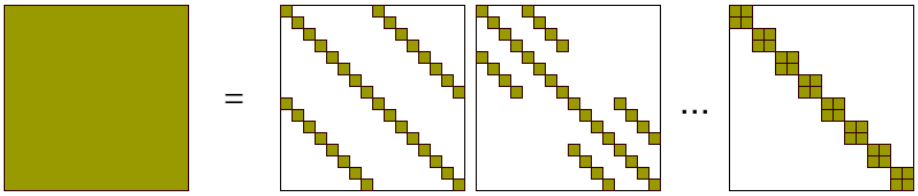
M2 internship subject: **Quantized Butterflies**

Supervisors: Elisa Riccietti (MCF ENSL), Rémi Gribonval (DR Inria),
Pascal Carrivain (Ingénieur de Recherche, Inria) – Ockham team
Theo Mary (CR CNRS, Sorbonne)

September 2024



Context: machine learning and scientific computing. Butterfly matrices are an important class of structured sparse matrices, well known for their expressivity, i.e., many matrices can be approximated with a low error as a product of such matrices. Thanks to their extremely sparse structure, such a factorization allows for a significant reduction of storage costs as well as energy and time cost for matrix–vector products.

Because of this important property they appear in many applications, from classical problems such as the factorization of the Hadamard or Fourier matrices to more recent developments in machine learning such as the fine-tuning of large language models [1, 2] where they are used as a mean for model compression.

Another tool widely used for model compression in machine learning is low precision quantization. A timely challenge related to butterflies is to couple their sparsity structure with low precision quantization. Even though butterfly matrices and quantization have been widely used separately, their combination poses significant challenges that have never been addressed in the literature.

A first step in this direction has been done in [4, 5], where we propose an efficient heuristic quantization schemes for butterfly matrices built exploiting natural scaling invariances that arise in the problem of quantizing the rank-one blocks. Butterfly matrices present indeed a peculiar structure, so that certain partial products of their factors can be decomposed into blocks admitting an exact representation as rank-one matrices. This allowed for instance the recent development of efficient algorithms to compute a general butterfly factorization of a given matrix; such algorithms accelerate gradient descent by several orders of magnitude, and are endowed with reconstruction guarantees if the target matrix admits exactly or approximately a butterfly factorization [7, 10, 6, 9]. Moreover, the development of a CUDA kernel for efficient matrix-vector products [3] allows to observe in practice the good theoretical speed-ups guaranteed by such a structured factorization.

Project description. *The goal of this internship will be to investigate the use of quantized butterflies for neural networks compression.* This will involve in priority:

- setting up an approach to compress a neural network by combining a butterfly factorization of the weights matrices and a low precision quantization;
- investigating the potential of the approach through en empirical study;

- studying the link between the precision of the factorization and the generalization error of the network;
- implementing, testing, documenting and illustrating with examples the proposed approach, and integrating it in `pyfaust` [8] .

References

[1] T. Dao, B. Chen, N. S. Sohoni, A. D. Desai, M. Poli, J. Grogan, A. Liu, A. Rao, A. Rudra, and C. Ré. Monarch: Expressive structured matrices for efficient and accurate training. In *International Conference on Machine Learning*, pages 4690–4721. PMLR, 2022.

[2] T. Dao, A. Gu, M. Eichhorn, A. Rudra, and C. Ré. Learning fast algorithms for linear transforms using butterfly factorizations. In *International Conference on Machine Learning*, pages 1517–1527. PMLR, 2019.

[3] Antoine Gonon, Léon Zheng, Pascal Carrivain, and Quoc-Tung Le. Make Inference Faster: Efficient GPU Memory Management for Butterfly Sparse Matrix Multiplication. working paper or preprint, May 2024.

[4] R. Gribonval, T. Mary, and E. Riccietti. Optimal quantization of rank-one matrices in floating-point arithmetic—with applications to butterfly factorizations. Preprint, Under review, June 2023.

[5] R. Gribonval, T. Mary, and E. Riccietti. Scaling is all you need: quantization of butterfly matrix products via optimal rank-one quantization. In *29ème Colloque sur le traitement du signal et des images (GRETSI)*, number 2023-1193 in Actes du GRETSI 2023, pages 497–500, Grenoble, France, August 2023. GRETSI - Groupe de Recherche en Traitement du Signal et des Images.

[6] Quoc-Tung Le. *Algorithmic and theoretical aspects of sparse deep neural networks.* Theses, Ecole normale supérieure de lyon - ENS LYON, December 2023.

[7] Quoc-Tung Le, Léon Zheng, Elisa Riccietti, and Rémi Gribonval. Fast learning of fast transforms, with guarantees. In *ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, May 2022. This paper is associated to code for reproducible research available at https://hal.inria.fr/hal-03552956.

[8] Ockham team. `pyfaust` python package.
`https://faust.inria.fr`.

[9] Léon Zheng. *Data frugality and computational efficiency in deep learning.* Theses, Ecole normale supérieure de lyon - ENS LYON, May 2024.

[10] Léon Zheng, Elisa Riccietti, and Rémi Gribonval. Efficient Identification of Butterfly Sparse Matrix Factorizations. *SIAM Journal on Mathematics of Data Science*, 5(1):22–49, 2023.