CROQuant: Complex Rank-One Quantization Algorithm

Maël Chaumette Rémi Gribonval Elisa Riccietti

GRETSI 2025

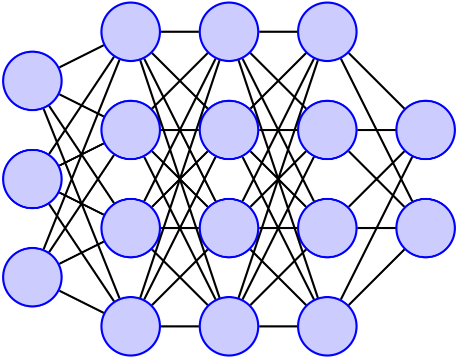
26/08/2025

ENS DE LYON

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

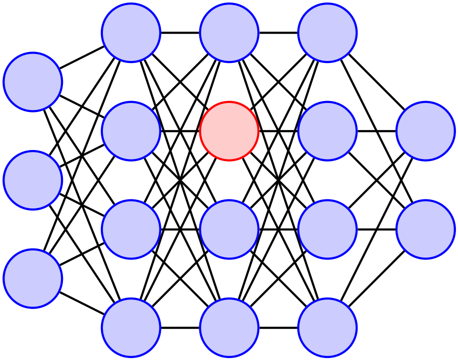
Context

Main goal: Quantize the weights of a neural network by using rescaling invariance property [Neyshabur et al., 2015].



Context

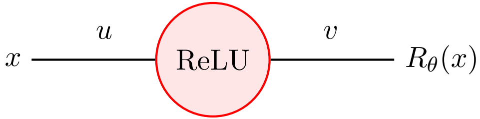
Main goal: Quantize the weights of a neural network by using rescaling invariance property [Neyshabur et al., 2015].



Context

We keep one neuron where

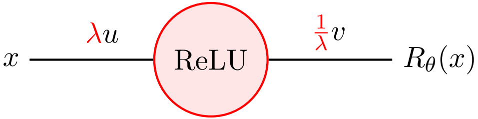
- $\theta = (u, v) \in \mathbb{R}^{d_{\text{in}} + d_{\text{out}}}$
- $R_\theta : x \in \mathbb{R}^{d_{\text{in}}} \mapsto \mathrm{ReLU}(\langle u, x \rangle) v = \mathbb{1}_{\langle u,x \rangle > 0} u v^\top x \in \mathbb{R}^{d_{\text{out}}}$



Context

We keep one neuron where

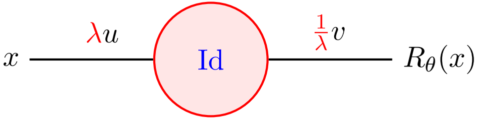
- $\theta = (u, v) \in \mathbb{R}^{d_{\text{in}}+d_{\text{out}}}$
- $R_\theta : x \in \mathbb{R}^{d_{\text{in}}} \mapsto \mathrm{ReLU}(\langle u, x\rangle)v = \mathbb{1}_{\langle \lambda u, x\rangle > 0} \lambda u \frac{1}{\lambda} v^\top x \in \mathbb{R}^{d_{\text{out}}}$



Context

We keep one neuron where

- $\theta = (u, v) \in \mathbb{R}^{d_{\text{in}}+d_{\text{out}}}$
- $R_\theta : x \in \mathbb{R}^{d_{\text{in}}} \mapsto \langle u, x\rangle v = \lambda u \frac{1}{\lambda} v^\top x \in \mathbb{R}^{d_{\text{out}}}$



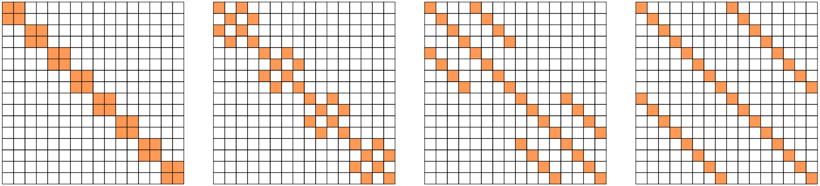
Context

1. Optimal quantization of real-valued rank-one matrices is done in Gribonval et al. [2023] by using rescaling invariance

Context

1. Optimal quantization of real-valued rank-one matrices is done in Gribonval et al. [2023] by using rescaling invariance

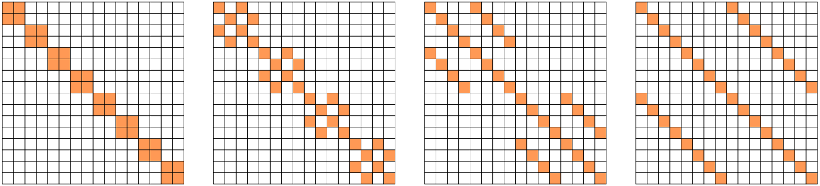
2. Application to quantization of butterfly matrices



Context

1. Optimal quantization of real-valued rank-one matrices is done in Gribonval et al. [2023] by using rescaling invariance

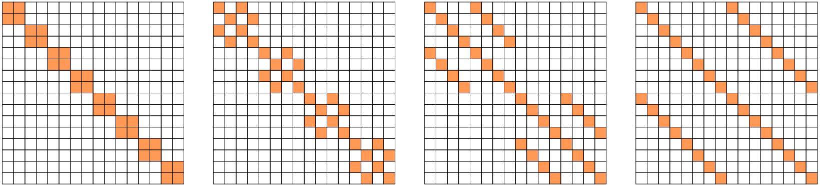
2. Application to quantization of butterfly matrices



3. Butterfly matrices appear in factorization of dense matrices, e.g., Fast Fourier Transform (FFT) [Cooley and Tukey, 1965]

Context

1. Optimal quantization of real-valued rank-one matrices is done in Gribonval et al. [2023] by using rescaling invariance
2. Application to quantization of butterfly matrices



3. Butterfly matrices appear in factorization of dense matrices, e.g., Fast Fourier Transform (FFT) [Cooley and Tukey, 1965]

$\Rightarrow$ Quantization of complex rank-one matrices to quantize butterfly matrices and see the impact on the FFT.

Quantization of complex rank-one matrices

$$\forall \lambda \in \mathbb{C}^*, \;\; (\lambda x) \left(\frac{1}{\bar{\lambda}} y \right)^H$$

Complex-valued rank-one matrices

Problem formulation

Given $(x, y) \in \mathbb{C}^m \times \mathbb{C}^n$ and letting $\mathbb{CF}_t := \mathbb{F}_t + i\mathbb{F}_t$ (with $\mathbb{F}_t$: floats with t-bit significand), we want to solve:

$$x^*, y^* \in \arg \min_{\hat{x} \in \mathbb{CF}_t^m, \hat{y} \in \mathbb{CF}_t^n} \|xy^H - \hat{x}\hat{y}^H\|^2$$

Complex-valued rank-one matrices

Problem formulation

Given $(x, y) \in \mathbb{C}^m \times \mathbb{C}^n$ and letting $\mathbb{CF}_t := \mathbb{F}_t + i\mathbb{F}_t$ (with $\mathbb{F}_t$: floats with t-bit significand), we want to solve:

$$x^*, y^* \in \arg \min_{\hat{x} \in \mathbb{CF}_t^m, \hat{y} \in \mathbb{CF}_t^n} \|xy^H - \hat{x}\hat{y}^H\|^2$$

Potential approaches

- **Naive:** Map x and y to their nearest neighbor in $\mathbb{CF}_t$ with $\mathrm{round}(\cdot)$.
- **Real-valued:** Use optimal quantization algorithm for *real-valued* rank-one matrices [Gribonval et al., 2023].

Problem resolution

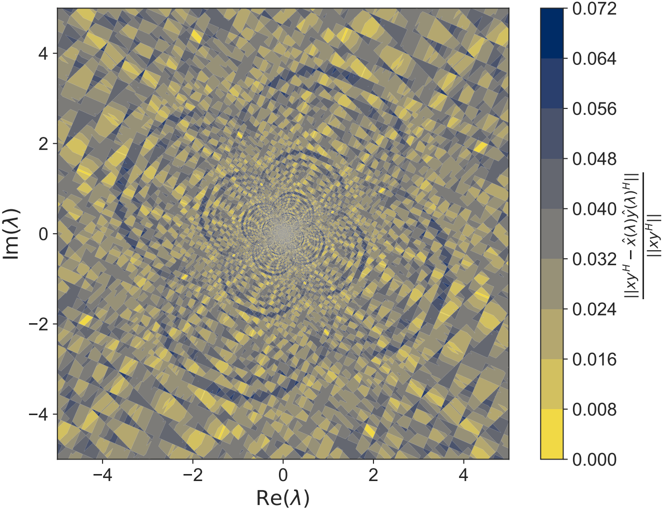
Lemma: problem characterization

$$\inf_{\hat{x}\in\mathbb{CF}_t^m,\hat{y}\in\mathbb{CF}_t^n} \|xy^H - \hat{x}\hat{y}^H\|^2 = \inf_{\lambda\in\mathbb{C}} f(\lambda)$$

where λ is the scaling parameter.

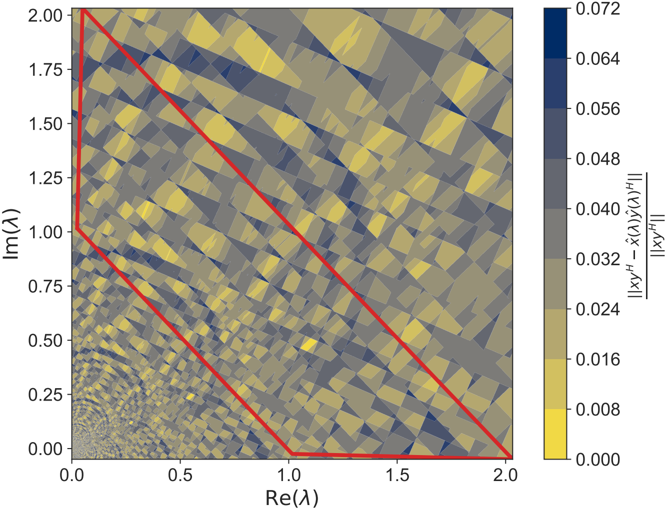
$\rightarrow$ Reduction of a problem with $4mn$ variables to a one scalar problem.

Study of the function f with $x, y \in \mathbb{C}^2$



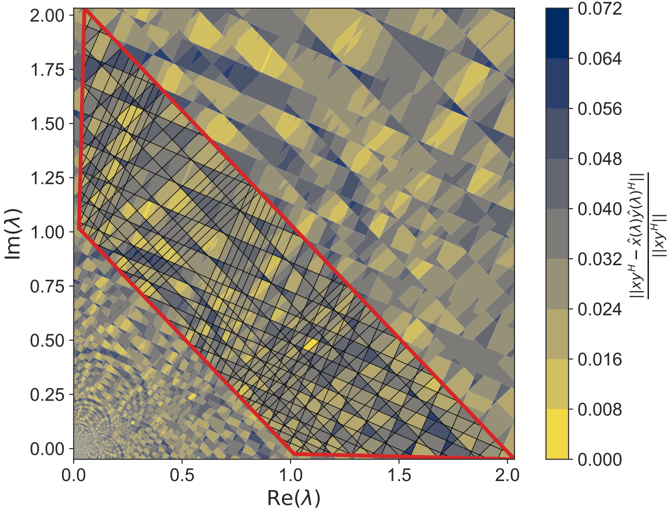
Shape of f with $t = 3$.

Study of the function f with $x, y \in \mathbb{C}^2$



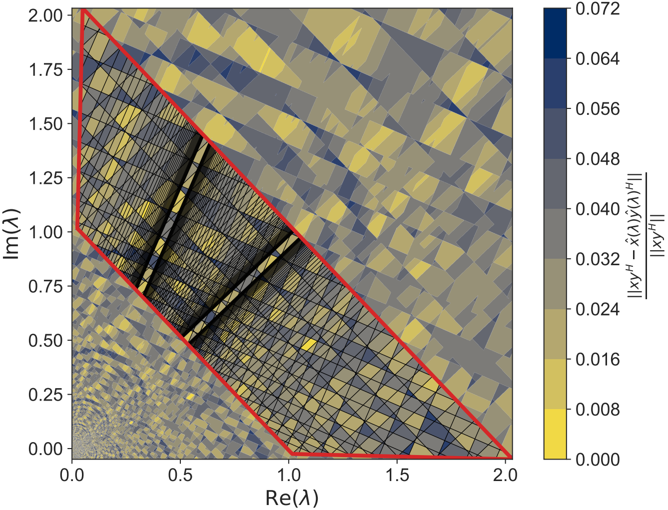
Shape of f with $t = 3$.

Study of the function f with $x, y \in \mathbb{C}^2$



Shape of f with $t = 3$ with $b \leq 3$ (cf. Lemma 2)

Study of the function f with $x, y \in \mathbb{C}^2$



Shape of f with $t = 3$ with $b \leq 7$ (cf. Lemma 2)

Definition of the algorithm

- Introduction of a parameter $b_m \in \mathbb{N}$ to control the number of discontinuity lines.

Definition of the algorithm

- Introduction of a parameter $b_m \in \mathbb{N}$ to control the number of discontinuity lines.
- Algorithm steps:
  1. compute all the centroids from the discontinuity lines
  2. evaluate f on all these centroids
  3. keep the **best** scaling factor, λ_{b_m}
  4. return $\hat{x}_{b_m} := \mathrm{round}(\lambda_{b_m} x)$ and $\hat{y}_{b_m} := \mathrm{round}(\mu_{b_m} y)$

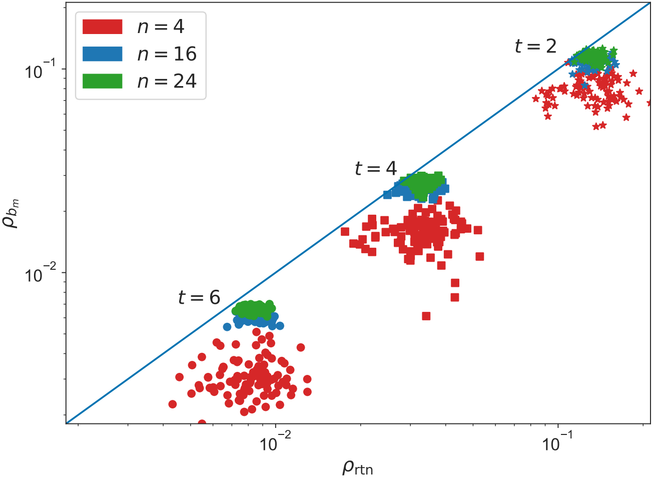
Definition of the algorithm

- Introduction of a parameter $b_m \in \mathbb{N}$ to control the number of discontinuity lines.

- Algorithm steps:
  1. compute all the centroids from the discontinuity lines
  2. evaluate f on all these centroids
  3. keep the **best** scaling factor, λ_{b_m}
  4. return $\hat{x}_{b_m} := \mathrm{round}(\lambda_{b_m} x)$ and $\hat{y}_{b_m} := \mathrm{round}(\mu_{b_m} y)$

Evaluation metric: $\rho := \dfrac{\|xy^H - \hat{x}\hat{y}^H\|}{\|xy^H\|}$

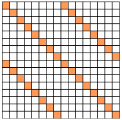
Results with 100 pairs $x, y \in \mathbb{C}^n$

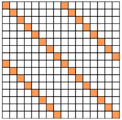
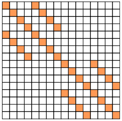


ρ_{b_m} in terms of ρ_{rtn} for different values of n and t with $b_m = 3$.

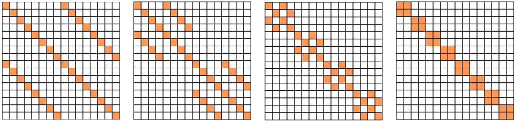
Application to butterfly matrices

What's FFT?

$F_n =$  $\begin{pmatrix} F_{n/2} & 0 \\ 0 & F_{n/2} \end{pmatrix}$ $\left(\begin{array}{c}\text{Sort the even} \\ \text{and odd indices}\end{array}\right)$

$=$   $\begin{pmatrix} F_{n/4} & 0 & 0 & 0 \\ 0 & F_{n/4} & 0 & 0 \\ 0 & 0 & F_{n/4} & 0 \\ 0 & 0 & 0 & F_{n/4} \end{pmatrix}$ (Permutation)

$\vdots$

$=$  (Permutation)

$\underbrace{B_1 \quad B_2 \quad B_3 \quad B_4}_{L := \log_2(n) \text{ butterfly factors}}$

Complex butterfly quantization

New problem formulation

Consider $B_1, ..., B_L \in \mathbb{C}^{n \times n}$. The new quantization problem is

$$B_1^*, ..., B_L^* \in \arg \min_{\hat{B}_1, ..., \hat{B}_L \in \mathbb{C}^{n \times n}} \|B_1 \cdots B_L - \hat{B}_1 \cdots \hat{B}_L\|$$

Complex butterfly quantization

New problem formulation

Consider $B_1, ..., B_L \in \mathbb{C}^{n \times n}$. The new quantization problem is

$$B_1^*, ..., B_L^* \in \arg \min_{\hat{B}_1, ..., \hat{B}_L \in \mathbb{C}^{n \times n}} \|B_1 \cdots B_L - \hat{B}_1 \cdots \hat{B}_L\|$$

Solvable problem with $L = 2$: the problem can be written as n independant **rank-one quantization problems**.

Complex butterfly quantization

New problem formulation

Consider $B_1, ..., B_L \in \mathbb{C}^{n\times n}$. The new quantization problem is

$$B_1^*, ..., B_L^* \in \arg \min_{\hat{B}_1,...,\hat{B}_L \in \mathbb{C}^{n\times n}} \|B_1 \cdots B_L - \hat{B}_1 \cdots \hat{B}_L\|$$

Solvable problem with $L = 2$: the problem can be written as n independant **rank-one quantization problems**.

Heuristic for the parenthesis decomposition

Pairwise: writing $(B_1 B_2)(B_3 B_4) \cdots (B_{L-1} B_L)$

Complex butterfly quantization

New problem formulation

Consider $B_1, ..., B_L \in \mathbb{C}^{n\times n}$. The new quantization problem is

$$B_1^*, ..., B_L^* \in \arg\min_{\hat{B}_1,...,\hat{B}_L \in \mathbb{C}^{n\times n}} \|B_1 \cdots B_L - \hat{B}_1 \cdots \hat{B}_L\|$$

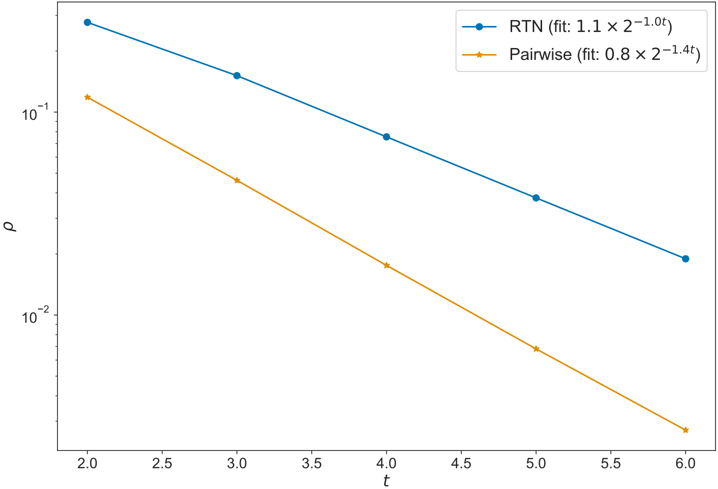
Solvable problem with $L = 2$: the problem can be written as n independant **rank-one quantization problems**.

Heuristic for the parenthesis decomposition

Pairwise: writing $(B_1 B_2)(B_3 B_4)\cdots(B_{L-1} B_L)$

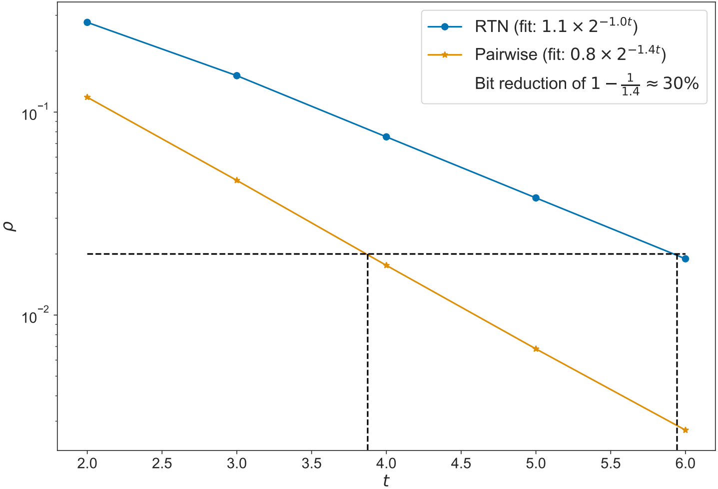
The metric is $\rho := \dfrac{\|B_1 \cdots B_L - \hat{B}_1 \cdots \hat{B}_L\|}{\|B_1 \cdots B_L\|}$

Quantization error on the butterfly decomposition



Average on 10 gaussian matrices of ρ in terms of t with $n = 256$ and $b_m = 3$.

Quantization error on the butterfly decomposition



Average on 10 gaussian matrices of ρ in terms of t with $n = 256$ and $b_m = 3$.

Quantization error on the FFT

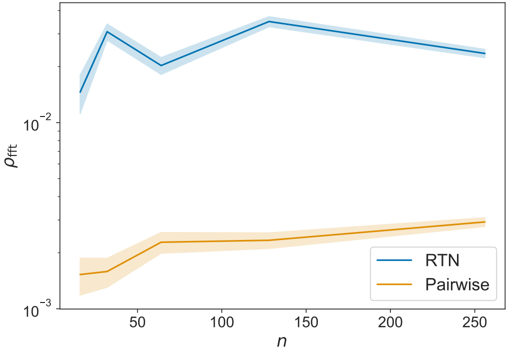
Let

- $x \in \mathbb{R}^n$ be the signal
- $y := Fx = B_1 \cdots B_L x \in \mathbb{C}^n$ its Fourier transform
- $\hat{y} := \hat{F}x = \hat{B}_1 \cdots \hat{B}_L x$
- $\rho_{\text{fft}} := \frac{\|y - \hat{y}\|}{\|y\|}$ the comparison metric

Quantization error on the FFT

Let

- $x \in \mathbb{R}^n$ be the signal
- $y := Fx = B_1 \cdots B_L x \in \mathbb{C}^n$ its Fourier transform
- $\hat{y} := \hat{F}x = \hat{B}_1 \cdots \hat{B}_L x$
- $\rho_{\text{fft}} := \frac{\|y-\hat{y}\|}{\|y\|}$ the comparison metric



Average on 10 gaussian signals of ρ_{fft} in terms of n with $t = 5$ and $b_m = 3$.

Conclusion

Wrap-up:

- High-performance complex-valued rank-one quantization algorithm
- Compared to RTN, the number of bits is reduced by 30% for a given precision on butterfly matrices

What's next?

- Working on an extended version
- Quantization of a product of matrices of any rank
- Extend this work to quantize ReLU networks

Thanks for your attention

Références

[1] B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pages 1376–1401. PMLR, 2015.

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

[3] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19 (90):297–301, 1965.

Appendix: on Lemma 1 and Lemma 2

Expression of f

$$f : \lambda \in \mathbb{C} \mapsto \max_{\hat{x} \in \mathrm{round}(\lambda x)} \| x y^H - \hat{x}\, \mathrm{round}(\mu(\hat{x}) y)^H \|$$

where $\mu(\hat{x}) := \frac{\langle x, \hat{x} \rangle}{\|x\|^2}$ if $x \neq 0$ and 0 otherwise.

Lemma: Discontinuity points of f

Let $x \in \mathbb{C}^m$. For each $x_j := u + iv$, $j = 1, ..., m$, the discontinuity points of the function $\lambda \in \Omega \mapsto \mathrm{round}(\lambda x_j)$ have for equations

$$\begin{cases} u\mathrm{Im}(\lambda) = -v\mathrm{Re}(\lambda) + (k + \frac{1}{2}) 2^{2-b-t} \\ v\mathrm{Im}(\lambda) = u\mathrm{Re}(\lambda) + (k + \frac{1}{2}) 2^{2-b-t} \\ v\mathrm{Im}(\lambda) = u\mathrm{Re}(\lambda) - (k + \frac{1}{2}) 2^{2-b-t} \end{cases} \qquad \forall k \in [\![2^{t-1}, 2^t - 1]\!], \forall b \in \mathbb{N}$$

Appendix: more results

Proposition: on the infimum of f

We can prove that:

- f is continuous on the accumulation lines.
- f admits a minimizer on $\mathbb{C}$.