Efficient identification of sparse neural networks with butterfly structure

Léon Zheng, Elisa Riccietti, Rémi Gribonval
Université de Lyon, ENS de Lyon, CNRS, INRIA, Laboratoire de l'Informatique du Parallélisme
valeo.ai, Paris

BUTTERFLY FACTORIZATION

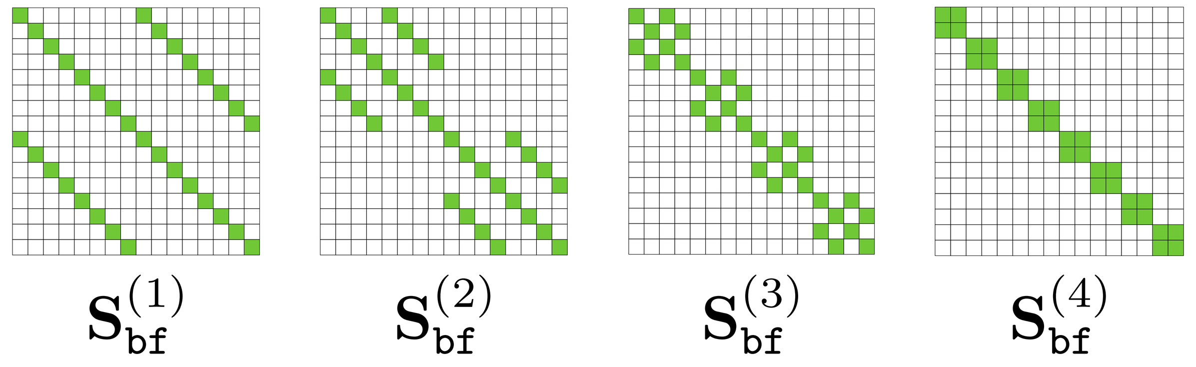
Given $\mathbf{Z} \in \mathbb{C}^{N \times N}$ with $N = 2^J$:

$$\min_{\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(J)}} \|\mathbf{Z} - \mathbf{X}^{(1)} \ldots \mathbf{X}^{(J)}\|_F^2$$

such that $\mathrm{supp}(\mathbf{X}^{(\ell)}) \subseteq \mathbf{S}_{\mathtt{bf}}^{(\ell)} := \mathbf{I}_{2^{\ell-1}} \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \otimes \mathbf{I}_{N/2^{\ell}}$

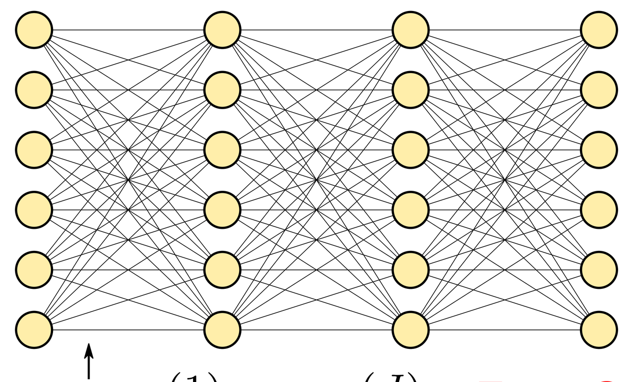
where $\mathrm{supp}(\mathbf{M}) := \{(i,j) \mid \mathbf{M}_{i,j} \neq 0\}$.

Example with J=4 factors:



MOTIVATIONS

REPLACE DENSE MATRICES BY BUTTERFLY MATRICES



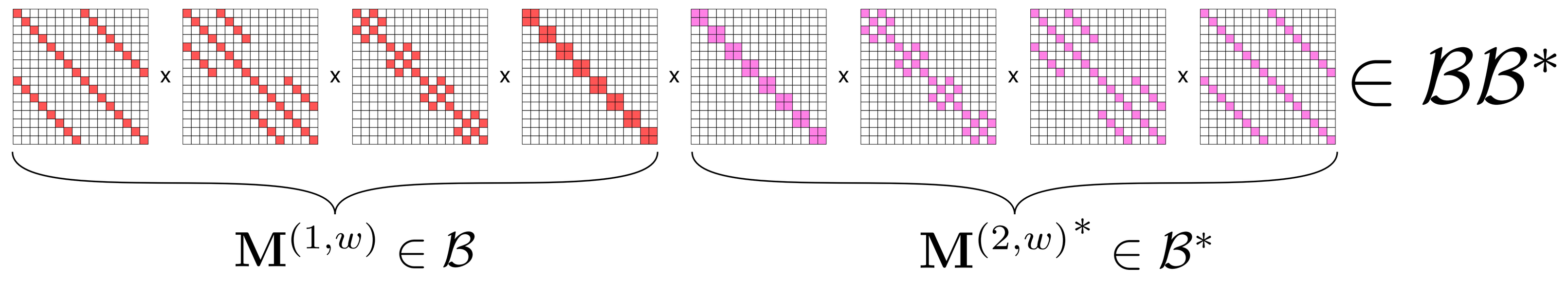
Less parameters
Faster training / inference time

$\mathbf{W} \approx \mathbf{X}^{(1)} \ldots \mathbf{X}^{(J)}$ Fast O(N log N) matrix-vector multiplications

THE BUTTERFLY STRUCTURE IS EXPRESSIVE

$$\mathcal{B} := \{\mathbf{X}^{(1)} \ldots \mathbf{X}^{(J)} \mid \mathrm{supp}(\mathbf{X}^{(\ell)}) \subseteq \mathrm{supp}(\mathbf{S}_{\mathtt{bf}}^{(\ell)})\}$$

Kaleidoscope representation [2]:



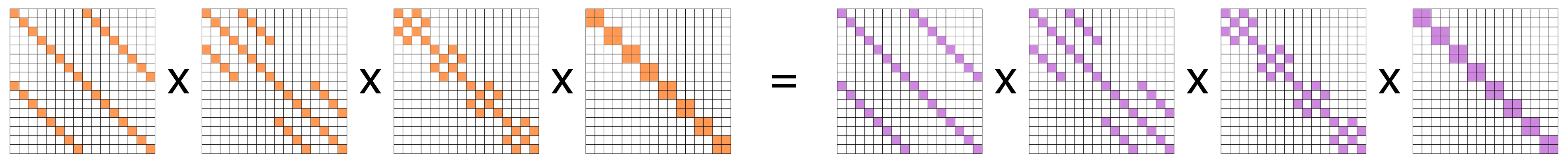
$\prod_{w=1}^{W} \mathbf{M}^{(1,w)}\mathbf{M}^{(2,w)*}$ captures

DFT, DSC, DCT, Hadamard
Convolutions, Toeplitz
Orthogonal random features
Fastfood, ACDC

CONTRIBUTIONS

IDENTIFIABILITY

If $\mathbf{Z} = \mathbf{X}^{(1)} \times \mathbf{X}^{(2)} \times \ldots \times \mathbf{X}^{(J)} = \bar{\mathbf{X}}^{(1)} \times \bar{\mathbf{X}}^{(2)} \times \ldots \times \bar{\mathbf{X}}^{(J)}$

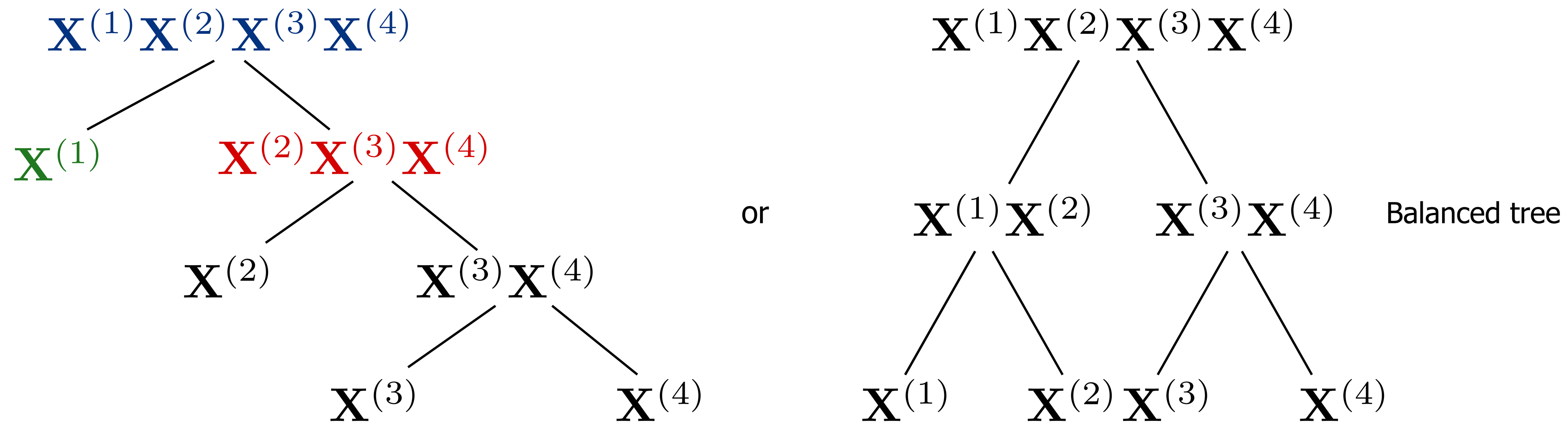


then $(\mathbf{X}^{(\ell)})_{\ell=1}^{J} \sim (\bar{\mathbf{X}}^{(\ell)})_{\ell=1}^{J}$ up to natural scaling ambiguity.

HIERARCHICAL ALGORITHM

Let $\mathbf{Z} := \mathbf{X}^{(1)} \ldots \mathbf{X}^{(J)}$ with exact butterfly factorization.

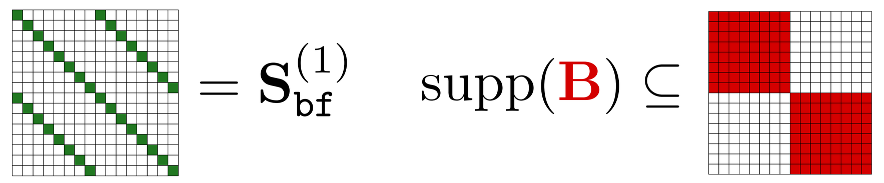
Principle: perform successive two-layer fixed-support factorizations, until obtaining J factors.



Exact recovery guarantee: the butterfly factors are recovered from $\mathbf{Z}$ by the algorithm, in O(N²) time.

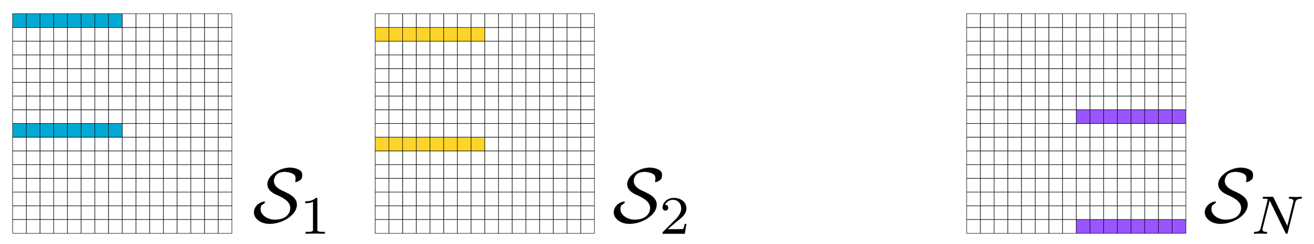
AT THE FIRST LEVEL of the unbalanced tree:

(FSMF) $\min_{\mathbf{A},\mathbf{B}} \|\mathbf{Z} - \mathbf{A}\mathbf{B}\|_F^2$ such that $\mathrm{supp}(\mathbf{A}) \subseteq$ [pattern] $= \mathbf{S}_{\mathtt{bf}}^{(1)}$ $\mathrm{supp}(\mathbf{B}) \subseteq$ [pattern] $= \mathbf{S}_{\mathtt{bf}}^{(2)}\mathbf{S}_{\mathtt{bf}}^{(3)}\mathbf{S}_{\mathtt{bf}}^{(4)}$



Decomposition into rank-one matrices: $\mathbf{A}\mathbf{B} = \mathbf{A}_{\bullet,1}\mathbf{B}_{1,\bullet} + \mathbf{A}_{\bullet,2}\mathbf{B}_{2,\bullet} + \ldots + \mathbf{A}_{\bullet,N}\mathbf{B}_{N,\bullet}$

Observation: the rank-one supports are pairwise disjoint.



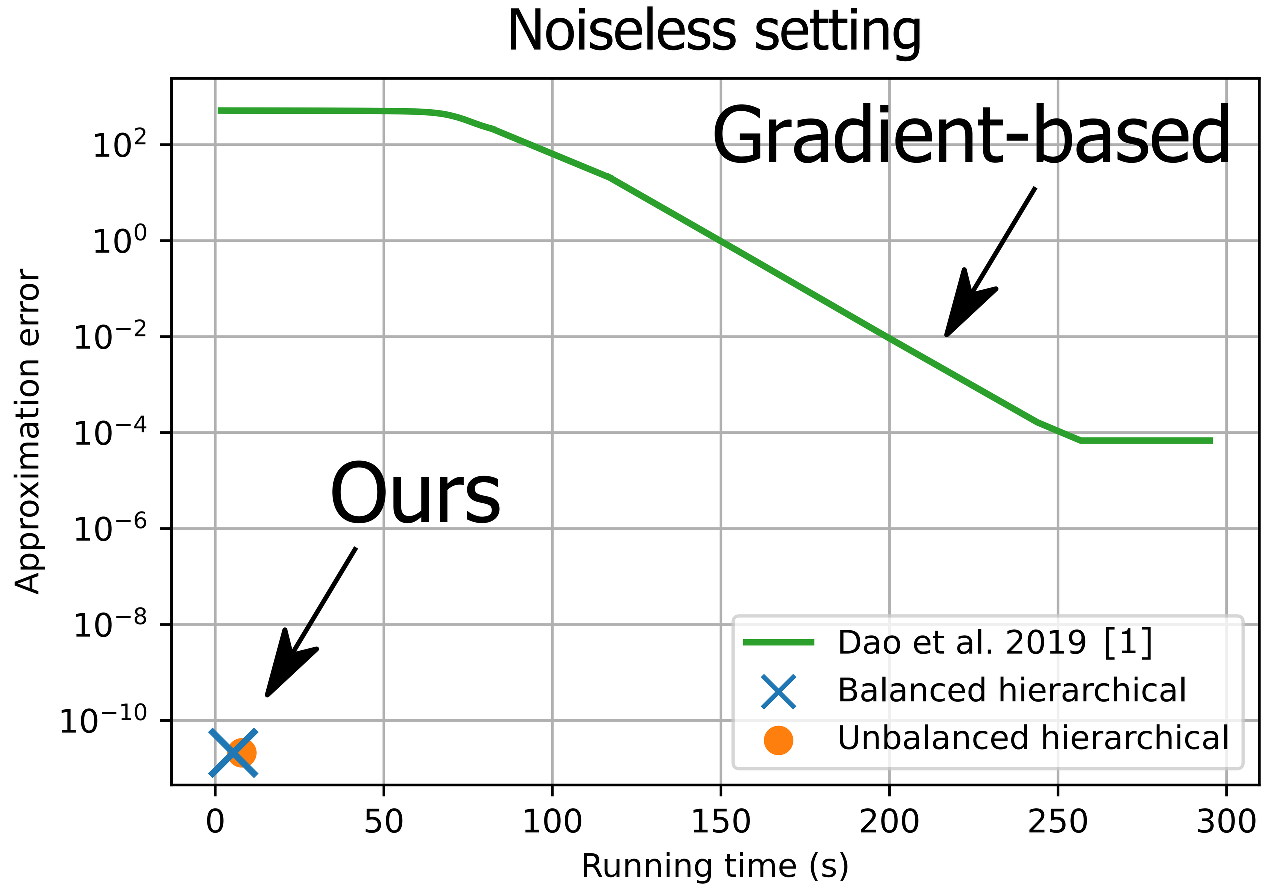
THEOREM:
(FSMF) is polynomially solvable [3] and admits an essentially unique solution [5].

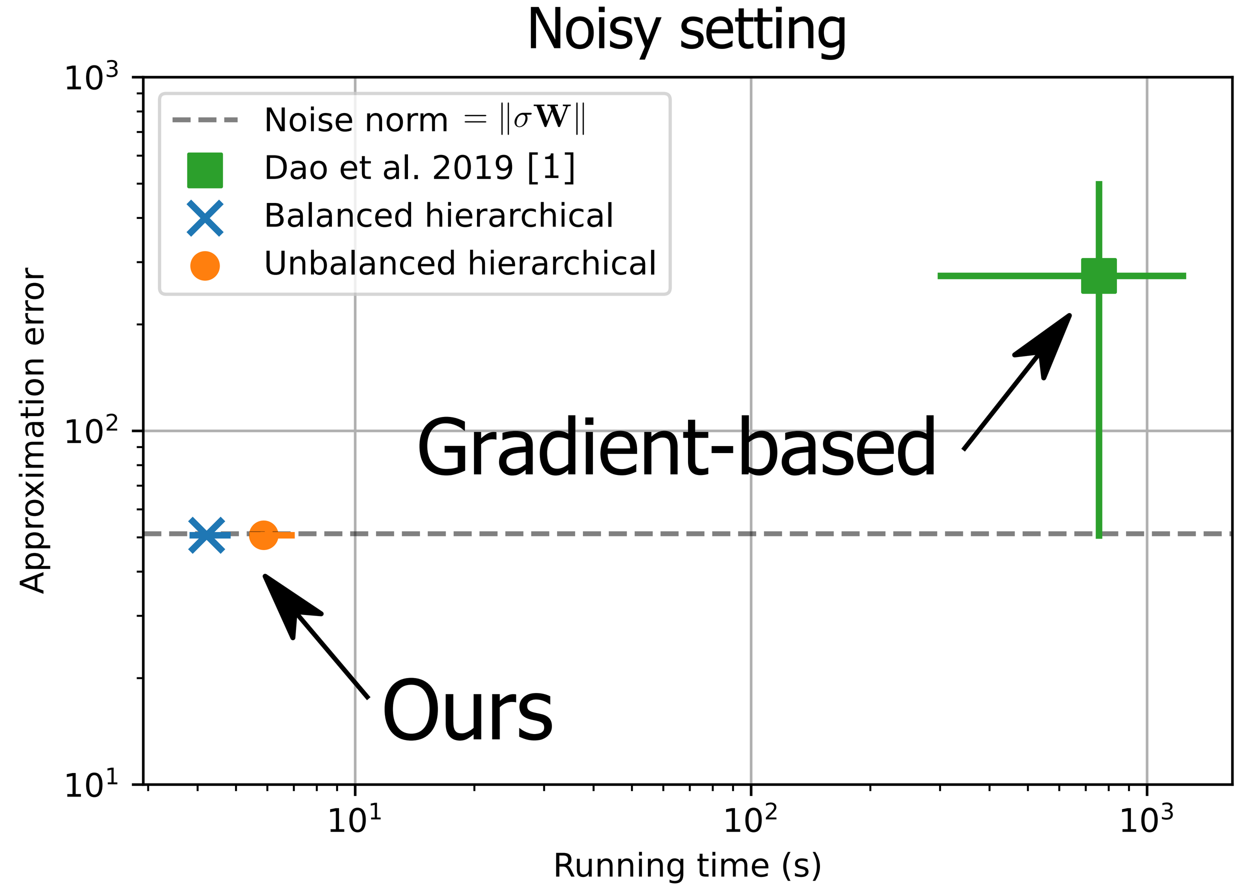
ALGORITHM to solve (FSMF) [3]:

1. Extract the submatrices $\mathbf{Z} \odot \mathcal{S}_i$ for $i = 1, \ldots, N$
2. Perform best rank-one approximation for each submatrix

NUMERICAL EXPERIMENTS

Recovery of the butterfly factors of the DFT matrix:





Experiments are taken from [4].

REFERENCES

[1] T. Dao, A. Gu, M. Eichhorn, A. Rudra, and C. Ré, Learning fast algorithms for linear transforms using butterfly factorizations. In ICML, 2019.
[2] T. Dao, N. Sohoni, A. Gu, M. Eichhorn, A. Blonder, M. Leszczynski, A. Rudra, and C. Ré, Kaleidoscope: An efficient, learnable representation for all structured linear maps. In ICLR, 2020.
[3] Q.-T. Le, E. Riccietti, and R. Gribonval, Spurious Valleys, Spurious Minima and NP-hardness of Sparse Matrix Factorization With Fixed Support. arXiv preprint, arXiv:2112.00386, 2021.
[4] Q.-T. Le, L. Zheng, E. Riccietti, and R. Gribonval, Fast learning of fast transforms, with guarantees. In ICASSP, 2022.
[5] L. Zheng, E. Riccietti, and R. Gribonval, Efficient Identification of Butterfly Sparse Matrix Factorizations. arXiv preprint, arXiv:2110.01230, 2022.