CS 59300 – Algorithms for Data Science

Classical and Quantum approaches

Lecture 3 (09/04)

Tensor Methods (III)

https://ruizhezhang.com/course_fall_2025.html

Slides are based on Sitan Chen's lecture

Recap: under-complete tensor decomposition

Let $T \in \mathbb{R}^{d \times d \times d}$ be a (symmetric) 3-tensor of the following form:

$$T = \sum_{i=1}^{k} \lambda_i u_i \otimes u_i \otimes u_i$$

- Jennrich's algorithm has good theoretical properties (exact recovery, stability, …) as well as some practical concerns (not noise robust in practice, efficiency, …)
- Tensor power method is a more practical approach while also has some theoretical guarantees
- However, there's still a big gap between theory and practice
  - → Theory requires $k \leq d$ (under-complete regime)
  - → Tensor power methods still seem to work for $d < k < d^{1.5}$, at least when $\{u_i\}$ are random

Over-complete tensor decomposition

$$T = \sum_{i=1}^{k} \lambda_i u_i \otimes u_i \otimes u_i$$

Can we find the decomposition of a tensor of rank $k \gg n$ in polynomial time?

A more basic question: when is a rank-k decomposition unique?

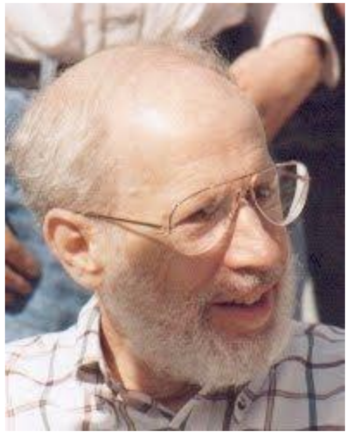
- **Jennrich, Harshman:** when $\{u_i\}$ are linearly independent ($k \leq d$)
- **Kruskal:** if every d columns of U are linearly independent, then the uniqueness holds when

$$k \leq \frac{3}{2}d - 1$$

  However, this result is non-algorithmic

- **Chiantini-Ottaviani:** Uniqueness of 3-tensors of rank $k \leq d^2/3$ generically

all except a measure zero set



Joseph Kruskal (1928-2010)

Over-complete tensor decomposition

From computational complexity perspective,

- It is **NP**-hard to decompose a tensor with rank $k \geq 6d$ in the worst-case
- Constructing an explicit 3-tensor with rank $\Omega(d^{1+\epsilon})$ will imply breakthrough in circuit lower bounds. The best-known rank bound for an explicit 3-tensor is only $3d - O(\log d)$

Today's plan:

- Algorithm for decomposing higher-order tensors

- Beyond worst-case analysis for over-complete tensor decomposition

Higher-order tensors decomposition

Suppose

$$T = \sum_{i=1}^{k} \lambda_i \underbrace{u_i \otimes u_i}_{} \otimes \underbrace{u_i \otimes u_i}_{} \otimes u_i \in \mathbb{R}^{d \times d \times d \times d \times d}$$

$$T = \sum_{i=1}^{k} \lambda_i \mathrm{vec}(u_i \otimes u_i) \otimes \mathrm{vec}(u_i \otimes u_i) \otimes u_i \in \mathbb{R}^{d^2 \times d^2 \times d}$$

Why do higher-order tensors help with decomposition?

Higher-order tensors decomposition

Suppose

$$T = \sum_{i=1}^{k} \lambda_i \underbrace{u_i \otimes u_i} \otimes \underbrace{u_i \otimes u_i} \otimes u_i \in \mathbb{R}^{d\times d\times d\times d\times d}$$

$$\Rightarrow$$

$$T = \sum_{i=1}^{k} \lambda_i \mathrm{vec}(u_i \otimes u_i) \otimes \mathrm{vec}(u_i \otimes u_i) \otimes u_i \in \mathbb{R}^{d^2\times d^2\times d}$$

Observation:

- Jennrich's algorithm requires $\{\mathrm{vec}(u_i \otimes u_i)\}$ are linearly independent
- $\mathrm{vec}(u_i \otimes u_i)$ is a d^2-dimensional vector, so it is possible to handle even $k \sim d^2$

Counterexample

Hope: $\mathrm{vec}(u_i \otimes u_i)$ is a d^2-dimensional vector, so it is possible to handle even $k \sim d^2$

Claim. Let $\{a_i\}_{i\in[d]}$ and $\{b_i\}_{i\in[d]}$ be two sets of orthonormal basis for $\mathbb{R}^d$. Then,

$$\{\mathrm{vec}(a_i \otimes a_i), \mathrm{vec}(b_i \otimes b_i)\}_{i\in[d]}$$

are linearly dependent.

Proof.

Note that

$$\sum_i \mathrm{vec}(a_i \otimes a_i) = \sum_i a_i a_i^\top = I = \sum_i \mathrm{vec}(b_i \otimes b_i)$$

Counterexample

Hope: $\mathrm{vec}(u_i \otimes u_i)$ is a d^2-dimensional vector, so it is possible to handle even $k \sim d^2$

Claim. Let $\{a_i\}_{i\in[d]}$ and $\{b_i\}_{i\in[d]}$ be two sets of orthonormal basis for $\mathbb{R}^d$. Then,

$$\{\mathrm{vec}(a_i \otimes a_i), \mathrm{vec}(b_i \otimes b_i)\}_{i\in[d]}$$

are linearly dependent.

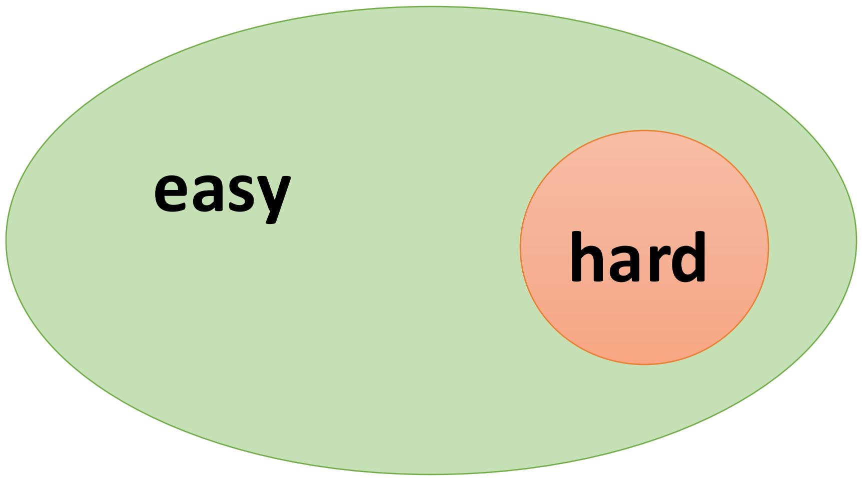
- Dimension does not grow multiplicatively in worst case
- But bad examples are pathological and hard to construct

Beyond worst-case analysis

NP-hardness results for the worst-case instances are too pessimistic

Average-case analysis:

- Showing that for a random instance, the probability that it is hard is small
- Examples: 3-SAT, graph coloring, clique finding, compressed sensing, etc
- Random tensor decomposition (u_i sampled from $\mathbb{S}^{d-1}$)
  - **Chiantini-Ottaviani:** Unique decomposition for rank $k \lesssim d^2$
  - **Ma et al, Ding et al:** Polynomial time algorithms for $k \sim d^{1.5}$



Problem instance space

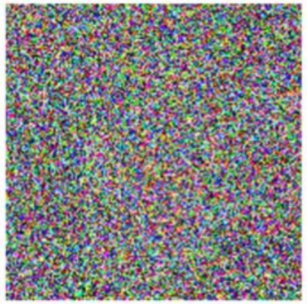
Beyond worst-case analysis

"However, average-case analysis may be unconvincing as the inputs encountered in many application domains may bear little resemblance to the random inputs that dominate the analysis."

(Spielman-Teng, 2003)

Smoothed analysis

- To explain why Simplex algorithm solves LPs efficiently in practice
- Worst-case instances + Random noise perturbation



Worst-case : $\max_x T(x)$

Smoothed analysis: $\max_x \operatorname{avg}_r T(x + \epsilon r)$

Average-case: $\operatorname{avg}_r T(r)$

Beyond worst-case analysis

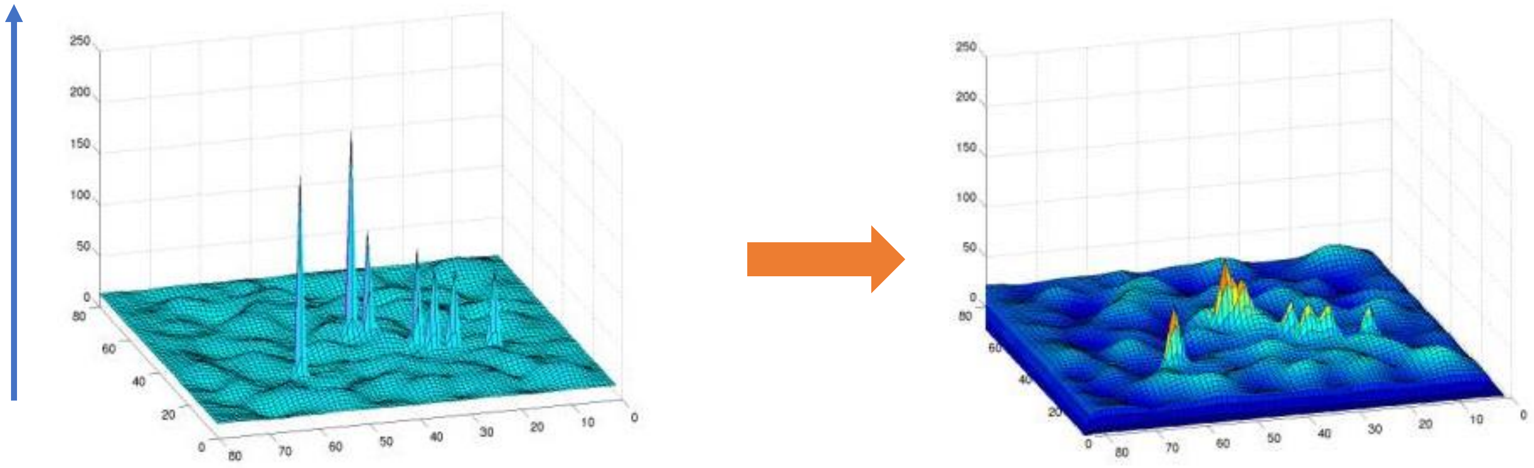
“However, average-case analysis may be unconvincing as the inputs encountered in many application domains may bear little resemblance to the random inputs that dominate the analysis.”

(Spielman-Teng, 2003)

Smoothed analysis

- To explain why Simplex algorithm solves LPs efficiently in practice
- Worst-case instances + Random noise perturbation

hardness



Beyond worst-case analysis

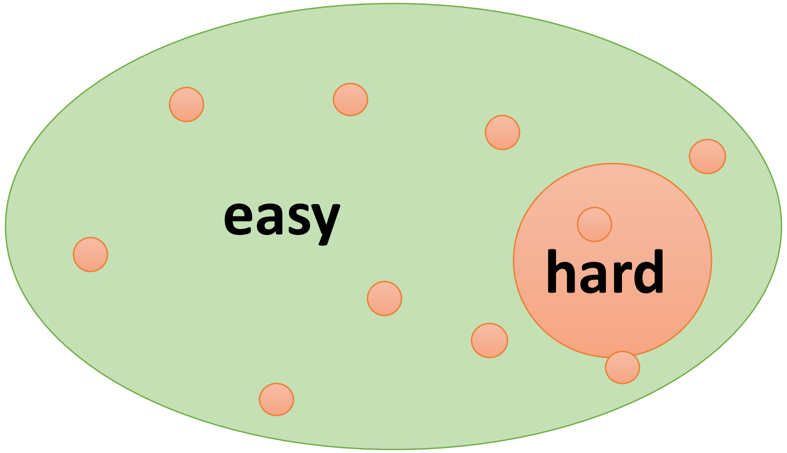
"However, average-case analysis may be unconvincing as the inputs encountered in many application domains may bear little resemblance to the random inputs that dominate the analysis."

(Spielman-Teng, 2003)

Smoothed analysis

- To explain why Simplex algorithm solves LPs efficiently in practice
- Worst-case instances + Random noise perturbation

Problem instance space



Smoothed analysis of tensor decomposition

Smoothed analysis model:

- $\rho > 0$ the smoothing parameter, k the rank, ℓ the order of tensor
- Let $u_i'^{(j)} \in \mathbb{R}^d$ be an arbitrary vector for $i \in [k], j \in [\ell]$ (picked by nature)
- Sample $u_i^{(j)} = u_i'^{(j)} + \frac{\rho}{\sqrt{d}} g_i^{(j)}$ for $g_i^{(j)} \sim \mathcal{N}(0, I)$ ρ-perturbation
- Observe $T = \sum_{i \in [k]} u_i^{(1)} \otimes \cdots \otimes u_i^{(\ell)} + \text{small noise}$

This is different from elements of T being randomly perturbed

- Smoothed model uses $\mathcal{O}(kld)$ bits of randomness, while randomly perturbing T uses $\mathcal{O}\left(d^{\ell}\right)$ bits of randomness
- An efficient algorithm for the element-wise perturbation model would imply a randomized algorithm for worst-case instances — which is considered very unlikely.

Main theorem of this lecture

Theorem (Bhaskara-Charikar-Moitra-Vijayraghavan, 2014).

Let $k \leq d^{\lfloor(\ell-1)/2\rfloor}/2$. There exists an algorithm that takes as input an ℓ-tensor in smoothed analysis model and runs in time $(d/\rho)^{\mathcal{O}(\ell)}$ to recover the decomposition, with probability $1 - 1/\mathrm{superpoly}(d)$ over the randomness of $\left\{g_i^{(j)}\right\}$.

- This result becomes non-trivial when $\ell \geq 5$
- When ρ is small, it is close to a worst-case instance; when ρ is large, it is close to an average-case instance
- The failure probability is important in smoothed analysis, since it essentially describes the fraction of points around any given point that are bad

Using higher order tensors

$$T' = \sum_i \text{vec}(u_i \otimes v_i) \otimes \text{vec}(w_i \otimes x_i) \otimes y_i$$

Stability guarantee of Jennrich's algorithm

Theorem 3.1. *Suppose we are given tensor $\widetilde{T} = T + E \in \mathbb{R}^{m\times n\times p}$, where T has a decomposition $T = \sum_{i=1}^{k} u_i \otimes v_i \otimes w_i$ satisfying the following conditions:*

1. *Matrices $U = (u_i : i \in [k]), V = (v_i : i \in [k])$ have condition number at most κ,*
2. *For all $i \neq j$, $\|\frac{w_i}{\|w_i\|} - \frac{w_j}{\|w_j\|}\|_2 \geq \delta$.*
3. *Each entry of E is bounded by $\|T\|_F \cdot \varepsilon/poly(\kappa, \max\{n, m, p\}, \frac{1}{\delta})$.*

Then the Algorithm 1 on input $\widetilde{T}$ runs in polynomial time and returns a decomposition $\{(\widetilde{u}_i, \widetilde{v}_i, \widetilde{w}_i) : i \in [k]\}$ s.t. there is a permutation $\pi : [k] \to [k]$ with

$$\forall i \in [k], \quad \|\widetilde{u}_i \otimes \widetilde{v}_i \otimes \widetilde{w}_i - u_{\pi(i)} \otimes v_{\pi(i)} \otimes w_{\pi(i)}\|_F \leq \varepsilon\|T\|_F.$$

Using higher order tensors

$$T' = \sum_i \mathrm{vec}(u_i \otimes v_i) \otimes \mathrm{vec}(w_i \otimes x_i) \otimes y_i$$

Stability guarantee of Jennrich's algorithm

- We need to show that $\{\mathrm{vec}(u_i \otimes v_i)\}$ are robustly linearly independent

Using higher order tensors

Khatri-Rao product

- $U, V \in \mathbb{R}^{d \times k}$
- $U \odot V \in \mathbb{R}^{d^2 \times k}$

$$U \odot V := \begin{pmatrix} \mathrm{vec}(u_1 \otimes v_1) & \cdots & \mathrm{vec}(u_k \otimes v_k) \end{pmatrix}$$

Main step

$$\left(\widetilde{U}^{(a)}\right)_{ij} \coloneqq \left(U^{(a)}\right)_{ij} + \frac{\rho}{\sqrt{d}} \cdot \mathcal{N}(0,1)$$

Proposition. Let $k \leq (1-\delta)d^{\ell}$. Given any $U^{(1)}, U^{(2)}, \dots, U^{(\ell)} \in \mathbb{R}^{d \times k}$ then for their random ρ-perturbations, we have

$$\Pr\left[\sigma_k\left(\widetilde{U}^{(1)} \odot \cdots \odot \widetilde{U}^{(\ell)}\right) < (\rho/d)^{\mathcal{O}(\ell)}\right] \leq k \exp\left(-\Omega_{\ell}(d)\right)$$

Theorem (Bhaskara-Charikar-Moitra-Vijayraghavan, 2014).

Let $k \leq d^{\lfloor(\ell-1)/2\rfloor}/2$. There exists an algorithm that takes as input an ℓ-tensor in smoothed analysis model and runs in time $(d/\rho)^{\mathcal{O}(\ell)}$ to recover the decomposition, with probability $1 - 1/\mathrm{superpoly}(d)$ over the randomness of $\{g_{i,j}\}$.

$$T = \sum_i \underbrace{\tilde{u}_i^{(1)} \otimes \cdots \otimes \tilde{u}_i^{((\ell-1)/2)}}_{d^{\lfloor(\ell-1)/2\rfloor} \times k} \otimes \underbrace{\tilde{u}_i^{((\ell-1)/2+1)} \otimes \cdots \otimes \tilde{u}_i^{(\ell-1)}}_{d^{\lfloor(\ell-1)/2\rfloor} \times k} \otimes \tilde{u}_i^{(\ell)}$$

Main step

$$\left(\widetilde{U}^{(a)}\right)_{ij} \coloneqq \left(U^{(a)}\right)_{ij} + \frac{\rho}{\sqrt{d}} \cdot \mathcal{N}(0,1)$$

Proposition. Let $k \leq (1-\delta)d^{\ell}$. Given any $U^{(1)}, U^{(2)}, \ldots, U^{(\ell)} \in \mathbb{R}^{d\times k}$ then for their random ρ-perturbations, we have

$$\Pr\left[\sigma_k\left(\widetilde{U}^{(1)} \odot \cdots \odot \widetilde{U}^{(\ell)}\right) < (\rho/d)^{\mathcal{O}(\ell)}\right] \leq k \exp\left(-\Omega_{\ell}(d)\right)$$

Proof strategy:

- The least singular value can be hard to handle directly
- We can bound leave-one-out distance as an alternative

Leave-one-out distance

Given a matrix $M \in \mathbb{R}^{d \times k}$, the leave-one-out distance of M is

$$\ell(M) = \min_{i \in [k]} \left\| \Pi_{-i}^{\perp} M_i \right\|$$

where $\Pi_{-i}^{\perp}$ is the orthogonal projection to $\operatorname{span}\left(\{M_j : j \neq i\}\right)$

The leave-one-out distance is closely related to the least singular value:

Lemma. For any matrix $M \in \mathbb{R}^{d \times k}$, we have

$$\frac{\ell(M)}{\sqrt{k}} \leq \sigma_{\min}(M) \leq \ell(M)$$

Leave-one-out distance

Lemma. For any matrix $M \in \mathbb{R}^{d\times k}$, we have

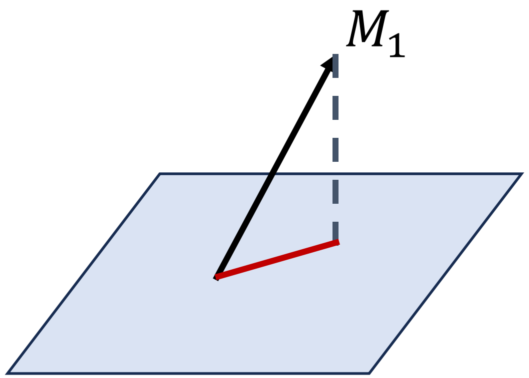
$$\frac{\ell(M)}{\sqrt{k}} \leq \sigma_{\min}(M) \leq \ell(M)$$

Proof.

- Let u be the least singular vector so that $\|Mu\| = \sigma_{\min}(M)$
- Wlog, suppose u_1 is the entry with the largest magnitude, so $|u_1| \geq \frac{1}{\sqrt{k}}$

$$\ell(M) \leq \left\|\Pi_{-1}^{\perp} M_1\right\| = \inf_{v\in \operatorname{span}(M_2,\ldots,M_k)} \|M_1 - v\|$$

$$\leq \left\| M_1 + \sum_{i>1} \frac{u_i}{u_1} M_i \right\| = \frac{1}{|u_1|}\|Mu\| \leq \frac{\sigma_{\min}(M)}{\sqrt{k}}$$



- The lemma is then proved

Main step

Proposition. Let $k \leq (1-\delta)d^{\ell}$. Given any $U^{(1)}, U^{(2)}, \dots, U^{(\ell)} \in \mathbb{R}^{d\times k}$ then for their random ρ-perturbations, we have

$$\Pr\left[\sigma_k\left(\widetilde{U}^{(1)} \odot \cdots \odot \widetilde{U}^{(\ell)}\right) < (\rho/d)^{\mathcal{O}(\ell)}\right] \leq k \exp\left(-\Omega_{\ell}(d)\right)$$

- Using the lemma, it suffices to prove:

$$\Pr\left[\ell\left(\widetilde{U}^{(1)} \odot \cdots \odot \widetilde{U}^{(\ell)}\right) < \sqrt{k} \cdot (\rho/d)^{\mathcal{O}(\ell)}\right] \leq k \exp\left(-\Omega_{\ell}(d)\right)$$

$$\Longrightarrow \quad \Pr\left[\ell\left(\widetilde{U}^{(1)} \odot \cdots \odot \widetilde{U}^{(\ell)}\right) < (\rho/d)^{\ell}\right] \leq k \exp\left(-\Omega_{\ell}(d)\right)$$

- Our goal:

$$\ell\left(\widetilde{U}^{(1)} \odot \cdots \odot \widetilde{U}^{(\ell)}\right) < (\rho/d)^{\ell}$$

- By the definition of the leave-one-out distance, we can consider each column:

$$\left\|\Pi_{-i}^{\perp}\left(\tilde{u}_i^{(1)} \otimes \tilde{u}_i^{(2)} \otimes \cdots \otimes \tilde{u}_i^{(\ell)}\right)\right\| \leq (\rho/d)^{\ell} \quad \forall i \in [k]$$

- Both $\Pi_{-i}^{\perp}$ and $\tilde{u}_i^{(1)} \otimes \tilde{u}_i^{(2)} \otimes \cdots \otimes \tilde{u}_i^{(\ell)}$ are random, but independent!

Projection lemma. Let $W \subset \mathbb{R}^{d^{\times \ell}}$ be an arbitrary subspace of dimension at least δd^{ℓ}. Given any $x_1, \ldots, x_{\ell} \in \mathbb{R}^d$, then their random ρ-perturbations $\tilde{x}_1, \ldots, \tilde{x}_{\ell}$ satisfy

$$\Pr\left[\|\Pi_W(\tilde{x}_1 \otimes \cdots \otimes \tilde{x}_{\ell})\| \leq (\rho/d)^{\ell}\right] \leq \exp\left(-\Omega(d)\right)$$

The proposition follows from Projection Lemma + union bound over all columns

Baby lemma

Projection lemma ($\ell = 1$). Let $W \subset \mathbb{R}^d$ be a subspace of dimension at least δd. If $\tilde{u} = u + \frac{\rho}{\sqrt{d}}\mathcal{N}(0, I)$, then

$$\Pr[\|\Pi_W \tilde{u}\| < \mathcal{O}(\rho/d)] \leq \exp\big(-\Omega(d)\big)$$

Proof (v1).

- Let $w_1, \dots, w_D$ be an orthonormal basis for W
- Then

$$\|\Pi_W \tilde{u}\| = \|(\langle w_1, \tilde{u}\rangle, \dots, \langle w_D, \tilde{u}\rangle)\| \geq \max_{i \in [D]} |\langle w_i, \tilde{u}\rangle|$$

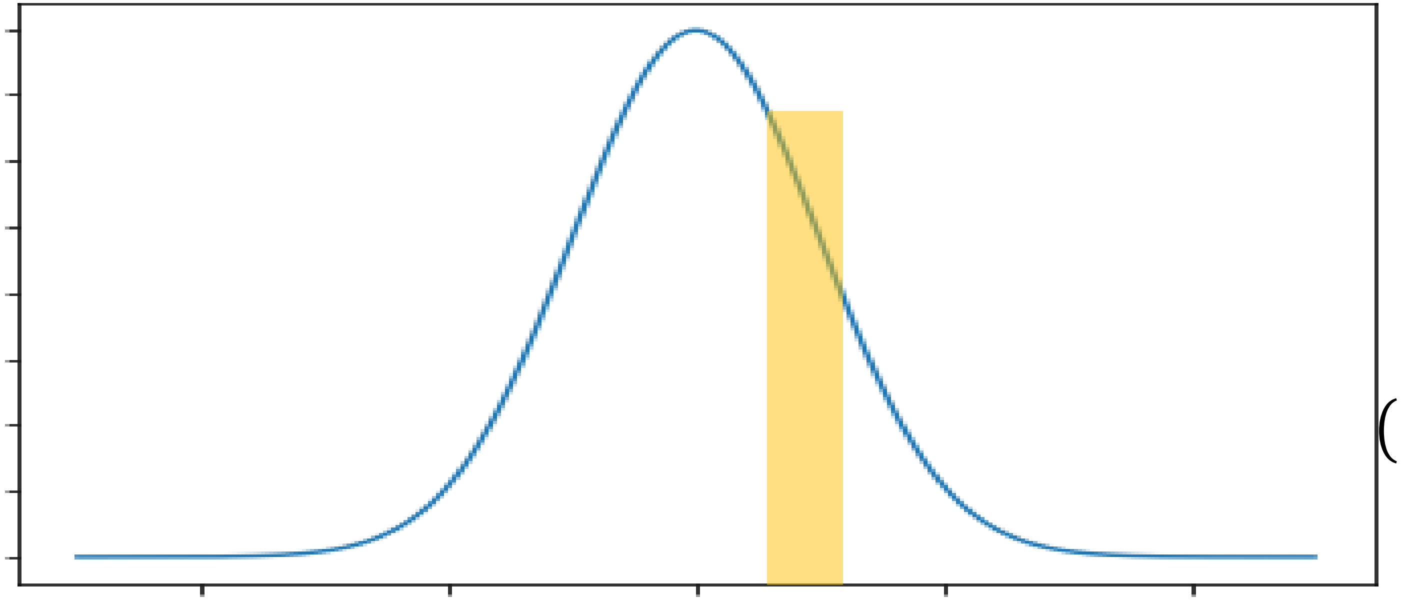
- $\langle w_i, \tilde{u}\rangle = \langle w_i, u\rangle + \frac{\rho}{\sqrt{d}}\mathcal{N}(0,1)$ are independent Gaussians with arbitrary means and variance $\frac{\rho^2}{d}$

Baby lemma

Fact (Gaussian anti-concentration). For $g \sim \mathcal{N}(0,1)$ and for any interval $I \subset \mathbb{R}$ of length t,

$$\Pr[g \in I] \leq O(t)$$

- We have
- Thus,



$$\Pr\left[\max_{i \in [D]} \quad \right. \qquad \left(-\Omega(d)\right)$$

However, this approach does not generalize to higher order case.

Baby lemma

Lemma ($\ell = 1$). Let $W \subset \mathbb{R}^d$ be a subspace of dimension at least δd. If $\tilde{u} = u + \frac{\rho}{\sqrt{d}}\mathcal{N}(0, I)$, then

$$\Pr[\|\Pi_W \tilde{u}\| < \mathcal{O}(\rho/d)] \leq \exp\big(-\Omega(d)\big)$$

Proof (v2).

$$\|\Pi_W \tilde{u}\| = \sup_{w \in W : \|w\| = 1} |\langle w, \tilde{u} \rangle|$$

- Instead of choosing an orthonormal basis, we choose a "row echelon" basis for W:
  - All $|\star|$'s ≤ 1
  - $\|w_i\| \leq \sqrt{d}$
  - Construction is similar to Gaussian elimination (permuting the coordinates if needed)

$$\left.\begin{aligned} w_1 &= [\ 1 \quad \star \quad \star \quad \star \quad \star\ ] \\ w_2 &= [\ 0 \quad 1 \quad \star \quad \star \quad \star\ ] \\ w_3 &= [\ 0 \quad 0 \quad 1 \quad \star \quad \star\ ] \\ w_4 &= [\ 0 \quad 0 \quad 0 \quad 1 \quad \star\ ] \end{aligned}\right\} \in W$$

Baby lemma

$$\left.\begin{array}{l} w_1 = [\ 1 \quad \star \quad \star \quad \star \quad \star\ ] \\ w_2 = [\ 0 \quad 1 \quad \star \quad \star \quad \star\ ] \\ w_3 = [\ 0 \quad 0 \quad 1 \quad \star \quad \star\ ] \\ w_4 = [\ 0 \quad 0 \quad 0 \quad 1 \quad \star\ ] \end{array}\right\} \in W$$

- Each w_i has a non-negligible component orthogonal to the span of $w_{i+1}, \dots, w_D$
- We will "reveal" $\langle w_D, \tilde{u}\rangle, \langle w_{D-1}, \tilde{u}\rangle, \dots, \langle w_1, \tilde{u}\rangle$ one at a time
- $\langle w_i, \tilde{u}\rangle = \langle w_i, u\rangle + \frac{\rho}{\sqrt{d}}\left(g_i + \sum_{j>i}(w_i)_j g_j\right)$

"left-over" randomness

$$\Pr\left[|\langle w_i, \tilde{u}\rangle| < \mathcal{O}\left(\frac{\rho}{\sqrt{d}}\right) \,\middle|\, \langle w_{i+1}, \tilde{u}\rangle, \dots, \langle w_D, \tilde{u}\rangle\right] \le \sup_{t\in\mathbb{R}} \Pr_{g_i \sim \mathcal{N}(0,1)}\left[\frac{\rho}{\sqrt{d}}|g_i - t| \le \mathcal{O}\left(\frac{\rho}{\sqrt{d}}\right)\right] = \mathcal{O}(1)$$

- Hence,

$$\Pr\left[|\langle w_i, \tilde{u}\rangle| < \mathcal{O}(\rho/\sqrt{d})\ \forall i \in [D]\right] = \exp(-\Omega(d))$$

- Since $\|w_i\| \le \sqrt{d}$, we get that

$$\Pr[\|\Pi_W \tilde{u}\| \le \mathcal{O}(\rho/d)] = \exp(-\Omega(d))$$

We loss a factor of $\sqrt{d}$, but this approach can generalize to $\ell > 1$

General case

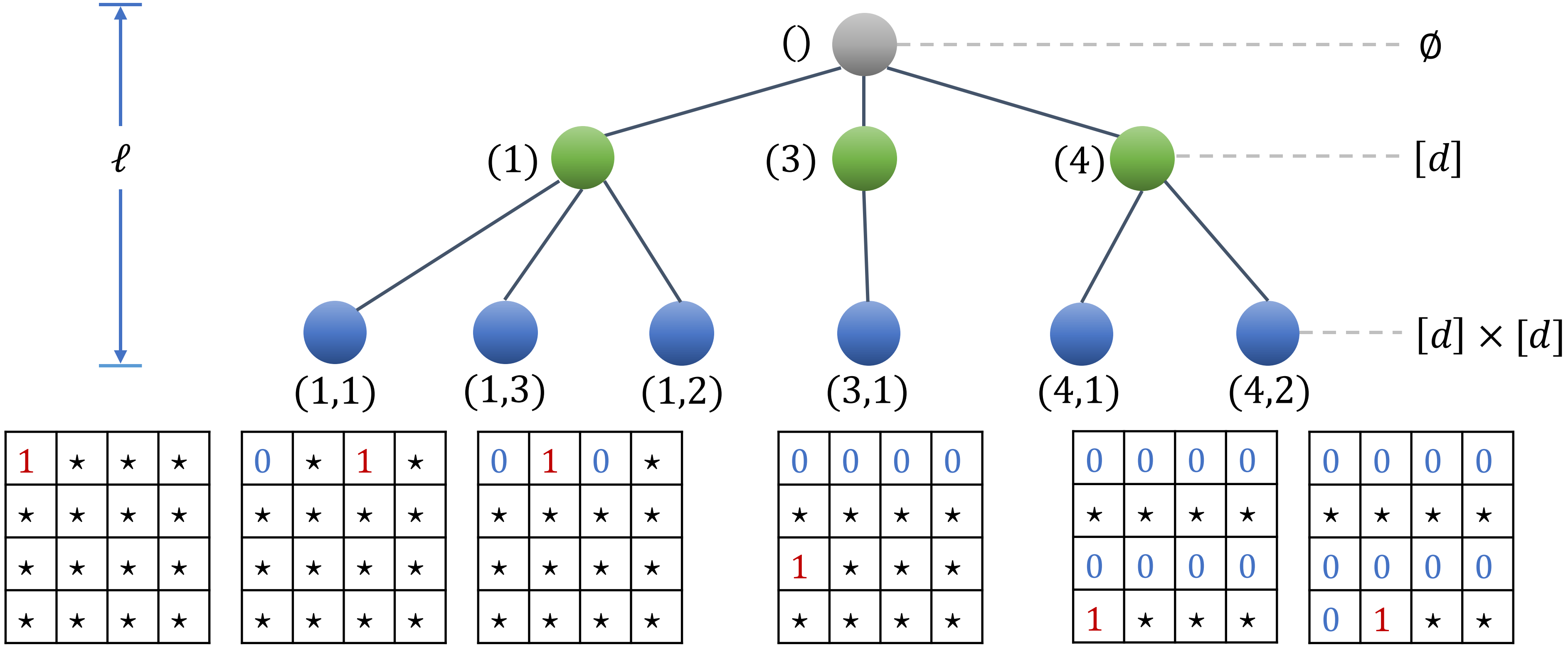
Projection lemma. Let $W \subset \mathbb{R}^{d^{\times \ell}}$ be an arbitrary subspace of dimension at least δd^{ℓ}. Given any $x_1, \ldots, x_\ell \in \mathbb{R}^d$, then their random ρ-perturbations $\tilde{x}_1, \ldots, \tilde{x}_\ell$ satisfy

$$\Pr\left[\|\Pi_W(\tilde{x}_1 \otimes \cdots \otimes \tilde{x}_\ell)\| \leq (\rho/d)^\ell\right] \leq \exp\left(-\Omega(d)\right)$$

Proof strategy:

- We need to construct tensor version of "row echelon" basis $\{T_I\}$ for W
- Show that $|T_I(\tilde{x}_1, \cdots, \tilde{x}_\ell)|$ is large with high probability

Index tree



post-traversal ordering: $(1,1) \prec (1,3) \prec (1,2) \prec (1) \prec (3,1) \prec (3) \prec (4,1) \prec (4,2) \prec (4)$

An echelon tree for W is an index tree where each leaf I is additionally labeled by an element $T_I \in W$ such that

- $(T_I)_{I_1,\ldots,I_\ell} = 1$
- For every $J \prec I$, $T_I(e_J,:) = 0$
- All $|\star|$'s ≤ 1

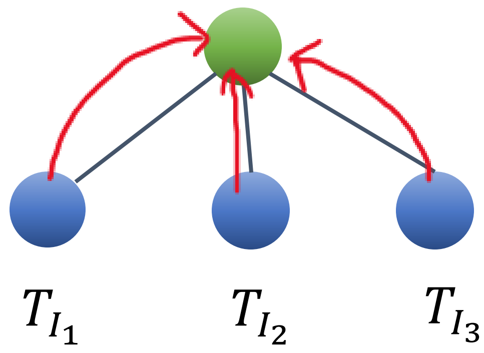
$T_I(e_{j_1}, \ldots, e_{j_{|J|}}, :, \ldots, :)$

Proof of the projection lemma via echelon tree

Claim (Echelon tree construction). Let $W \subset \mathbb{R}^{d^{\times \ell}}$ be a subspace of dimension at least δd^{ℓ}. Then, there exists an echelon tree for W such that every non-leaf node has at least $\frac{\delta}{2\ell} d$ children.

We'll show that this claim implies the projection lemma for a general $\ell > 1$

- $\|T_I\|_F \leq d^{\ell/2}$ for every leaf I. So it suffices to show that $|T_I(\tilde{x}_1, \dots, \tilde{x}_\ell)| \geq \left(\rho/\sqrt{d}\right)^{\ell}$ for some I w.h.p.
- We will fix $\tilde{x}_\ell, \tilde{x}_{\ell-1}, \dots, \tilde{x}_1$ one at a time, and simultaneously reduce the height of the tree by 1



$$T_J \coloneqq \arg\max \left|T(e_J)\right|$$

$$T = T_I(:, \tilde{x}_\ell)$$

Proof of the projection lemma via echelon tree

We say an echelon tree is x-large if $|T_I(e_I)| \geq x$ for every leaf I

Claim. If we start with an x-large echelon tree, then the next echelon tree is $\frac{\rho}{\sqrt{d}}x$-large

- For a fixed node J of level $\ell - 1$, we want to prove that there exists a child node I such that

$$\left|T_I(e_J, \tilde{x}_\ell)\right| \geq \frac{\rho}{\sqrt{d}} x$$

- By the previous claim, J has $m \geq \frac{\delta}{2\ell} d$ children, with labels:

$$(J, i_1), (J, i_2), \dots, (J, i_m)$$

- Then, it is almost the same as baby lemma for $\ell = 1$ and $d' = m$. Thus, we have

$$\Pr\left[\forall j \in [m]: \left|T_{I_j}(e_J, \tilde{x}_\ell)\right| \leq \rho x/\sqrt{d}\right] \leq \exp\left(-\Omega(m)\right) = \exp\left(-\Omega(d)\right)$$

- There are at most $d^{(\ell-1)}$ nodes at level $\ell - 1$. By union bound, w.p. $\geq 1 - d^{\ell-1}\exp\left(-\Omega(d)\right)$, the next echelon tree is $\frac{\rho}{\sqrt{d}}x$-large, and the claim is proved

Proof of the projection lemma via echelon tree

Claim. If we start with an x-large echelon tree, then the next echelon tree is $\frac{\rho}{\sqrt{d}}x$-large

Inductively, this implies that with probability at least

$$1 - \left(1 + d + \cdots + d^{\ell-1}\right)\exp\bigl(-\Omega(d)\bigr) \geq 1 - d^{\ell}\exp\bigl(-\Omega(d)\bigr),$$

there exists some I in the echelon tree (which is also in W) such that

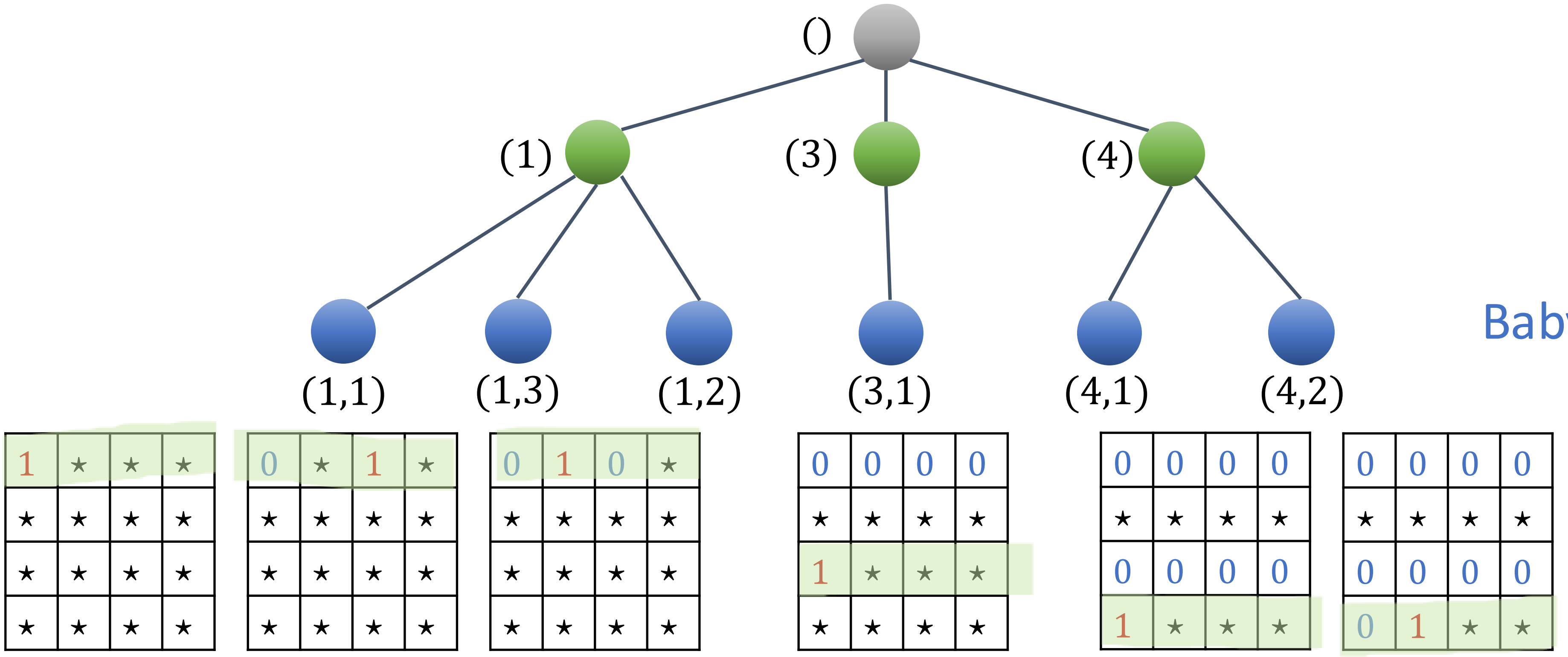
$$|T_I(\tilde{x}_1, \dots, \tilde{x}_\ell)| \geq \left(\rho/\sqrt{d}\right)^{\ell}$$

Hence,

$$\Pr\left[\|\Pi_W(\tilde{x}_1 \otimes \cdots \otimes \tilde{x}_\ell)\| \leq (\rho/d)^{\ell}\right] \leq \exp\bigl(-\Omega(d)\bigr)$$

which proves the projection lemma

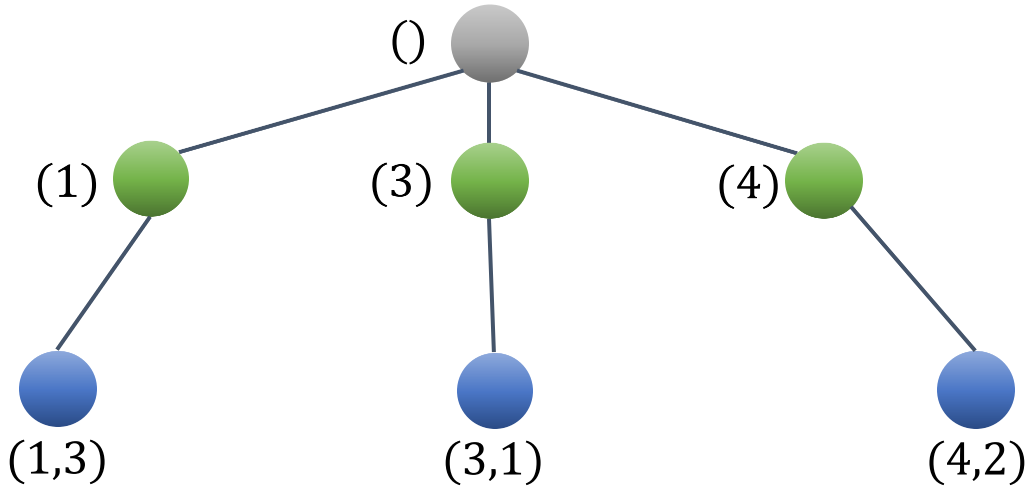
$\ell = 2$ case



Baby lemma × 3

$|e_1^\top T_I \tilde{x}_2|$ $\quad$ $|e_3^\top T_I \tilde{x}_2|$ $\quad$ $|e_4^\top T_I \tilde{x}_2|$

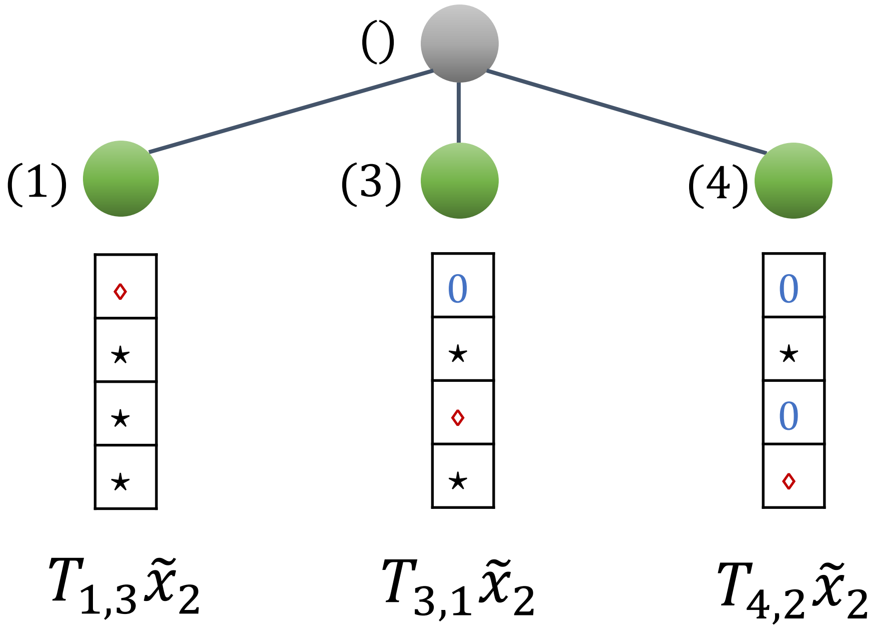
$\ell = 2$ case



$$\begin{array}{|c|c|c|c|}\hline 0 & \star & 1 & \star \\ \hline \star & \star & \star & \star \\ \hline \star & \star & \star & \star \\ \hline \star & \star & \star & \star \\ \hline \end{array} \qquad \begin{array}{|c|c|c|c|}\hline 0 & 0 & 0 & 0 \\ \hline \star & \star & \star & \star \\ \hline 1 & \star & \star & \star \\ \hline \star & \star & \star & \star \\ \hline \end{array} \qquad \begin{array}{|c|c|c|c|}\hline 0 & 0 & 0 & 0 \\ \hline \star & \star & \star & \star \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 1 & \star & \star \\ \hline \end{array}$$

$$\left|e_1^\top T_{1,3}\tilde{x}_2\right| \qquad \left|e_3^\top T_{3,1}\tilde{x}_2\right| \qquad \left|e_4^\top T_{4,2}\tilde{x}_2\right| \quad \geq \frac{\rho}{\sqrt{d}}$$

$\ell = 2$ case



Baby lemma: $$\left|\tilde{x}_1^\top T_{3,1}\tilde{x}_2\right| \geq \frac{\rho}{\sqrt{d}} \cdot \frac{\rho}{\sqrt{d}} = \frac{\rho^2}{d}$$

Constructing the echelon tree

Echelon tree construction

Claim (Echelon tree construction). Let $W \subset \mathbb{R}^{d^{\times \ell}}$ be a subspace of dimension at least δd^{ℓ}. Then, there exists an echelon tree for W such that every non-leaf node has at least $\frac{\delta}{2\ell} d$ children.

Claim (stronger version). Let $W \subset \mathbb{R}^{d_1 \times \cdots \times d_\ell}$ be a subspace. Let $\alpha \in (0,1]$ be such that

$$(1-\alpha)^{\ell} \geq 1 - \dim(W)/d_1 \cdots d_\ell$$

Then, there exists an echelon tree for W such that every node at level i has at least αn_i children.

Claim (stronger version). Let $W \subset \mathbb{R}^{d_1 \times \cdots \times d_\ell}$ be a subspace. Let $\alpha \in (0,1]$ be such that

$$(1-\alpha)^\ell \geq 1 - \dim(W)/d_1 \cdots d_\ell$$

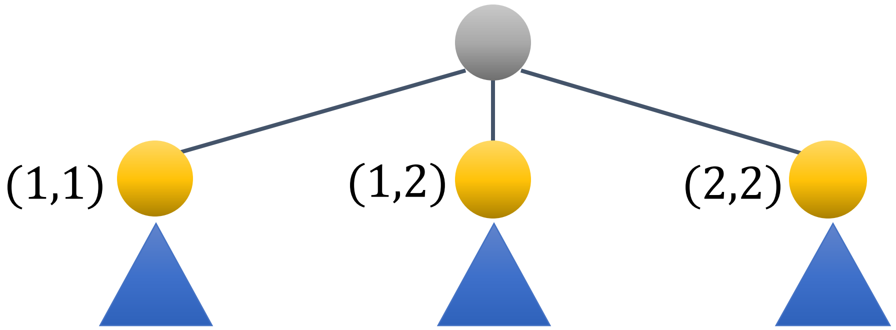
Then, there exists an echelon tree for W such that every node at level i has at least αn_i children.

Proof by induction over the height ℓ

- $\ell = 1$ is trivial (as proved in the baby lemma)
- Suppose $\ell - 1$ holds. To "grow" one more level, we first flatten the first two dimensions, i.e.,

$$W \cong W' \subset \mathbb{R}^{\underbrace{d_1 d_2 \times d_3 \times \cdots \times d_\ell}_{\ell - 1 \text{ levels}}}$$

- Level-1 has $\geq \alpha d_1 d_2$ nodes (by induction hypothesis)



$d = 2$

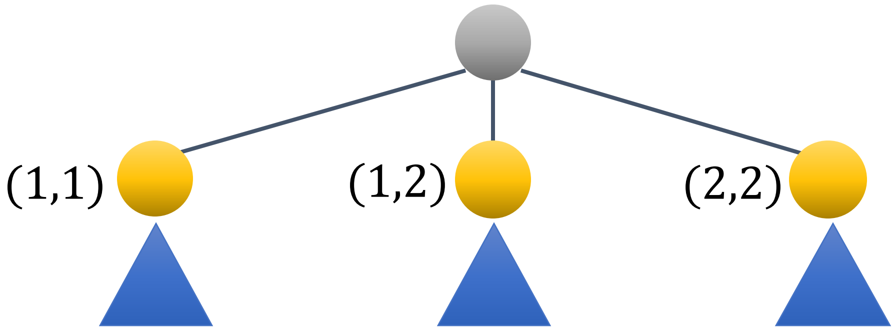
Claim (stronger version). Let $W \subset \mathbb{R}^{d_1 \times \cdots \times d_\ell}$ be a subspace. Let $\alpha \in (0,1]$ be such that

$$(1-\alpha)^\ell \geq 1 - \dim(W)/d_1 \cdots d_\ell$$

Then, there exists an echelon tree for W such that every node at level i has at least αn_i children.

Proof by induction over the height ℓ

- $\ell = 1$ is trivial (as proved in the baby lemma)
- Suppose $\ell - 1$ holds. To "grow" one more level, we first flatten the first two dimensions, i.e.,
$$W \cong W' \subset \mathbb{R}^{d_1 d_2 \times d_3 \times \cdots \times d_\ell}$$
- Level-1 has $\geq \alpha d_1 d_2$ nodes (by induction hypothesis).
At least αd_2 of them has the same first coordinate (by pigeonhole principle)



$d = 2$

Claim (stronger version). Let $W \subset \mathbb{R}^{d_1 \times \cdots \times d_\ell}$ be a subspace. Let $\alpha \in (0,1]$ be such that

$$(1-\alpha)^\ell \geq 1 - \dim(W)/d_1 \cdots d_\ell$$

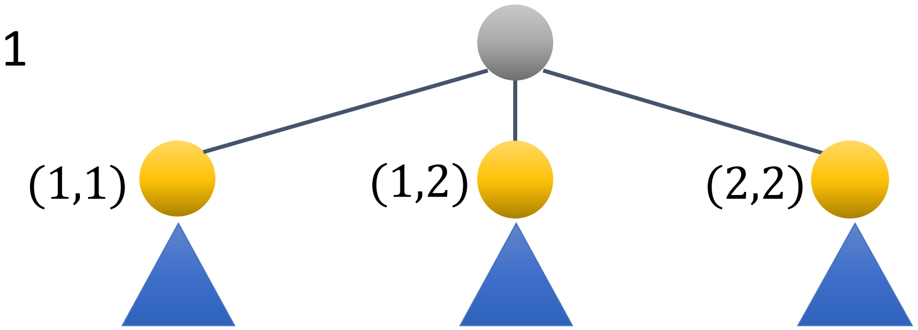
Then, there exists an echelon tree for W such that every node at level i has at least αn_i children.

Proof by induction over the height ℓ

- $\ell = 1$ is trivial (as proved in the baby lemma)
- Suppose $\ell - 1$ holds. To "grow" one more level, we first flatten the first two dimensions, i.e.,

$$W \cong W' \subset \mathbb{R}^{d_1 d_2 \times d_3 \times \cdots \times d_\ell}$$

- Level-1 has $\geq \alpha d_1 d_2$ nodes (by induction hypothesis).
  At least αd_2 of them has the same first coordinate (by pigeonhole principle)
- Remove the other nodes at level-1



$d = 2$

Claim (stronger version). Let $W \subset \mathbb{R}^{d_1 \times \cdots \times d_\ell}$ be a subspace. Let $\alpha \in (0,1]$ be such that

$$(1-\alpha)^\ell \geq 1 - \dim(W)/d_1 \cdots d_\ell$$

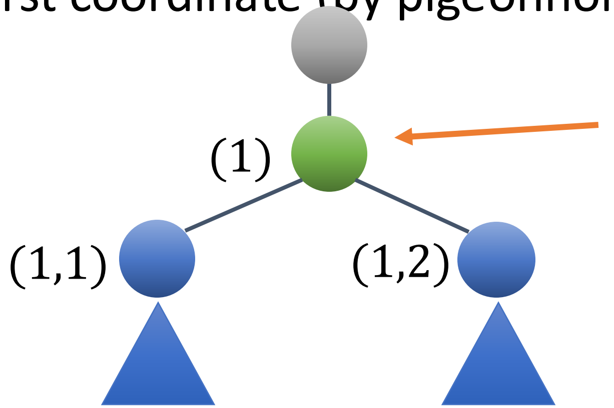
Then, there exists an echelon tree for W such that every node at level i has at least αn_i children.

Proof by induction over the height ℓ

- $\ell = 1$ is trivial (as proved in the baby lemma)
- Suppose $\ell - 1$ holds. To "grow" one more level, we first flatten the first two dimensions, i.e.,

$$W \cong W' \subset \mathbb{R}^{d_1 d_2 \times d_3 \times \cdots \times d_\ell}$$

- Level-1 has $\geq \alpha d_1 d_2$ nodes (by induction hypothesis).
  At least αd_2 of them has the same first coordinate (by pigeonhole principle)
- Remove the other nodes at level-1
- Extract the first coordinate



This node has $\geq \alpha d_2$ children 👍

$d = 2$

Claim (stronger version). Let $W \subset \mathbb{R}^{d_1 \times \cdots \times d_\ell}$ be a subspace. Let $\alpha \in (0,1]$ be such that

$$(1-\alpha)^\ell \geq 1 - \dim(W)/d_1 \cdots d_\ell$$

Then, there exists an echelon tree for W such that every node at level i has at least αn_i children.

- Consider the subspace

$$W_1 \coloneqq \{T \in W \,|\, T(e_{i_1}, :, \ldots, :) = 0\} \cong W_1' \subset \mathbb{R}^{(d_1-1)d_2 \times d_3 \times \cdots \times d_\ell}$$

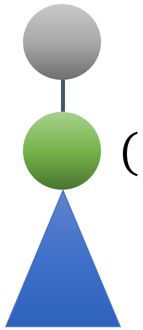
$d_2 \cdots d_\ell$ constraints

- $\dim(W_1) = \dim(W) - d_2 \cdots d_\ell$ and

$$1 - \frac{\dim(W_1)}{(d_1-1)d_2 \cdots d_\ell} = 1 - \frac{\dim(W) - d_2 \cdots d_\ell}{(d_1-1)d_2 \cdots d_\ell}$$

$$= \frac{1 - \dim(W)/d_1 \cdots d_\ell}{1 - 1/d_1} \leq \frac{(1-\alpha)^\ell}{1-1/d_1} \leq (1-\alpha)^{\ell-1}$$

- By induction hypothesis, there is an echelon tree for W_1



(i_1)

Claim (stronger version). Let $W \subset \mathbb{R}^{d_1 \times \cdots \times d_\ell}$ be a subspace. Let $\alpha \in (0,1]$ be such that

$$(1-\alpha)^\ell \geq 1 - \dim(W)/d_1 \cdots d_\ell$$

Then, there exists an echelon tree for W such that every node at level i has at least αn_i children.

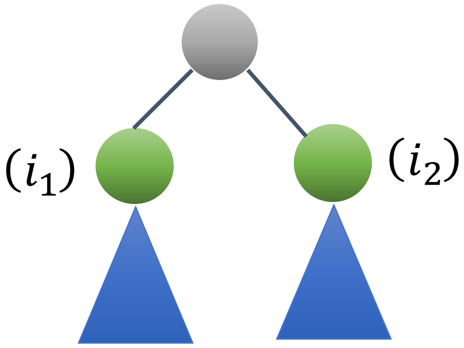
- Consider the subspace

$$W_1 \coloneqq \left\{ T \in W \middle| T\left(e_{i_1}, :, \dots, :\right) = 0 \right\} \cong W_1' \subset \mathbb{R}^{(d_1-1)d_2 \times d_3 \times \cdots \times d_\ell}$$

- $\dim(W_1) = \dim(W) - d_2 \cdots d_\ell$ and

$$1 - \frac{\dim(W_1)}{(d_1-1)d_2 \cdots d_\ell} = 1 - \frac{\dim(W) - d_2 \cdots d_\ell}{(d_1-1)d_2 \cdots d_\ell}$$
$$= \frac{1 - \dim(W)/d_1 \cdots d_\ell}{1 - 1/d_1} \leq \frac{(1-\alpha)^\ell}{1-1/d_1} \leq (1-\alpha)^{\ell-1}$$

- By induction hypothesis, there is an echelon tree for W_1
- Repeating the previous argument, we obtain the second subtree:



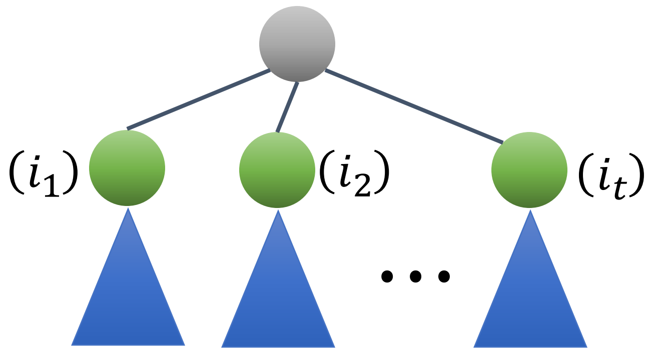
Claim (stronger version). Let $W \subset \mathbb{R}^{d_1 \times \cdots \times d_\ell}$ be a subspace. Let $\alpha \in (0,1]$ be such that

$$(1-\alpha)^\ell \geq 1 - \dim(W)/d_1 \cdots d_\ell$$

Then, there exists an echelon tree for W such that every node at level i has at least αn_i children.

- Suppose we apply this procedure for t times
- The subspace becomes

$$W_{t+1} \coloneqq \left\{T \in W \middle| T\left(e_{i_j}, :, \ldots, :\right) = 0 \; \forall j \in [t]\right\}$$
$$\cong W'_{t+1} \subset \mathbb{R}^{(d_1 - t)d_2 \times d_3 \times \cdots \times d_\ell}$$



- And we can check the condition:

$$1 - \frac{\dim(W_{t+1})}{(d_1 - t)d_2 \cdots d_\ell} = \frac{1 - \dim(W)/d_1 \cdots d_\ell}{1 - t/d_1} \leq \frac{(1-\alpha)^\ell}{1 - t/d_1} \leq (1-\alpha)^{\ell-1}$$

- Hence, we can add new subtrees until $t > \alpha d_1$. Then, the root has αd_1 children and it is an echelon tree of height ℓ 👍