SIAM LA 2024
13-17 May 2024

Optimal quantization of rank-one matrices for butterfly factorizations

Theo Mary
Sorbonne Université, CNRS, LIP6

Joint work with **Rémi Gribonval** and **Elisa Riccietti** (ENS Lyon, Inria)

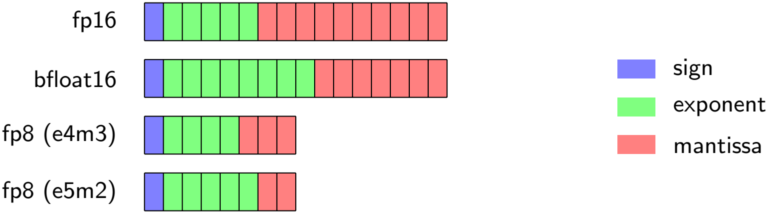
Slides
https://bit.ly/SIAMLA24

Preprint
https://bit.ly/rank1quant

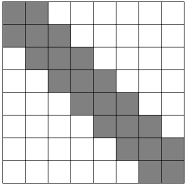
Motivation

Growing size of models and datasets $\rightarrow$ approximate computing

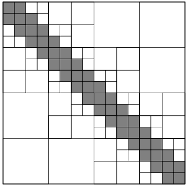
- Quantization to low precision floating-point arithmetic



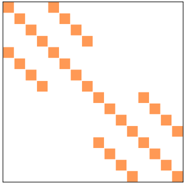
- Low-rank, structured, data sparse matrices



BLR matrix



$\mathcal{H}$-matrix



Butterfly matrix

Rank-one matrices

Goal: quantize the rank-one matrix

$$xy^T \to \widehat{x}\widehat{y}^T \qquad (x \in \mathbb{R}^m, \; y \in \mathbb{R}^n)$$

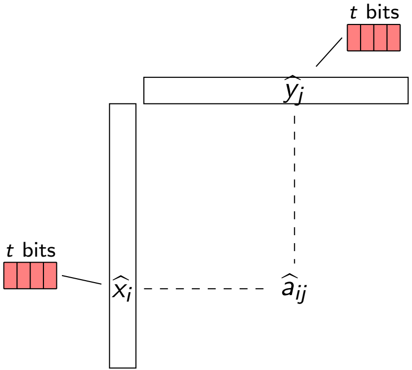
where the coefficients of $\widehat{x}$, $\widehat{y}$ have t bits of mantissa

- The standard approach uses round-to-nearest (RTN) and leads to an error of order $u = 2^{-t}$: if $\widehat{x} = \mathrm{round}(x)$, $\widehat{y} = \mathrm{round}(y)$ then

$$\begin{aligned} \|\widehat{x} - x\| &\le u\|x\| \\ \|\widehat{y} - y\| &\le u\|y\| \\ \Rightarrow \|\widehat{x}\widehat{y}^T - xy^T\| &\le (2u + u^2)\|x\|\|y\| \end{aligned}$$

- We will show this is far from optimal!

An analogy

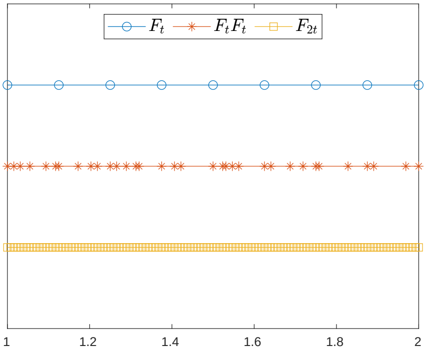


- What we really care about is the accuracy of $\widehat{a}_{ij} = \widehat{x}_i \widehat{y}_j$
- Think of multiword arithmetic: $a \approx \widehat{x} + \widehat{y}$ with $\widehat{x} = \text{round}(a)$ and $\widehat{y} = \text{round}(a - \widehat{x}) \rightarrow 2t$-bit accuracy
- What about $a = xy$? (Which $\widehat{x}, \widehat{y}$ yields the best approximation $\widehat{x}\widehat{y}$?)

The set $\mathbb{F}_t\mathbb{F}_t$

- Let $\mathbb{F}_t$ be the set of t-bit floating-point numbers. We are interested in the set

$$\mathbb{F}_t\mathbb{F}_t = \{a = xy,\ x \in \mathbb{F}_t,\ y \in \mathbb{F}_t\}$$

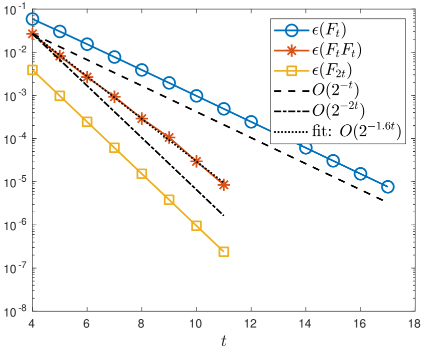


- No closed form expression of its elements, but we can simply enumerate all of them for small t

The set $\mathbb{F}_t\mathbb{F}_t$

- Let $\mathbb{F}_t$ be the set of t-bit floating-point numbers. We are interested in the set

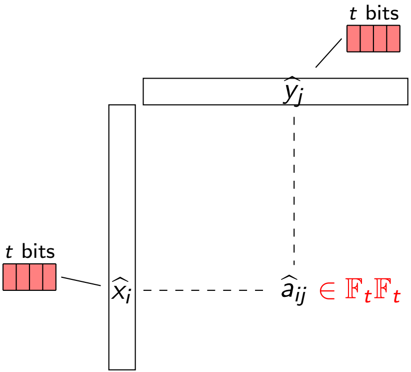
$$\mathbb{F}_t\mathbb{F}_t = \{a = xy,\ x \in \mathbb{F}_t,\ y \in \mathbb{F}_t\}$$



- No closed form expression of its elements, but we can simply enumerate all of them for small t

$\Rightarrow$ Worst-case error of order $2^{-1.6t}$

A constrained combinatorial problem



We don't just have one scalar, but a rank-one matrix $\Rightarrow$ two issues:

- We have constraints: $\widehat{x}_i$ must be the same in $\widehat{a}_{ij} = \widehat{x}_i\widehat{y}_j$ and $\widehat{a}_{ik} = \widehat{x}_i\widehat{y}_k$
- How can we find the optimal quantization? Combinatorial problem!

$$\min_{\widehat{x}\in\mathbb{F}_t^m, \widehat{y}\in\mathbb{F}_t^n} \|xy^T - \widehat{x}\widehat{y}^T\|$$

Characterization of the optimum

Theorem

$$\min_{\widehat{x}\in\mathbb{F}_t^m,\widehat{y}\in\mathbb{F}_t^n} \|xy^T - \widehat{x}\widehat{y}^T\| = \min_{\lambda\in\mathbb{R}} \|xy^T - \mathrm{round}(\lambda x)\,\mathrm{round}(\mu(\lambda)y)^T\|$$

The optimal quantization $\widehat{x}\widehat{y}^T$ is given by

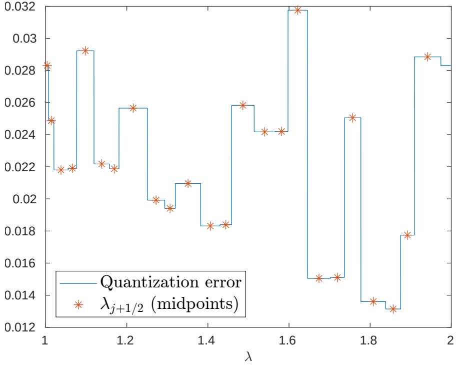
$$\widehat{x} = \mathrm{round}(\lambda x)$$
$$\widehat{y} = \mathrm{round}(\mu(\lambda)y^T)$$

where $\lambda \in \mathbb{R}$ and $\mu(\lambda) = \frac{x^T\widehat{x}}{\|\widehat{x}\|^2}$.

- It suffices to find the optimal λ to find the optimal $\widehat{x}\widehat{y}^T$!

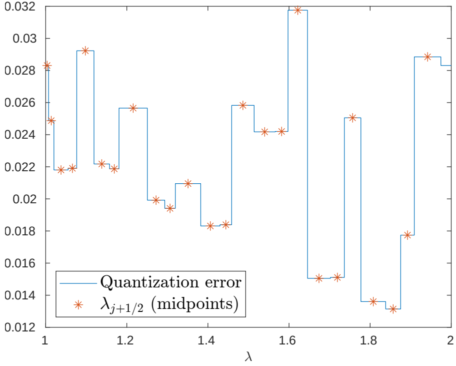
Finding λ

- How do we find the optimal $\lambda \in \mathbb{R}$?
- The optimum is stable under sign flip and multiplication by powers of two $\rightarrow$ restrict the search to $\lambda \in [1, 2]$
- Only a finite number of values of λ change the value of round(λx). Denoting these "breakpoints" as λ_j, we can enumerate the midpoints $\lambda_{j+1/2} = (\lambda_j + \lambda_{j+1})/2$



Finding λ

- How do we find the optimal $\lambda \in \mathbb{R}$?
- The optimum is stable under sign flip and multiplication by powers of two $\rightarrow$ restrict the search to $\lambda \in [1, 2]$
- Only a finite number of values of λ change the value of round(λx). Denoting these "breakpoints" as λ_j, we can enumerate the midpoints $\lambda_{j+1/2} = (\lambda_j + \lambda_{j+1})/2$

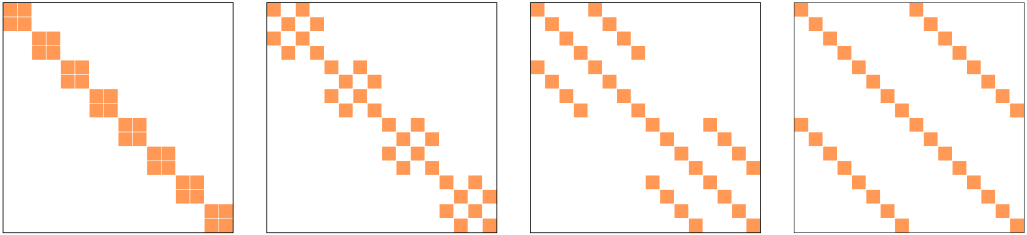


Algorithm:

- Build the set of midpoints
- For each midpoint $\lambda_{j+1/2}$:
  - Build $\widehat{x} = \text{round}(\lambda_{j+1/2} x)$
  - Compute $\mu(\widehat{x}) = x^T \widehat{x} / \|\widehat{x}\|^2$
  - Build $\widehat{y} = \text{round}(\mu y)$
  - Test the accuracy of $\widehat{x}\widehat{y}^T$

$O(mn2^t)$ complexity $\Rightarrow$ tractable for large matrices and low precisions

Butterfly matrices



- Butterfly matrices are extremely sparse yet highly expressive, they appear in many fast linear transforms
- Butterfly factorization: decompose dense $n \times n$ matrix as $B_1 \ldots B_L$, with $L = \log_2 n \Rightarrow O(n \log n)$ complexity

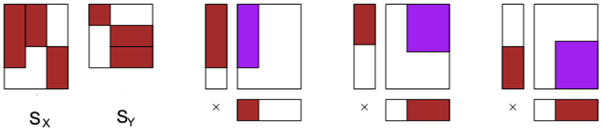
Optimal two-factor quantization

- Key property[1]: for any partial product XY^T of consecutive factors

$$B_1 \dots B_j \underbrace{B_{j+1} \dots B_k}_{X} \underbrace{B_{k+1} \dots B_\ell}_{Y^T} B_{\ell+1} \dots B_L$$

$$XY^T = \sum_{i=1}^{n} x_i y_i^T$$

where the rank-one matrices $x_i y_i^T$ have disjoint support.



[1] QT. Le, E. Riccietti, R. Gribonval, SIMAX (2023).

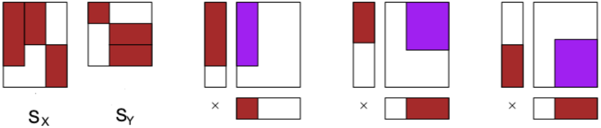
Optimal two-factor quantization

- Key property[1]: for any partial product XY^T of consecutive factors

$$B_1 \dots B_j \underbrace{B_{j+1} \dots B_k}_{X} \underbrace{B_{k+1} \dots B_\ell}_{Y^T} B_{\ell+1} \dots B_L$$

$$XY^T = \sum_{i=1}^{n} x_i y_i^T$$

where the rank-one matrices $x_i y_i^T$ have disjoint support.



- We can optimally quantize two factors X and Y by quantizing each $x_i y_i^T$ optimally and independently: $\widehat{x_i} = \mathrm{round}(\lambda_i x_i)$, $\widehat{y_i} = \mathrm{round}(\mu_i y_i)$ yields

$$\widehat{X} = \mathrm{round}(X\Lambda), \quad \Lambda = \mathrm{diag}(\lambda_i)$$
$$\widehat{Y} = \mathrm{round}(YM), \quad M = \mathrm{diag}(\mu_i)$$

[1] QT. Le, E. Riccietti, R. Gribonval, SIMAX (2023).

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$B_1 \, B_2 \, B_3 \, B_4 \ldots B_L$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{B_1 \, B_2}_{XY^T} \, B_3 \, B_4 \ldots B_L$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{\widehat{B}_1\,\widehat{B}_2}_{XY^T}\,B_3\,B_4\ldots B_L$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{\widehat{B}_1 \, \widehat{B}_2}_{XY^T} \, \underbrace{B_3 \, B_4}_{XY^T} \dots B_L$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{\widehat{B}_1 \widehat{B}_2}_{XY^T} \underbrace{\widehat{B}_3 \widehat{B}_4}_{XY^T} \dots B_L$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{\widehat{B}_1\,\widehat{B}_2}_{XY^T}\,\underbrace{\widehat{B}_3\,\widehat{B}_4}_{XY^T}\ldots\underbrace{\widehat{B}_L}_{RTN}$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{\widehat{B}_1 \, \widehat{B}_2}_{XY^T} \underbrace{\widehat{B}_3 \, \widehat{B}_4}_{XY^T} \dots \underbrace{\widehat{B}_L}_{RTN}$$

- Left-to-right heuristic:

$$B_1 \, B_2 \, B_3 \dots B_{L-1} B_L$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{\widehat{B}_1\,\widehat{B}_2}_{XY^T}\,\underbrace{\widehat{B}_3\,\widehat{B}_4}_{XY^T}\ldots\underbrace{\widehat{B}_L}_{RTN}$$

- Left-to-right heuristic:

$$\underbrace{B_1}_{X}\,\underbrace{B_2\,B_3\ldots B_{L-1}B_L}_{Y^T}$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{\widehat{B}_1\,\widehat{B}_2}_{XY^T}\,\underbrace{\widehat{B}_3\,\widehat{B}_4}_{XY^T}\ldots\underbrace{\widehat{B}_L}_{RTN}$$

- Left-to-right heuristic:

$$\underbrace{\widehat{B}_1}_{X}\,\underbrace{\color{red}{M_2}\color{black}B_2\,B_3\ldots B_{L-1}B_L}_{Y^T}$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{\widehat{B}_1 \widehat{B}_2}_{XY^T} \underbrace{\widehat{B}_3 \widehat{B}_4}_{XY^T} \dots \underbrace{\widehat{B}_L}_{RTN}$$

- Left-to-right heuristic:

$$\underbrace{\widehat{B}_1}_{X} \underbrace{\underbrace{{\color{red}M_2} B_2}_{X} \underbrace{B_3 \dots B_{L-1} B_L}_{Y^T}}_{Y^T}$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{\widehat{B}_1 \, \widehat{B}_2}_{XY^T} \, \underbrace{\widehat{B}_3 \, \widehat{B}_4}_{XY^T} \dots \underbrace{\widehat{B}_L}_{RTN}$$

- Left-to-right heuristic:

$$\underbrace{\widehat{B}_1}_{X} \; \underbrace{\underbrace{\widehat{B}_2}_{X} \, \underbrace{{\color{red}M_3} B_3 \dots B_{L-1} B_L}_{Y^T}}_{Y^T}$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{\widehat{B}_1 \widehat{B}_2}_{XY^T} \underbrace{\widehat{B}_3 \widehat{B}_4}_{XY^T} \ldots \underbrace{\widehat{B}_L}_{RTN}$$

- Left-to-right heuristic:

$$\underbrace{\widehat{B}_1}_{X} \underbrace{\underbrace{\widehat{B}_2}_{X} \underbrace{\widehat{B}_3 \ldots \underbrace{\color{red}{M_{L-1}} B_{L-1}}_{X} \underbrace{B_L}_{Y^T}}_{Y^T}}_{Y^T}$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

$$\underbrace{\widehat{B}_1\,\widehat{B}_2}_{XY^T}\,\underbrace{\widehat{B}_3\,\widehat{B}_4}_{XY^T}\dots\underbrace{\widehat{B}_L}_{RTN}$$

- Left-to-right heuristic:

$$\underbrace{\widehat{B}_1}_{X}\;\underbrace{\underbrace{\widehat{B}_2}_{X}\;\underbrace{\widehat{B}_3\dots\underbrace{\widehat{B}_{L-1}}_{X}\,\underbrace{\widehat{B}_L}_{Y^T}}_{Y^T}}_{Y^T}$$

Heuristics for the L-factor quantization

When $L > 2$, need heuristics to decide how to order/group the factors

- Pairwise heuristic:

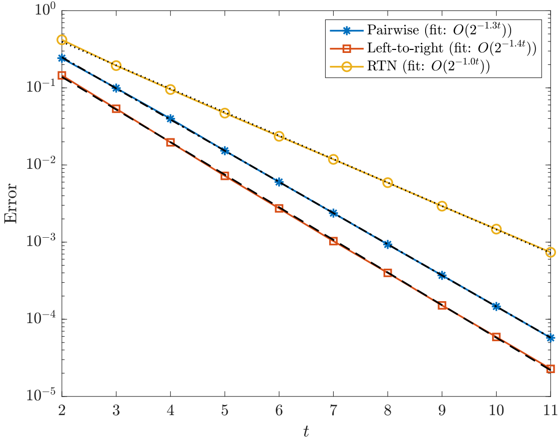
$$\underbrace{\widehat{B}_1 \, \widehat{B}_2}_{XY^T} \underbrace{\widehat{B}_3 \, \widehat{B}_4}_{XY^T} \ldots \underbrace{\widehat{B}_L}_{RTN}$$

- Left-to-right heuristic:

$$\underbrace{\widehat{B}_1}_{X} \; \underbrace{\underbrace{\widehat{B}_2}_{X} \; \underbrace{\widehat{B}_3 \ldots \underbrace{\widehat{B}_{L-1}}_{X} \, \underbrace{\widehat{B}_L}_{Y^T}}_{Y^T}}_{Y^T}$$

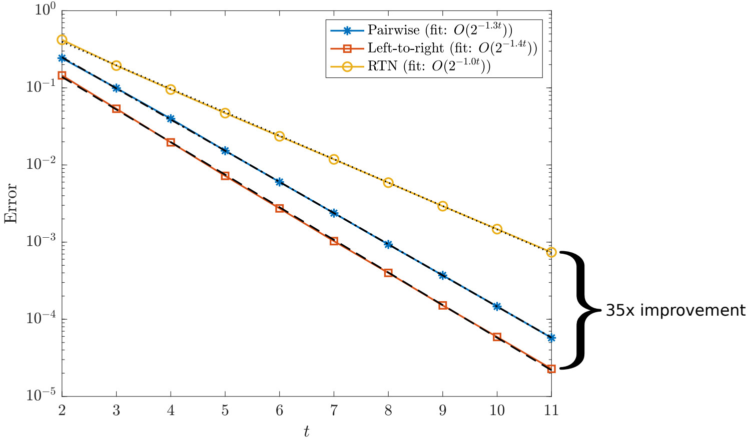
- L2R more expensive because it densifies the factors

Experimental results



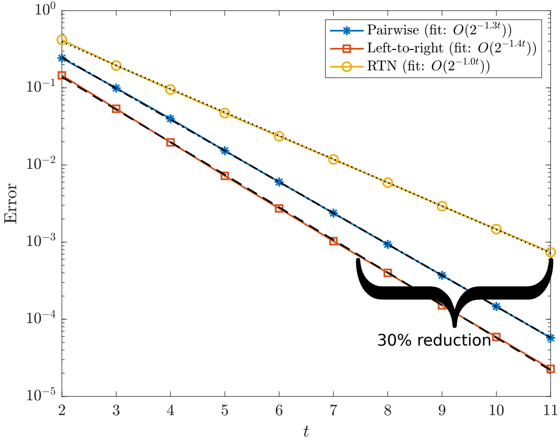
- Randomly generated butterfly factors

Experimental results



- Randomly generated butterfly factors
- Significant accuracy improvement...

Experimental results



- Randomly generated butterfly factors
- Significant accuracy improvement...
- ...or, equivalently, can reduce storage by about 30% with no loss of accuracy

Conclusion

Key results:

- Characterized optimal quantization of xy^T as $\mathrm{round}(\lambda x)\,\mathrm{round}(\mu y)^T$
- Proposed algorithm to find the optimal λ in $O(mn2^t)$ complexity
- Proposed two heuristics to apply method to butterfly factorization and obtained storage reductions of 30% with no loss of accuracy

Butterfly matrices are only one possible application, many other perspectives:
rank-r matrices, tensors, DNNs, ...

Slides

https://bit.ly/SIAMLA24

Preprint

https://bit.ly/rank1quant

Thanks! Questions?