

- In eqn.(1) if $\deg g =: d'$, then the largest degree g_{ij} in any summand has deg $\geq d'/5$ (by homogeneity). Moreover, the second largest degree is $\geq \frac{1}{4}\cdot\left(\frac{d'}{2}\right)$, as $\deg g_{ij} \leq d/2$.

$\frac{d'-d'/2}{5-1}$

$\Rightarrow$ in each new summand there are two factors of degree $\geq d'/8$.

$\Rightarrow$ Whenever we expand by eqn.(1), a factor of deg $\geq t$, we introduce at least one more factor of deg $\geq t/8$ (in each new summand).

- Note that, by homogeneity, there can be $\leq 8d/t$ factors (in a summand) of deg $\geq t/8$.

$\Rightarrow$ the number of iterations is $\leq 8d/t$.

$\Rightarrow$ The eventual #summands $= s^{O(d/t)}$.

- Note that the factors in a summand can have at most n^t many monomials.
  $\Rightarrow$ Eventually, C converts to a $\Sigma\Pi\Sigma\Pi^t$ circuit with top fanin $s^{O(d/t)}$ & size $s^{O(t+d/t)}$. $\square$

<u>Corollary</u>: An n-var. d-deg polynomial f requires homogeneous $\Sigma\Pi^{O(d/t)}\Sigma\Pi^t$ circuits of top fanin $n^{\omega(d/t)}$

$\Rightarrow$ f requires arithmetic circuits of size $n^{\omega(1)}$.

<u>Proof</u>:
- We just proved its contrapositive! $\square$

- Could we reduce this $\Sigma\Pi\Sigma\Pi^t$ circuit to a $\Sigma\Pi\Sigma$ one (nontrivially)?

- YES, over zero characteristic fields.

Depth-3 Chasm

Theorem [Gupta, Kamath, Kayal, Saptharishi, 2013]: Let f be a deg-d polynomial computed by a size-s circuit over $\mathbb{F}$ (char $\mathbb{F} = 0$). Then, there is a $\underline{\Sigma\Pi\Sigma^{\sqrt{d}}}$ circuit of size $s^{O(\sqrt{d})}$ computing f.

[$\Sigma\Pi\Sigma^{m}$ circuit looks like $\sum_{i\in[k]}\prod_{j\in[d_i]} l_{ij}$, where k is the <u>top fanin</u> & each l_{ij} is a <u>linear</u> polynomial in some m variables.]

[In the above thm. we get inhomogeneous $\Sigma\Pi\Sigma$ where both k & d_i could be $s^{\sqrt{d}}$.]

Corollary: Over $\mathbb{Q}$, $\det_n$ has a $n^{O(\sqrt{n})}$-size $\Sigma\Pi\Sigma^{\sqrt{n}}$ circuit.

Conjecture: 1) $\det_n$ requires $n^{\Omega(\sqrt{n})}$-size $\Sigma\Pi\Sigma$.

*2) per_n requires $n^{\Omega(n)}$-size $\Sigma\Pi\Sigma^{\sqrt{n}}$.

[Weaker: optimality of Ryser's formula $\Rightarrow$ VP $\neq$ VNP.]

- The proof requires a host of ideas. One common feature is to use powers basis, instead of the standard basis of monomials, to express polynomials.

- Outline: circuit $\xrightarrow{\text{Step 0}}$ $\Sigma\Pi\Sigma\Pi$ $\xrightarrow{\text{Step 1}}$ $\Sigma\wedge\Sigma\wedge\Sigma$ circuits $\xrightarrow{\text{Step 2}}$ $\Sigma\Pi\Sigma$ (over $\mathbb{C}$) $\xrightarrow{\text{Step 3}}$ $\Sigma\Pi\Sigma$ (over $\mathbb{Q}$).

Step 0: • Let f have a size-s_0 circuit $C_0(x_1,\ldots,x_n)$.
• By depth-4 reduction we get a size $s_1 = s_0^{O(\sqrt{d})}$ homogeneous $\Sigma\Pi^{O(\sqrt{d})}\Sigma\Pi^{\sqrt{d}}$ circuit C_1.

Step 1: • First, we show a general way to "change basis" that converts "Π" to "$\Sigma\wedge\Sigma$":

Lemma (Fischer's trick '94): Over $ch(\mathbb{F}) > r$ or zero, any expression $g = \sum_{i\in[k]} \prod_{j\in[r]} g_{ij}$, $\deg g_{ij} \le \delta$, can be rewritten as $g = \sum_{i=1}^{k'} c_i \cdot g_i^r$, where $k' = k \cdot 2^r$ & $\deg g_i \leq \delta$. $c_i \in \mathbb{F}$

Proof:

- Recall Ryser's formula for permanent.
- $r! \cdot y_1 \cdots y_r = \text{per}\begin{pmatrix} y_1 & \cdots & y_r \\ \vdots & & \vdots \\ y_1 & \cdots & y_r \end{pmatrix}$

$$= \sum_{S \subseteq [r]} \Big(\sum_{j \in S} y_j\Big)^r \cdot (-1)^{r-|S|}.$$

- We can apply this on each product $g_{i1} \cdots g_{ir}$ to rewrite g as a sum of powers of g_j's. $\square$

– Eg. Over $\mathbb{F}_2$, $x_1 x_2$ cannot be written as a sum of powers. (Exercise)

- Using Fischer's trick on all the product gates of $\Sigma\Pi^{O(\sqrt{d})}\Sigma\Pi^{\sqrt{d}}$ circuit C_1, we get a $\Sigma\Lambda^{O(\sqrt{d})}\Sigma\Lambda^{\sqrt{d}}\Sigma^{\sqrt{d}}$ circuit C_2 of size $s_2 = s_1 \cdot 2^{O(\sqrt{d})} = s_0^{O(\sqrt{d})}$.

Step 2: • First, we show a general transformation from $\wedge\Sigma$ to $\Sigma\Pi\Sigma$ (over $\mathbb{C}$): duality trick (S. '08).

• Before that we recall the classic interpolation formula.

Fact (Interpolation) [Waring 1779]: Let $F(x)$ be a deg-D polynomial & $\alpha_0, \ldots, \alpha_D \in \mathbb{F}$ be distinct. Then, $\forall\, 0 \le i \le D$, $\exists\, \beta_0(\bar{\alpha}), \ldots, \beta_D(\bar{\alpha}) \in \mathbb{F}$ s.t.

$$\mathrm{coef}(x^i)(F) = \sum_{0 \le j \le D} \beta_j \cdot F(\alpha_j) .$$

Proof: • Let $F(x) = \sum_{0 \le j \le D} c_j \cdot x^j$. Thus, as a matrix:

$$\begin{pmatrix} 1 & \alpha_0 & \cdots & \alpha_0^D \\ 1 & \alpha_1 & \cdots & \alpha_1^D \\ \vdots & \vdots & & \vdots \\ 1 & \alpha_D & \cdots & \alpha_D^D \end{pmatrix} \cdot \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_D \end{pmatrix} = \begin{pmatrix} F(\alpha_0) \\ F(\alpha_1) \\ \vdots \\ F(\alpha_D) \end{pmatrix} .$$

• The Vandermonde matrix is invertible.

(Exercise) $\square$

The duality trick

Theorem [S. '08]: There exists a deg-b polynomial f_i s.t. $(z_1+\dots+z_s)^b = \sum_{i\in[sb(b+1)]} c_i \cdot f_i(z_1)\cdots f_i(z_s)$.

[This transforms $\Sigma\wedge\Sigma$ circuit to a sum-of-product of univariates. The latter is $\Sigma\Pi\Sigma$ over $\mathbb{Q}$.]

Proof: (We see a simpler pf by Shpilka.)

- Consider the polynomial $F(t) := \prod_{i\in[s]} (t+z_i)$.

- Using interpolation (at points $\alpha_1,\dots,\alpha_{sb}$) we can extract the $\mathrm{coef}(t^{(s-1)b})((F-t^s)^b)$ as:

$$\Big(\sum_{i\in[s]} z_i\Big)^b = \sum_{i\in[sb]} \beta_i \cdot (F(\alpha_i)-\alpha_i^s)^b$$

$$\Rightarrow \Big(\sum_{i=1}^{s} z_i\Big)^b = \sum_{\substack{i\in[sb] \\ 0\le j\le b}} \gamma_{ij} \cdot F(\alpha_i)^j$$

$$=: \sum_{i,j} \gamma_{ij} \cdot (\alpha_i+z_1)^j \cdots\cdots (\alpha_i+z_s)^j \,. \qquad \square$$

- Thus, a homogeneous $\wedge\Sigma\wedge$ circuit can be transformed as:

$$(z_{\gamma_1}^a + \cdots + z_{\gamma_s}^a)^b = \sum_{i,j} \gamma_{ij} \cdot (\alpha_i + z_{\gamma_1}^a)^j \cdots (\alpha_i + z_{\gamma_s}^a)^j .$$

- The summand will factor nicely over $\mathbb{C}$. In fact, we can choose (α_i) to be an integral a-power, for all i. Then, the factors would live over $\mathbb{Q}(\zeta_a)$. $[\zeta_a := 1^{1/a} \in \mathbb{C}]$

$\Rightarrow$ $\Sigma\wedge^b\Sigma\wedge^a\Sigma^1$ circuit can be expressed as a $\Sigma\Pi\Sigma^2$ circuit, over $\mathbb{Q}(\zeta_a)$, of $O(s^3ab^2)$-size.

$\Rightarrow$ We have **obtained** a $\Sigma\Pi\Sigma^{a+1}$ circuit, over $\mathbb{Q}(\zeta_a)$ for $a := \lceil\sqrt{d}\rceil$, denoted by C_3 of size $s_3 = \tilde{O}(s_2^3) = s_0^{O(\sqrt{d})}$, that also computes C_2.

(intermediate deg in C_3 is extremely high!)

<u>Step 3:</u> Note that C_3 has coefficients in $\mathbb{Q}(\zeta_a)$, but eventually it computes C_2 which is free of ζ_a. We can utilize this to eliminate ζ_a from C_3.

<u>Lemma:</u> Let $f(\bar{x}) \in \mathbb{Q}(\zeta_a)[\bar{x}]$ be a $\Sigma\Pi\Sigma$ circuit of deg-d, size s computing a poly. in $\mathbb{Q}[\bar{x}]$. Then, $\exists$ equivalent $\Sigma\Pi\Sigma$ circuit $g \in \underline{\mathbb{Q}[\bar{x}]}$ of deg-d, size-$\underline{O(sda)}$.

<u>Proof:</u>

- Replace each occurrence of ζ_a^i, in the circuit f, by y^i to get a circuit $\tilde{f} \in \mathbb{Q}[\bar{x}, y]$.

$\tilde{f}$ is $\Sigma\Pi\Sigma\Pi$ because of y.

- $\deg_y \tilde{f} \leq (d+1)a$, since f is $\Sigma\Pi^d\Sigma$.
- Also, $\tilde{f}(\bar{x}, \zeta_a) = f(\bar{x})$.

- Note that $\sum_{0 \leq i \leq d+1} \text{coef}(y^{ia})(\tilde{f}) = f$.

[Taking $tr_{\mathbb{Q}(\zeta_a)/\mathbb{Q}}$ both sides.]

- Thus, we could interpolate f by evaluating $\tilde{f}(\bar{x}, y)$ on $1+(d+1)a$ distinct points in $\mathbb{Q}$.
- This yields an $O(sda)$-size $\Sigma\Pi\Sigma$ circuit, for f, over $\mathbb{Q}$.

$\square$

- Thus, we get a $\Sigma\Pi\Sigma^{\sqrt{d}}$ circuit C_4 computing C_3, over $\mathbb{Q}$, which is of size $s_4 = \tilde{O}(s_3) = s_0^{O(\sqrt{d})}$.

This completes the depth-3 chasm.

$\square$

Width reduction in ABP

- We now explore the power of constant-width ABP.

Theorem [Ben-Or, Cleve '88]: Formulas & width-3 ABP are equivalent up to poly-size.

Proof:

- Let F be a formula of size-s.
  Wlog we can assume it to be of fanin two & depth $d = O(\lg s)$. [Brent's formula reduction]

- We intend to compute F by IMM of 3×3 matrices using bottom-up induction.
- Gate $E \in \mathbb{F} \cup \bar{x}$ can be computed as (base case): $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ E & 0 & 1 \end{pmatrix}$ Idea: $(R_1, R_2, R_3) \mapsto (R_1 + R_3 \cdot E, R_2, R_3)$
- Gate $E = E_1 + E_2$ can be computed as:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ E_1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ E_2 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ E & 0 & 1 \end{pmatrix}$$

$\triangleright \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ E & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -E & 0 & 1 \end{pmatrix}.$

- Gate $E = E_1 \cdot E_2$ computed as:

$$\begin{pmatrix} 1 & 0 & 0 \\ -E_2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & E_1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ E_2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -E_1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ -E_2 & 1 & 0 \\ 0 & E_1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ E_2 & 1 & 0 \\ 0 & -E_1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ E_1 \cdot E_2 & 0 & 1 \end{pmatrix}.$$

- Note that $\begin{pmatrix} 1 & 0 & 0 \\ f & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ refers to the step

$(R_1, R_2, R_3) \mapsto (R_1 + R_2 \cdot f, R_2, R_3)$. Thus, its IMM form mirrors that for $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ f & 0 & 1 \end{pmatrix}$.

The latter we have constructed by induction.

$\Rightarrow$ The IMM for $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ F & 0 & 1 \end{pmatrix}$ can be found by induction & is of size $\leq 4^d$ $= \mathrm{poly}(s)$.

▷ IMM for F is poly(s)-size & involves permuted triangular 3×3 matrices of det $= \pm 1$.
unimodular

- For the converse, let $A = A_1 \cdots A_s$ be a product of 3×3 symbolic matrices.

- Suppose we have size $T(s/2)$ formulas for $L = A_1 \cdots A_{s/2}$ & $A_{s/2+1} \cdots A_s = R$.

As L & R are 3×3 matrices we can make 3 copies of each of their entries & get a formula for $L \cdot R$ of size:

$T(s) = 6 \cdot T(s/2) + O(1)$

$\Rightarrow T(s) = O(6^{\lg s}) = o(s^3)$.

▷ constant-width IMM A has poly(s)-size formula.

$\square$

▷ $O(1)$-depth ckt. $\leqslant$ formula $\equiv$ $O(1)$-width ABP

▷ quasipoly formula $\geqslant$ ABP $\leqslant$ low-deg circuit.

Width-2 Chasm

- We now show that even triangular 2×2 matrices give a strong ABP.

Theorem [Saha, Saptharishi, S. '09]: Let f be a $\Sigma^k \Pi^d \Sigma^{n+1}$ polynomial. There is a size $O(dk^2)$ width-2 ABP that computes $f\cdot L$, where L is a product of nonzero linear polynomials.

[The proof gives a poly(dkn)-time algorithm to compute a width-2 ABP with upper triangular matrices.

In this sense the ABP uses the minimum amount of non-commutativity.]

Proof:

- Let $f = \sum_{i=1}^{k} P_i$, where $P_i = \prod_{j=1}^{d} \ell_{ij}$ is a product of linear polynomials.

- Observe that P_i can be computed by

the length-d matrix product:

$$\begin{pmatrix} \ell_{i1} & 0 \\ 0 & 1 \end{pmatrix} \cdots\cdots \begin{pmatrix} \ell_{i,d-1} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \ell_{id} \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} P_i' & P_i \\ 0 & 1 \end{pmatrix} \quad , \quad P_i' := P_i / \ell_{id} .$$

- Now we need to express their sum, i.e. $P_1 + P_2$.
- More generally, suppose we have IMM for multiples of g & h as:

$$\begin{pmatrix} L_1 & L_2 g \\ 0 & L_3 \end{pmatrix} \ \& \ \begin{pmatrix} M_1 & M_2 h \\ 0 & M_3 \end{pmatrix} .$$

- Can we express $g+h$ in a similar way?
- We can try scaling them by A, B as:

$$\begin{pmatrix} L_1 & L_2 g \\ 0 & L_3 \end{pmatrix} \cdot \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \cdot \begin{pmatrix} M_1 & M_2 h \\ 0 & M_3 \end{pmatrix} =$$

$$\begin{pmatrix} AL_1 M_1 & AL_1 M_2 h + BL_2 M_3 g \\ 0 & BL_3 M_3 \end{pmatrix}$$

- Clearly, this suggests picking $A = L_2 M_3$ & $B = L_1 M_2$ to get
$$\begin{pmatrix} L_1 L_2 M_1 M_3 & L_1 L_2 M_2 M_3 (g+h) \\ 0 & L_1 L_3 M_2 M_3 \end{pmatrix}$$
as an IMM.

- The size of the IMM has increased by at most a multiple of 4.

- We can apply this recursive step to $\sum_{i=1}^{k} P_i$ at most $\lg k + 1$ times.

$\Rightarrow$ We get an IMM of length $\leq d \cdot 4^{\lg k+1} = O(dk^2)$ for
$$\begin{pmatrix} L' & L \cdot f \\ 0 & L'' \end{pmatrix}$$
where, L is a product of $O(dk^2)$ linear polynomials (that appear in the circuit).

$\square$

- width-2 is an incomplete model (Allender, Wang'16).
- Recently, Bringmann, Ikenmeyer & Zuiddam (CCC'17) showed that formulas could be simulated by width-2, if we allow approximative computations!

Thm [BIZ'17]: $\overline{VF} = $ width-2-$\overline{VBP}$, if ch $\mathbb{F} \neq 2$.

Defn: In approximative models, we compute a polynomial $f \in \mathbb{F}[\bar{x}_n]$ by computing a $g \in \mathbb{F}(\varepsilon)[\bar{x}_n]$ s.t.

$$g = f + O(\varepsilon),$$ where notation $O(\varepsilon)$ refers to an element in $\varepsilon \cdot \mathbb{F}[\varepsilon, \bar{x}_n]$.

→ This defn. allows us to use constants like $1/\varepsilon$ in the model, yet set $\varepsilon = 0$ in the end!

Proof: • Use the matrix $Q(f) := \begin{pmatrix} f & 1 \\ 1 & 0 \end{pmatrix}$ to store $f \in \mathbb{F}[\varepsilon, \bar{x}_n]$.

• If we have width-2 ABP for $Q(f) + O(\varepsilon^k)$ & $Q(g) + O(\varepsilon^k)$ then we have IMM:

$$Q(f+g) + O(\varepsilon^k) = \big(Q(f)+O(\varepsilon^k)\big)\cdot Q(0)\cdot\big(Q(g)+O(\varepsilon^k)\big).$$

$$\Big[\because \text{RHS} = \begin{pmatrix} f & 1 \\ 1 & 0 \end{pmatrix}\cdot\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\cdot\begin{pmatrix} g & 1 \\ 1 & 0 \end{pmatrix} + O(\varepsilon^k)$$

$$= \begin{pmatrix} 1 & f \\ 0 & 1 \end{pmatrix}\begin{pmatrix} g & 1 \\ 1 & 0 \end{pmatrix} + O(\varepsilon^k) = \begin{pmatrix} f+g & 1 \\ 1 & 0 \end{pmatrix} + O(\varepsilon^k).\Big]$$

- A trickier identity shows that if we have an IMM for $Q(f)+O(\varepsilon^3)$ then we have an IMM:

$$Q(-f^2)+O(\varepsilon) =$$

$$\begin{pmatrix} -1/\varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}\cdot\big(Q(f)+O(\varepsilon^3)\big)\cdot\begin{pmatrix} \varepsilon^2 & 1 \\ -1 & 0 \end{pmatrix}\cdot\big(Q(f)+O(\varepsilon^3)\big)\cdot\begin{pmatrix} 1/\varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$$

& one for $Q(f^2)+O(\varepsilon)$.

- Finally, for $Q(fg)+O(\varepsilon)$ we can use the identity:

$$fg = -(f/2)^2 + (-g^2) + (f/2+g)^2.$$

- Induction will now convert a formula to width-2-ABP. $\square$