**Title:** Tensorization of neural networks for improved privacy and interpretability

**Speakers:** José Ramón Pareja Monturiol

**Collection/Series:** Machine Learning Initiative

**Subject:** Other

**Date:** February 07, 2025 - 2:30 PM

**URL:** https://pirsa.org/25020035


**Abstract:**

We present a tensorization algorithm for constructing tensor train representations of functions, drawing on sketching and cross interpolation ideas. The method only requires black-box access to the target function and a small set of sample points defining the domain of interest. Thus, it is particularly well-suited for machine learning models, where the domain of interest is naturally defined by the training dataset. We show that this approach can be used to enhance the privacy and interpretability of neural network models. Specifically, we apply our decomposition to (i) obfuscate neural networks whose parameters encode patterns tied to the training data distribution, and (ii) estimate topological phases of matter that are easily accessible from the tensor train representation. Additionally, we show that this tensorization can serve as an efficient initialization method for optimizing tensor trains in general settings, and that, for model compression, our algorithm achieves a superior trade-off between memory and time complexity compared to conventional tensorization methods of neural networks.

# Tensorization of neural networks for improved privacy and interpretability

**José Ramón Pareja Monturiol** (UCM - ICMAT)
Alejandro Pozas-Kerstjens (UNIGE)
David Pérez-García (UCM - ICMAT)

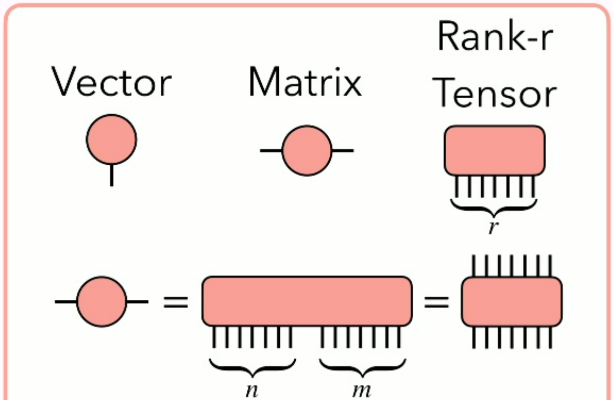
PIQuIL seminar
Feb. 7, 2025

arxiv:2501.06300

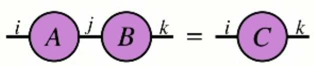
**Outline:**

1. Tensor Networks
2. TNs for Machine Learning
3. Privacy with TNs
4. Tensorization (TT-RSS)
5. Applications:
    - Privacy
    - Interpretability

# Tensor Networks

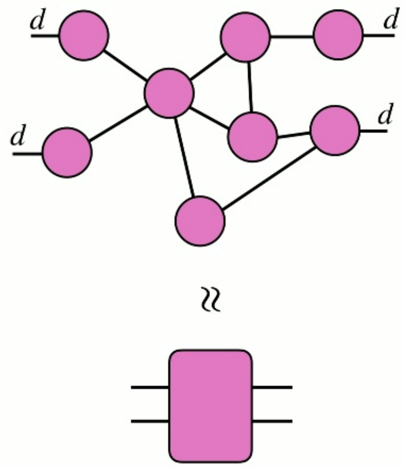



$$\mathbb{C}^{d^n} \to \mathbb{C}^{d^m}$$

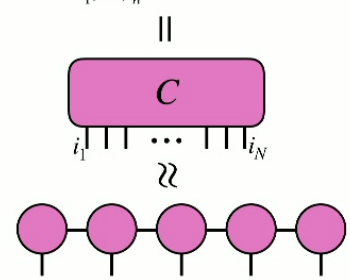



$$\sum_{i,j,k} A_{ij}B_{jk}\,|i\rangle\langle j|j\rangle\langle k| = \sum_{ij} C_{ik}\,|i\rangle\langle k|$$

$$A_{ij}B_{jk} = C_{ik}$$

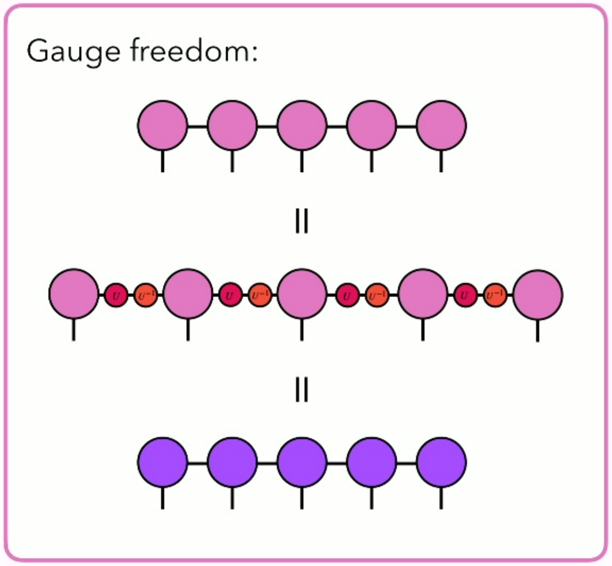


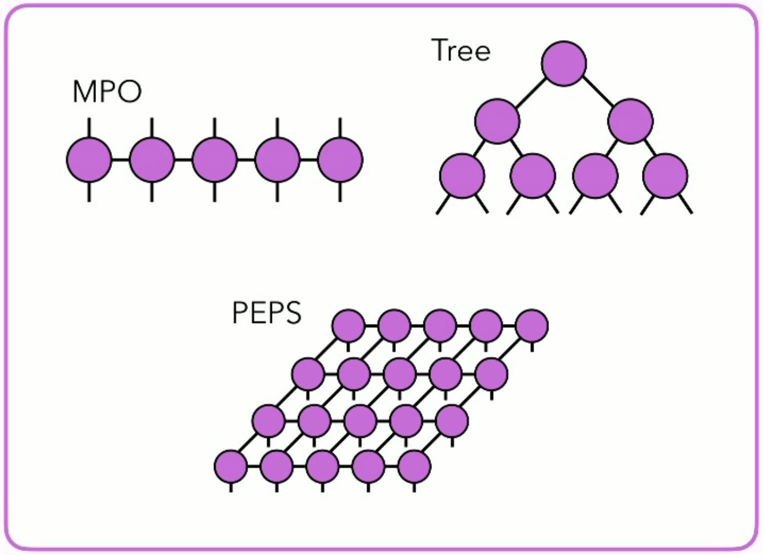

$$|\Psi\rangle = \sum_{i_1,\ldots,i_n} C_{i_1,\ldots,i_N}|i_1\rangle\ldots|i_N\rangle$$




Matrix Product State / Tensor Train

# Tensor Networks



# TNs for Machine Learning

## Problem

| Date collected | Plot | Species | Sex | Weight |
|---|---|---|---|---|
| 1/9/78 | 1 | DM | M | 40 |
| 1/9/78 | 1 | DM | F | 36 |
| 1/9/78 | 1 | DS | F | 135 |
| 1/20/78 | 1 | DM | F | 39 |
| 1/20/78 | 2 | DM | M | 43 |
| 1/20/78 | 2 | DS | F | 144 |
| 3/13/78 | 2 | DM | F | 51 |
| 3/13/78 | 2 | DM | F | 44 |
| 3/13/78 | 2 | DS | F | 146 |

```
def __init__(self,
             num: int,
             name: Text,
             node: Optional['AbstractNode'] = None,
             node1: bool = True) -> None:

    # Check types
    if not isinstance(num, int):
        raise TypeError('`num` should be int type')

    if not isinstance(name, str):
        raise TypeError('`name` should be str type')

    if node is not None:
        if not isinstance(node, AbstractNode):
            raise TypeError('`node` should be AbstractNode type')
```

## Machine

- Linear Regression
- Decision Tree
- Support Vector Machine
- Clustering
- Neural Network

⋮

## Learning

- Define loss function:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \|y_i - f_\theta(x_i)\|^2$$

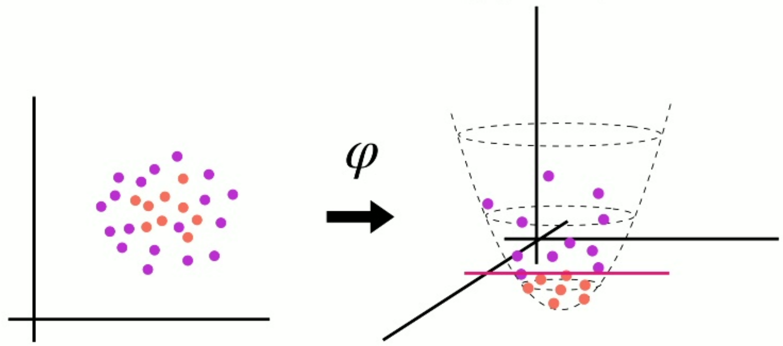
$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(f_\theta(x_i))$$

- Minimize:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \mathcal{L}(\theta_t)$$

# TNs for Machine Learning

Embed data into bigger space

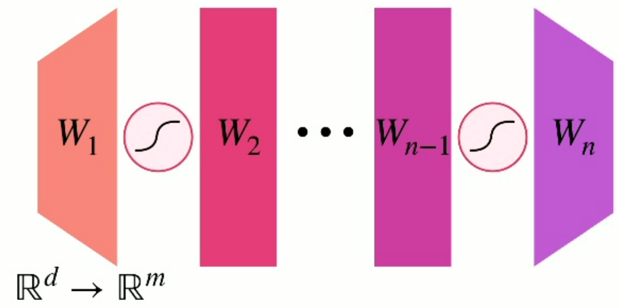



## Examples

- Kernel Support Vector Machine

$$\langle w, \varphi(x) \rangle = \sum_i \alpha_i y_i k(x_i, x)$$

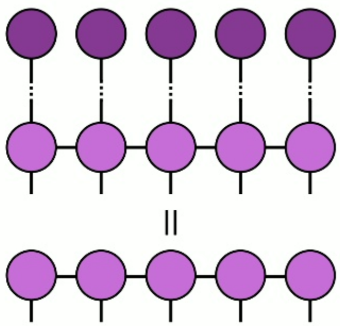
- Neural Networks



# TNs for Machine Learning

$[x_1, \ldots, x_N]$

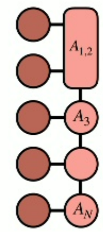
$\varphi \downarrow$ Embed into tensor product space



**Supervised Learning with Quantum-Inspired Tensor Networks** (Stoudenmire and Schwab, 2016)

$$\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\varphi} \quad \varphi(x) = \phi(x_1) \otimes \cdots \otimes \phi(x_N)$$

$$\phi(x_i) = \left[ cos\left(\frac{\pi}{2} x_i\right), sin\left(\frac{\pi}{2} x_i\right) \right]$$



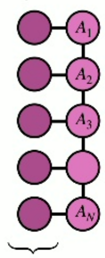

DMRG-like training

$$A_{1,2}^{t+1} = A_{1,2}^{t} - \eta \nabla_t \mathscr{L}$$

**Exponential machines** (Novikov et al., 2016)

$$\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\varphi} \quad \varphi(x) = \phi(x_1) \otimes \cdots \otimes \phi(x_N)$$

$$\phi(x_i) = [1, x_i]$$




Train all sites

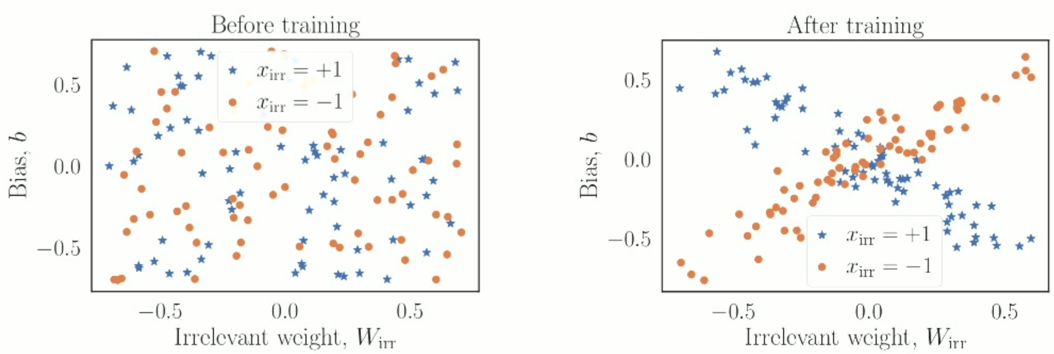
$$A_i^{t+1} = A_i^{t} - \eta \nabla_t \mathscr{L}$$

$[1, x_1, \cdots, x_N, x_1 x_2, \cdots, x_{N-1} x_N]$

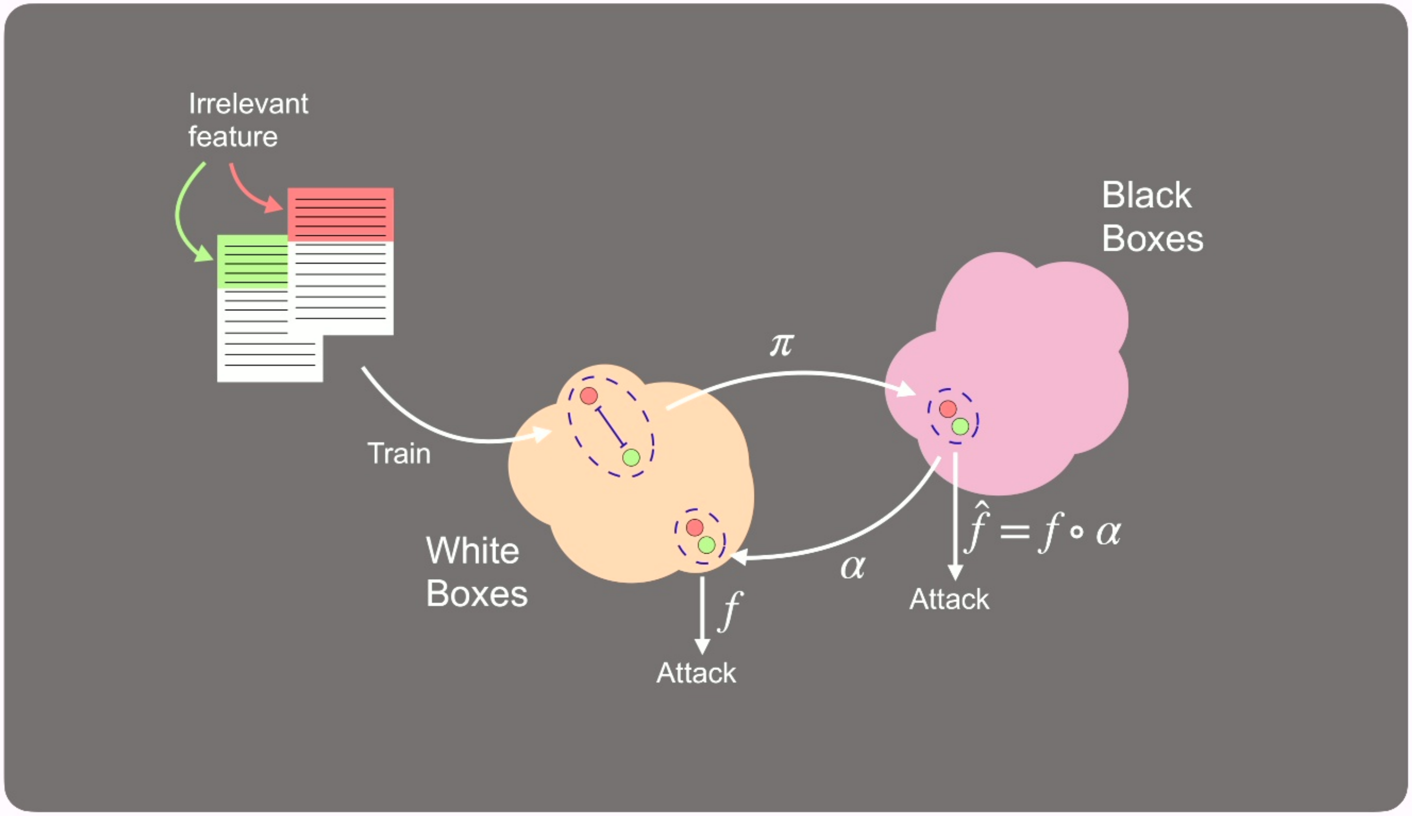
# Privacy with TNs

**White-box privacy vulnerability:** arXiv:2202.12319

- Approximate function $f(x_{rel}, x_{irr}) = sign(x_{rel})$
- With model $NN(x_{rel}, x_{irr}) = \phi(W_{rel}x_{rel} + W_{irr}x_{irr} + b)$
- $\mathscr{L}$ loss function: $\partial_{W_{irr}}\mathscr{L} = x_{irr}\phi'\partial_\phi\mathscr{L}$ and $\partial_b\mathscr{L} = \phi'\partial_\phi\mathscr{L}$, implying that $\partial_{W_{irr}}\mathscr{L} = x_{irr}\partial_b\mathscr{L}$



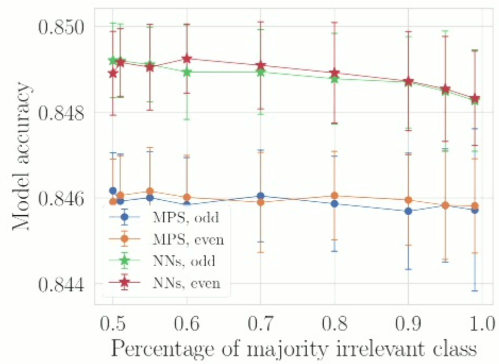
# Privacy with TNs

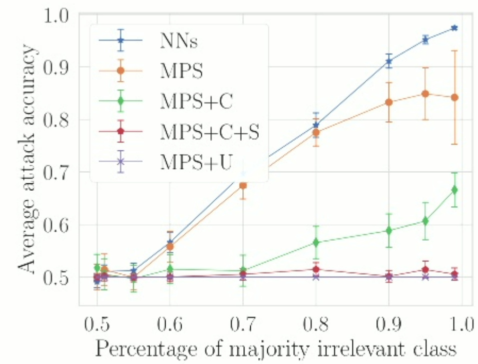


# Privacy with TNs

**Experiment:** arXiv:2202.12319

- **Target:** Predict outcome of COVID-19 cases given demographics and symptoms.
- **Irrelevant feature:** Parity of the day of registration of the record.
- **Attack goal:** Extract the majority value of the irrelevant feature.



# Privacy with TNs

**Next step: go bigger**

- Bigger networks (Trees, PEPS, NNs+TNs)
- Bigger datasets (more dimensions: images, audio, text, etc.)

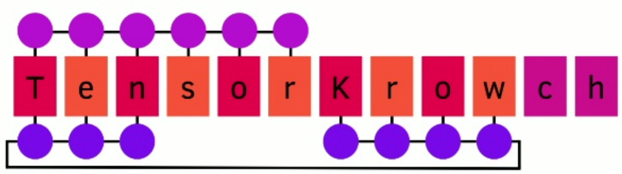
**But... many variables in TNs:**

- Topology of the network
- Initialization of tensors
- Embeddings
- Optimization routines
- Appropriate hyperparameters
- ...

# Privacy with TNs

**Next step: go bigger**

- Bigger networks (Trees, PEPS, NNs+TNs)
- Bigger datasets (more dimensions: images, audio, text, etc.)



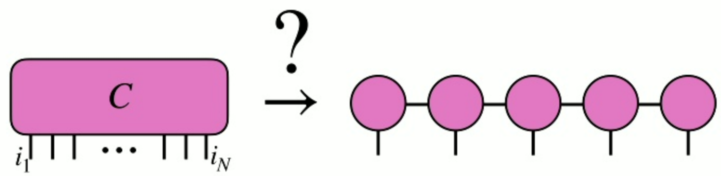

arxiv:2306.08595
https://github.com/joserapa98/tensorkrowch

**But… many variables in TNs:**

- Topology of the network
- Initialization of tensors
- Embeddings
- Optimization routines
- Appropriate hyperparameters
- …

# Objective

$f(i_1, \ldots, i_N) \; =$




**Restrictions:**

- Without optimization

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \mathscr{L}(\theta_t)$$

✗

- High dimensionality

$$f(x_1, \ldots, x_N)$$

$$N = 500,\ 1000,\ \ldots$$

- High sparsity

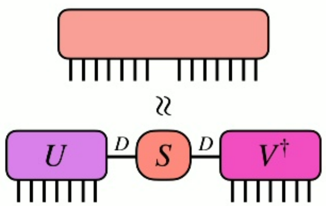
$$\begin{bmatrix} 0 & \cdots & 0 & 0.016 & 0 & \cdots & 0 \\ 0 & 0.002 & 0 & & \cdots & & 0 \\ 0 & & & \cdots & & & 0 \\ 0 & & \cdots & 0 & 0.07 & 0 & 0 \end{bmatrix}$$

**Cases of interest:**

- Ground states of quantum many-body systems:
  - Entanglement structure
  - Symmetries
  - Topological order
- Machine Learning models:
  - Efficiency
  - Privacy
  - Interpretability

# Tools

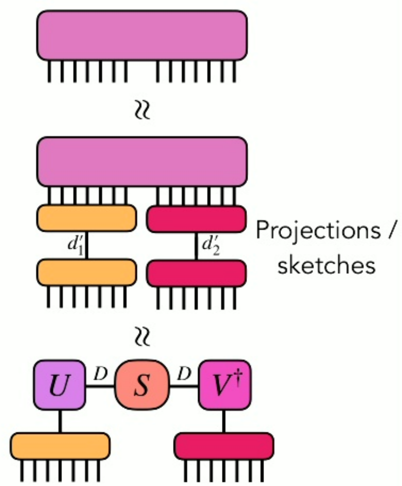
## Singular Value Decomposition




$O(d^n d^m D)$

**Inefficient** for high-dimensional tensors

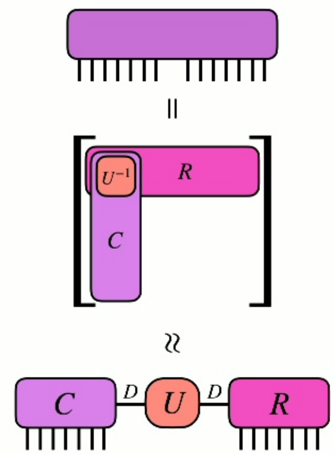
## Randomized SVD




$O(d'_1 d'_2 D)$

Efficient **if** projection can be made efficiently

## Cross Interpolation



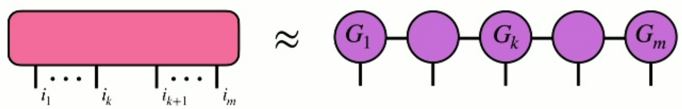

A ***good*** set of rows/ columns can cover the whole span

# Tensorization

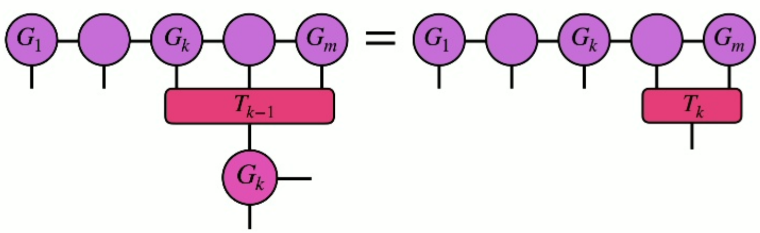
**Tensor Train via Recursive Sketching:**

arXiv:2202.11788

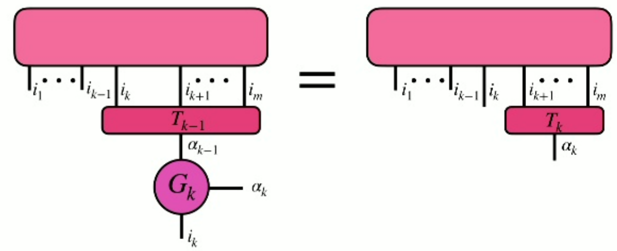
Assuming



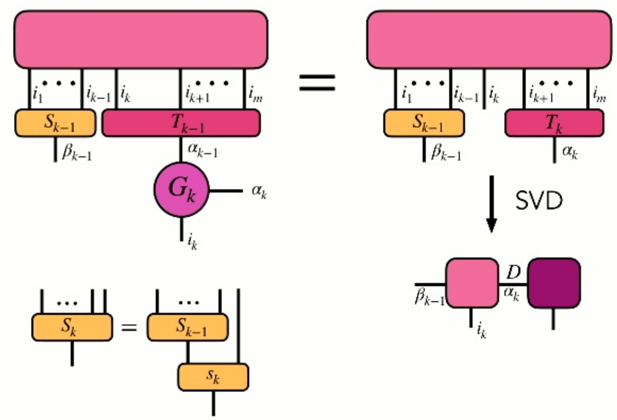
We can solve for $G_k$



Core Determining Equations (overdetermined):



Project to reduce equations:



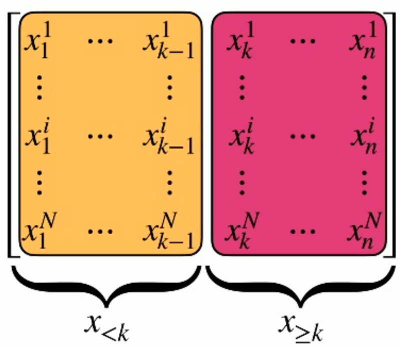
# Tensorization

## Tensor Train via Recursive Sketching *from Samples:*

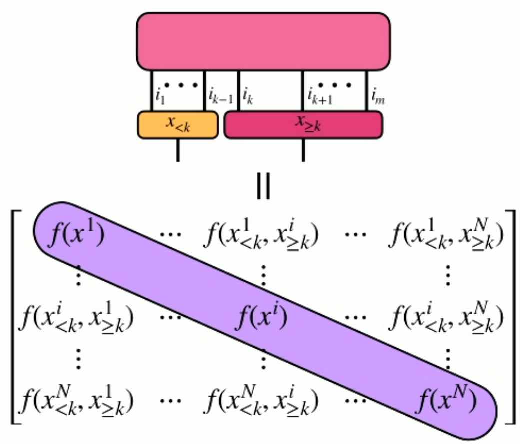
Take a set of *sketch samples:*

- Ground state: sample configurations
- ML model: subset of training points

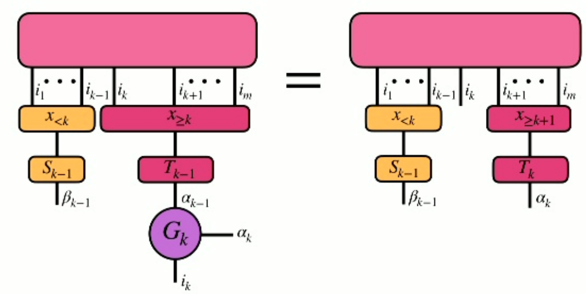
$N$ samples

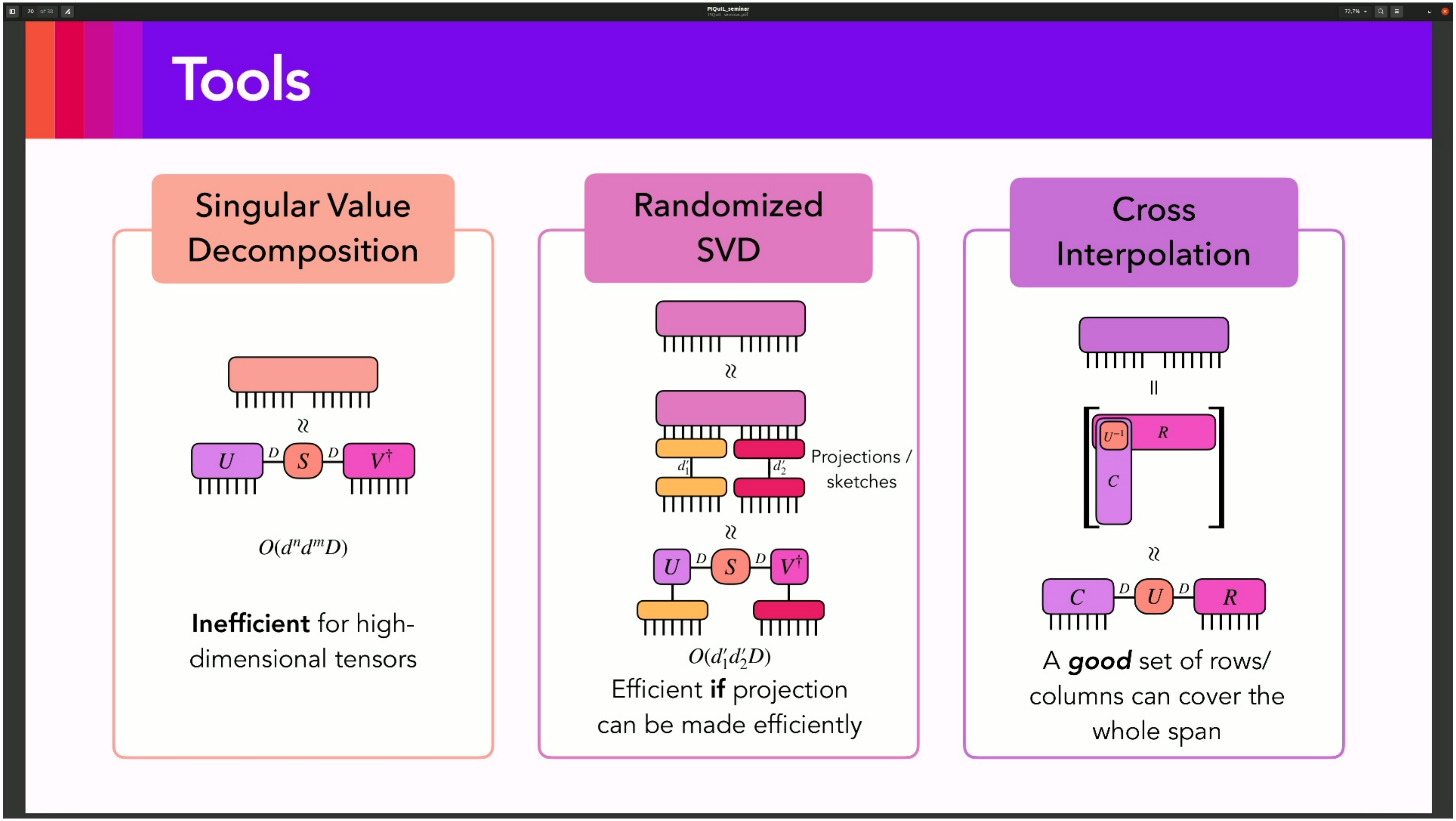



Project to *high volume subspace:*



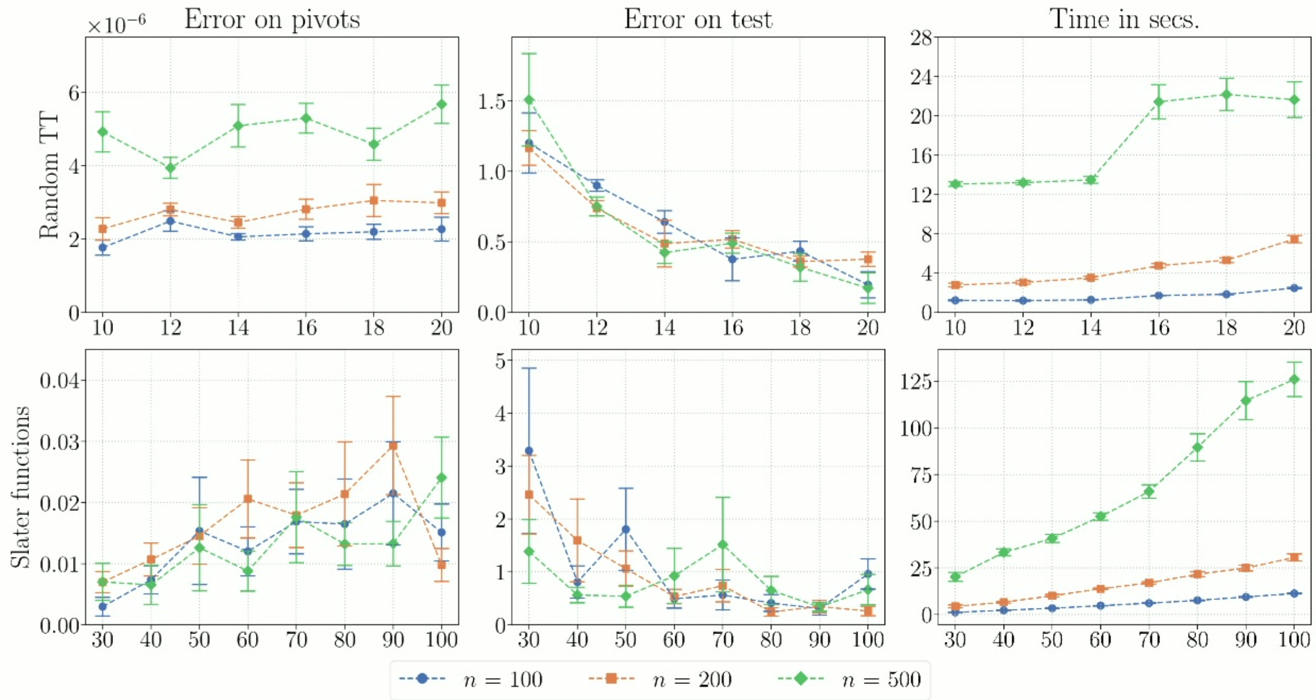
Set equations:



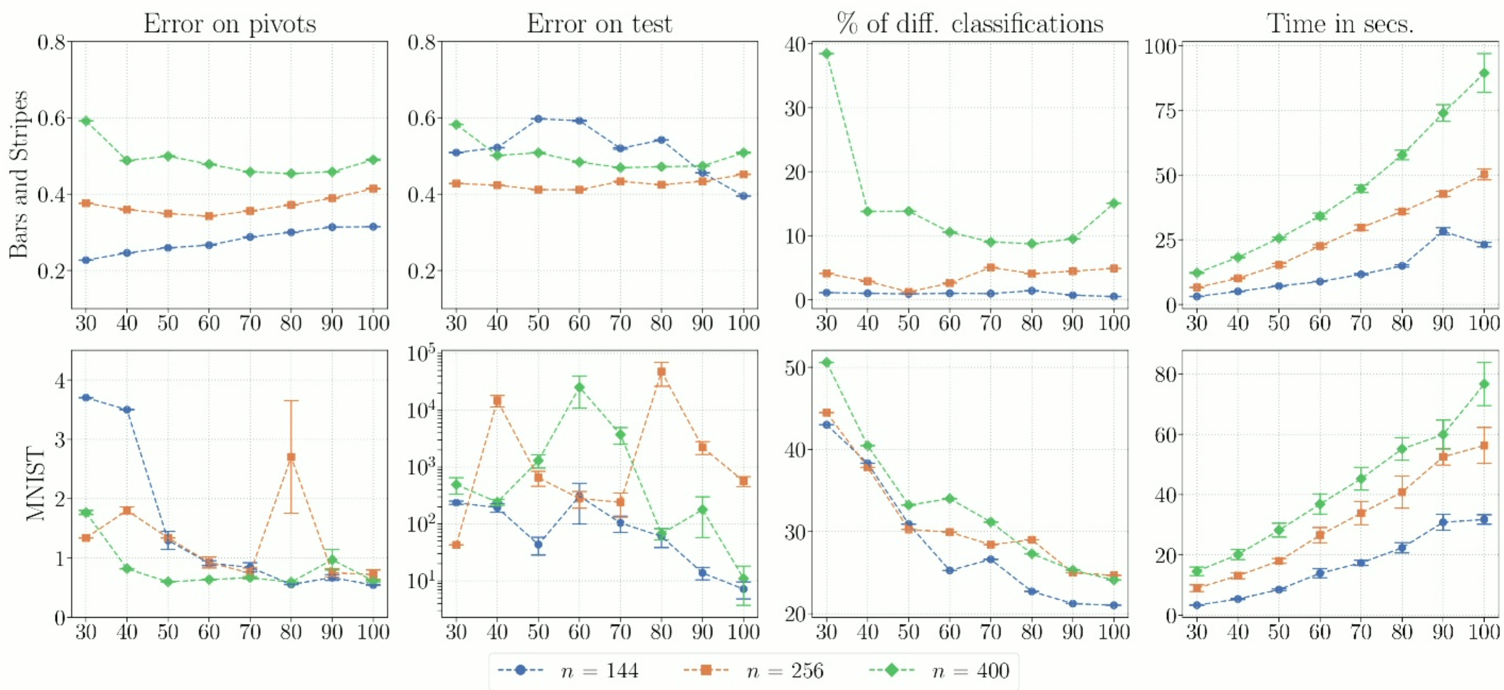

Tools
Singular Value Decomposition
U
D
S
D
V†
O(dⁿdᵐD)
Inefficient for high-dimensional tensors
Randomized SVD
d'₁
d'₂
Projections / sketches
U
D
S
D
V†
O(d'₁d'₂D)
Efficient if projection can be made efficiently
Cross Interpolation
U⁻¹
R
C
C
D
U
D
R
A good set of rows/ columns can cover the whole span

# Performance

- **Random-TT:** bond dim. = 10
- **Slater functions:** $\frac{e^{-\|x\|}}{\|x\|}$, with $x \in [0, L]^m$, each $x_i$ discretized in $d$ variables.

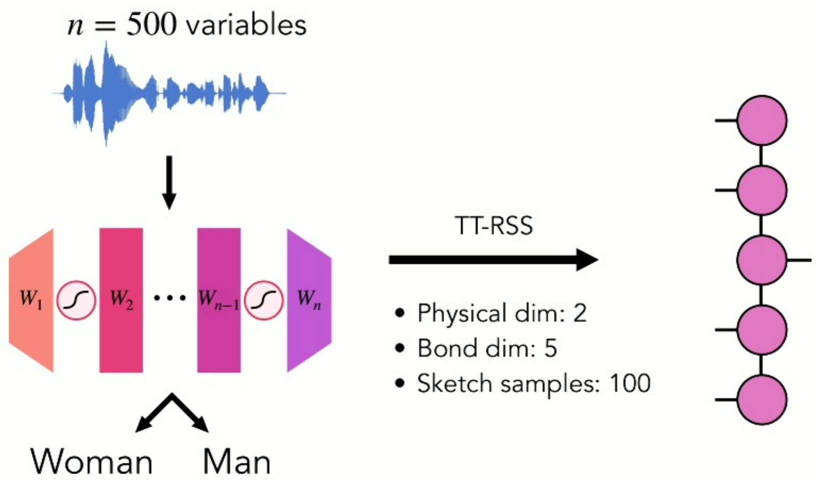


# Performance



# Applications: Privacy

**Voice classification:**

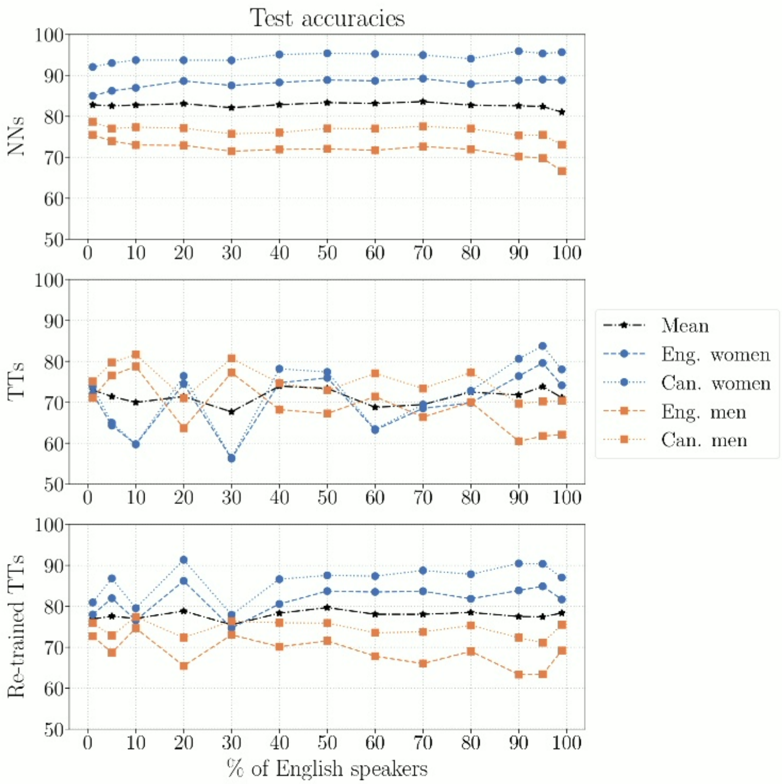



- ~82% accuracy
- ~25k parameters

- ~78% accuracy
- ~25k parameters

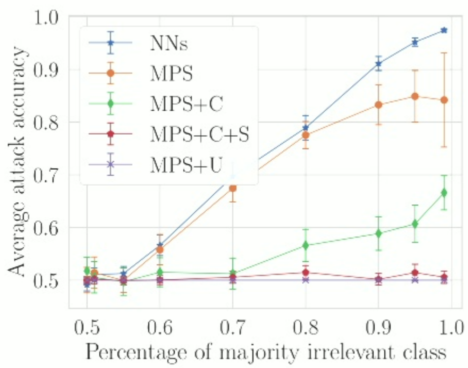
# Applications: Privacy

- Voices are from people with **English** or **Canadian** accents (**irrelevant** feature)
- We repeat experiments for different proportions of imbalance of the accent (hidden) feature.

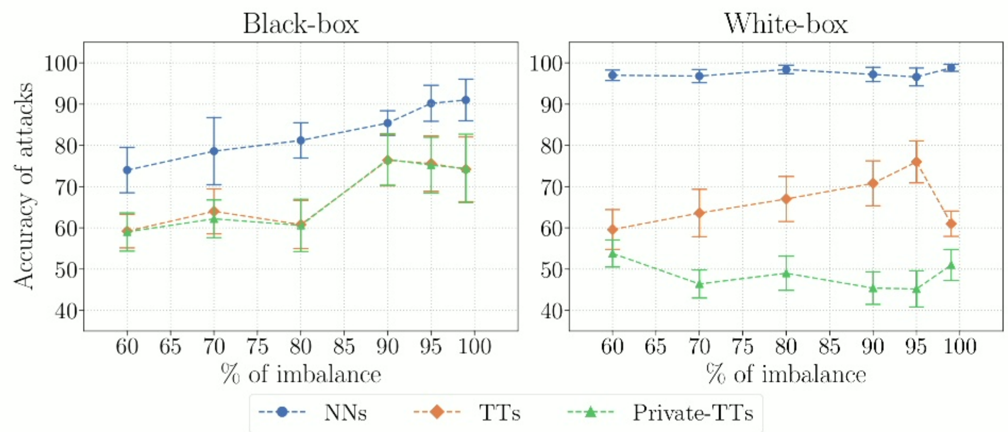


# Applications: Privacy

**Attacks:**




arXiv:2202.12319




arXiv:2501.06300

# Applications: Interpretability

**AKLT model:**

$$\hat{H} = \sum_{\langle ij \rangle} P^{(2)}_{\langle ij \rangle} \sim \sum_j \vec{S}_j \cdot \vec{S}_{j+1} + \frac{1}{3}(\vec{S}_j \cdot \vec{S}_{j+1})^2$$

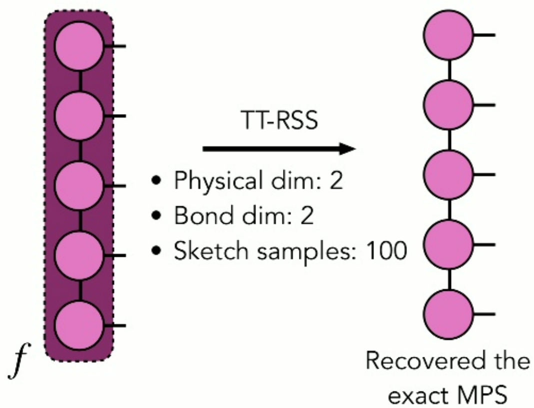
Ground state with exact MPS representation

$$|\Psi\rangle = \sum_{\{s\}} \mathrm{Tr}[A^{s_1} A^{s_2} \ldots A^{s_N}] |s_1 s_2 \ldots s_N\rangle$$

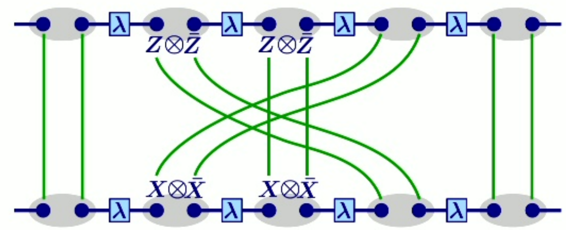
$$A^{+} = +\sqrt{\tfrac{2}{3}}\,\sigma^{+}$$

$$A^{0} = -\sqrt{\tfrac{1}{3}}\,\sigma^{z}$$

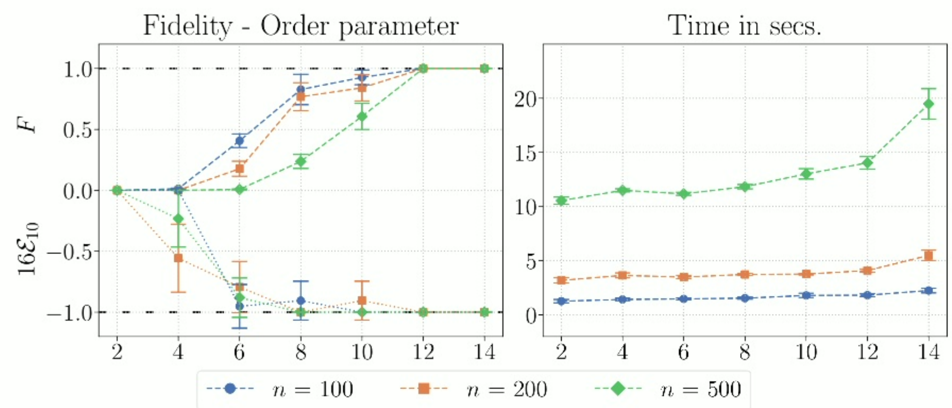
$$A^{-} = -\sqrt{\tfrac{2}{3}}\,\sigma^{-}$$




**Compute topological order parameter from TN:**




arXiv:1201.4174



arxiv:2501.06300

# Thank you!

ICMAT INSTITUTO DE CIENCIAS MATEMÁTICAS

EXCELENCIA SEVERO OCHOA

GOBIERNO DE ESPAÑA

MINISTERIO DE CIENCIA E INNOVACIÓN

Comunidad de Madrid

Dirección General de Investigación e Innovación Tecnológica

CONSEJERÍA DE CIENCIA, UNIVERSIDADES E INNOVACIÓN

# Applications: Interpretability

**AKLT model:**

$$\hat{H} = \sum_{\langle ij \rangle} P^{(2)}_{\langle ij \rangle} \sim \sum_j \vec{S}_j \cdot \vec{S}_{j+1} + \frac{1}{3}(\vec{S}_j \cdot \vec{S}_{j+1})^2$$

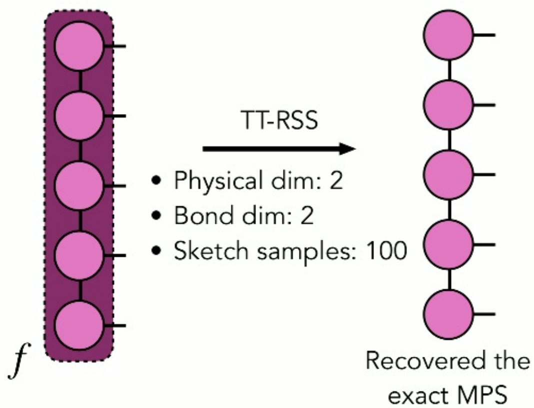
Ground state with exact MPS representation

$$|\Psi\rangle = \sum_{\{s\}} \mathrm{Tr}[A^{s_1} A^{s_2} \ldots A^{s_N}] |s_1 s_2 \ldots s_N\rangle$$

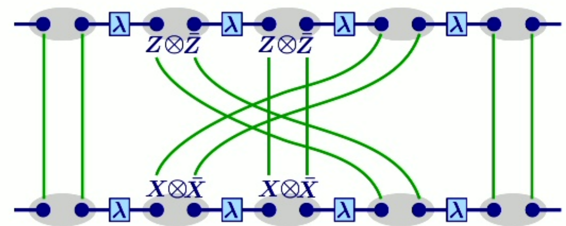
$$A^+ = +\sqrt{\tfrac{2}{3}}\, \sigma^+$$

$$A^0 = -\sqrt{\tfrac{1}{3}}\, \sigma^z$$

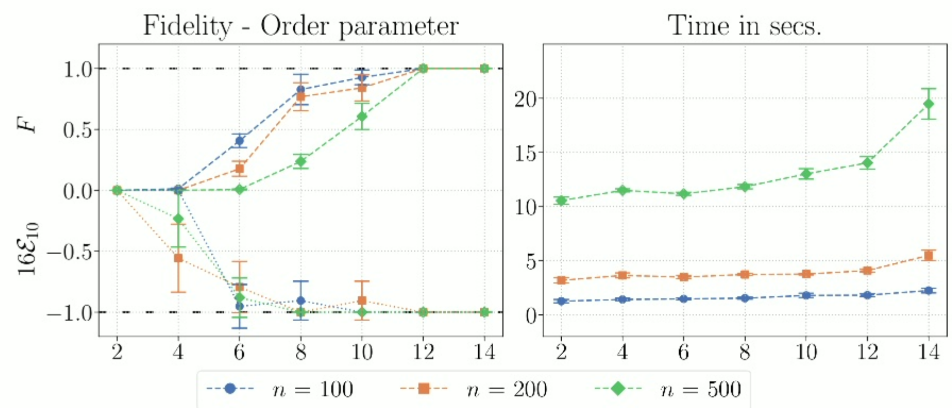
$$A^- = -\sqrt{\tfrac{2}{3}}\, \sigma^-$$




**Compute topological order parameter from TN:**

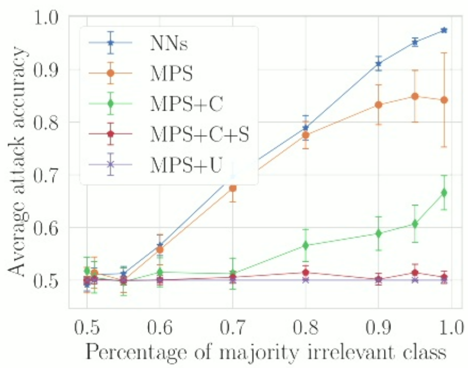



arXiv:1201.4174

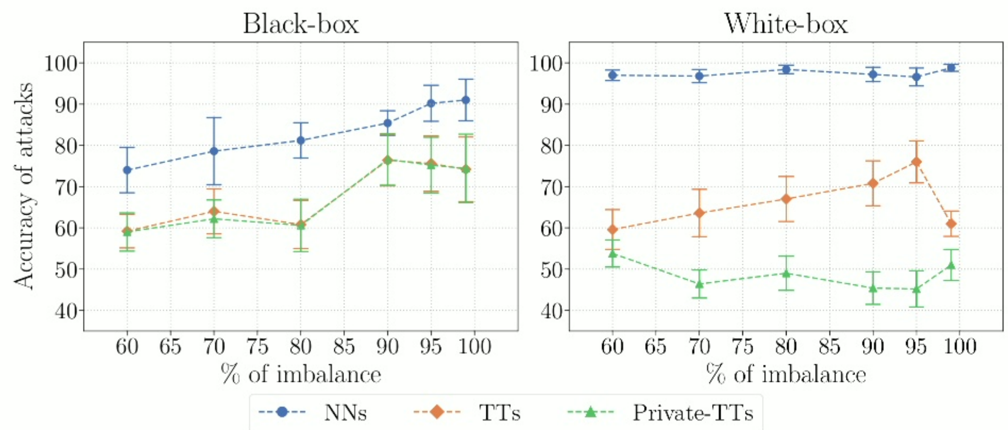


# Applications: Privacy

**Attacks:**




arXiv:2202.12319




arXiv:2501.06300